Σχολή Θετικών Επιστημών και Τεχνολογίας

Μεταπτυχιακή Εξειδίκευση στα Πληροφοριακά Συστήματα

Πτυχιακή / Διπλωματική Εργασία

Αυτόματη περίληψη κειμένων

Ευαγγελία Μουργή

Επιβλέπουσα καθηγήτρια: Αμαλία Φωκά

Πάτρα, Σεπτέμβριος 2023

Αυτόματη περίληψη κειμένων

Ευαγγελία Μουργή

ΑΜ 147759

Επιτροπή Επίβλεψης Πτυχιακής / Διπλωματικής Εργασίας

Επιβλέπουσα Καθηγήτρια:
Αμαλία Φωκά

Συν-Επιβλέπων Καθηγητής:
Εμμανουήλ Τζαγκαράκης

Πάτρα, Ιούνιος 2023

Περίληψη

Η παρούσα εργασία ασχολείται με την υλοποίηση μιας εφαρμογής για την αυτόματη συλλογή ειδησεογραφικών άρθρων από διαδικτυακές πηγές και την αυτόματη περίληψη αυτών των άρθρων. Ο καταιγισμός πληροφορίας στο διαδίκτυο έχει προκαλέσει υπερφόρτωση πληροφορίας, με αποτέλεσμα η εύρεση, η απορρόφηση και η επεξεργασία της να είναι πολύ δύσκολη και χρονοβόρα. Για την ενημέρωση μας για ένα θέμα μπορεί να χρειαστεί να ανατρέξουμε σε πολλές διαδικτυακές πηγές. Σε πολλές δε περιπτώσεις βρισκόμαστε αντιμέτωποι με αναδημοσιεύσεις, με αποτέλεσμα να σπαταλάμε πολύτιμο χρόνο.

Από την άλλη, η δημιουργία χειροκίνητα περιλήψεων ειδησεογραφικών άρθρων είναι μια χρονοβόρα διαδικασία, η οποία απαιτεί την ενασχόληση εξειδικευμένων επαγγελματιών, και είναι αδύνατο να είναι διαθέσιμη τη στιγμή που τη χρειαζόμαστε.

Η εφαρμογή αυτή έχει σκοπό να μειώσει το χρόνο που απαιτείται για την αναζήτηση της πληροφορίας για ένα θέμα μέσω της αυτόματης αναζήτησης κειμένων, αλλά και του χρόνου που απαιτείται για την πρόσληψη της, αφού προσφέρεται σε μορφή περίληψης. Με αυτόν τον τρόπο αποφεύγεται η ανάγκη για επίσκεψη πολλών σελίδων και την ανάγνωση πολλών κειμένων, αρκετά από τα οποία περιέχουν αναδημοσιεύσεις.

Η υλοποίηση γίνεται με την υλοποίηση ενός μηχανισμού για την αυτόματη αναζήτηση, συγκέντρωση και ανάκτηση των άρθρων. Στη συνέχεια, αφού διασφαλίσουμε ότι τα άρθρα που συλλέγουμε είναι μοναδικά, χρησιμοποιούμε τεχνικές μηχανικής μάθησης και επεξεργασίας φυσικής γλώσσας για να δημιουργήσουμε μία μοναδική περίληψη η οποία περιέχει τις σημαντικές πληροφορίες όλων των άρθρων.

Έτσι, συμβάλλουμε στη μείωση του χρόνου και της ενέργειας που απαιτείται για την ενημέρωση μας σχετικά με ένα θέμα.

Λέξεις – Κλειδιά

Αυτόματη περίληψη, μηχανική μάθηση.

Automatic Text Summarization

Evangelia Mourgi

Abstract

The present work deals with the implementation of an application for automatically collecting news articles from online sources and automatically summarizing these articles. The flood of information on the internet has caused information overload, making finding, absorbing and processing it very difficult and time-consuming. In order to inform us about a topic we may need to refer to many online sources. And in many cases we are faced with republishing, as a result of which we waste valuable time.

On the other hand, manually creating summaries of news articles is a time-consuming process, which requires the involvement of specialized professionals, and is impossible to have available at the moment we need it.

This application aims to reduce the time required to search for information on a topic through automatic text search, but also the time required to receive it, since it is offered in summary form. This avoids the need to visit many pages and read many texts, many of which contain reposts.

The implementation is done by implementing a mechanism for the automatic search, aggregation and retrieval of the articles. Then, after ensuring that the articles we collect are unique, we use machine learning and natural language processing techniques to create a single summary that contains the important information of all the articles.

Thus, we help reduce the time and energy required to inform us about a topic.

Keywords

Automatic text summarization, Machine Learning.

Περιεχόμενα

Κατάλογος Εικόνων / Σχημάτων

Κατάλογος Πινάκων

1. Εισαγωγή

Η πληθώρα ειδησεογραφικών άρθρων που υπάρχει διαθέσιμη στο διαδίκτυο, καθιστά την ανάκτηση πληροφορίας δύσκολη και χρονοβόρα. Καθημερινά δημοσιεύονται εκατοντάδες άρθρα, με αποτέλεσμα να ερχόμαστε αντιμέτωποι με υπερφόρτωση πληροφορίας.

Την έννοια της υπερφόρτωσης πληροφορίας τη συναντάμε ιστορικά από τον 3 ή 4 αιώνα πΧ,, με τους ανθρώπους της εποχής να αποδοκιμάζουν την πληθώρα βιβλίων.

Στα μέσα του 15^{ου} αιώνα η ανάδυση της τυπογραφίας και της μαζικής παραγωγής εντύπων οδήγησε στη δυσαρέσκεια των αναγνωστών, οι οποίοι διατύπωναν ότι είναι δύσκολο να υποδεχθούν, επεξεργαστούν και διαχειριστούν το σύνολο της πληροφορίας.

Τα τελευταία χρόνια, με την ανάδυση του διαδικτύου, την ανάδειξη των ειδησεογραφικών ιστοσελίδων και των ΜΚΔ, η ποσότητα της πληροφορίας αυξάνεται εκθετικά και μαζί της αυξάνεται και η ποσότητα της ανεπιθύμητης ή επαναλαμβανόμενης πληροφορίας.

Αν θα θέλαμε να προσδιορίσουμε το πρόβλημα, θα μπορούσαμε να πούμε ότι η παρουσία υπερβολικής πληροφορίας δημιουργεί στους αποδέκτες της δυσκολία στην επεξεργασία και την αξιοποίησή της (Sharkar, 2019).

Η ενημέρωση για ένα θέμα απαιτεί πολύ χρόνο, καθώς περιλαμβάνει την αναζήτηση του θέματος ενδιαφέροντος, την πλοήγηση σε πολλαπλά άρθρα τα οποία δεν είναι πάντα χρήσιμα ή και πολλές φορές επαναλαμβανόμενα, καθιστώντας αναγκαία την ανάγνωση και κατανόηση πολλών διαφορετικών και συχνά μακροσκελών άρθρων. Η διαδικασία αυτή είναι χρονοβόρα και πολλές φορές εγκαταλείπεται.

Το παραπάνω πρόβλημα δε θα μπορούσε να επιλυθεί με τη δημιουργία χειροκίνητων περιλήψεων, ειδικά όταν πρόκειται για κείμενα που συλλέγονται με ad hoc αναζήτηση από το χρήστη. Η χειροκίνητη δημιουργία περιλήψεων προϋποθέτει τη συμμετοχή εξειδικευμένων ατόμων τα οποία, θα πρέπει προηγουμένως να μελετήσουν το σχετικό υλικό.

Μια τέτοια λογική έχει εξαιρετικά μεγάλο κόστος, καθώς η συγγραφή περίληψης είναι χρονοβόρα διαδικασία και δεν αξιοποιεί τη δυνατότητα αναζήτησης και συλλογής άρθρων από το διαδίκτυο και στη συνέχεια άμεσης παρουσίασης τους σε μορφή περίληψης.

Η ανάγκη για αποφυγή υπερφόρτωσης πληροφορίας αλλά και μείωσης του δαπανώμενου χρόνου για την ενημέρωση για ένα θέμα, απαιτεί τεχνικές όπως η θεματική ομαδοποίηση ή

η αυτόματη περίληψη, οι οποίες συμβάλλουν στο μετριασμό των δύο αυτών παραμέτρων. Στόχος των τεχνικών αυτών είναι η συμπύκνωση των βασικών νοημάτων των κειμένων σε λίγες γραμμές, ώστε η κατανόηση και αξιοποίηση τους να είναι εύκολη και γρήγορη.

Σκοπός αυτής της εργασίας είναι η ελαχιστοποίηση της απαιτούμενης προσπάθειας και χρόνου που απαιτείται για την ενημέρωση για ένα θέμα με χρήση του διαδικτύου. Στοχεύουμε στην αυτοματοποίηση της διαδικασίας αναζήτησης, ανάκτησης και αξιολόγησης ειδησεογραφικών άρθρων και στη δημιουργία μιας, σχετικά μικρού μεγέθους, ενιαίας περίληψης των ανακτώμενων άρθρων.

Η διαδικασία αναζήτησης και ανάκτησης άρθρων, θα είναι πλήρως αυτοματοποιημένη.

Η αναζήτηση θα γίνεται με βάση το επιθυμητό θέμα, και η ανάκτηση θα έχει ως αποτέλεσμα έναν αριθμό διαφορετικών μεταξύ τους άρθρων, τα οποία θα περιέχουν κείμενο σχετικό με το θέμα, απαλείφοντας κατά το δυνατό αποτελέσματα που δεν αναφέρονται σε αυτό.

Η περίληψη που θα δημιουργηθεί, επιδιώκουμε να περιέχει το σύνολο των σημαντικών νοημάτων των άρθρων, να είναι συνεκτική και να είναι κατά το δυνατό πιο κοντά στην περίληψη που θα δημιουργούσε ένας άνθρωπος.

Για την εξυπηρέτηση των παραπάνω, επιλέξαμε την υλοποίηση συνδυασμού έτοιμων μοντέλων για παραγωγή αυτόματης περίληψης, προκειμένου να παραχθεί ένα κατά το δυνατό συνεκτικό αποτέλεσμα.

Στο δεύτερο κεφάλαιο, επιχειρείται μια παρουσίαση του θεωρητικού υποβάθρου της διαδικασίας και ανασκόπηση της σχετικής βιβλιογραφίας. Το τρίτο κεφάλαιο περιγράφει τη μεθοδολογία που ακολουθήθηκε. Στο τέταρτο κεφάλαιο παρουσιάζεται αναλυτικά η υλοποίηση. Στο πέμπτο κεφάλαιο παρουσιάζεται συνοπτική ανάλυση και αξιολόγηση των αποτελεσμάτων. Στο τελευταίο κεφάλαιο παρουσιάζονται τα συμπεράσματα και τα μελλοντικά βήματα.

2. Θεωρητικό υπόβαθρο

Σε αυτό το κεφάλαιο θα παρουσιάσουμε αναλυτικά τις βασικές έννοιες και τεχνικές που θα χρησιμοποιήσουμε. Πιο συγκεκριμένα, θα περιγράψουμε το αντικείμενο της αυτόματης περίληψης κειμένων, των ελέγχων ομοιότητας, των μεθόδων αξιολόγησης και του σταδίου συλλογής των άρθρων.

2.1. Ανασκόπηση σχετικών τεχνικών

Η δημιουργία μιας περίληψης είναι εξαιρετικά απαιτητική εργασία για τους υπολογιστές καθώς πρέπει να επεξεργαστούν τη φυσική γλώσσα, να κατανοήσουν τη σημαντικότητα των πληροφοριών, να συγχωνεύσουν τα νοήματα και στη συνέχεια να δημιουργήσουν ένα νέο κείμενο, το οποίο να μπορεί όχι μόνο να γίνει κατανοητό, αλλά να μεταδώσει την πληροφορία στον αναγνώστη του.

Σε μεγάλο βαθμό οι προηγούμενες έρευνες ασχολήθηκαν με την εξαγωγική περίληψη, δηλαδή την επιλογή των σημαντικών προτάσεων των κειμένων και τη χρήση τους ως έχουν για τη δημιουργία της περίληψης. Από το τέλος της δεκαετίας του 1950 ακόμα ο Luhn πρότεινε τη χρήση της συχνότητας εμφάνισης λέξεων για την εύρεση εκείνων των λέξεων μέσα σε ένα κείμενο οι οποίες μεταφέρουν τις πιο σημαντικές πληροφορίες του (Nenkova & McKeown, 2011), και στη συνέχεια την επιλογή εκείνων των προτάσεων που έχουν μεγάλη πυκνότητα τέτοιων λέξεων.

Από τότε μέχρι σήμερα έχουν δημιουργηθεί πολλές διαφορετικές προσεγγίσεις όσον αφορά την εξαγωγική περίληψη κειμένων, οι οποίες έχουν βελτιώσει πολύ την αντιπροσωπευτικότητα του παραγόμενου αποτελέσματος.

Οι τελευταίες περιλαμβάνουν μεθόδους που βασίζονται σε γράφους όπως ο LexRank, μεθόδους που βασίζονται στην κεντρικότητα όπως το Centroid και μεθόδους που βασίζονται σε συλλογές κειμένων όπως το TsSum (Zangh, M., Gang, Wanting, Ningbo, & Wenfen, 2022).

Τα μοντέλα αφαιρετικής περίληψης ανήκουν στην επεξεργασία φυσικής γλώσσας καθώς απαιτούν την κατανόηση αλλά και την παραγωγή κειμένου, οπότε για αρκετό καιρό η πρόοδος ήταν αργή.

Όμως η εξέλιξη της τεχνολογίας των νευρωνικών δικτύων και της βαθιάς μάθησης, έδωσε το απαραίτητο υπόβαθρο για την εξέλιξη των μοντέλων αφαιρετικής περίληψης.

Το 2015, οι Rush et al. εφάρμοσαν τεχνολογίες βαθιάς μάθησης για τη δημιουργία αυτόματης περίληψης (Rush, 2015). Από τότε και μετά, έχουν αναδυθεί πολλές μέθοδοι για την αυτόματη αφαιρετική περίληψη κειμένου που βασίζονται σε τεχνολογίες βαθιάς μάθησης (Zangh, M., Gang, Wanting, Ningbo, & Wenfen, 2022).

Πλέον έχουν δημοσιευτεί αρκετά άρθρα, τα οποία, πέρα από τη βελτιστοποίηση ενός μοντέλου για την παραγωγή αυτόματων περιλήψεων, επιχειρούν το συνδυασμό μεθόδων και τεχνικών.

Τα παραπάνω επικεντρώνονται σχεδόν αποκλειστικά στην αυτόματη περίληψη ενός μοναδικού εγγράφου. Η ανάγκη για εξαγωγή μίας περίληψης από πολλαπλά κείμενα προέκυψε μεταγενέστερα (Nenkova & McKeown, 2011), λόγω της πληθώρας κειμένων στο διαδίκτυο. Με αυτό το δεδομένο, αποδείχθηκε πιο χρήσιμη η δημιουργία μιας περίληψης από πολλά κείμενα που αναφέρονται στο ίδιο θέμα. Η έρευνα και σε αυτό το αντικείμενο έχει προχωρήσει αρκετά, με εφαρμογές οι οποίες εκμεταλλεύονται διάφορες μεθόδους και τους συνδυασμούς τους.

Οι Banerjee, S, Mitra, P, & Sugiyama, K (Banerjee, S, Mitra, P, & Sugiyama, K, 2016) δημιούργησαν ένα μοντέλο αφαιρετικής περίληψης, το οποίο πριν προχωρήσει σε περίληψη των κειμένων τα ομαδοποιεί με βάση τις προτάσεις του πιο σημαντικού κειμένου της συλλογής. Μια άλλη προσέγγιση είναι ο συνδυασμός συστημάτων για την αυτόματη περίληψη, που περιλαμβάνει τη δημιουργία υποψήφιων περιλήψεων από διαφορετικά συστήματα, οι οποίες συνδυάζονται προκειμένου να βελτιωθεί το περιεχόμενο τους (Hong, Marcus, & Nenkova, 2015)

Επιπλέον κερδίζει έδαφος η εξειδίκευση των μηχανισμών περίληψης κειμένων ανάλογα με το είδος των κειμένων αυτών. Έτσι, αναδύονται μοντέλα τα οποία ρυθμίζονται για την περίληψη συγκεκριμένου είδους εγγράφων όπως οι επιστημονικές δημοσιεύσεις, τα ειδησεογραφικά άρθρα, ή τα λογοτεχνικά κείμενα. Μια εστιασμένη στα ειδησεογραφικά άρθρα προσέγγιση είναι αυτή των (Mishra & Gayen, 2018), ενώ οι Fabri (Fabbri, Li, She, Li, & Radev, 2019) δημιούργησαν ένα σύνολο δεδομένων (dataset) για την εκπαίδευση και ρύθμιση μοντέλων για τη δημιουργία αφαιρετικών περιλήψεων πολλαπλών ειδησεογραφικών άρθρων.

2.2. Γενικές έννοιες

2.2.1. Επεξεργασία φυσικής γλώσσας

Η επεξεργασία φυσικής γλώσσας (Natural Language Processing), είναι εκείνο το πεδίο της επιστήμης των υπολογιστών και της τεχνητής νοημοσύνης που ασχολείται με το χειρισμό και την κατανόηση της φυσικής γλώσσας από τους τελευταίους.

Ο όρος φυσική γλώσσα περιλαμβάνει όλες εκείνες τις γλώσσες οι οποίες αναπτύχθηκαν φυσικά, μέσω της ανθρώπινης επικοινωνίας σε αντίθεση με τις κατασκευασμένες γλώσσες, όπως οι γλώσσες προγραμματισμού (Sharkar, 2019).

Η επεξεργασία φυσικής γλώσσας έχει ρίζες στην υπολογιστική γλωσσολογία και κύριο αντικείμενο της είναι η ανάπτυξη τεχνικών και εφαρμογών που δίνουν τη δυνατότητα στους υπολογιστές να κατανοήσουν και να επεξεργαστούν τη φυσική γλώσσα, επιστρέφοντας χρήσιμα αποτελέσματα.

Η επεξεργασία φυσικής γλώσσας περιλαμβάνει μεταξύ άλλων μεθόδους που μπορούν να χρησιμοποιηθούν για:

- **Αυτόματη μετάφραση κειμένων.**
  Είναι μια από τις πιο δημοφιλείς εφαρμογές της επεξεργασίας φυσικής γλώσσας. Αρχικά βασίστηκε στην απλή αντικατάσταση των λέξεων από τη γλώσσα πηγή με τις αντίστοιχες στη γλώσσα στόχο. Στη συνέχεια αναπτύχθηκαν πιο εξελιγμένες τεχνικές, οι οποίες μπορούσαν να λαμβάνουν υπόψη γραμματικούς συντακτικούς και σημασιολογικούς κανόνες, δίνοντας πιο κατανοητό και ορθό αποτέλεσμα (Sharkar, 2019).
- **Αναγνώριση λόγου.**
  Πρόκειται για εφαρμογές που έχουν στόχο να αναγνωρίσουν τον ανθρώπινο λόγο συνήθως στη μορφή ερωτήματος και στη συνέχεια να παράξουν αυτόματα την κατάλληλη απάντηση. Χρησιμοποιούν τεχνικές όπως η σύνθεση και η ανάλυση λόγου και η ανάλυση δομής (parsing) οι οποίες έχουν βελτιώσει σημαντικά το παραγόμενο αποτέλεσμα. Ειδικότερα, τα συστήματα απάντησης ερωτήσεων, εκτός από τις τεχνικές επεξεργασίας της φυσικής γλώσσας, χρησιμοποιούν και τεχνικές ανάκτησης πληροφοριών (Information Retrieval) (Sharkar, 2019).
- **Αναγνώριση και ανάλυση περιεχομένου**

Αφορά μεταξύ άλλων με τη συντακτική και σημασιολογική ανάλυση του λόγου. Μια από τις εφαρμογές του είναι η αναγνώριση του πραγματικού νοήματος μιας λέξης η οποία βρίσκεται σε μία πρόταση. Για την επίτευξη του συγκεκριμένου στόχου, λαμβάνει υπόψη τη σημασία της λέξης, το αν είναι απομονωμένη, αλλά και το νόημα που λαμβάνει βάσει της θέσης της στην πρόταση. Μια άλλη εφαρμογή της συγκεκριμένης μεθόδου, είναι η εύρεση των διαφορετικών λέξεων, που όμως αναφέρονται στην ίδια οντότητα (Sharkar, 2019).

- **Περίληψη κειμένου:**
  Στόχος αυτών των εφαρμογών είναι η μείωση του μεγέθους ενός κειμένου με τέτοιο τρόπο ώστε να διατηρούνται τα σημαντικά νοήματα και πληροφορίες του αρχικού. Οι τεχνικές περίληψης κειμένου κατηγοριοποιούνται σε δύο τύπους, την αποσπασματική ή εξαγωγική περίληψη (extractive summarization) και την αφαιρετική περίληψη (abstractive summarization).

Καθώς τα δεδομένα φυσικής γλώσσας στα κείμενα βρίσκονται σε αδόμητη μορφή, για την εκτέλεση των παραπάνω ενεργειών απαιτούνται στάδια προ-επεξεργασίας τους (Sharkar, 2019).

Η προ-επεξεργασία είναι μια διαδικασία μετατροπής του κειμένου από συλλογή αδόμητων δεδομένων σε δεδομένα με συγκεκριμένη δομή. Η προ-επεξεργασία κειμένου περιγράφεται αναλυτικά σε επόμενη παράγραφο.

2.2.2. Μηχανική μάθηση

Με τον όρο μηχανική μάθηση, αναφερόμαστε σε εκείνο το πεδίο της τεχνητής νοημοσύνης που ασχολείται με την ανάπτυξη τεχνικών που επιτρέπουν στις μηχανές να μαθαίνουν αυτόματα με βάση συγκεκριμένα μοντέλα. Αυτό έχει ως αποτέλεσμα τη βελτίωση τους με την πάροδο του χρόνου χωρίς την ανάγκη εξαντλητικού προγραμματισμού για κάθε πιθανή περίπτωση. Οι μέθοδοί της περιλαμβάνουν την επιβλεπόμενη μάθηση (supervised learning), τη μη επιβλεπόμενη μάθηση (unsupervised learning) και την ενισχυτική μάθηση (reinforcement learning) (Sharkar, 2019).

2.2.3. Νευρωνικά δίκτυα

Τα τεχνητά νευρωνικά δίκτυα είναι μοντέλα τα οποία έχουν τη δυνατότητα να μαθαίνουν μέσω εκπαίδευσης. Ο σχεδιασμός τους βασίζεται στη λειτουργία του ανθρώπινου

εγκεφάλου. Ο ανθρώπινος εγκέφαλος αποτελείται από κυρίως από νευρώνες (νευρικά κύτταρα) που διασυνδέονται μεταξύ τους με νευρίτες (νηματοειδείς ίνες).

Όταν ένας νευρώνας διεγερθεί, μεταφέρει τη διέγερση του σε άλλους νευρώνες, οι οποίοι παραλαμβάνουν τη διέγερση μέσω των δενδριτών (προεκτάσεις των νευρώνων) τους. Η σύναψη (ισχύς στο σημείο επαφής μεταξύ νευρίτη και δενδρίτη) καθορίζει τη συνεκτικότητα ανάμεσα στους νευρώνες. Η εκπαίδευση του ανθρώπινου εγκεφάλου αλλάζει την ισχύ της σύναψης μεταξύ νευρώνων κατά την επαναλαμβανόμενη διέγερση από το ίδιο ερέθισμα. Τα περίπου 100 δισεκατομμύρια νευρώνες του ανθρώπινου εγκεφάλου που είναι διασυνδεδεμένοι με πολύπλοκους τρόπους, καθιστά δυνατή τη μάθηση. Μιμούμενο τον ανθρώπινο εγκέφαλο, το τεχνητό νευρωνικό δίκτυο περιλαμβάνει έναν αριθμό μονάδων εργασίας (κόμβοι) που συνδέονται μεταξύ τους μέσω συνδέσμων, και επιδιώκουν την προσαρμογή των βαρών των συνδέσμων (ισχύς σύναψης στον ανθρώπινο εγκέφαλο) ώστε να ταιριάζουν με τις σχέσεις εισόδων-εξόδων που λαμβάνουν (Tan, Steinbach, Karpatne, & Kumar, 2020).

Τα τεχνητά νευρωνικά δίκτυα είναι οργανωμένα σε επίπεδα, δηλαδή οι κόμβοι τους είναι οργανωμένοι σε ομάδες.

Οι κόμβοι των τεχνητών νευρωνικών δικτύων είναι διατεταγμένοι σε ομάδες οι οποίες καλούνται επίπεδα. Τα επίπεδα συνήθως έχουν ως είσοδο την έξοδο του προηγούμενου επιπέδου, και επιτελούν το καθένα διαφορετική εργασία στα δεδομένα εισόδου του. Όταν τα σήματα διαδίδονται μόνο προς τα εμπρός, τότε έχουμε δίκτυο προσοτροφοδότησης.

Η χρήση των τεχνητών νευρωνικών δικτύων βασίζεται στο ότι τα πολύπλοκα χαρακτηριστικά των υψηλών επιπέδων μπορούν να συντεθούν από τα απλούστερα χαρακτηριστικά των χαμηλών επιπέδων. Τα δίκτυα με μεγάλο αριθμό κρυφών επιπέδων, δηλαδή των επιπέδων που βρίσκονται μεταξύ του πρώτου (είσοδος) και του τελευταίου (έξοδος) ονομάζονται βαθιά νευρωνικά δίκτυα και η διαδικασία εκπαίδευσης τους καλείται βαθιά μάθηση (Tan, Steinbach, Karpatne, & Kumar, 2020).

Με βάση τα παραπάνω, το ισχυρό σημείο των νευρωνικών δικτύων είναι η ικανότητα τους να μαθαίνουν μέσω της εκπαίδευσης, κατά την οποία ανακαλύπτουν πολύπλοκες συσχετίσεις και πρότυπα, τα οποία στη συνέχεια θα χρησιμοποιήσουν για να εκτελέσουν πολύπλοκες εργασίες.

Η αρχιτεκτονική των νευρωνικών δίκτυών έχει εξελιχθεί σημαντικά τα τελευταία χρόνια, με την εισαγωγή προηγμένων μοντέλων όπως τα συνελικτικά νευρωνικά δίκτυα

(convolutional neural networks), τα αναδραστικά νευρωνικά δίκτυα (recurent neural networks) και τα συνελικτικά νευρωνικά δίκτυα (convolutional neural networks) (Zangh, M., Gang, Wanting, Ningbo, & Wenfen, 2022).

Τα αναδραστικά νευρωνικά δίκτυα υλοποιούνται με βάση το ότι η ανθρώπινη γνωστική λειτουργία βασίζεται στην εμπειρία και τη μνήμη, και επεξεργάζονται ακολουθίες δεδομένων. Το δίκτυο σε κάθε βήμα συνδυάζει τα χαρακτηριστικά εισόδου του τρέχοντος βήματος με αυτά του προηγουμένου και με αυτό τον τρόπο προβλέπει το αποτέλεσμα (Zangh, M., Gang, Wanting, Ningbo, & Wenfen, 2022).

Τα αναδραστικά νευρωνικά δίκτυα έχουν εφαρμογή σε συστήματα εξαγωγής πληροφορίας, αυτόματης περίληψης και μηχανικής μάθησης. Τα RNN παρουσιάζουν πρόβλημα όταν η ακολουθία είναι πολύ μεγάλη, το οποίο όμως αντιμετωπίστηκε με τα νευρωνικά δίκτυα Long Short-Term Memory (LSTM), τα οποία αποθηκεύουν ή διαγράφουν πληροφορίες επιλεκτικά.

Τα συνελικτικά νευρωνικά δίκτυα είναι βαθιά δίκτυα προσοτροφοδότησης τα οποία αποτελούνται από πολλές συνελικτικές λειτουργίες. Τα πιο σημαντικά χαρακτηριστικά τους είναι η αραιή διάδραση, ο διαμοιρασμός παραμέτρων και οι ισοδύναμες αναπαραστάσεις, που δίνουν τη δυνατότητα στα μοντέλα να χειριστούν εισόδους διαφορετικών μεγεθών. Η συνέλιξη δίνει τη δυνατότητα αναγνώρισης των σημαντικών χαρακτηριστικών της εισόδου, τη σημαντική απλοποίηση της και την μείωση των απαιτούμενων παραμέτρων. Τα CNN βρίσκουν εφαρμογή πλέον στην αναγνώριση προσώπου, την αυτόματη μετάφραση, την ανάλυση κίνησης και την επεξεργασία φυσικής γλώσσας με καλά αποτελέσματα (Zangh, M., Gang, Wanting, Ningbo, & Wenfen, 2022).

2.3. Προεπεξεργασία κειμένου

Η προεπεξεργασία ή κανονικοποίηση κειμένου είναι η διαδικασία καθαρισμού και κανονικοποίησης του κειμένου προκειμένου αυτό να έχει μία μορφή κατάλληλη για διεργασίες επεξεργασίας φυσικής γλώσσας. Με δεδομένο ότι το κείμενο είναι η πλέον αδόμητη μορφή δεδομένων και περιέχει αρκετά στοιχεία που είναι περιττά ή επιβλαβή για την περαιτέρω επεξεργασία του και την παραγωγή ποιοτικών αποτελεσμάτων, το στάδιο της προεπεξεργασίας είναι απαραίτητο. Η προεπεξεργασία είναι μια διαδικασία η οποία

περιλαμβάνει διάφορα στάδια και τεχνικές για την προετοιμασία του κειμένου. Συνήθως περιλαμβάνονται τα παρακάτω στάδια (Sharkar, 2019):

- Σε πρώτο στάδιο το κείμενο καθαρίζεται από εκείνο το περιεχόμενο το οποίο είναι περιττό ή ακόμα και επιβλαβές κατά την εφαρμογή τεχνικών επεξεργασίας φυσικής γλώσσας.

  Ο καθαρισμός του θορύβου μπορεί να περιλαμβάνει ετικέτες (tags) όπως HTML ή Javascript , αφαίρεση χαρακτήρων με τόνους και αντικατάστασή τους από χαρακτήρες χωρίς τόνους αλλά και αφαίρεση συντμήσεων και αντικατάστασή τους με ολόκληρη τη φράση (για παράδειγμα μετατροπή του «απ'το» σε «από το») (Sharkar, 2019).
- Στη συνέχεια το κείμενο διαχωρίζεται σε μικρότερες μονάδες (tokenization). Τα tokens είναι τα ελάχιστα σε μέγεθος, ανεξάρτητα μεταξύ τους συστατικά του τελευταίου, τα οποία έχουν σαφή σύνταξη και σημασία. Για παράδειγμα ένα κείμενο αποτελείται από παραγράφους. Οι παράγραφοι μπορούν να χωριστούν σε προτάσεις, και αυτές με τη σειρά στους σε φράσεις ή λέξεις. Το μέχρι ποιο σημείο θα φτάσει η κατάτμηση του κειμένου εξαρτάται από την τεχνική που θα εφαρμοστεί στη συνέχεια, αν και συνήθως αυτό είναι στο επίπεδο των προτάσεων, φράσεων ή λέξεων (Sharkar, 2019).
- Επόμενο βήμα είναι η αφαίρεση λέξεων με μικρή ή χωρίς σημασία (stopwords) όπως είναι τα άρθρα και οι σύνδεσμοι, αφού συνήθως δεν προσφέρουν κάτι στην τεχνική που πρόκειται να εφαρμοστεί αλλά και η απαλοιφή ειδικών (μη αλφαριθμητικών) χαρακτήρων. Συχνά εφαρμόζεται μετατροπή των κεφαλαίων σε πεζά και διορθώσεις στο κείμενο όπως η απαλοιφή επαναλαμβανόμενων χαρακτήρων ή η διόρθωση ορθογραφικών λαθών (Sharkar, 2019).
- Στη συνέχεια το κείμενο κανονικοποιείται, δηλαδή αφαιρούνται οι πολλαπλές αναπαραστάσεις της ίδιας λέξης και αντικαθίστανται από μια κανονική της μορφή. Μία διαδικασία που μπορούμε να εφαρμόσουμε είναι αυτή του stemming, κατά την οποία διατηρείται μόνο το στέλεχος κάθε λέξης αφαιρώντας την κατάληξη. Για παράδειγμα, οι λέξεις «γράφω, γράφεις, γράφουμε» μπορούν να αντικατασταθούν από το «γραφ» με την αφαίρεση των καταλήξεων τους. Η δεύτερη διαδικασία που μπορεί να εφαρμοστεί είναι το lemmatization, κατά το οποίο αντικαθιστούμε τις πολλαπλές αναπαραστάσεις με την βασική τους μορφή, δηλαδή το λήμμα. Στο παραπάνω παράδειγμα, όλες οι λέξεις μπορούν να αντικατασταθούν με το «γράφω».

Αυτή η διαδικασία είναι πιο αργή, καθώς πρέπει να εντοπιστεί το λήμμα (Sharkar, 2019).

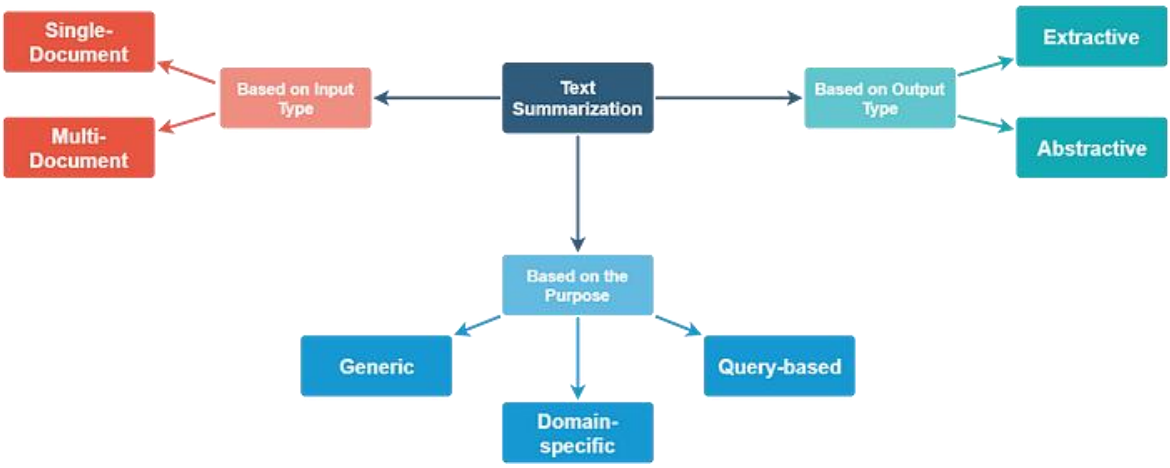
2.4. Αυτόματη περίληψη κειμένου

Η αυτόματη περίληψη κειμένων είναι μία τεχνική του πεδίου Επεξεργασίας Φυσικής Γλώσσας, η οποία έχει στόχο τη δημιουργία μιας συνοπτικής περίληψης ενός κειμένου ή συνόλου κειμένων (Sharkar, 2019). Η παραγόμενη περίληψη θα πρέπει να παρουσιάζει σε συνοπτική μορφή τις βασικές πληροφορίες και νοήματα του αρχικού κειμένου.

Η δημιουργία περιλήψεων που να αποτυπώνει πλήρως το περιεχόμενο των αρχικών κειμένων, είναι μια εργασία που αποδεικνύεται δύσκολη ακόμα και για τους ανθρώπους. Προϋποθέτει τη δυνατότητα πλήρους κατανόησης του κειμένου, αλλά και της αναδιοργάνωσης και συγχώνευσης των πληροφοριών (Nenkova & McKeown, 2011). Τέλος, προϋποθέτει τη δυνατότητα συγγραφής ενός νέου κειμένου με μικρή έκταση, το οποίο να περιγράφει συνοπτικά το περιεχόμενο του αρχικού κειμένου.

Όπως αναφέρθηκε και παραπάνω, οι τεχνικές αυτόματης περίληψης κειμένων χωρίζονται σε δύο βασικές κατηγορίες, ανάλογα με τον τρόπο παραγωγής της περίληψης, την αποσπασματική περίληψη (extractive summarization) και την αφαιρετική περίληψη (abstractive summarization). Οι συγκεκριμένες κατηγορίες μπορούν να υλοποιηθούν με χρήση διαφορετικών τεχνικών, τις οποίες και θα παρουσιάσουμε παρακάτω.

Το παρακάτω διάγραμμα παρουσιάζει σχηματικά τις διαφορετικές κατηγοριοποιήσεις των αυτόματων περιλήψεων. Τα μοντέλα αυτόματης περίληψης κειμένων, χαρακτηρίζονται από το αν δέχονται ως είσοδο ένα ή πολλαπλά έγγραφα. Με βάση το σκοπό τους, δηλαδή αν πρόκειται να χρησιμοποιηθούν για γενική, επικεντρωμένη σε έναν τομέα, ή βασισμένη σε συγκεκριμένο ερώτημα περίληψη. Τέλος, με βάση το αποτέλεσμα τους και κατ’επέκταση τη μέθοδο με την οποία θα παραχθεί αυτό, σε μοντέλα αποσπασματικής ή αφαιρετικής περίληψης.



Εικόνα 1 Κατηγοριοποιήσεις αυτομάτων περιλήψεων (Chauhan, 2018)

2.4.1. Αποσπασματική περίληψη

Η αποσπασματική περίληψη κειμένου είναι η πρώτη μέθοδος αυτόματης περίληψης κειμένων που αναπτύχθηκε. Η παραγόμενη περίληψη δημιουργείται με την επιλογή αυτούσιων προτάσεων που βρίσκονται στο αρχικό κείμενο. Κατά τη δημιουργία της περίληψης, επιλέγονται οι πιο σημαντικές προτάσεις του κειμένου, προκειμένου το αποτέλεσμα να αποτυπώνει τα κύρια σημεία του πρωτότυπου.

Η πρώτη τεχνική υλοποίησης αφαιρετικής περίληψης κειμένου έγινε στα τέλη της δεκαετίας του 1950 από το Luhn. Για την εύρεση των σημαντικών προτάσεων ενός κειμένου, οι οποίες αποτελούν και την περίληψη του, πρότεινε τη χρήση της συχνότητας εμφάνισης λέξεων για την εύρεση εκείνων των λέξεων μέσα σε ένα κείμενο οι οποίες μεταφέρουν τις πιο σημαντικές πληροφορίες του (Nenkova & McKeown, 2011) και την επιλογή εκείνων των προτάσεων που έχουν μεγάλη πυκνότητα τέτοιων λέξεων.

Οι αλγόριθμοι υλοποιούνται με διάφορες τεχνικές, όπως ο υπολογισμός συχνότητας λέξεων, αλγόριθμοι ομαδοποίησης ή μηχανικής μάθησης, οι οποίες στοχεύουν στην αξιολόγηση των προτάσεων και την επιλογή των σημαντικότερων από αυτές. Για τις τεχνικές αυτές, το σημαντικό ερώτημα είναι το ποιες είναι οι σημαντικές προτάσεις ενός κειμένου (Nenkova & McKeown, 2011).

Όπως είδαμε παραπάνω, η φιλοσοφία της αποσπασματικής περίληψης βασίζεται στην αναγνώριση των σημαντικών προτάσεων που περιέχονται σε ένα κείμενο, οι οποίες συνιστούν μια επισκόπηση του κειμένου. Για το σκοπό αυτό, οι σύγχρονες τεχνικές, χρησιμοποιούν μεθόδους πολύ πιο εξελιγμένες από την αρχική πρόταση του Luhn, όπως

στατιστικές μεθόδους προκειμένου να μη χρειάζονται αυθαίρετους κανόνες για να εντοπίσουν τις αντιπροσωπευτικές λέξεις του κειμένου (Nenkova & McKeown, 2011). Θα μπορούσαμε να κατηγοριοποιήσουμε τις μεθόδους της αποσπασματικής περίληψης ως εξής:

Η πρώτη κατηγορία αφορά την παραγωγή των περιλήψεων με επιλογή των σημαντικών προτάσεων ενός κειμένου. Ως κριτήρια για την εύρεση των σημαντικών προτάσεων μπορούν να χρησιμοποιηθεί η συχνότητα εμφάνισης των λέξεων, η σημασιολογική τους συνάφεια ή η εμφάνιση μέσα σε αυτές λέξεων κλειδιών.

Ένα τέτοιο παράδειγμα είναι το SumBasic (Nenkova A. &., 2005), το οποίο βασίζεται αποκλειστικά στη συχνότητα εμφάνισης συγκεκριμένων λέξεων προκειμένου να καθορίσει τις σημαντικές προτάσεις που πρόκειται να συμπεριληφθούν στην περίληψη, βασιζόμενο στην υπόθεση ότι όσο πιο συχνά εμφανίζεται μια λέξη μέσα σε ένα κείμενο, τόσο πιο πιθανό είναι να εμφανιστεί και στην περίληψη του.

Μια άλλη προσέγγιση, επικεντρώνεται στις θεματικές που εμφανίζονται σε ένα κείμενο και επιλέγει προτάσεις οι οποίες είναι αντιπροσωπευτικές των θεματικών αυτών. Ένα παράδειγμα τέτοιας προσέγγισης είναι το TSsum (Lin, Chin-Yew & Hovy, 2000). Δημιουργούνται θεματικές ενότητες οι οποίες περιλαμβάνουν λέξεις-κλειδιά προκειμένου να εντοπίσουν πιο πολύπλοκες έννοιες οι οποίες μπορεί και να μην περιγράφονται στο αρχικό κείμενο, και στη συνέχεια, με βάση αυτές αξιολογούνται οι προτάσεις που θα συμμετέχουν στην περίληψη.

Οι παραπάνω προσεγγίσεις χρησιμοποιούν στατιστικές μεθόδους για να εντοπίσουν τις σημαντικές προτάσεις ενός κειμένου. Είναι σχετικά εύκολες στην εφαρμογή τους και δεν απαιτούν εκπαίδευση με συγκεκριμένα δεδομένα. Όμως, ενδέχεται να μη συμπεριλάβουν σημαντικές πληροφορίες οι οποίες είναι διατυπωμένες με διαφορετικό τρόπο, και άρα στατιστικά μη σημαντικές. Επίσης δεν λαμβάνουν υπόψη τις σχέσεις μεταξύ των προτάσεων στο σύνολο του κειμένου αφού επεξεργάζονται τις προτάσεις ανεξάρτητα.

Η δεύτερη κατηγορία βασίζεται σε γράφους για τη σύνδεση των προτάσεων (Sentence connection graphs). Τεχνικές όπως οι αλγόριθμοι LexRank και TextRank (Mihalcea & Tarau, 2004) χρησιμοποιούν γράφους για να εντοπίσουν τη συνάφεια μεταξύ των προτάσεων και να επιλέξουν τις πιο σημαντικές για την περίληψη. Οι γράφοι αναπαριστούν τις προτάσεις του κειμένου οι οποίες στη συνέχεια αξιολογούνται ως προς τη σημασία τους με βάση τη δομή του γράφου. Και οι δύο χρησιμοποιούν τον αλγόριθμο PageRank της

Google για την αξιολόγηση της σημασίας των προτάσεων. Η βασική τους διαφορά είναι στον υπολογισμό της ομοιότητας των προτάσεων, όπου ο LexRank χρησιμοποιεί την ομοιότητα συνημιτόνου αποκλειστικά, ενώ ο TextRank χρησιμοποιεί και άλλα μέτρα ομοιότητας. Ο υπολογισμός της ομοιότητας μεταξύ προτάσεων μέσα στο κείμενο χρησιμεύει για τη στάθμιση των προτάσεων προκειμένου στη συνέχεια να επιλεγούν οι πιο αντιπροσωπευτικές προτάσεις του κειμένου.

Η χρήση γράφων λαμβάνει υπόψη τη συνολική δομή του κειμένου, αφού ολόκληρο το κείμενο αποτυπώνεται στο γράφο πριν την αξιολόγηση των προτάσεων. Δεν απαιτούν εκπαίδευση με συγκεκριμένα δεδομένα, το οποίο τις κάνει ευέλικτες και εύκολες στη χρήση. Με δεδομένο όμως ότι, όπως και οι μέθοδοι της πρώτης κατηγορίας δεν περιλαμβάνουν κάποιου είδους σημασιολογική ανάλυση, επίσης ενδέχεται να μην έχουν ένα συνεκτικό αποτέλεσμα.

Μια τρίτη κατηγορία περιλαμβάνει προσεγγίσεις που βασίζονται σε μαθηματικές μεθόδους όπως η παραγοντοποίηση πινάκων. Τέτοια προσέγγιση είναι η Λανθάνουσα Σημασιολογική Ανάλυση (LSA, Latent Scemantics Analysis) (Liu, Xin & Gong, Yihong, 2001) . Για να αναγνωριστούν οι πιο σημαντικές έννοιες ενός κειμένου, δημιουργείται ένας πίνακας όπου οι γραμμές αναπαριστούν όρους και οι στήλες αναπαριστούν προτάσεις. Στη συνέχεια, μειώνονται με μαθηματικές μεθόδους οι διαστάσεις του πίνακα ώστε να μείνουν μόνο οι βασικές έννοιες και οι προτάσεις που τις περιέχουν και τέλος επιλέγονται από αυτές οι πλέον αντιπροσωπευτικές προτάσεις, οι οποίες και σχηματίζουν την περίληψη.

Πλεονεκτική σε σχέση με τις προηγούμενες κατηγορίες είναι η δυνατότητα αυτών των μεθόδων να λαμβάνουν υπόψη τις σχέσεις μεταξύ των λέξεων και των προτάσεων. Έτσι μπορούν να αναγνωριστούν καλύτερα οι έννοιες που περιέχει το έγγραφο και να παράγουν πιο συνεκτικές περιλήψεις. Όμως για την εφαρμογή τους απαιτούνται πολλοί πόροι και ενδέχεται να μην συμπεριλάβουν σημαντικές πληροφορίες οι οποίες όμως δεν είναι αντιπροσωπευτικές των εννοιών που έχουν αναγνωριστεί.

Πρόσφατα έχουν αναπτυχθεί τεχνικές αποσπασματικής περίληψης που βασίζονται στη Μηχανική και τη Βαθιά μάθηση Τα μοντέλα αυτά εκπαιδεύονται να αναγνωρίζουν τα σημαντικά χαρακτηριστικά ενός κειμένου με μια συλλογή κειμένων και των περιλήψεων τους, και να παράγουν περιλήψεις με βάση αυτά (Liu, 2019). Ένα παράδειγμα είναι η εκπαίδευση του BERT για την αποσπασματική περίληψη διαλέξεων (Miller, 2019). Οι τεχνικές αυτές έχουν τη δυνατότητα να διαχειρίζονται πιο πολύπλοκες σχέσεις μεταξύ των

δεδομένων, και να παράγουν πιο συνεκτικές περιλήψεις. Όμως, το αποτέλεσμα εξαρτάται σε μεγάλο βαθμό από τα δεδομένα με τα οποία εκπαιδεύονται.

2.4.2. Αφαιρετική περίληψη

Η αφαιρετική περίληψη (abstractive summarization) αναπτύχθηκε πολύ αργότερα από την αποσπασματική. Ο βασικός λόγος για αυτό είναι ότι για τη δημιουργία αφαιρετικής περίληψης απαιτείται η δυνατότητα κατανόησης της φυσικής γλώσσας, αλλά και η δυνατότητα παραγωγής κειμένου.

Το τελευταίο διάστημα, με την ανάπτυξη της τεχνολογίας των νευρωνικών δικτύων, η βαθιά μάθηση έχει γίνει μία από τις πιο αποτελεσματικές μεθόδους στην επεξεργασία φυσικής γλώσσας. Το 2015, οι Rush et al. εφάρμοσαν τεχνολογίες βαθιάς μάθησης, για να δημιουργήσουν ένα μοντέλο για τη δημιουργία αυτόματης περίληψης το οποίο βασίζεται σε αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή. (Rush, 2015).

Η περίληψη που παράγεται με αυτόν τον τρόπο, δε δημιουργείται με επιλογή προτάσεων από το κείμενο, αλλά προκύπτει από μια δημιουργική διαδικασία κατά την οποία δημιουργούνται νέες προτάσεις οι οποίες μεταδίδουν το περιεχόμενο του πρωτότυπου κειμένου.

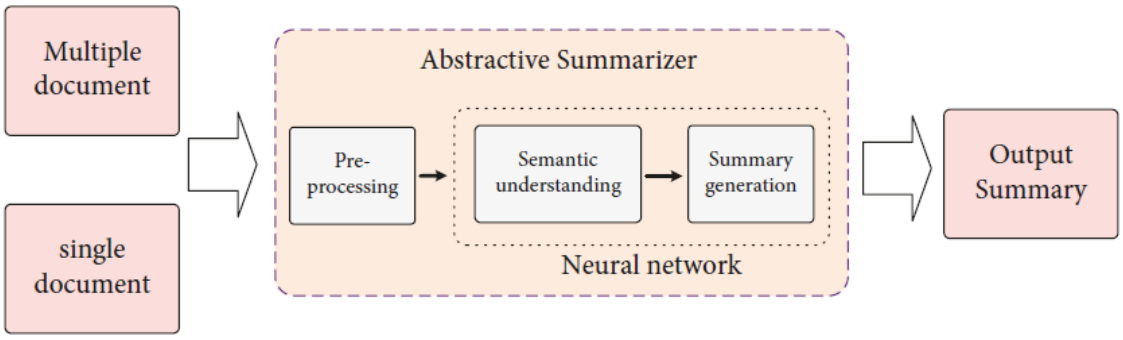
Για το σκοπό αυτό απαιτείται πρώτα να γίνει κατανοητή η σημασία του κειμένου και στη συνέχεια να εφαρμοστεί κάποια τεχνική παραγωγής φυσικής γλώσσας ή οποία και θα παράξει μια συνεκτική περίληψη, χρησιμοποιώντας παράφραση, συνώνυμα, συμπίεση προτάσεων, κλπ (Zangh, M., Gang, Wanting, Ningbo, & Wenfen, 2022).

Το αποτέλεσμα προσομοιάζει περισσότερο αυτό της δημιουργίας περίληψης από άνθρωπο (Nenkova & McKeown, 2011). Έτσι, τα μοντέλα αφαιρετικής περίληψης κειμένων παράγουν ένα αποτέλεσμα το οποίο είναι εύκολα αναγνώσιμο και γραμματικά σωστό (Banerjee, S, Mitra, P, & Sugiyama, K, 2016), χαρακτηριστικά τα οποία είναι και κριτήρια αξιολόγησης των περιλήψεων.

Παρ' όλα αυτά, τα τελευταία χρόνια έχουν γίνει σημαντικές πρόοδοι σε αυτόν τον τομέα. Μοντέλα όπως το PEGASUS (Zhang et al., 2019) και το BART (Lewis, et al., 2019) χρησιμοποιούν τεχνικές μεταφοράς μάθησης για να επιτύχουν αυτή την κατανόηση του κειμένου και να παράγουν περιλήψεις. Ωστόσο, παρόλο που αυτά τα μοντέλα επιτυγχάνουν

αξιόλογα αποτελέσματα, παραμένουν σημαντικές προκλήσεις. Η αφαιρετική περίληψη εξακολουθεί να είναι ένα ανοιχτό πεδίο έρευνας με πολλές δυνατότητες για περαιτέρω βελτιώσεις και προσαρμογές (See, Liu, & Manning, 2017).

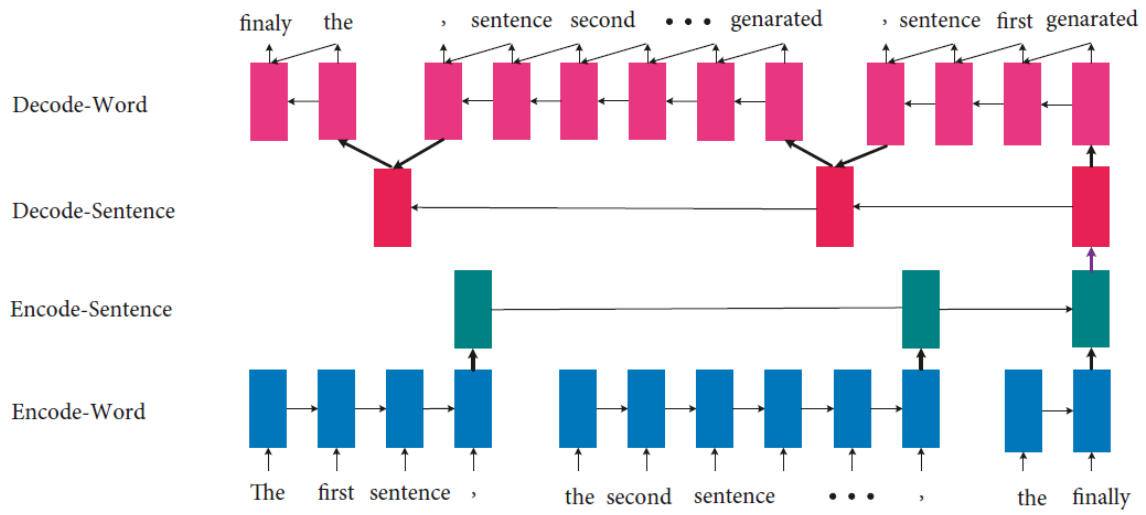
Όπως είδαμε, σκοπός της αφαιρετικής περίληψης είναι η κατανόηση του αρχικού κειμένου και η δημιουργία ενός νέου. Αυτό επιτυγχάνεται με ένα ευρύ φάσμα τεχνικών, οι οποίες παρουσιάζουν συνεχή εξέλιξη. Αρκετές από αυτές βασίζονται στα νευρωνικά δίκτυα, και συνήθως ακολουθούν τη γενική αρχιτεκτονική (Zangh, M., Gang, Wanting, Ningbo, & Wenfen, 2022) που παρουσιάζεται στην παρακάτω εικόνα.



Εικόνα 2 Αφαιρετική περίληψη κειμένων με νευρωνικά δίκτυα (Zangh, M., Gang, Wanting, Ningbo, & Wenfen, 2022)

Τα συστήματα αυτά μπορούν να δέχονται ως είσοδο ένα ή πολλαπλά έγγραφα. Το πρώτο στάδιο είναι προεπεξεργασία με στόχο την κανονικοποίηση του εγγράφου, χρησιμοποιώντας τεχνικές όπως αυτές που περιγράψαμε σε προηγούμενη παράγραφο, δηλαδή tokenization, αφαίρεση λέξεων μικρής σημασίας, κλπ. Ακολουθεί η δημιουργία ενός νευρωνικού δικτύου, το οποίο αναλαμβάνει την κατανόηση του κειμένου. Το αποτέλεσμα αυτού του σταδίου χρησιμοποιείται στο τρίτο στάδιο, στο οποίο το νευρωνικό δίκτυο παράγει την περίληψη του κειμένου.

Τεχνικές αυτής της κατηγορίας είναι τα μοντέλα κωδικοποιητή-αποκωδικοποιητή (encoder-decoder), επίσης γνωστά ως μοντέλα ακολουθίας σε ακολουθία (Seq2Seq). Τα μοντέλα αυτά χρησιμοποιούν τον κωδικοποιητή για να αναγνωρίσουν και να κατανοήσουν το κείμενο, και στη συνέχεια ο αποκωδικοποιητής αναλαμβάνει την δημιουργία της περίληψης. Στα μοντέλα αυτά μπορούν να εφαρμόζονται μηχανισμοί προσοχής (Attention mechanisms) για την καλύτερη ενσωμάτωση της πληροφορίας του κειμένου και την αποφυγή επαναλήψεων στην τελική περίληψη (Zangh, M., Gang, Wanting, Ningbo, & Wenfen, 2022).



Εικόνα 3 Μοντέλο κωδικοποιητή - αποκωδικοποιητή (Zangh, M., Gang, Wanting, Ningbo, & Wenfen, 2022).

Αυτή τη φιλοσοφία ακολουθεί το μοντέλο που πρότειναν οι (Chopra, S., Auli, M, & Rush, A., 2016), στο οποίο εφαρμόζουν ένα συνελικτικό μηχανισμό προσοχής προκειμένου να διασφαλίσουν ότι ο αποκωδικοποιητής εναρμονίζεται με την αντίστοιχη είσοδο σε κάθε βήμα αποκωδικοποίησης ώστε να υπάρχει μια διαδικασία προσαρμογής του αποτελέσματος.

Παρόλη την καλύτερη απόδοση σε σχέση με προγενέστερες υλοποιήσεις, τα μοντέλα αυτά αντιμετωπίζουν προβλήματα όπως η επανάληψη φράσεων στην περίληψη, η αναπαραγωγή ανακριβών πληροφοριών (factual errors) και η αδυναμία χειρισμού λέξεων που δε βρίσκονται στο λεξιλόγιο τους (See, Liu, & Manning, 2017).

Για να αντιμετωπίσουν τα παραπάνω τρία ζητήματα, οι (See, Liu, & Manning, 2017), κυρίως σε πιο μακροσκελή κείμενα, χρησιμοποιούν ένα μοντέλο Seq2Seq με μηχανισμό προσοχής στο οποίο εφαρμόζουν ένα υβριδικό δίκτυο pointer-generator το οποίο, δίνει τη δυνατότητα αντιγραφής λέξεων από το αρχικό κείμενο μέσω του δείκτη (pointer) ενώ ταυτόχρονα διατηρεί τη δυνατότητα παραγωγής (generation) νέων λέξεων. Ουσιαστικά, εισάγουν στο μοντέλο τους λειτουργίες αφαιρετικής περίληψης με την αντιγραφή συγκεκριμένων λέξεων από το αρχικό κείμενο. Έτσι, αντιμετωπίζουν προβλήματα όπως η διαχείριση των άγνωστων λέξεων (out of vocabulary words) και μειώνουν την παραγωγή ανακριβών πληροφοριών. Παρόλο που το πρόβλημα των μη συνεκτικών ή ανακριβών περιλήψεων βελτιώνεται σημαντικά, συνεχίζει να παραμένει.

Το 2017,προτάθηκε ένα νέο μοντέλο κωδικοποιητή αποκωδικοποιητή, οι μετασχηματιστές (Transformers) το οποίο εκμεταλλεύεται πολλαπλούς μηχανισμούς προσοχής για να εκτελέσει διάφορες εργασίες που ανήκουν στην επεξεργασία φυσικής γλώσσας (Vaswani, και συν., 2017). Με τη βοήθεια των μηχανισμών προσοχής, το Transformer μπορεί να μοντελοποιήσει πολύπλοκες συσχετίσεις μεταξύ των λέξεων και να παράγει πολύ ακριβείς προβλέψεις. Τα μοντέλα PEGASUS (Zhang et al., 2019) και BART (Lewis, et al., 2019) που αναφέραμε παραπάνω βασίζονται σε transformers και χρησιμοποιούν μηχανισμούς προσοχής για να κατανοήσουν το κείμενο και να δημιουργήσουν περιλήψεις. Αυτά τα μοντέλα προεκπαιδεύονται σε μεγάλα σύνολα δεδομένων (datasets), προκειμένου να δημιουργούν πιο ποιοτικές περιλήψεις.

Το μοντέλο PEGASUS (Zhang et al., 2019), διαφοροποιείται από άλλα μοντέλα βασισμένα σε transformers στη διαδικασία εκπαίδευσης του. Οι δημιουργοί του έκαναν την υπόθεση ότι αν εκπαιδεύσουν το μοντέλο με τρόπο που να προσομοιάζει καλύτερα το αποτέλεσμα το οποίο θέλουν να παράξουν στη συνέχεια, αυτό θα έχει καλύτερα αποτελέσματα. Έτσι, η εκπαίδευση βασίστηκε στην παραγωγή κειμένου από ένα αρχικό κείμενο. Για το λόγο αυτό αφαίρεσαν τις σημαντικές προτάσεις από το αρχικό κείμενο και τις συμπεριέλαβαν σε μια ψευδο-περίληψη. Ο στόχος της εκπαίδευσης ήταν το μοντέλο να μπορέσει να δημιουργήσει τις προτάσεις που λείπουν κατανοώντας το υπόλοιπο αρχικό κείμενο.

Η διαφοροποίηση στην αρχική εκπαίδευση του PEGASUS, η οποία είναι στοχευμένη στη δημιουργία περιλήψεων και όχι γενική για εργασίες NLP, το καθιστά πιο εξειδικευμένο στο αντικείμενο. Ταυτόχρονα όμως, ενδέχεται οι περιλήψεις που παράγει να περιέχουν αυτούσιες προτάσεις από το αρχικό κείμενο.

Το μοντέλο BART (Lewis, et al., 2019) ακολουθεί την ίδια αρχιτεκτονική, αλλά διαφορετική διαδικασία εκπαίδευσης από το PEGASUS. Εδώ, τα αρχικά κείμενα τροποποιούνται όχι με αφαίρεση προτάσεων, αλλά με αλλοίωση του περιεχομένου τους με αυθαίρετη προσθήκη θορύβου και το μοντέλο μαθαίνει να αναδομεί το αρχικό κείμενο. Το BART είναι ένα μοντέλο το οποίο μπορεί να βρει εφαρμογή σε εργασίες NLP που απαιτούν κατανόηση και παραγωγή κειμένου, μεταξύ των οποίων η δημιουργία αφαιρετικών περιλήψεων. Η εκπαίδευση του στην αφαίρεση των περιττών στοιχείων ενός κειμένου (θόρυβος) του δίνει τη δυνατότητα ποικιλόμορφων αναπαραστάσεων των δεδομένων, και άρα δημιουργίας πιο αφαιρετικών περιλήψεων. Όμως, πολλές φορές παράγει περιλήψεις οι οποίες περιέχουν μεγάλο μέρος νέου κειμένου, το οποίο μπορεί να είναι εκτός θέματος.

Ένας περιορισμός των Transformers είναι η αδυναμία τους να επεξεργαστούν πολύ μεγάλα κείμενα. Αυτό οφείλεται στο ότι η εφαρμογή των μηχανισμών προσοχής απαιτεί την επεξεργασία ολόκληρου του εγγράφου ταυτόχρονα. έχει λοιπόν παρατηρηθεί ότι όσο προχωράει η δημιουργία της περίληψης, η ποιότητα του αποτελέσματος υποβαθμίζεται (Shearing, Gertner, & Wellner, 2020). Για τη διαχείριση του περιορισμού στο μέγεθος του κειμένου, τα προ-εκπαιδευμένα μοντέλα συνήθως ορίζουν ένα μέγιστο όριο για το προς περίληψη κείμενο.

2.4.3. Εκπαίδευση και εξειδίκευση

Για την αποτελεσματική λειτουργία ενός μοντέλου επεξεργασίας φυσικής γλώσσας άρα και των μοντέλων για τη δημιουργία περιλήψεων, είναι σημαντική η εκπαίδευση του. Η εκπαίδευση είναι η διαδικασία κατά την οποία το μοντέλο τροφοδοτείται με μεγάλα σύνολα δεδομένων (dataset), κατά την επεξεργασία των οποίων «μαθαίνει» να εκτελεί συγκεκριμένες εργασίες, χωρίς να απαιτείται εξαντλητικός προγραμματισμός για την πρόβλεψη κάθε πιθανής περίπτωσης.

Οι τεχνικές εκπαίδευσης περιλαμβάνουν την επιβλεπόμενη μάθηση (supervised learning), τη μη επιβλεπόμενη μάθηση (unsupervised learning) και την ενισχυτική μάθηση (reinforcement learning) (Sharkar, 2019).

Τα σύνολα δεδομένων (datasets) συνήθως περιέχουν τα αρχικά κείμενα και το αποτέλεσμα τους. Για παράδειγμα, όταν στόχος είναι η δημιουργία αφαιρετικής περίληψης το dataset θα περιέχει το αρχικό κείμενο και την πρότυπη περίληψη του. Τα σύνολα αυτά χωρίζονται σε ομάδες και στη συνέχεια τροφοδοτούνται στο μοντέλο. Με την επεξεργασία τους, το μοντέλο μαθαίνει να παράγει το επιθυμητό αποτέλεσμα.

Για παράδειγμα, τα μοντέλα PEGASUS και BART έχουν το καθένα ακολουθήσει διαφορετικό τρόπο εκπαίδευσης. Στην περίπτωση του PEGASUS, έχει επιλεγεί ένα σύνολο δεδομένων στο οποίο έχουν αφαιρεθεί οι σημαντικές προτάσεις του αρχικού κειμένου, και έχουν συμπεριληφθεί στην πρότυπη περίληψη του. Στόχος της εκπαίδευσης είναι το μοντέλο να μπορεί να δημιουργήσει τις προτάσεις αυτές με βάση το υπόλοιπο κείμενο (Zhang, Zhao, Saleh, & Liu, 2016).

Το BART έχει εκπαιδευτεί ώστε να μπορεί να εκτελεί διάφορες εργασίες επεξεργασίας φυσικής γλώσσας, σε ένα σύνολο δεδομένων, στο οποίο τα αρχικά κείμενα έχουν αλλοιωθεί

με διάφορες τεχνικές. Στόχος της εκπαίδευσης του είναι το μοντέλο να μπορεί να αποκαταστήσει αυτή την αλλοίωση (Lewis, και συν., 2019).

Το σημαντικότερο ίσως συστατικό για την αποτελεσματική εκπαίδευση ενός μοντέλου είναι το dataset που θα χρησιμοποιηθεί για την εκπαίδευση του, καθώς στα περιεχόμενα του θα βασιστεί η μετέπειτα λειτουργία του. Τα dataset αξιολογούνται με βάση χαρακτηριστικά όπως η εξειδίκευση τους με την εργασία που πρέπει να επιτελεστεί, το μέγεθος, η ποικιλομορφία τους, η αντικειμενικότητα τους κλπ. Για παράδειγμα, το dataset MultiNews (Fabbri, Li, She, Li, & Radev, 2019) έχει δημιουργηθεί με στόχο την περίληψη πολλαπλών κειμένων, οπότε οι πρότυπες περιλήψεις που περιέχει είναι προσαρμοσμένες σε αυτό το σκοπό.

Σε κάθε περίπτωση, αλλά περισσότερο όταν η αρχική εκπαίδευση του μοντέλου δεν είναι απολύτως προσανατολισμένη σε συγκεκριμένη εργασία, απαιτείται να ακολουθηθεί ένας δεύτερος, μικρότερος κύκλος εκπαίδευσης που αφορά στην εξειδίκευση (fine-tuning) του μοντέλου με τη χρήση κάποιου πιο εξειδικευμένου dataset (Zhang, Zhao, Saleh, & Liu, 2016). Για παράδειγμα, για την περίληψη ειδησεογραφικών άρθρων, θα επιλέξουμε ένα dataset για την εξειδίκευση το οποίο ιδανικά θα έχει δημιουργηθεί για αυτό το σκοπό.

2.4.4. Πλεονεκτήματα και μειονεκτήματα των μεθόδων

Στις προηγούμενες παραγράφους, περιγράψαμε την αφαιρετική και την αποσπασματική αυτόματη περίληψη και παρουσιάσαμε κάποιες από τις τεχνικές που χρησιμοποιούνται στην εφαρμογή τους. Η επιλογή μιας από τις μεθόδους εξαρτάται από τα χαρακτηριστικά της κάθε υλοποίησης. Σε αυτή την παράγραφο θα προσπαθήσουμε να συνοψίσουμε τα πλεονεκτήματα και τα μειονεκτήματα κάθε μιας από αυτές.

Κατά την αφαιρετική περίληψη ουσιαστικά δημιουργείται νέο κείμενο. Αυτό μπορεί να βελτιώσει τη συνοχή και την κατανόηση της παραγόμενης περίληψης, όμως μπορεί να οδηγήσει σε απώλεια πληροφορίας λόγω της αναδιατύπωσης. Για τον ίδιο λόγο, τα μοντέλα αφαιρετικής περίληψης απαιτούν κατά κανόνα πιο εκτεταμένη εκπαίδευση.

Κατά την αποσπασματική περίληψη, το αποτέλεσμα παράγεται με επιλογή προτάσεων από το αρχικό κείμενο. Αυτό έχει ως αποτέλεσμα την μείωση του κινδύνου απώλειας

πληροφορίας, όμως έχει επίπτωση στη συνοχή του κειμένου. Τα μοντέλα εξαγωγικής περίληψης είναι κατά κανόνα πιο απλά και ταχύτερα από τα αφαιρετικά.

2.4.5. Περίληψη πολλαπλών κειμένων

Στις προηγούμενες ενότητες, είδαμε τις βασικές τεχνικές των δύο μεθόδων της αυτόματης περίληψης. Οι δύο μέθοδοι και οι τεχνικές τους αρχικά εφαρμόστηκαν στην περίληψη ενός μοναδικού κειμένου. Η ανάγκη για περίληψη πολλαπλών κειμένων προέκυψε μεταγενέστερα (Nenkova & McKeown, 2011), λόγω της πληθώρας κειμένων στο διαδίκτυο, οπότε και αποδείχθηκε πιο χρήσιμη η δημιουργία μιας περίληψης πολλών κειμένων που αναφέρονται στο ίδιο θέμα. Έχουν πλέον αναπτυχθεί εφαρμογές οι οποίες εκμεταλλεύονται διάφορες τεχνικές αυτόματης περίληψης και γενικότερα επεξεργασίας φυσικής γλώσσας, τις οποίες συνδυάζουν για να παράξουν συνεκτικές περιλήψεις από πολλαπλά κείμενα.

Η μεγαλύτερη πρόκληση κατά την περίληψη πολλαπλών κειμένων, είναι ότι πρέπει να πραγματοποιηθεί επεξεργασία ολόκληρης της συλλογής και στη συνέχεια να δημιουργηθεί μία περίληψη, αντιπροσωπευτική ολόκληρης της συλλογής (Shearing, Gertner, & Wellner, 2020). Αυτό αναλύεται από τη μία στο πως θα εντοπίσουμε τα σημαντικά νοήματα των κειμένων, και από την άλλη με ποιον τρόπο και ποια από αυτά θα αποτυπωθούν σε μία συνεκτική περίληψη.

Μια προσέγγιση σε αυτό είναι μια διαδικασία δύο σταδίων, η οποία μειώνει με κάποιον τρόπο το περιεχόμενο κάθε κειμένου ξεχωριστά, στη συνέχεια συγχωνεύει τα κείμενα και εξάγει την περίληψη του κειμένου που προκύπτει από τη συγχώνευση.

Την παραπάνω προσέγγιση ακολουθεί το μοντέλο Multinews των Fabri (Fabbri, Li, She, Li, & Radev, 2019) , οι οποίοι δημιούργησαν ένα μοντέλο για την περίληψη πολλαπλών κειμένων και μια συλλογή κειμένων και περιλήψεων αναφοράς (dataset) για την εκπαίδευση και ρύθμιση του μοντέλου αυτού.

Για να μειώσουν το περιεχόμενο των άρθρων, κρατούν τις πρώτες προτάσεις κάθε άρθρου, κάνοντας την παραδοχή ότι αυτές είναι αντιπροσωπευτικές για το σύνολο του περιεχομένου του. Στη συνέχεια συνενώνουν τα παραγόμενα και δημιουργούν ένα κείμενο. Τέλος, εφαρμόζουν στο τελικό κείμενο αφαιρετική περίληψη (Fabbri, Li, She, Li, & Radev, 2019).

Το dataset που συνοδεύει το μοντέλο, είναι ένα από τα λίγα διαθέσιμα dataset για την εκπαίδευση μοντέλων στην περίληψη πολλαπλών κειμένων (Shearing, Gertner, & Wellner, 2020).

Αυτή η προσέγγιση από τη μία μειώνει αποτελεσματικά το μέγεθος των προς περίληψη κειμένων προκειμένου να μπορέσει στη συνέχεια να δημιουργηθεί μία ενιαία αφαιρετική περίληψη. Όμως, η επιλογή της διατήρησης των πρώτων προτάσεων κάθε άρθρου μπορεί να οδηγήσει σε απώλεια σημαντικών πληροφοριών που αναφέρονται στο υπόλοιπο άρθρο.

Μια δεύτερη προσέγγιση είναι η εφαρμογή κάποιας μορφής ομαδοποίησης στα προς περίληψη κείμενα και στη συνέχεια η περίληψη τους.

Αυτή ακολούθησαν οι Banerjee, S, Mitra, P, & Sugiyama, K (Banerjee, S, Mitra, P, & Sugiyama, K, 2016) και δημιούργησαν ένα μοντέλο αφαιρετικής περίληψης, το οποίο πριν προχωρήσει σε περίληψη των κειμένων τα ομαδοποιεί με βάση τις προτάσεις του πιο σημαντικού κειμένου της συλλογής. Η λογική είναι να συγκεντρωθούν ομάδες προτάσεων με παρεμφερές νόημα και στη συνέχεια να δημιουργηθεί μία πρόταση για κάθε ομάδα. Τα αποτελέσματα είναι σε μεγάλο βαθμό αποδεκτά, όμως έχουν δεν έχουν καλή γλωσσική ποιότητα στο παραγόμενο κείμενο, όπως αυτή αξιολογήθηκε από ειδικούς (στο ίδιο).

Μια άλλη προσέγγιση είναι ο συνδυασμός συστημάτων για την αυτόματη περίληψη, με δημιουργία υποψήφιων περιλήψεων, οι οποίες συνδυάζουν περιλήψεις που παράγονται από διαφορετικά συστήματα, προκειμένου να βελτιώσουν το περιεχόμενο τους (Hong, Marcus, & Nenkova, 2015).

Σε κάθε περίπτωση, οι προκλήσεις που πρέπει να αντιμετωπιστούν παραμένουν κοινές στις προσεγγίσεις:

- Η πρώτη πρόκληση είναι να εντοπιστούν τα σημαντικότερα νοήματα και πληροφορίες κάθε κειμένου. Αυτά μπορεί να είναι υπερβολικά ετερογενή ή πολύ αλληλεπικαλυπτόμενα όταν συγκεντρωθούν όλα μαζί.
- Η δεύτερη είναι κατά το δυνατό να συγχωνευτούν ώστε να μειωθεί το μέγεθος του τελικού κειμένου προς περίληψη.
- Η τρίτη είναι η ίδια η δημιουργία της περίληψης, και το πως αυτή θα παρουσιάζει τις σημαντικές πληροφορίες των αρχικών κειμένων σε μια συνεκτική και κατανοητή περίληψη.

2.4.6. Ομοιότητα κειμένων

Τα μέτρα ομοιότητας κειμένων βρίσκουν εφαρμογή στην ανάλυση και ομαδοποίηση κειμένων με βάση την ομοιότητα. Η ομοιότητα κειμένων η διεργασία της χρήσης μιας απόστασης ή μετρικής που βασίζεται στην ομοιότητα που μπορεί να αναγνωρίσει πόσο όμοιο είναι ένα έγγραφο με ένα ή πολλά άλλα (Sharkar, 2019). Η ομοιότητα κειμένων μπορεί να είναι σε επίπεδο λεξιλογικό ή σημασιολογικό. Το πρώτο ασχολείται με τις λέξεις και τη σύνταξη ενός κειμένου, ενώ το δεύτερο με το πόσο μοιάζουν δύο κείμενα σε επίπεδο νοήματος. Ο υπολογισμός της ομοιότητας δύο κειμένων, μπορεί να πραγματοποιηθεί, είτε με απλές μαθηματικές τεχνικές όπως η ομοιότητα συνημίτονου ή η Ευκλείδια απόσταση, οι οποίες επιστρέφουν ένα αριθμητικό μέτρο ομοιότητας συγκρίνοντας τα χαρακτηριστικά των δύο κειμένων, ή με πιο πολύπλοκες τεχνικές που βασίζονται στα νευρωνικά δίκτυα, ή σε μηχανική μάθηση. (Gahman & Vinayak, 2023).

Οι μαθηματικές τεχνικές είναι εύκολες στον υπολογισμό, όμως δεν λαμβάνουν υπόψη τη σημασιολογική ομοιότητα. Οι τεχνικές που βασίζονται σε νευρωνικά δίκτυα, επιτυγχάνουν πολύ καλά αποτελέσματα και στα δύο επίπεδα, όμως απαιτούν πολλούς πόρους. Οι τεχνικές υπολογισμού της ομοιότητας που βασίζονται στη μάθηση μπορούν να εντοπίσουν τη σημασιολογική ομοιότητα, όμως βασίζονται πολύ στα δεδομένα στα οποία εκπαιδεύτηκαν. (Gahman & Vinayak, 2023).

Όπως αναφέραμε παραπάνω, οι μαθηματικές τεχνικές βασίζονται στον υπολογισμό της απόστασης μεταξύ των δύο κειμένων. Για να είναι αυτό εφικτό θα πρέπει να εφαρμοστούν στάδια προεπεξεργασίας στα κείμενα (Gahman & Vinayak, 2023), και στη συνέχεια αυτά να μετατραπούν στη διανυσματική αναπαράστασή τους, προκειμένου να είναι εφικτός ο υπολογισμός της απόστασης (Sharkar, 2019).

Παρακάτω παραθέτουμε τον τρόπο υπολογισμού μερικών από τις μετρικές ομοιότητας, της Απόστασης Hamming, της Απόστασης Manhattan, και της απόστασης - ομοιότητας συνημίτονου (Sharkar, 2019).

- Απόσταση Hamming

Η απόσταση Hamming μετρά την απόσταση μεταξύ δύο συμβολοσειρών με την παραδοχή ότι αυτές είναι ίσες σε μήκος. Ο μαθηματικός τύπος για τον υπολογισμό της είναι:

$$hd(u, v) = \sum_{i=1}^{n} (u_i \neq v_i)$$

όπου u και v είναι δύο όροι μήκους n.

- Απόσταση Manhattan

Στον υπολογισμό της απόστασης Manhattan μετράμε τον αριθμό των ασυμφωνιών (mismatches) και αφαιρούμε τη διαφορά μεταξύ κάθε ζεύγους χαρακτήρων που βρίσκονται στην ίδια θέση των δύο συμβολοσειρών. Ο μαθηματικός τύπος για τον υπολογισμό της είναι:

$$md(u,v) = \sum_{i=1}^{n} |u_i - v_i|$$

όπου u και v είναι δύο όροι μήκους n.

- Ευκλείδεια απόσταση

Η ευκλείδια απόσταση ορίζεται ως η μικρότερη ευθεία που ενώνει δύο σημεία. Ο μαθηματικός τύπος για τον υπολογισμό της είναι:

$$ed(u,v) = \sum_{i=1}^{n} (u_i - v_i)^2$$

όπου u και v είναι δύο όροι μήκους n.

- Απόσταση - ομοιότητα συνημίτονου

Η απόσταση συνημίτονου βασίζεται στη μέτρηση του συνημίτονου της γωνίας μεταξύ των διανυσματικών αναπαραστάσεων των κειμένων. Έτσι, τα κείμενα με μεγάλη ομοιότητα έχουν τιμή κοντά στο 1, ενώ τα κείμενα με μικρή ομοιότητα έχουν τιμή πιο κοντά στο 0. Ο υπολογισμός ομοιότητας συνημίτονου έχει καλά αποτελέσματα στη σύγκριση μεγάλων και διαφορετικού μήκους κειμένων.

$$cd(u,v) = 1 - \cos(\theta) = 1 - \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2 \sum_{i=1}^{n} v_i^2}}$$

2.5. Αξιολόγηση περιλήψεων και μετρικές

Η αξιολόγηση της ποιότητας μιας περίληψης είναι άλλη μία πρόκληση για την αυτόματη περίληψη των εγγράφων. Μια περίληψη μπορεί να αξιολογηθεί ως προς το κατά πόσο

εξυπηρετεί το σκοπό για τον οποίο δημιουργήθηκε, ή ως προς την ποιότητα του περιεχομένου της.

Στη δεύτερη περίπτωση, αυτό συνήθως γίνεται με τη σύγκριση της περίληψης με μια περίληψη αναφοράς, δηλαδή αυτή που δημιουργεί ένας άνθρωπος. Σκοπός της σύγκρισης είναι να αξιολογηθεί το περιεχόμενο, αλλά και η ποιότητα του κειμένου της περίληψης (Steinberger & Ježek, 2012).

Η αξιολόγηση αυτή μπορεί να πραγματοποιηθεί είτε χειροκίνητα από ανθρώπους, είτε αυτόματα με σύγκριση του περιεχομένου της περίληψης με περίληψης αναφοράς (Nenkova & McKeown, 2011).

Κατά τη χειροκίνητη αξιολόγηση υπολογίζονται δείκτες όπως η κάλυψη των σημαντικών τμημάτων του αρχικού κειμένου από την περίληψη (Precision and Recall), η κάλυψη των «χρήσιμων» τμημάτων του κειμένου από αυτή (Relative Usability), η κάλυψη του περιεχομένου του κειμένου από την περίληψη (DUC Manual Evaluation) και η γλωσσική ποιότητα (Nenkova & McKeown, 2011)

Για την αυτόματη αξιολόγηση, μπορούν να χρησιμοποιηθούν μετρικές βασισμένες στην ακρίβεια και την ανάκληση (F-score), στην ομοιότητα συνημίτονου, ή στην επικάλυψη μονάδων (Steinberger & Ježek, 2012).

Άλλες μετρικές αξιολογούν με βάση το περιεχόμενο, όπως η στατιστική συνεμφάνιση n-gram (N-gram Co-occurrence Statistics – ROUGE). Πρόκειται για ένα σύνολο μετρήσεων που βασίζεται στην ομοιότητα των n-gram. Για τον υπολογισμό τους χρησιμοποιούνται περιλήψεις αναφοράς, οι οποίες συγκρίνονται με τις αυτόματα παραγόμενες (Steinberger & Ježek, 2012).

Όταν πρόκειται για πολλαπλά κείμενα το θέμα της αξιολόγησης γίνεται αρκετά πολύπλοκο, καθώς η απόκτηση ικανού αριθμού περιλήψεων αναφοράς για τον όγκο των κειμένων είναι χρονοβόρα και ακριβή διαδικασία. Οπότε, η αξιολόγηση με τους παραπάνω τρόπους δεν είναι εύκολη στην εφαρμογή.

Για την αντιμετώπιση του προβλήματος της αξιολόγησης περιλήψεων από πολλαπλά κείμενα, οι Wolhalander, Cattan, Ernst, & Dagan (Wolhalander, Cattan, Ernst, & Dagan, 2022) προτείνουν έναν τρόπο αξιολόγησης περιλήψεων πολλαπλών κειμένων κατά τον οποίο αξιολογείται η διασπορά της πληροφορίας της περίληψη στα αρχικά κείμενα, δηλαδή το από πόσα αρχικά κείμενα προέρχονται οι πληροφορίες που υπάρχουν στην τελική περίληψη.

Συνοψίζοντας, οι αυτόματα παραγόμενες περιλήψεις μπορούν να αξιολογηθούν είτε χειροκίνητα είτε αυτόματα, με την εφαρμογή διαφόρων μεθόδων. Οι τεχνικές αυτόματης αξιολόγησης απαιτούν την ύπαρξη περιλήψεων αναφοράς. Για την περίληψη πολλαπλών κειμένων, οι περισσότερες μετρικές είναι δύσκολες στην εφαρμογή γιατί απαιτούν την εντατική ενασχόληση ανθρώπων, είτε για τη δημιουργία πρότυπων περιλήψεων, είτε για την αξιολόγηση του παραγόμενου αποτελέσματος.

3. Μεθοδολογία

Σκοπός του παρόντος είναι η δημιουργία μιας εφαρμογής για την αυτόματη συγκέντρωση και περίληψη πολλαπλών άρθρων από πηγές του διαδικτύου. Για την υλοποίηση, επιλέχθηκε η χρήση προεκπαιδευμένων μοντέλων.

Το πρώτο στάδιο της διαδικασίας είναι η συγκέντρωση των άρθρων από διαδικτυακές πήγες. Η συλλογή των άρθρων πραγματοποιείται σε 2 βήματα, τα οποία περιλαμβάνουν και τις απαιτούμενες διαδικασίες ελέγχου.

- Το πρώτο βήμα, είναι η αναζήτηση των άρθρων και περιλαμβάνει έλεγχο για την εγκυρότητα των url. Το αποτέλεσμα του βήματος είναι ένα σύνολο που περιλαμβάνει τόσα μοναδικά url όσα όρισε ο χρήστης.
- Το δεύτερο βήμα, είναι η εξαγωγή του κειμένου των άρθρων. Τα άρθρα εξάγονται και αποθηκεύονται σε dataframe. Σε αυτό το βήμα πραγματοποιείται αρχικά έλεγχος για την ύπαρξη περιεχομένου στα άρθρα
- Στη συνέχεια έλεγχοι διπλοτύπων και ομοιότητας στα άρθρα, ώστε η τελική συλλογή να μην περιέχει διπλότυπα, τα οποία θα αλλοιώσουν το παραγόμενο αποτέλεσμα.
- Τα παραπάνω βήματα επαναλαμβάνονται μέχρι τη συγκέντρωση του απαιτούμενου αριθμού άρθρων.

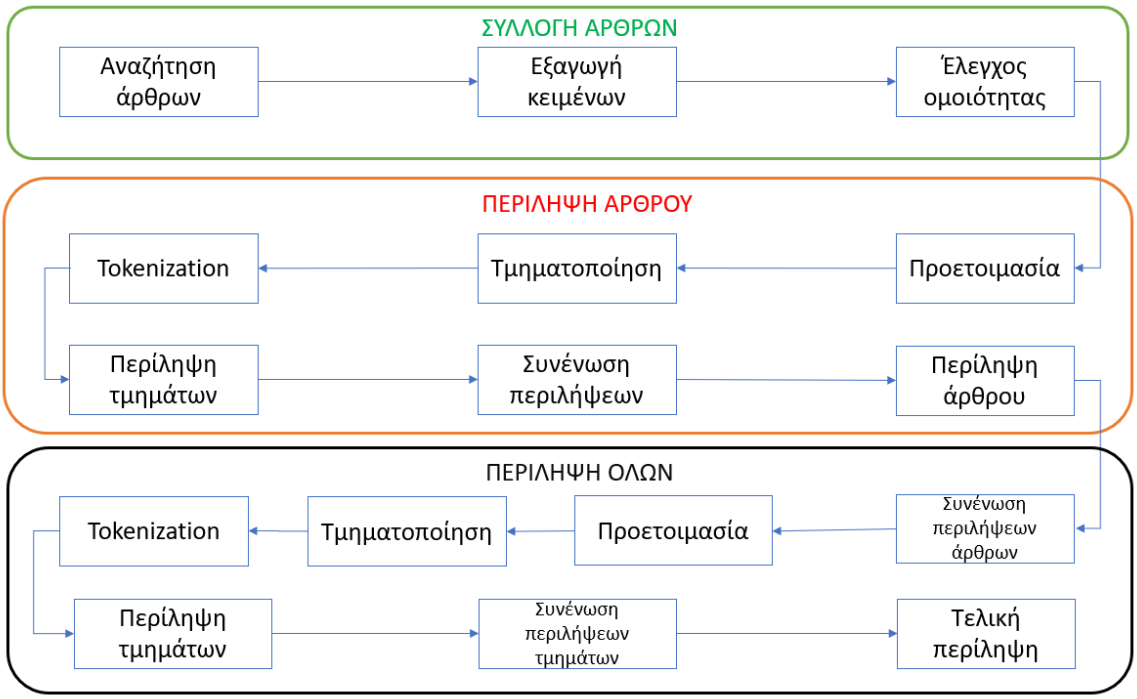
Μετά τη συγκέντρωση των άρθρων, περνάμε στο δεύτερο στάδιο της διαδικασίας ή οποία αφορά την προεπεξεργασία και την περίληψη κάθε άρθρου ξεχωριστά. Αυτό το στάδιο αποτελείται από 6 βήματα και εκτελείται για κάθε άρθρο.

- Το πρώτο βήμα αφορά την προετοιμασία και περιλαμβάνει την αφαίρεση των κενών γραμμών, και εμβόλιμων ενοτήτων στο κείμενο, όπως το τμήμα των Σχετικών άρθρων.
- Στο δεύτερο βήμα πραγματοποιείται τμηματοποίηση του άρθρου. Η τμηματοποίηση απαιτείται σε αρκετές περιπτώσεις, καθώς τα μοντέλα περίληψης έχουν όριο στον αριθμό των tokens που μπορούν να επεξεργαστούν.
- Το τρίτο βήμα αφορά το tokenization. Τα μοντέλα που χρησιμοποιούνται περιλαμβάνουν μέθοδο tokenization, η οποία εκτελεί και τις απαιτούμενες για τη δημιουργία περίληψης προεπεξεργασίες.
- Στο επόμενο βήμα εξάγεται η περίληψη κάθε τμήματος ξεχωριστά.

- Στο πέμπτο βήμα, οι περιλήψεις κάθε τμήματος συνενώνονται σε ένα κείμενο
- Στο τελευταίο βήμα αυτού του σταδίου δημιουργείται η περίληψη για το συνολικό κείμενο του άρθρου

Το τελευταίο στάδιο της διαδικασίας περιλαμβάνει τη συγκέντρωση όλων των περιλήψεων σε ένα κείμενο, και στη συνέχεια την επανάληψη των βημάτων του προηγούμενου σταδίου σε αυτό για την παραγωγή της τελικής περίληψης:

- Συνένωση περιλήψεων
- Προετοιμασία του κειμένου
- Τμηματοποίηση του συνολικού κειμένου
- Tokenization
- Εξαγωγή περίληψης για κάθε τμήμα ξεχωριστά.
- Συνένωση των περιλήψεων κάθε τμήματος σε ένα κείμενο
- Δημιουργία η περίληψη τελικής περίληψης για το σύνολο των άρθρων



Εικόνα 4 Βήματα υλοποίησης

Καταλήξαμε στην παραπάνω μεθοδολογία για την περίληψη των κειμένων προκειμένου να αντιμετωπίσουμε τις προκλήσεις που θέτει ο στόχος της εργασίας, δηλαδή η δημιουργία μίας περίληψης από πολλαπλά ειδησεογραφικά άρθρα, τα οποία έχουν συγκεντρωθεί αυτόματα.

Τα ζητήματα που έπρεπε να αντιμετωπιστούν είναι:

- Πιθανή ετερογένεια στο περιεχόμενο των άρθρων, η οποία κάνει πιο δύσκολη ή πιο ευρεία την επιλογή των σημαντικών πληροφοριών.
- Αλληλεπικάλυψη των σημαντικών πληροφοριών
- Μείωση του μεγέθους του τελικού κειμένου προς περίληψη, ειδικά με το δεδομένο ότι τα μοντέλα που αξιολογήθηκαν έχουν περιορισμό ως προς το μέγεθος του κειμένου που μπορούν να επεξεργαστούν.
- Δημιουργία μιας συνεκτικής και κατανοητής περίληψης από όλα τα άρθρα

Για να αντιμετωπιστούν τα παραπάνω, επιλέχθηκε:

- Η αρχική σύγκριση των κειμένων ως προς την ομοιότητα προκειμένου να αποφύγουμε τα διπλότυπα κείμενα.
- Για να αντιμετωπίσουμε τον περιορισμό σε σχέση με το μέγεθος του κειμένου, τα κείμενα που είναι μεγαλύτερα από το όριο του μοντέλου τμηματοποιούνται και πραγματοποιείται περίληψη κάθε τμήματος ξεχωριστά. Το κάθε τμήμα περιλαμβάνει μία ή περισσότερες ολόκληρες παραγράφους.
- Για να διατηρήσουμε τις σημαντικές πληροφορίες κάθε άρθρου ξεχωριστά μειώνοντας παράλληλα το μέγεθος του, επιλέξαμε να διαχειριστούμε το κάθε άρθρο ανεξάρτητα δημιουργώντας μια περίληψη για το καθένα, κάνοντας την παραδοχή ότι η περίληψη είναι καλή μέθοδος για την αναπαράσταση των σημαντικών πληροφοριών ενός άρθρου σε μικρότερο μέγεθος.
- Τέλος, αντιμετωπίσαμε το σύνολο των περιλήψεων ως ένα ενιαίο κείμενο, και δημιουργήσαμε την περίληψη του.

4. Υλοποίηση

Στόχος είναι να δημιουργηθεί μια συλλογή, που αποτελείται από διαφορετικά μεταξύ τους άρθρα στα οποία στη συνέχεια θα εφαρμοστούν οι μηχανισμοί περίληψης.

Ως γλώσσα εργασίας επιλέξαμε τα αγγλικά. Στην αρχή, διερευνήθηκε η δυνατότητα υλοποίησης στην ελληνική γλώσσα, όμως αυτό δεν ήταν εφικτό. Εξετάσαμε κάποια διαθέσιμα μοντέλα για την ελληνική γλώσσα, τα οποία όμως δεν είχαν ικανοποιητικά αποτελέσματα.

Το πρώτο μοντέλο που εξετάστηκε ήταν το GreekBERT (https://github.com/nlpaueb/greek-bert) (Koutsikakis, Chalkidis, Malakasiotis, P., & Androutsopoulos, 2020) το οποίο όμως δεν ήταν κατάλληλο για της ανάγκες μας, καθώς ήταν αναγκαία η εκπαίδευσή του. Το GreekBert είναι ένα μοντέλο επεξεργασίας φυσικής γλώσσας εκπαιδευμένο στην ελληνική γλώσσα, όμως δεν είναι εξειδικευμένο για τη δημιουργία περιλήψεων. Αυτό σημαίνει ότι απαιτείται η εξειδίκευση του (fine-tuning) για εργασίες αφαιρετικής περίληψης. Η εξειδίκευση αυτή απαιτεί dataset το οποίο να περιέχει αρχικά κείμενα και πρότυπες περιλήψεις ιδανικά γραμμένες από επαγγελματίες. Τέτοιο dataset δυστυχώς δεν είναι διαθέσιμο στην ελληνική γλώσσα, ούτε υπάρχουν οι απαραίτητοι πόροι για τη δημιουργία του στο πλαίσιο αυτής της εργασίας.

Στη συνέχεια εξετάστηκε το μοντέλο GreekBART (https://github.com/iakovosevdaimon/greekbart), το οποίο όμως δεν είχε καλά αποτελέσματα όπως αναφέρουν και οι δημιουργοί του (Evdaimon, et al., 2023). Το GreekBART βασίζεται στο BART (Lewis, et al., 2019), αλλά έχει εκπαιδευτεί για την εκτέλεση διάφορων εργασιών NLP με ελληνικά dataset όπως η ελληνική Wikipedia, τα ελληνικά πρακτικά του Ευρωκοινοβουλίου και άλλα, καλύπτοντας έτσι ένα ευρύ φάσμα γλωσσικών περιοχών.

Για την εξειδίκευση στη δημιουργία αφαιρετικών περιλήψεων, οι δημιουργοί του διαπίστωσαν ότι δεν υπάρχουν datasets για τη δημιουργία περιλήψεων στα ελληνικά. Έτσι δημιούργησαν το GreekSUM, ένα dataset που δημιουργήθηκε με αυτόματο τρόπο, συλλέγοντας άρθρα που συνοδεύονται από τις περιλήψεις και τον τίτλο τους από το ειδησεογραφικό site News24/7. Κατά την αξιολόγηση του dataset διαπίστωσαν ότι οι πρότυπες περιλήψεις των κειμένων δεν ήταν αρκετά αφαιρετικές, χαρακτηριστικό που

αναμένεται να μεταφερθεί και στην παραγωγή περιλήψεων από το μοντέλο. Επίσης οι περιλήψεις δεν αξιολογήθηκαν θετικά κατά την αξιολόγηση από ανθρώπους.

Τέλος δοκιμάσαμε το μοντέλο Kriton (https://huggingface.co/kriton/greek-text-summarization), το οποίο βασίζεται στο mT5-small και έχει εκπαιδευτεί χρησιμοποιώντας με το dataset «News Articles in Greek» (https://www.kaggle.com/datasets/kpittos/news-articles), μια συλλογή άρθρων από το tvxs.gr. Το dataset δεν περιέχει περιλήψεις αλλά μόνο τους τίτλους και το κείμενο των άρθρων. Το αποτέλεσμα είναι να παράγονται πολύ συνοπτικές περιλήψεις, οι οποίες δεν αποδίδουν το νόημα των κειμένων.

Ένα παράδειγμα της λειτουργίας του Kriton είναι το παρακάτω[1]:

«Στη ΔΕΘ πρόκειται να ανακοινωθεί και με τον πλέον επίσημο τρόπο το Food Pass - το τρίτο επίδομα που εξετάσει η κυβέρνηση έπειτα από τις επιδοτήσει σε καύσιμα και ηλεκτρικό ρεύμα. Η επιταγή ακρίβειας για τρόφιμα θα δοθεί εφάπαξ σε οικογένειες, που πληρούν συγκεκριμένα κριτήρια, με στόχο να στηριχτούν όσοι πλήττονται περισσότερο από την αύξηση του πληθωρισμού και την ακρίβεια, όπως: χαμηλοσυνταξιούχοι, άνεργοι και ευάλωτες κοινωνικές ομάδες. Σύμφωνα με τις διαθέσιμες πληροφορίες, η πληρωμή θα γίνει κοντά στις γιορτές των Χριστουγέννων, όταν - σύμφωνα με εκτιμήσεις - θα έχει βαθύνει η ενεργειακή κρίση και θα έχουν αυξηθεί ακόμη παραπάνω οι ανάγκες των νοικοκυριών. Ουσιαστικά, με τον τρόπο αυτό, η κυβέρνηση θα επιδοτήσει τις αγορές του σούπερ μάρκετ για έναν μήνα για ευάλωτες ομάδες. Αντιδράσεις από τους κρεοπώλες Υπενθυμίζεται πως την Πέμπτη η Πανελλήνια Ομοσπονδία Καταστηματαρχών Κρεοπωλών (ΠΟΚΚ) Οικονομικών - με επιστολή της προς τα συναρμόδια υπουργεία - διαμαρτυρήθηκε σχετικά με τη χορήγηση του Food Pass στα σούπερ μάρκετ. Στην επιστολή της ομοσπονδίας, την οποία κοινοποίησε και στην Κεντρική Ένωση Επιμελητηρίων και στη ΓΣΕΒΕΕ, σημειώνεται ότι εάν το μέτρο εξαργύρωσης του επιδόματος τροφίμων αφορά μόνο στις αλυσίδες τροφίμων τότε απαξιώνεται ο κλάδος των κρεοπωλών και «εντείνεται ο αθέμιτος ανταγωνισμός, εφόσον κατευθύνεται ο καταναλωτή σε συγκεκριμένες επαγγελματικές κατηγορίες». Στο πλαίσιο αυτό, οι κρεοπώλες ζητούν να διευθυνθεί η λίστα των επαγγελματιών όπου ο καταναλωτής θα μπορεί να εξαργυρώσει το εν λόγω βοήθημα.»

Η παραγόμενη περίληψη:

[1] Το παράδειγμα δίνεται αυτούσιο από την σελίδα παρουσίασης του μοντέλου https://huggingface.co/kriton/greek-text-summarization

«Το Food Pass - το τρίτο επίδομα που εξετάζει η κυβέρνηση έπειτα από τις επιδοτήσεις σε καύσιμα και ηλεκτρικό ρεύμα για έναν μήνα για ευάλωτες ομάδες. Σύμφωνα με πληροφορίες της Πανελλήνια Ομοσπονδίας Καταστηματαρχών Κρεοπωλών (ΠΟΚΚ) Οικονομικών, οι κρεοπώλες διαμαρτυρήθηκαν σχετικά με τη χορήγηση του «fast food pass» προκειμένου να αυξάνουν ακόμη περισσότερες κοινωνικές ανάγκους στο κλάδο»

Συνοψίζοντας, κατά τη διερεύνηση της δυνατότητας υλοποίησης στα ελληνικά, διαπιστώσαμε ότι, παρόλο που ήδη υπάρχουν τέτοιες υλοποιήσεις, δεν υπάρχουν διαθέσιμα ακόμα dataset κατάλληλα για τη δημιουργία αφαιρετικών περιλήψεων στα ελληνικά. Με δεδομένο ότι δεν υπήρχε η δυνατότητα δημιουργίας ενός νέου dataset για την εξειδίκευση του μοντέλου μας, η προσπάθεια εγκαταλείφθηκε και προτιμήσαμε την αγγλική γλώσσα.

Στην αρχή της υλοποίησης πραγματοποιήσαμε αυτόματη αναζήτηση ειδησεογραφικών άρθρων με συγκεκριμένα κριτήρια. Στη συνέχεια, πραγματοποιήσαμε εξαγωγή του κειμένου των άρθρων από τις ιστοσελίδες, και αφαιρέσαμε τα διπλότυπα.

Προχωρήσαμε στην προετοιμασία των άρθρων για περίληψη, αν και με δεδομένο ότι χρησιμοποιούμε έτοιμα μοντέλα, μεγάλο μέρος της προεπεξεργασίας γίνεται από το ίδιο το μοντέλο.

Την περισσότερη προσοχή, απαίτησε η διαδικασία τμηματοποίησης των άρθρων, καθώς τα μοντέλα που χρησιμοποιήθηκαν έχουν περιορισμό στον αριθμό των tokens που μπορούν να επεξεργαστούν.

Για την περίληψη των κειμένων ακολουθήθηκε διαδικασία σταδίων, κατά την οποία παράγεται η περίληψη των άρθρων, στη συνέχεια γίνεται συνένωση των περιλήψεων και με βάση αυτή δημιουργείται μια νέα περίληψη.

4.1. Αφαιρετική περίληψη με PEGASUS

Πριν ξεκινήσουμε την περιγραφή της υλοποίησης με αναφορά στα επιμέρους στάδια της θεωρούμε σκόπιμο να παρουσιάσουμε το μοντέλο αφαιρετικής περίληψης το οποίο επιλέξαμε, καθώς θεωρούμε ότι είναι το σημαντικότερο τμήμα της υλοποίησης μας.

Στην παράγραφο 2.4.2 παρουσιάσαμε επιλεκτικά κάποιες ενδιαφέρουσες υλοποιήσεις για την αφαιρετική περίληψη κειμένων. Είδαμε ότι τα μοντέλα Seq2Seq με εφαρμογή μηχανισμού προσοχής, όπως αυτό των (Chopra, S., Auli, M, & Rush, A., 2016), αντιμετωπίζουν προβλήματα όπως η επανάληψη φράσεων στην περίληψη, η αναπαραγωγή

ανακριβών πληροφοριών (factual errors) και η αδυναμία χειρισμού λέξεων που δε βρίσκονται στο λεξιλόγιο τους (See, Liu, & Manning, 2017).

Στη συνέχεια είδαμε την υλοποίηση των (See, Liu, & Manning, 2017), οι οποίοι με την εφαρμογή ενός δικτύου pointer-generator, προσπάθησαν να αντιμετωπίσουν την ανακρίβεια και τη διαχείριση των άγνωστων λέξεων (out of vocabulary words), ουσιαστικά εκμεταλλευόμενοι πλεονεκτήματα της αποσπασματικής περίληψης. Όμως, παρόλη τη σημαντική βελτίωση, τα προβλήματα στη συνοχή και την ακρίβεια των περιλήψεων παραμένουν.

Στη συνέχεια, παρουσιάσαμε 2 δημοφιλή μοντέλα που βασίζονται σε Transformers. Τα PEGASUS (Zhang et al., 2019) και BART (Lewis, et al., 2019). Είδαμε ότι το μεν BART είναι ένα μοντέλο που μπορεί να βρει εφαρμογή σε εργασίες NLP που απαιτούν κατανόηση και παραγωγή κειμένου, μεταξύ των οποίων η δημιουργία αφαιρετικών περιλήψεων με μεγάλο βαθμό παραγωγής νέου κειμένου, το οποίο όμως πολλές φορές μπορεί να είναι παραπλανητικό και να περιέχει κατασκευασμένες πληροφορίες.

Αντίστοιχα το PEGASUS είναι ένα μοντέλο το οποίο είναι εξ αρχής εκπαιδευμένο για την παραγωγή αφαιρετικών περιλήψεων, έχει όμως το μειονέκτημα ότι μπορεί οι παραγόμενες περιλήψεις να περιέχουν αρκετές αυτούσιες φράσεις από το αρχικό κείμενο. Τα παραπάνω αποτελέσματα επιβεβαιώθηκαν με τη δοκιμή fine-tuned υλοποιήσεων των δύο μοντέλων στο ίδιο dataset.

Στην ενότητα 2.4.4 παρουσιάσαμε υλοποιήσεις για την περίληψη πολλαπλών κειμένων. Για την υλοποίηση εξετάσαμε το μοντέλο Multinews (Fabbri, Li, She, Li, & Radev, 2019), όμως θεωρήσαμε ότι η προσέγγιση της διατήρησης μόνο των πρώτων προτάσεων κάθε κειμένου δεν καλύπτει το εύρος της πληροφορίας που μπορεί να περιέχει ένα ειδησεογραφικό άρθρο.

Σύμφωνα με τα παραπάνω, επιλέξαμε να χρησιμοποιήσουμε το PEGASUS για την υλοποίηση μας. Στη συνέχεια, αποφασίσαμε να χρησιμοποιήσουμε μια από τις διαθέσιμες fine-tuned υλοποίηση του PEGASUS σε συγκεκριμένο dataset. Εξετάσαμε fine-tuned υλοποιήσεις του PEGASUS σε datasets τα οποία είναι κατάλληλα για την περίληψη ειδησεογραφικών άρθρων[2]. Συγκεκριμένα εξετάστηκαν οι υλοποιήσεις με τα datasets multi_news (Fabbri, Li, She, Li, & Radev, 2019) και cnn_dailymail (Hermann, K. M., et al., 2015). Κατά τις δοκιμές που πραγματοποιήσαμε, διαπιστώσαμε ότι το PEGASUS/

[2] Όλες οι υλοποιήσεις είναι διαθέσιμες στο https://huggingface.co/

multi_news αγνοούσε μέρος των προς περίληψη άρθρων, όπως άλλωστε ήταν αναμενόμενο. Έτσι, καταλήξαμε στην επιλογή του PEGASUS/ cnn_dailymail.

Η χρήση του μοντέλου κατά την υλοποίηση είναι αρκετά απλή, καθώς περιλαμβάνει tokenizer ο οποίος, πραγματοποιεί και τα σημαντικά στάδια της προεπεξεργασίας (https://huggingface.co/docs/transformers/main/model_doc/pegasus). Για την καλύτερη ρύθμιση του tokenization όπως truncation (ορίζεται το αν θα αποκοπούν τα tokens που ξεπερνούν το μέγιστο όριο του μοντέλου, 1024 tokens, padding (αναφέρεται στο μέγεθος των tokens), return_tensors (το είδος των επιστρεφόμενων tensors), κλπ.

Στη συνέχεια για τη δημιουργία της περίληψης καλείται η μέθοδος summarizer.generate, η οποία δέχεται ως είσοδο τα tokens που έχουν παραχθεί παραπάνω και η κλήση της περιλαμβάνει παραμέτρους όπως max_length (μέγιστο μέγεθος περίληψης κλπ.). Η δημιουργία της περίληψης γίνεται σε tokens.

Στο τελευταίο βήμα καλείται η μέθοδος tokenizer.decode για τη μετατροπή των tokens σε κείμενο..

4.2. Συλλογή άρθρων από διαδικτυακές πηγές

Στόχος της υλοποίησης είναι να πραγματοποιείται η συλλογή χωρίς να απαιτείται παρέμβαση του χρήστη, ο ρόλος του οποίου θα πρέπει να περιορίζεται στον ορισμό των παραμέτρων αναζήτησης.

Για το λόγο αυτό, χρησιμοποιούμε τη βιβλιοθήκη GoogleNews (https://github.com/Iceloof/GoogleNews) η οποία προσφέρει τη δυνατότητα αυτόματης αναζήτησης με βάση συγκεκριμένα κριτήρια. Επιστρέφει ένα dataframe, το οποίο περιλαμβάνει τον τίτλο, το μέσο δημοσίευσης, το url του άρθρου, μια συνοπτική περιγραφή του.

Στόχος της συλλογής ειδήσεων, είναι τελικά το dataframe να περιλαμβάνει τόσα ενεργά url και μοναδικά άρθρα όσα ζήτησε ο χρήστης ορίζοντας τον αριθμό των αποτελεσμάτων. Για να έχουμε μια συλλογή που να ικανοποιεί τα παραπάνω, αρχικά ελέγξαμε την εγκυρότητα των url χρησιμοποιώντας τη βιβλιοθήκη requests (https://github.com/psf/requests). Επίσης, αφαιρέσαμε τις διπλότυπες εγγραφές με βάση διαφορετικά κριτήρια (τίτλος, περιγραφή).

Το δεύτερο βήμα της διαδικασίας είναι να ανακτήσουμε τα άρθρα από την παραπάνω συλλογή url. Για το σκοπό αυτό, χρησιμοποιούμε τη βιβλιοθήκη newspaper3k (https://github.com/codelucas/newspaper), η οποία επιστρέφει μεταξύ άλλων τον τίτλο, το κείμενο, ή άλλα μεταδεδομένα που μπορεί να επιλέξουμε.

Η μέθοδος αυτή για την ανάκτηση των άρθρων προτιμήθηκε σε σχέση με τις εναλλακτικές, όπως η BeautifulSoup γιατί επιστρέφει το άρθρο έχοντας αφαιρέσει τα HTML tags, μειώνοντας με αυτό τον τρόπο τις απαιτήσεις για προεπεξεργασία. Επίσης δεν απαιτεί το συγκεκριμένο ορισμό των HTML tags που πρέπει να αφαιρεθούν. Μειονέκτημα της βιβλιοθήκης είναι ότι σε μερικές περιπτώσεις δεν μπορεί να ανακτήσει ολόκληρο το άρθρο. Συναντήσαμε μία τέτοια περίπτωση, σε άρθρο του bbc.co.uk.

4.2.1. Ανάλυση κώδικα

Η εκτέλεση ξεκινά από τη μέθοδο main, όπου ορίζονται οι παράμετροι της διαδικασίας.

Παράμετρος	Περιγραφή
lang	Γλώσσα αναζήτησης
period	Χρονικό διάστημα αναζήτησης σε ημέρες (πχ. 7d)
topic	Θέμα αναζήτησης
icount	Αριθμός άρθρων που θέλουμε να ανακτήσουμε
mode	Επιλογή τρόπου υλοποίησης της περίληψης.
UseExistinDF	Με τιμή True για τη χρήση άρθρων που είναι αποθηκευμένα στο αρχείο Articles.csv ή False για την αναζήτηση νέων άρθρων και την αποθήκευση τους στο Articles.csv

Πίνακας 1: Παράμετροι αναζήτησης άρθρων

Η διαδικασία έχει υλοποιηθεί με χρήση των παρακάτω εναλλακτικών συνδυασμών:

- Αφαιρετική περίληψη των άρθρων με χρήση του μοντέλου "google/pegasus-cnn_dailymail" (mode 5)
- Πρώτα εφαρμογή αποσπασματικής περίληψης με χρήση του sumy-TextRank των άρθρων και στη συνέχεια αφαιρετική περίληψη με χρήση του μοντέλου "google/pegasus-cnn_dailymail" (mode 15)

Για την αναζήτηση και ανάκτηση των άρθρων δημιουργήθηκε η κλάση NewsFinder η οποία περιέχει τις μεθόδους για την αναζήτηση και ανάκτηση των άρθρων.

```
class NewsFinder:
    # η κλάση που δημιουργεί τη συλλογή άρθρων. τα instances έχουν ως attributes τα κριτήρια αναζήτησης
    def __init__(self, lang, period, topic, icount):
        self.lang = lang
        self.period = period
        self.topic = topic
        self.icount = icount
```

Μετά τη δημιουργία του instance της κλάσης εκτελούμε τη μέθοδο iterate της κλάσης NewsFinder

```
#δημιουργώ instance της κλάσης NewsFinder με τις παραμέτρους αναζήτησης και τον αριθμό των αποτελεσμάτων
    nf = NewsFinder(lang, period, topic, icount)
#η μέθοδος iterate δημιουργεί τη συλλογή των άρθρων
    news_df = nf.iterate()
```

Η μέθοδος iterate είναι αυτή που θα πραγματοποιήσει και στη συνέχεια θα ανακτήσει τα άρθρα. Αρχικά καλεί τη μέθοδο getgooglenews, η οποία θα εκτελέσει την αρχική αναζήτηση και θα επιστρέψει τα αποτελέσματα σε μορφή dataframe, το df. Στη συνέχεια, καλεί τη μέθοδο getarticles για την ανάκτηση των άρθρων από τα url που επέστρεψε η αναζήτηση, η οποία επιστρέφει ένα dataframe, το news_df που περιέχει τα άρθρα που ανακτήθηκαν και το df_new το οποίο είναι το ίδιο με το αρχικό dataframe, όμως έχουν αφαιρεθεί όλα τα url που απορρίφθηκαν κατά την ανάκτηση των άρθρων. Στη συνέχεια θα επαναλάβει τη διαδικασία με νέες παραμέτρους μέχρι να συμπληρωθεί ο απαιτούμενος αριθμός έγκυρων άρθρων.

```
def iterate(self):
    # πρώτη αναζήτηση και ανάκτηση άρθρων
    df = self.getgooglenews(self.icount)
    news_df, df_new = self.getarticlesdf(df)

    while df_new.shape[0] < self.icount:
    # αν τα άρθρα είναι λιγότερα, η διαδικασία επαναλαμβάνεται
        print("Count is")
        print(df_new.shape[0])
        # ο νέος αριθμός άρθρων που θέλουμε να ανακτήσουμε
        now_count = self.icount-df_new.shape[0]
        df_new = self.getgooglenews(now_count, True, df_new)
        news_df, df_new = self.getarticlesdf(df_new)

    return news_df
```

Η αναζήτηση των άρθρων γίνεται με τη μέθοδο getgooglenews της κλάσης NewsFinder. Αρχικά δημιουργείται ένα instance της κλάσης GoogleNews (βιβλιοθήκη GoogleNews).

Τροφοδοτείται με τις παραμέτρους αναζήτησης και στη συνέχεια πραγματοποιείται η αναζήτηση. Τα αποτελέσματα που επιστρέφονται αποθηκεύονται σε dataframe για και στη συνέχεια γίνεται έλεγχος για την εγκυρότητα του url. Αρκετές φορές, τα url είναι invalid, κυρίως λόγω πολιτικής των ιστοτόπων.

Τα μη έγκυρα url αφαιρούνται και στη συνέχεια το dataframe καθαρίζεται από διπλοεγγραφές. Η πρώτη επιστροφή αποτελεσμάτων από τη googlenews περιέχει τα αποτελέσματα της πρώτης μόνο σελίδας. Πραγματοποιούμε έλεγχο για το αν τα αποτελέσματα είναι αρκετά, αλλιώς επιστρέφουμε τα αποτελέσματα της δεύτερης σελίδας, τα ελέγχουμε ως προς την εγκυρότητα και στη συνέχεια τα συνενώνουμε με αυτά της πρώτης.

```
def getgooglenews(self, gncount, bexclude=False, df_exclude=None):
    # η μέθοδος κάνει την αναζήτηση στο Googlenews
    googlenews = GoogleNews()
    # οι τιμές των παραμέτρων αναζήτησης
    googlenews.set_lang(self.lang)
    googlenews.set_period(self.period)
    googlenews.set_encode('utf-8')
    googlenews.search(self.topic)
    # επιστοφή αποτελεσμάτων
    result = googlenews.results()
    # μετατροπή σε dataframe
    df = pd.DataFrame(result)
    df = self.checkforvalid(df)

    # το googlenews αρχικά επιστρέφει τα αποτελέσματα της πρώτης σελίδας.
    # αν ο αριθμός τους δεν επαρκεί, θα πρέπει να πάρουμε και επόμενων
    ipage = 1

    # το bexclude είναι μια συνθήκη, για τα επόμενα iterations
    # δείχνει ότι έχω ήδη προχωρήσει στην εξαγωγή άρθρων
    if bexclude:
        # τα αποτελέσματα που έχω ήδη συγκεντρώσει συγχωνεύονται με τα νέα
        df = pd.concat([df, df_exclude], ignore_index=True)
        # αφαιρούνται τα διπλά αποτελέσματα
        df.drop_duplicates(inplace=True)
    # εδω γίνεται ουσιαστικά η αναζήτηση.
    # Επαναλαμβάνεται μέχρι να συμπληρωθεί ο απαιτούμενος αριθμός αποτελεσμάτων
    while df.shape[0] < gncount:
        # ανάκτηση αποτελεσμάτων από την επόμενη σελίδα
        ipage += 1
        googlenews.getpage(ipage)
        result = googlenews.results()
        # προσωρινό dataframe για την αποθήκευση των αποτελεσμάτων της δεύτερης σελίδας
        df_temp = pd.DataFrame(result)
        df_temp = self.checkforvalid(df_temp)
        # ενοποίηση των dataframes και αφαίρεση διπλότυπων
        df = pd.concat([df, df_temp], ignore_index=True)
        df.drop_duplicates(inplace=True)
        df = df.drop_duplicates(subset='title', keep="first")
        df = df.drop_duplicates(subset='desc', keep="first")
    df.reset_index(drop=True, inplace=True)
```

```
    print(df.loc[:, "link"])
    print(df.shape[0])
    # αποθήκευση σε csv
    df.to_csv("dataframe_initial.csv")
    return df
```

Η iterate θα καλέσει τη getgooglenews επαναληπτικά μέχρι να ανακτήσουμε τον επιθυμητό αριθμό έγκυρων άρθρων.

Ο έλεγχος εγκυρότητας γίνεται με τη μέθοδο checkforvalid η οποία πραγματοποιεί request (βιβλιοθήκη requests) για κάθε ένα από τα url και απορρίπτει όσα δεν απαντήσουν με status_code = 200 (valid). Τα url που απορρίπτονται αφαιρούνται από το dataframe με τα αποτελέσματα της αναζήτησης.

Η ανάκτηση των άρθρων γίνεται με τη μέθοδο getarticlesdf, η οποία χρησιμοποιεί τις μεθόδους της βιβλιοθήκης newspaper3k για την ανάκτηση της ημερομηνίας, του μέσου δημοσίευσης, του τίτλου και του κειμένου του άρθρου. Αυτή η μέθοδος εφαρμόζει και τις λειτουργίες που περιγράφονται στην επόμενη παράγραφο σχετικά με την αξιολόγηση των ανακτημένων άρθρων, όπως την απόρριψη όμοιων ή άρθρων χωρίς περιεχόμενο.

```
def getarticlesdf(self, df):
    # εδώ γίνεται η ανάκτηση των κειμένων από τις ιστοσελίδες
    # ρυθμίσεις browser user agent
    config = Config()
    config.browser_user_agent = "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/113.0.0.0 Safari/537.36"
    config.request_timeout = 20
    news_list = []
    excluded = []
    # ανάκτηση άρθρων. από κάθε link του dataframe ανακτώ το άρθρο
    for ind in df.index:
        dict = {}
        try:
            a = Article(df['link'][ind], language='en', config=config)
            a.url.strip()
            a.download()
            a.parse()
            # δημιουργία λεξικού για εισαγωγή στο νέο dataframe
            dict['Link'] = df['link'][ind]
            dict['Date'] = df['date'][ind]
            dict['Media'] = df['media'][ind]
            dict['Title'] = a.title
            dict['Article'] = a.text
            # έλεγχος μεγέθους άρθρου. αν δεν έχει περιεχόμενο θα εξαιρεθεί
            if len(dict['Article']) > 400:
                news_list.append(dict)
            else:
                excluded.append(ind)

        except:
            # αν η διαδικασία δώσει σφάλμα, η εγγραφή τηρείται σε λίστα exclude για να αφαιρεθεί
            print(f"{df['link'][ind]} is invalid")
            excluded.append(ind)
            continue
```

```
#δημιουργία dataframe άρθρων news_df
news_df = pd.DataFrame(news_list)
# ότι δε γύρισε αποτέλεσμα αφαιρείται από το dataframe με τα αποτελέσματα απο το googlenews
for i in excluded:
    df.drop(i, inplace=True)
df.drop_duplicates(inplace=True)
df.reset_index(drop=True)


#αφαίρεση διπλοτύπων. αφαιρούνται και από τα δύο dataframes
news_df = news_df.drop_duplicates(subset='Title', keep="first")
df = df.drop_duplicates(subset='title', keep="first")
if news_df.shape[0] > 0:


    news_df.reset_index(drop=True, inplace=True)
    if news_df.shape[0] > 1:
        sc = SimilarityChecker()
        exclude = sc.compare(news_df)
        for i in exclude:
            link = news_df['Link'][i]
            df.drop(df[df['link'] == link].index, inplace=True)
            news_df.drop(i, inplace=True)
    df.reset_index(drop=True)
    news_df.reset_index(drop=True)
    # τα τελικά αποτελέσματα αποθηκεύονται σε csv
    df.to_csv("dataframe_new.csv")
    news_df.to_csv("articles.csv")
return news_df, df
```

4.3. Αξιολόγηση περιεχομένου άρθρων

Κατά την ανάκτηση των άρθρων παρατηρήσαμε ότι αρκετά από αυτά δεν περιείχαν περιεχόμενο που να είναι ικανοποιητικού μεγέθους, ότι τα url εμφανίζονταν μη έγκυρα, οπότε δε μπορούσε να πραγματοποιηθεί ανάκτηση, υποθέτουμε λόγω πολιτικής των ιστοτόπων. Τέλος παρατηρήσαμε ότι σε κάποιες περιπτώσεις τα άρθρα ήταν πανομοιότυπα, παρόλο που προέρχονταν από διαφορετικές σελίδες.

Αρχικά προχωρήσαμε στην αφαίρεση των μη έγκυρων url και αυτών χωρίς περιεχόμενο και από το αρχικό dataframe που δημιουργήσαμε με τη χρήση του GoogleNews και από αυτό που δημιουργήθηκε με το newspaper3k. Στη συνέχεια, προχωρήσαμε σε καθαρισμό του dataframe από τα διπλότυπα άρθρα.

Για την εύρεση των όμοιων άρθρων υλοποιήσαμε έλεγχο ομοιότητας συνημιτόνου κατά ζεύγη άρθρων με χρήση συναρτήσεων της βιβλιοθήκης scikit-learn (https://github.com/scikit-learn/scikit-learn) και nltk (https://github.com/nltk/nltk). Η βιβλιοθήκη nltk χρησιμοποιήθηκε για την προεπεξεργασία των κειμένων (αφαίρεση περιττών λέξεων και σημείων στίξης), ενώ η scikit-learn χρησιμοποιήθηκε για τον

υπολογισμό της ομοιότητας συνημιτόνου. Τα άρθρα ομοιότητα συνημιτόνου μεγαλύτερη από 0,95 αφαιρέθηκαν. Η επιλογή να χρησιμοποιηθούν και λειτουργίες από την nltk έγινε γιατί είναι πιο αποτελεσματική στα στάδια της προεργασίας.

Επιλέξαμε τη μέθοδο της ομοιότητας συνημιτόνου για τρεις λόγους. Ο πρώτος είναι ότι ουσιαστικά θέλουμε να αφαιρέσουμε τα άρθρα που έχουν πολύ μεγάλη ομοιότητα με κάποιο άλλο, οπότε αρκεί μία αριθμητική τιμή πάνω από κάποιο όριο. Ο δεύτερος είναι ότι φαίνεται να λειτουργεί αποτελεσματικά, ανεξάρτητα από το μέγεθος των κειμένων, αφού οι υπολογισμοί βασίζονται στη γωνία μεταξύ των διανυσμάτων τους. Τέλος, είναι εύκολη η υλοποίησή της με χρήση των δύο βιβλιοθηκών.

Οι διαδικασίες ελέγχου διπλοτύπων, εγκυρότητας url και επαρκούς μεγέθους άρθρων υλοποιούνται κατά την ανάκτηση των άρθρων.

4.3.1. Κώδικας για τον υπολογισμό ομοιότητας συνημιτόνου

Για τον έλεγχο ομοιότητας συνημιτόνου υλοποιήθηκε η κλάση SimilarityChecker με τη βοήθεια των βιβλιοθηκών scikit-learn και nltk, και πραγματοποιείται με τις μεθόδους tokenize για κανονικοποίηση και tokenization, cosine_sim για υπολογισμό της ομοιότητας συνημίτονου μεταξύ 2 κειμένων και compare για την κλήση της cosine_sim με κάθε ζεύγος άρθρων[3].

```
class SimilarityChecker:
    def __init__(self):
        # nltk.download('punkt')
        self.stemmer = nltk.stem.porter.PorterStemmer()
        self.remove_punctuation_map = dict((ord(char), None) for char in string.punctuation)
        self.stop_words = set(stopwords.words('english'))
        self.vectorizer = TfidfVectorizer(tokenizer=self.tokenize, stop_words=self.stopwords)


    # tokenization, stemming και αφαίρεση stopwords
    def tokenize(self, text):
        tokens = nltk.word_tokenize(text.lower().translate(self.remove_punctuation_map))
        stemmed_tokens = [self.stemmer.stem(token) for token in tokens if token not in self.stop_words]
        return stemmed_tokens

    # υπολογισμός ομοιότητας συνημιτόνου
    def cosine_sim(self, text1, text2):
        #μετατροπή σε διανύσματα
        tfidf = self.vectorizer.fit_transform([text1, text2])
        return ((tfidf * tfidf.T).A)[0, 1]
```

[3] Η υλοποίηση βασίστηκε στη συζήτηση https://stackoverflow.com/questions/8897593/how-to-compute-the-similarity-between-two-text-documents

```
def compare(self, df):
    imax = df[df.columns[0]].count()
    exclude = []
    # σύγκριση κειμένων ανά δύο και υπολογισμός ομοιότητας συνημιτόνου
    for i in df.index:
        for j in range(i + 1, imax):
            # υπολογισμός ομοιότητας
            similarity = self.cosine_sim(df['Article'][i], df['Article'][j])
            print(similarity)
            # όριο ομοιότητας για την απόρριψη ενός κειμένου
            if similarity > 0.95:
                exclude.append(j)
                exclude = list(set(exclude))

    return exclude
```

4.4. Προεπεξεργασία και τμηματοποίηση

Έχοντας πλέον στη διάθεση μας μια πλήρη συλλογή άρθρων, ξεκινήσαμε την προεπεξεργασία για τη δημιουργία αυτόματης περίληψης. Με δεδομένη τη χρήση έτοιμων μοντέλων τα οποία χειρίζονται το μεγαλύτερο μέρος της απαιτούμενης προεργασίας, δεν απαιτήθηκαν πολλά βήματα.

Συγκεκριμένα, απαιτήθηκε η προετοιμασία των κειμένων με αφαίρεση των κενών γραμμών, απαλοιφή κάποιων εμβόλιμων τμημάτων στα άρθρα όπως «Advertisement» ή «Related Articles».

Τα κύρια στάδια της προεπεξεργασίας ήταν ο χωρισμός σε tokens και η τμηματοποίηση των άρθρων πριν την περίληψή τους.

Όπως είδαμε στην ενότητα 2.4.2, λόγω των μηχανισμών προσοχής των μοντέλων που βασίζονται σε Transformers, τίθεται όριο ως προς το μέγιστο μέγεθος ενός κειμένου για επεξεργασία. Το μοντέλο αυτόματης περίληψης που χρησιμοποιήσαμε, έχει μέγιστο όριο στον αριθμό των tokens που μπορεί να επεξεργαστεί. Συγκεκριμένα, μπορεί να επεξεργαστεί μέχρι 1024 tokens. Οπότε, το μέγεθος του προς περίληψη εγγράφου μετά το tokenization δεν μπορεί να ξεπερνά αυτόν τον αριθμό tokens.

Δοκιμάστηκαν διάφορες διαδικασίες τμηματοποίησης των κειμένων, όπως η τμηματοποίηση με βάση ένα συγκεκριμένο αριθμό χαρακτήρων.

Η διαδικασία δεν πέτυχε το στόχο της, καθώς ο αριθμός χαρακτήρων δεν έχει σαφή σχέση με τον αριθμό των tokens που δημιουργούνται κατά το tokenization. Επίσης, με αυτόν τρόπο μπορεί να διασπαστούν τμήματα του κειμένου τα οποία αποτελούν μια νοηματική ενότητα.

Στη συνέχεια, δοκιμάσαμε να πραγματοποιήσουμε tokenization και να φτιάξουμε τα τμήματα του εγγράφου ανάλογα με τον αριθμό των tokens. Όμως και αυτό δεν κατέληξε σε καλό αποτέλεσμα καθώς αυτού του είδους η τμηματοποίηση είχε ως συνέπεια να μοιραστεί μια πρόταση μεταξύ δύο τμημάτων του κειμένου. Το αποτέλεσμα της αφαιρετικής περίληψης, το οποίο βασίζεται στο νόημα του κειμένου, άρα και στις σχέσεις μεταξύ των λέξεων δεν παρουσίαζε συνοχή, και σε πολλές περιπτώσεις το νόημα αλλοιωνόταν.

Τελικά, η τμηματοποίηση έγινε με βάση τις ίδιες τις νοηματικές ενότητες του κειμένου, χρησιμοποιώντας τη βιβλιοθήκη regex (https://regexlib.com/)[4]. Το κείμενο μετατρέπεται σε λίστα παραγράφων, όπου κάθε αντικείμενο της είναι μία παράγραφος.

Ακολουθεί η διαδικασία tokenization κάθε παραγράφου ξεχωριστά, χρησιμοποιώντας τον tokenizer του μοντέλου περίληψης. Στην περίπτωση που ο αριθμός των tokens της παραγράφου ξεπερνά το όριο του μοντέλου, η παράγραφος διασπάται σε προτάσεις. Ακολουθεί tokenization των προτάσεων, μέχρι το μέγεθος να φτάσει τα 1024 tokens. Έτσι, μία παράγραφος μπορεί να τμηματοποιηθεί, όμως ένα τμήμα δε θα αποτελείται από 2 παραγράφους ή προτάσεις τους.

Το αποτέλεσμα της διαδικασίας είναι τμήματα σε μορφή κειμένου, τα οποία δεν ξεπερνούν τα 1024 tokens, ενώ ταυτόχρονα διατηρείται η νοηματική συνάφεια.

Η διαδικασία του tokenization επαναλαμβάνεται για κάθε τμήμα χωριστά, προκειμένου να υπάρχει σωστή χαρτογράφηση του τμήματος για το βήμα περίληψης που ακολουθεί.

4.4.1. Υλοποίηση Τμηματοποίησης

Η τμηματοποίηση του κειμένου αφορά τη δημιουργία τμημάτων του κειμένου, τα οποία να είναι κατάλληλα για τη δημιουργία περιλήψεων με βάση τα όρια που θέτει το μοντέλο περίληψης. Η διαδικασία τμηματοποίησης υλοποιείται με τον παρακάτω κώδικα.

Η μέθοδος splittext ανήκει στην κλάση Summarizer, η οποία κάνει την τμηματοποίηση, το tokenization και υλοποιεί την αφαιρετική περίληψη.

[4] Η σχετική συζήτηση στο χώρο συζήτησης του huggingface.co (https://discuss.huggingface.co/t/summarization-on-long-documents/920/28)

Το κείμενο χωρίζεται σε παραγράφους με τη μέθοδο split της βιβλιοθήκης regex (https://regexlib.com/)[5] και γίνεται tokenization. Σε αυτή τη φάση αφαιρούνται και παράγραφοι με πολύ μικρό μήκος.

```
def SplitText(self, text, max_token_limit):
    # τμηματοποίηση κειμένου. Το μέγεθος του κειμένου δεν ξεπερνά τα 1024 tokens

    # χωρισμός παραγράφων (νέα γραμμή)
    paragraphs = text.split("/n")

    segments = []
    current_segment = ""
    for paragraph in paragraphs:
        # αν η παράγραφος είναι πολύ μικρή δεν κρατιέται
        if len(paragraph) < 50:
            continue
```

στη συνέχεια υπολογίζεται το μήκος της παραγράφου, προσθέτοντας διαδοχικά tokens. Αν το μήκος σε tokens ξεπερνά το 1024,

```
tokens = self.tokenizer.tokenize(paragraph)
paragraph_len = 0
# μήκος παραγράφου σε tokens
for token in tokens:
    paragraph_len += len(token)
    # αν το μήκος της παραγράφου ξεπερνάει το όριο
    #τότε χωρίζουμε σε προτάσεις
if paragraph_len > max_token_limit:
```

τότε η παράγραφος χωρίζεται σε προτάσεις

```
# χωρισμός της παραγράφου σε προτάσεις
sentences = re.split(r'(?<=[.!?])\s+', paragraph)
```

Η διαδικασία επαναλαμβάνεται για τις προτάσεις της παραγράφου και στη συνέχεια για τις υπόλοιπες παραγράφους, και δημιουργούνται Segments.

```
            sentence_len += len(token) + 1

        if sentence_len + len(current_segment) < max_token_limit:
        # αν το μήκος των προτάσεων του segment επαρκεί, προσθέτω την πρόταση στο segment
            for token in tokens:
                current_segment += token
        else:
            # αν όχι προσθέτω το segment σε λίστα και δημιουργώ νέο
            segments.append(current_segment.strip())
            current_segment = ""
    else:
```

[5] Η σχετική συζήτηση στο χώρο συζήτησης του huggingface.co (https://discuss.huggingface.co/t/summarization-on-long-documents/920/28)

```
        if paragraph_len + len(current_segment) < max_token_limit:
            # αν το μήκος δεν ξεπερνάει το όριο
            # τότε προστίθεται στο segment
            for token in tokens:
                current_segment += token
        else:
            segments.append(current_segment.strip())
            current_segment = ""

# αν έχει περισσέψει segment, το προσαρτώ στη λίστα
if current_segment:
    segments.append(current_segment.strip())
```

Η μέθοδος επιστρέφει μια λίστα η οποία περιέχει τα segments του κειμένου.

4.4.2. Υλοποίηση Tokenization

Για το tokenization των κειμένων, τόσο στο στάδιο της τμηματοποίησης, όσο και πριν την περίληψη του κειμένου, χρησιμοποιήθηκε ο tokenizer του μοντέλου, όπως άλλωστε προτείνεται και από τους δημιουργούς του (https://huggingface.co/docs/transformers/main/model_doc/pegasus)

```
self.tokenizer = PegasusTokenizer.from_pretrained("google/pegasus-cnn_dailymail")
```

η χρήση τους έχει το πλεονέκτημα ότι πραγματοποιεί και τα σημαντικά στάδια της προεπεξεργασίας.

4.5. Αφαιρετική περίληψη

Για την περίληψη χρησιμοποιήθηκε η μέθοδος του μοντέλου, μετά από πειραματισμό στο συνδυασμό των ρυθμίσεων του. Οι ρυθμίσεις αφορούν τον ορισμό length_penalty (μεγαλύτερη περίληψη μπορεί να έχει επίπτωση στη συνοχή), το μέγιστο μήκος της περίληψης max_length, το οποίο ορίζουμε στο 20% του μήκους του αρχικού κειμένου, κλπ. Το αποτέλεσμα της περίληψης κάθε τμήματος τηρήθηκε σε μια λίστα.

Στη συνέχεια, η λίστα συνενώθηκε σε ένα ενιαίο κείμενο, και με την ίδια μέθοδο πραγματοποιήθηκε περίληψη στις περιλήψεις. Σε περίπτωση που το προς περίληψη κείμενο ξεπερνούσε τα όρια του μοντέλου, η διαδικασία της τμηματοποίησης επαναλήφθηκε.

Για να ελέγξουμε το μέγεθος των παραγόμενων περιλήψεων, ορίσαμε ένα μέγιστο αριθμό χαρακτήρων. Αν το μήκος της περίληψης ξεπερνά αυτόν τον αριθμό, τότε η διαδικασία περίληψης επαναλαμβάνεται. Αυτό έχει επίπτωση στην ποιότητα της περίληψης, όμως

αποφεύγουμε τις μεγάλες περιλήψεις που δεν εξυπηρετούν το στόχο της εύκολης ανασκόπησης από το χρήστη.

4.5.1. Υλοποίηση αφαιρετικής περίληψης

Για την υλοποίηση του μηχανισμού της αφαιρετικής περίληψης, δημιουργήθηκε η κλάση PegasusSummarizer.

Κατά τη δημιουργία instance της κλάσης ορίζεται και ποιο μοντέλο αφαιρετικής περίληψης πρόκειται να χρησιμοποιηθεί.

```
lass PegasusSummarizer:
    def __init__(self, model):
        self.model = model
        self.tokenizer = PegasusTokenizer.from_pretrained("google/pegasus-cnn_dailymail")
        self.summarizer = PegasusForConditionalGeneration.from_pretrained("google/pegasus-cnn_dailymail")
```

Για την περίληψη του κειμένου καλείται η μέθοδος summarize με είσοδο το κείμενο, και το όριο χαρακτήρων της περίληψης. Θα κληθεί η Summarize_Chunks, η οποία παράγει την περίληψη του κειμένου. Στη συνέχεια γίνεται καθαρισμός της περίληψης από ειδικούς χαρακτήρες (μέθοδος TextProcessor.Replace) και στη συνέχεια ελέγχεται το αν πρέπει να επαναληφθεί η διαδικασία περίληψης, με βάση το όριο των χαρακτήρων.

```
def summarize(self, text, countchar):
    # περίληψη με τμηματοποίηση
    summary_text, icount = self.Summarize_Chunks(text)
    print(summary_text)
    # αφαίρεση <n> ή άλλων χαρακτήρων
    summary_text = TextProcessor.Replace(summary_text)
    # αν το μήκος της περίληψης > από το όριο και προέρχεται από πολλά κομμάτια
    # τότε ξανά περίληψη για ομογενοποίηση
    if len(summary_text) > countchar and icount > 1:
        summary_text, icount = self.Summarize_Chunks(summary_text)

    return summary_text
```

Αρχικά καλείται η μέθοδος Summarize_Chunks, η οποια περιέχει την κλήση της μεθόδους τμηματοποίησης

```
def Summarize_Chunks(self, text):

    # καλείται η μέθοδος για την τμηματοποίηση και επιστρέφει λίστα με τα τμήματα
    chunks = self.SplitText(text, 1024)
    print(len(chunks))
```

και στη συνέχεια την επαναληπτική κλήση της μεθόδου Summarize_Peg για την περίληψη κάθε τμήματος.

```
for chunk in chunks:
    print(len(chunk))
    # κάθε τμήμα γίνεται περίληψη
```

```
    summary_text = self.Summarize_Peg(chunk)
    Helpers.writetotxt("segmentsummary: ", summary_text)
    # οι περιλήψεις μπαίνουν σε λίστα
    summaries.append(summary_text)
```

το αποτέλεσμα αυτής της διαδικασίας είναι μια λίστα με τις περιλήψεις των τμημάτων, οι οποίες στο επόμενο στάδιο θα συνενωθούν σε ένα κείμενο.

```
for summary in summaries:
    # συνένωση σε ένα κείμενο
    summarized = TextProcessor.concatstrings(summarized,summary)

final = summarized

return final, len(chunks)
```

Στη μέθοδο Summarize_Peg επαναλαμβάνεται το tokenization για το συγκεκριμένο τμήμα εγγράφου, η δημιουργία της περίληψης και η αποκωδικοποίηση του αποτελέσματος σε κείμενο.

```
def Summarize_Peg(self, text):
    tokenizer = self.tokenizer
    summarizer = self.summarizer
    # ορισμός μέγιστου μήκους περίληψης
    max_length = int(0.2 * len(text))
    # ορισμός μέγιστων ορίων για κάθε μοντέλο,
    # ώστε να μην είναι μικρότερα από τα min
    if self.model == 1:
        if max_length < 35:
            max_length = 35
    elif self.model == 2:
        if max_length < 60:
            max_length = 60
    # ξαναγίνεται tokenization
    tokens = tokenizer(text, truncation=False, padding="longest", return_tensors="pt")
    # παραγωγή περίληψης
    summary = summarizer.generate(**tokens, length_penalty=2.0, max_length=int(max_length),
                                  early_stopping=True, num_beams=5, no_repeat_ngram_size=2)
    summary_text = tokenizer.decode(summary[0], skip_special_tokens=True)


    return summary_text
```

4.6. Αποσπασματική περίληψη

Για τη δημιουργία των αποσπασματικών περιλήψεων που χρησιμοποιείται στο πρώτο στάδιο της μίας από τις 2 εναλλακτικές υλοποιήσεις χρησιμοποιήθηκε η εφαρμογή του αλγόριθμού Text Rank της βιβλιοθήκης sumy (https://github.com/miso-belica/sumy) στο πρώτο στάδιο της διαδικασίας, δηλαδή στην περίληψη των άρθρων.

Η επιλογή έγινε μεταξύ των υλοποιήσεων αποσπασματικής περίληψης που περιγράφονται στην παράγραφο 2.4.1. Με δεδομένη τη δομή των ειδησεογραφικών άρθρων, δε θεωρούμε ότι μπορούμε να βασιστούμε στη συχνότητα εμφάνισης συγκεκριμένων λέξεων ή φράσεων ή θεματικών, γιατί με αυτό τον τρόπο ενδέχεται να χαθούν σημαντικές πληροφορίες που περιγράφονται σε ένα συνοπτικό ειδησεογραφικό άρθρο, ειδικά όταν αυτό είναι χωρισμένο σε νοηματικές ενότητες.

Επίσης, στη δημοσίευση των Harinatha, Sreeya, Tasara, Beauty, & Qomariyah, Nunung Nurul (Harinatha, Sreeya, Tasara, Beauty, & Qomariyah, Nunung Nurul, 2021), προκρίνεται η χρήση του TextRank για την αποσπασματική περίληψη ειδησεογραφικών άρθρων μεταξύ των LexRank TextRank και LSA, αλλά και κατόπιν σύγκρισης με το BERT, που όπως είδαμε παραπάνω χρησιμοποιεί βαθιά μάθηση. Το BERT παράγει μεν πιο συνεκτικές περιλήψεις, όμως ο TextRank συμπεριλαμβάνει περισσότερες πληροφορίες, πράγμα που αξιολογήσαμε ως πιο σημαντικό στο στάδιο της περίληψης κάθε άρθρου ξεχωριστά. Ένα ακόμα πλεονέκτημα του TextRank είναι ότι απαιτεί ελάχιστο χρόνο σε σχέση με το BERT.

Εδώ ενδέχεται να απαιτείται διόρθωση της υλοποίησης καθώς εσφαλμένα θεωρήσαμε ότι κατά το tokenization πραγματοποιούνται όλες οι υπόλοιπες εργασίες προεπεξεργασίας και καθαρισμού του κειμένου.

4.6.1. Υλοποίηση αποσπασματικής περίληψης

Για το σκοπό αυτό δημιουργήθηκε η κλάση ExtrSummarizer και η μέθοδος extractive_summarization_one, η οποία δέχεται ως είσοδο το κείμενο. Για να παραχθεί η περίληψη, πρέπει να ορίσουμε το μέγεθος της σε αριθμό προτάσεων, οπότε, χωρίζουμε το κείμενο σε προτάσεις, και το μέγεθος της περίληψης στο 20% του αρχικού αριθμού προτάσεων, με ελάχιστο όριο τις 5 προτάσεις, για να αποφύγουμε περιλήψεις της μίας πρότασης.

```
class ExtrSummarizer:
    #extractive summarization με TextRamk
    def __init__(self):
        self.summarizer = TextRankSummarizer()

    def extractive_summarization_one(self, text):
    # tokenization
        parser = PlaintextParser.from_string(text, Tokenizer('english'))

        # μέτρηση προτάσεων αρχικού κειμένου.
        # θέλουμε η περίληψη να περιέχει το 20% του αρχικού
        summary = ""
```

```
        sentences = re.split(r'(?<=[.!?])\s+', text)
        or_sentence_count = len(sentences)
        sentence_count = int(or_sentence_count * 0.2)
        # για να μην πάρουμε πολύ μικρή περίληψη
        if sentence_count < 5:
            sentence_count = 5
        # περίληψη κειμένου
        summary_sum = self.summarizer(parser.document, sentence_count)
        # συνένωση των προτάσεων σε κείμενο
        summary = ' '.join(map(str, summary_sum))

        return summary
```

4.7. Άλλες κλάσεις

Παραπάνω περιγράφηκαν οι πιο σημαντικές κλάσεις της υλοποίησης. Οι λειτουργία τους υποστηρίζεται από κάποιες κλάσεις ακόμα τις οποίες θα δούμε σε συντομία.

ModeSelector: Η κλάση αυτή είναι η πρώτη κλάση που καλείται μετά την ολοκλήρωση της διαδικασίας συλλογής των άρθρων. Ουσιαστικά εδώ ορίζονται οι επιμέρους παράμετροι όπως είναι αυτές της επιλογής μοντέλου για τη δημιουργία των περιλήψεων.

```
class ModeSelector:
    def __init__(self, df, mode):
        self.df = df
        self.mode = mode
        # δημιουργούμε τους summarizers που πρόκειται να χρησιμοποιήσουμε
        # 1 pegasus/cnn_dailymail, 2: facebook/bart-large-cnn
        if mode in (1, 5):
            self.summarizer = PegasusSummarizer(1)
        if mode in (11, 15):
            self.summarizer_ex = ExtrSummarizer()
            self.summarizer = PegasusSummarizer(1)
```

Ανάλογα με την επιλογή του mode καλείται η κατάλληλη μέθοδος για τη δημιουργία της περίληψης. Για παράδειγμα, για αποσπασματική περίληψη 2 σταδίων πρώτα κάθε άρθρου ξεχωριστά και στη συνέχεια συνένωση των περιλήψεων:

```
def mode_5(self, df, mode):
    Helpers.writetotxt("mode: ", mode)
    # περίληψη κάθε άρθρου ξεχωριστά
    text = ''
    new_rows = []
    for ind in df.index:
        Helpers.writetotxt("article: ", ind)
        # αφαίρεση κενών γραμμών
        article = TextProcessor.RemoveEmptyLines(df['Article'][ind])
        print(ind)
        # ορισμός αυθαίρετου ορίου κειμένου περίληψης
        # αν ξεπεραστεί, τότε ξαναγίνεται περίληψη με τον ίδιο μηχανισμό
```

```
        countchar = 4000
        summary = self.summarizer.summarize(article, countchar)
        # αντικατάσταση χαρακτήρων που παράγει ο summarizer μεταξύ των λέξεων ή των προτάσεων
        # <n>, Ġ, Âł)
        summary = TextProcessor.Replace(summary)
        # δημιουργία εγγραφής για το dataframe και τοποθέτηση της σε λίστα
        new_row = {'mode': mode, 'num': ind, 'title': df['Title'][ind], 'summary': summary}
        new_rows.append(new_row)
        print('Article summary')
        print(summary)
        # συνένωση περιλήψεων
        text = TextProcessor.concatstrings(text, summary)
    print('Summaries')
    print(text)
    # εγγραφή των αποτελεσμάτων στο dataset
    iret = Helpers.writetodataset_mul(new_rows)
    if iret == -1:
        print("Write to dataframe failed")
    countchar = 6000
    # περίληψη των περιλήψεων με όριο χαρακτήρων της παραγόμενης
    summary = self.summarizer.summarize(text, countchar)
    # αντικατάσταση χαρακτήρων που παράγει ο summarizer μεταξύ των λέξεων ή των προτάσεων
    summary = TextProcessor.Replace(summary)
    # εγγραφή της τελικής περίληψης στο dataframe
    iret = Helpers.writetodataset(mode, 1000, "SummaryofSummaries", summary)
    if iret == -1:
        print("Write to dataframe failed")
    print("Final summary")
    print(summary)
```

TextProcessor: Η κλάση αυτή περιλαμβάνει τις μεθόδους για την επεξεργασία των εγγράφων πριν την περίληψη τους, όπως την αφαίρεση κενών γραμμών (RemoveEmptyLines), τη συνένωση κειμένων (concatstrings) την αφαίρεση ανεπιθύμητων χαρακτήρων και ενοτήτων (Replace και remove_unwanted αντίστοιχα), κλπ.

Helpers: Η κλάση Helpers περιέχει τις μεθόδους για την εγγραφή των αποτελεσμάτων ή των αρχικών άρθρων σε αρχεία. Τα αρχεία που τηρούνται περιγράφονται στον παρακάτω πίνακα:

Αρχείο	**Περιεχόμενο**
Dataframe_initial.csv	Περιέχει τα αποτελέσματα που δίνει κατά την αναζήτηση το googlenews (Έχει χρησιμοποιηθεί μόνο για την επίβλεψη της διαδικασίας κατά την υλοποίηση)
Dataframe_new.csv	Περιέχει τα αποτελέσματα αναζήτησης που απομένουν μετά την ολοκλήρωση των ελέγχων. (Έχει χρησιμοποιηθεί μόνο για την επίβλεψη της διαδικασίας κατά την υλοποίηση)

Articles.csv	Περιέχει τα ανακτημένα άρθρα μαζί με τα μεταδεδομένα τους.
Segments.txt	Είναι μια μορφή log των ενδιάμεσων σταδίων τμηματοποίησης και περίληψης τμημάτων
Results.csv	Περιέχει τις περιλήψεις των άρθρων και τις τελικές περιλήψεις.

Πίνακας 2: Αρχεία που δημιουργούνται κατά την εκτέλεση

5. Ανάλυση αποτελεσμάτων

Σε αυτή την ενότητα θα παρουσιάσουμε την ανάλυση των αποτελεσμάτων της υλοποίησης. Το δείγμα μας είναι πέντε πρόσφατα άρθρα (παρατίθενται στο παράρτημα Β) με θέμα "Greek elections". Τα συγκεκριμένα άρθρα συγκεντρώθηκαν μετά τις εκλογές της 23ης Ιουνίου.

5.1. Συλλογή άρθρων

Η διαδικασία της συλλογής άρθρων λειτούργησε κανονικά, συγκεντρώνοντας πέντε μοναδικά φαινομενικά άρθρα, τα οποία ανακτήθηκαν και αποθηκεύτηκαν στο αρχείο των άρθρων. Από την ανάλυση των csv αρχείων, προκύπτει ότι από τα έξι άρθρα που συγκεντρώθηκαν από τη διαδικασία αναζήτησης και ανάκτησης διατηρήθηκαν τα πέντε, με το ένα να απορρίπτεται λόγω πολύ μικρού μεγέθους.

Όμως, όπως γίνεται φανερό από τη μελέτη των κειμένων των άρθρων, τα δύο από αυτά (άρθρα 3 και 4), παρόλο που είναι δημοσιευμένα σε διαφορετικά μέσα, με ελαφρώς διαφοροποιημένο τίτλο είναι κατά πολύ μικρότερες σε μέγεθος εκδόσεις ενός μεγαλύτερου άρθρου (άρθρο 1). Αυτό μας οδηγεί στο συμπέρασμα ότι θα πρέπει να τροποποιηθεί ο έλεγχος ομοιότητας, είτε μεταβάλλοντας την οριακή τιμή για την ομοιότητα συνημιτόνου, είτε επιλέγοντας μια άλλη μέθοδο, η οποία να μπορεί να εντοπίσει τα κοινά περιεχόμενα μεταξύ των άρθρων και να απορρίψει άρθρα που δεν προσφέρουν νέα πληροφορία.

5.2. Προεπεξεργασία και τμηματοποίηση

Οι προκαθορισμένες εργασίες προεπεξεργασίας λειτούργησαν αρκετά καλά, όπως θα γίνει εμφανές από τη σύγκριση των αρχικών άρθρων που δεν έχουν υποστεί προεπεξεργασία και τον τμημάτων τους. Δεν παρατηρήθηκαν απώλειες κατά την τμηματοποίηση, ούτε κατάτμηση προτάσεων.

Βελτίωση χρειάζεται η αφαίρεση περιεχομένου που δεν είναι σχετική με το άρθρο, αν και στο υπο αξιολόγηση παράδειγμα αυτά τα τμήματα δεν επηρέασαν το τελικό αποτέλεσμα.

5.3. Δημιουργια περίληψης

Δυστυχώς δεν υπήρχε η δυνατότητα εφαρμογής κάποιας μετρικής αξιολόγησης, γιατί οι δημοφιλείς μετρικές όπως οι Rouge δεν προορίζονται για την αξιολόγηση περιλήψεων πολλαπλών άρθρων. Επίσης δεν υπάρχουν επαρκή datasets για την περίληψη πολλαπλών ειδησεογραφικών άρθρων, οπότε η ανάλυση των αποτελεσμάτων γίνεται εμπειρικά.

Δεν θα παραθέσουμε εδώ ολόκληρα τα κείμενα των άρθρων, αλλά μόνο τα συμπεράσματα από τη διαδικασία. Τα πλήρη αποτελέσματα βρίσκονται στο παράρτημα Β.

Στον πίνακα που ακολουθεί, παρουσιάζονται τα σημαντικότερα προβλήματα για την περίληψη με το μοντέλο Pegasus.

Τρόπος περίληψης	**Αποτελέσματα**	**Παράδειγμα στο τελικό κείμενο**	**Αρχικό κείμενο παραδείγματος**
Χρήση Pegasus για περίληψη κάθε άρθρου, και για τη συνολική περίληψη	Αλλαγή θέσης φράσης σε σημείο που δεν έβγαζε νόημα σε μικρό βαθμό	Greek far-right groups swept up over 12% of the vote in Sunday's election. The surge of three parties with their ultra-nationalist views could swing public debate at home and prove a springboard in European elections, Ilias Kasidiaris from the banned Golden Dawn party endorsed it from his prison cell, mirroring the rise of populist politicians across Europe.	Greek far-right groups swept up over 12% of the vote in Sunday's election, mirroring the rise of populist and ultra-nationalist politicians across Europe. The surge of three parties with their ultra-nationalist views - including 'Spartans' which barely registered in polls until Ilias Kasidiaris from the banned Golden Dawn party endorsed it from his prison cell
	Αφαίρεση πληροφορίας που μεταφέρεται ως άποψη τρίτου και όχι του γράφοντος (όχι απαραίτητα πρόβλημα)		The deaths of hundreds of migrants when their boat sank off Greece while being tracked by the Greek coastguard has sharpened the debate over immigration. Despite mourning the tragedy, many Greeks want to halt the stream of migrants. "The re-emergence of the far right is a byproduct of a political strategy of the New Democracy government which tried to appeal to the centre with an agenda of economic liberalism and, at the same time to the far right with an agenda of law and order and anti-immigrant discourse," said Akritas Kaidatzis…
	Αφαίρεση τμημάτων που δεν έχουν σχέση με το κείμενο		Reporting By Michele Kambas and Renee Maltezou, writing by Michele Kambas, editing by Edmund Blair and Christina Fincher Our Standards: The Thomson Reuters Trust Principles.

	Λείπει η πληροφορία που βρίσκεται στο τέλος του τμήματος της περίληψης, όμως το υπόλοιπο κομμάτι μιλούσε για την άνοδο της ακροδεξιάς.		While the centre-right New Democracy party of Kyriakos Mitsotakis stormed to victory in the June 25 Greek vote, winning 158 seats in the 300-seat parliament, the Spartans emerged as the fifth largest group.
	Δημιουργία ανακριβούς πληροφορίας	Kasidiaris is serving a 13-year jail term for his leadership role in Golden Dawn.<n>Golden Dawn was declared a criminal gang linked to hate crimes in a 2020 court ruling, and was banned in Greece in 2011 <n>It was once Greece's third largest party, but has been on the wane since the financial crisis of 2008 The return to power for the Left is likely to be a slow process and avoiding the repetition of the fall of Syrizo should be the main objective, argues Pavlos Tsimas and his co-authors, in this special edition of The Left's Revolution: How to Win the Fight for a Just and Progressive Europe	Kasidiaris is serving a 13-year jail term for his leadership role in Golden Dawn, once Greece's third largest party. It was declared a criminal gang linked to hate crimes in a 2020 court ruling. The return to power for the Left is likely to be a slow process, and avoiding the repetition of the fall of Syriza should be the main objective. The Rise and the Fall of Syriza Syriza's 2015 rise to power was the most significant political shock in Europe in recent memory.

	Επανάληψη πληροφορίας	Tsipras, 48, served as Greece's prime minister from 2015 to 2019. He forged a more cohesive party, taking it from a small political group to general election victory in 2015. Tsipras is expected to stay on as leader for several weeks until his successor is elected by his rank-an-file membership. Alexis Tsipras, 48, served as Greece’s prime minister from 2015 to 2019.	

Πίνακας 3: Ανάλυση αποτελεσμάτων

Παρόλα αυτά το τελικό κείμενο, με εξαίρεση κάποια μικρά ζητήματα, μας δίνει μια καλή εικόνα για το περιεχόμενο των άρθρων.

Με μώβ σημειώνονται οι ανακριβείς πληροφορίες, με μπλε χρώμα η προσθήκη πληροφοριών, ενώ με πορτοκαλί σημειώνεται η επανάληψη.

«Greek far-right groups swept up over 12% of the vote in Sunday's election. The surge of three parties with their ultra-nationalist views could swing public debate at home and prove a springboard in European elections Ilias Kasidiaris from the banned Golden Dawn party endorsed it from his prison cell, mirroring the rise of populist politicians across Europe.

Tsipras, 48, served as Greece's prime minister from 2015 to 2019. He forged a more cohesive party, taking it from a small political group to general election victory in 2015. Tsipras is expected to stay on as leader for several weeks until his successor is elected by his rank-an-file membership.

After four years of major political scandals, New Democracy won the elections with a twenty-point margin against Syriza. The return to power for the Left is likely to be a slow process and avoiding the repetition of the fall of Syrizo should be the main objective, argues Pavlos Tsimas and his co-authors.

Syriza's narrative was that after the completion of certain austerity-oriented agreements, it was time to implement their own, more progressive agenda and return the accumulated budget surplus to society. Unsurprisingly, the majority of the voters were not convinced, and New Democracy won themajority of seats in the parliament, forming the first single-party government since 2009. New Democracy also spent twenty million euros on COVID-19 public Health media campaigns.

Evidence suggests Greek intelligence services spied on journalists and political opponents. Yanis Varoufakis: New Democracy took advantage of turn of events and argued that Syriza is not the party that cares about working-class people' Varoufikis: Financial markets would have reacted aggressively against the election of a left-wing government, leading to higher borrowing rates.

Alexis Tsipras, 48, served as Greece's prime minister from 2015 to 2019. His left-wing Syriza party received just under 18% of the vote, while the winning New Democracy party topped 40% in Sunday's general election Mr Tsipras is expected to stay on as leader until his successor is elected by the party membership.»

Τα ίδια προβλήματα φαίνεται να παρουσιάζει και η εφαρμογή Extractive Summarization σε συνδυασμό με το Pegasus, αυξημένα όμως ως προς την ανακριβή πληροφορία. Όμως σε αυτή την εφαρμογή συμπεριλήφθηκε στο κείμενο η μεταφορά γνώμης που δεν είχε συμπεριληφθεί με τον προηγούμενο τρόπο. Ακολουθεί το κείμενο της τελικής περίληψης:

Greek far-right groups swept up over 12% of the vote in Sunday's election. The surge of three parties with their ultra-nationalist views could swing public debate at home and prove a springboard in European elections, Ilias Kasidiaris from the banned Golden Dawn party endorsed it from his prison cell, mirroring the rise of populist politicians across Europe.

Tsipras served as Greece's prime minister from 2015 to 2019. He forged a more cohesive party, taking it from a small political group to general election victory in 2015 on a pledge to push back on harsh austerity measures demanded by bailout lenders from other euro zone members and the International Monetary Fund.

Commentators blamed Syriza's poor election result on the party's largely-negative campaign, the resurgence of the traditionally-strong Socialist party Pasok, and the appearance of splinter parties. After five years of irrelevant austerity policies, Greek people voted in favor of a potential rupture with the EU, writes Pavlos Tsimas, who argues that the return to power for the Left is likely to be a slow process, with a main objective of avoiding the repetition of this fall of Tsipras' party, is the main goal.

Tspras's Syriza re-won the majority of the parliament and formed a new coalition government. A few months after the reelection of New Democracy, the pandemic started and the new government had to abandon its austerity-oriented agenda out of necessity Social unrest caused by the party's authoritarian public health approach pushed the government to spend a big share of its surplus on wage subsidies and social benefits.

"The time has come to start a new cycle," Tsipras said in a televised address. Led by Tsipras, Syriza stormed to power in 2015 at the height of Greece's deep economic crisis, riding a wave of anti-austerity and Anti-bailout anger among Greeks.

Mr Tsipras, 48, served as Greece's prime minister from 2015 to 2019. During that time, the country struggled to remain in the euro zone and end a series of international bailouts.

Ο μηχανισμός Extractive Summarization θα πρέπει να συμπεριλάβει στάδια προεπεξεργασίας του κειμένου τα οποία θα μπορούσαν να βελτιώσουν το παραγόμενο αποτέλεσμα ως προς την ύπαρξη μη σχετικών τμημάτων στο κείμενο (πχ. ATHENS, June 29 (Reuters) -)

Όμως θεωρούμε ότι το πρόβλημα στην αύξηση της παραποιημένης πληροφορίας που βλέπουμε παραπάνω έγκειται στο ότι ακολουθεί βήμα αφαιρετικής περίληψης σε κείμενο που έχει παραχθεί με επιλογή αυτούσιων προτάσεων.

Παρακάτω παρατίθεται ένα παράδειγμα του αποτελέσματος της αποσπασματικής περίληψης του άρθρου 0, το οποίο είναι αρκετά ικανοποιητικό σε σχέση με το πρωτότυπο

ATHENS, June 29 (Reuters) - With backing from a politician jailed for leading a party declared a criminal gang, Greek far-right groups swept up over 12% of the vote in Sunday's election, mirroring the rise of populist and ultra-nationalist politicians across Europe. The surge of three parties with their ultra-nationalist views - including 'Spartans' which barely registered in polls until Ilias Kasidiaris from the banned Golden Dawn party endorsed it from his prison cell -- could swing public debate at home and prove a springboard in European elections. While the centre-right New Democracy party of Kyriakos Mitsotakis stormed to victory in the June 25 Greek vote, winning 158 seats in the 300-seat parliament, the Spartans emerged as the fifth largest group. "The re-emergence of the far right is a byproduct of a political strategy of the New Democracy government which tried to appeal to the centre with an agenda of economic liberalism and, at the same time to the far right with an agenda of law and order and anti-immigrant discourse," said Akritas Kaidatzis, assistant professor of constitutional law at the Aristotle University in Thessaloniki. Support for Germany's anti-immigrant AfD party is at its highest since the wake of Europe's migrant crisis in 2018, in Italy a former far-right activist is prime minister and in Sweden a far-right party has joined the government.

6. Συμπεράσματα και μελλοντική έρευνα

Σκοπός του παρόντος ήταν η υλοποίηση ενός μηχανισμού για την αυτόματη περίληψη πολλαπλών κειμένων, τα οποία συγκεντρώνονται αυτόματα από διαδικτυακές πηγές, με αυτόματη αναζήτηση με βάση συγκεκριμένα κριτήρια. Για την επίτευξη του δημιουργήσαμε μια διαδικασία αναζήτησης και ανάκτησης άρθρων από το διαδίκτυο. Στη συνέχεια δημιουργήσαμε μια διαδικασία αυτόματης περίληψης κειμένων με δύο διαφορετικούς τρόπους. Ο πρώτος αφορά την εφαρμογή αφαιρετικής περίληψης σε κάθε άρθρο, στη συνέχεια δημιουργία ενός ενιαίου κειμένου που περιλαμβάνει τις περιλήψεις και εκ νέου περίληψη. Ο δεύτερος ακολουθεί τα ίδια βήματα, μόνο που στην αρχή αντί αφαιρετικής περίληψης εφαρμόζεται αποσπασματική περίληψη.

Η συλλογή άρθρων από το διαδίκτυο και η ανάκτηση τους απαιτεί μια επαναληπτική διαδικασία η οποία περιλαμβάνει βήματα ελέγχου. Αυτό είναι αναγκαίο καθώς πολλές φορές τα επιστρεφόμενα αποτελέσματα είτε δεν αντιστοιχούν σε έγκυρα url είτε δεν επιτρέπουν την αυτόματη ανάκτηση του περιεχομένου τους.

Η συγκέντρωση μοναδικών άρθρων απαιτεί τη σύγκριση των άρθρων που ανακτώνται για να αποφύγουμε την ύπαρξη διπλοτύπων στη συλλογή. Η σύγκριση των άρθρων με χρήση της ομοιότητας συνημιτόνου που εφαρμόσαμε, ενώ αφαίρεσε τις πανομοιότυπες εγγραφές, δεν εντόπισε εκείνα τα κείμενα που αποτελούν υποσύνολα άλλων. Θα πρέπει λοιπόν να εφαρμοστεί μια διαφορετική τεχνική για τον εντοπισμό όλων των περιπτώσεων επανάληψης του ίδιου περιεχομένου.

Πριν την εφαρμογή του μηχανισμού περίληψης εφαρμόστηκε περιορισμένη προετοιμασία του κειμένου, με δεδομένο ότι τα περισσότερα στάδια της προεπεξεργασίας πραγματοποιούνται από το ίδιο το μοντέλο αφαιρετικής περίληψης. Αυτό θα πρέπει να αναθεωρηθεί, καθώς δεν ισχύει το ίδιο στην περίπτωση της εφαρμογής εξαγωγικής περίληψης.

Η περίληψη με χρήση του προεκπαιδευμένου μοντέλου transformers που χρησιμοποιήθηκε έχει περιορισμό ως προς το μέγεθος του αρχικού κειμένου. Απαιτείται η τμηματοποίηση του εγγράφου πριν την περίληψη του. Η τμηματοποίηση που εφαρμόστηκε ήταν αποτελεσματική και χωρίς απώλειες στο περιεχόμενο των άρθρων.

Στο αποτέλεσμα των παραγόμενων περιλήψεων παρατηρήθηκαν κάποια προβλήματα, όπως η αλλοίωση του νοήματος από αναδιάταξη φράσεων, ή η δημιουργία ανακριβούς

πληροφορίας. Επίσης, παρατηρήθηκε σε μικρό βαθμό επανάληψη σημείων του κειμένου, κατά την εφαρμογή δύο σταδίων αφαιρετικής περίληψης.

Το τελικό αποτέλεσμα ήταν πιο ακριβές με εφαρμογή δύο σταδίων αφαιρετικής περίληψης, αλλά περιείχε επανάληψη πληροφορίας. Το αποτέλεσμα της εφαρμογής αποσπασματικής περίληψης και στη συνέχεια αφαιρετικής, δεν είχε επαναλαμβανόμενη πληροφορία, και είχε περισσότερη πληροφορία, αλλά παρουσίασε αυξημένη ανακρίβεια.

Για τη βελτίωση της διαδικασίας, προτείνουμε κατ' αρχήν τη δημιουργία άλλου μηχανισμού σύγκρισης κειμένων που να μπορεί να εντοπίσει και να αφαιρέσει τα αλληλεπικαλυπτόμενα κείμενα και όχι μόνο τα πανομοιότυπα. Αυτό θα βελτιώσει σημαντικά την περιεκτικότητα και τη συνοχή του τελικού αποτελέσματος, άλλα και θα βοηθήσει στη βελτιστοποίηση της διαδικασίας συγκέντρωσης διαφορετικών μεταξύ τους άρθρων.

Επίσης σημαντική προσθήκη είναι ένας μηχανισμός κατηγοριοποίησης των άρθρων με βάση το περιεχόμενο τους. Αυτό θα μπορούσε να βοηθήσει στη συνοχή του τελικού αποτελέσματος, με τη δημιουργία ενδιάμεσων περιλήψεων των ομοειδών κειμένων πριν την τελική περίληψη.

Βιβλιογραφία

Ακολουθούν οι βιβλιογραφικές αναφορές (πηγές) της Εργασίας.

Banerjee, S, Mitra, P, & Sugiyama, K. (2016). Multi-document abstractive summarization using ILP based multi-sentence compression. *IJCAI'15 Proceedings of the 24th International Conference on Artificial Intelligence* (σσ. 1208-1214). AAAI Press. doi:https://doi.org/10.48550/arXiv.1609.07034

Chauhan, K. (2018, August 6). Ανάκτηση από Medium.com: https://medium.com/jatana/unsupervised-text-summarization-using-sentence-embeddings-adb15ce83db1

Chopra, S., Auli, M, & Rush, A. (2016). Abstractive Sentence Summarization with Attentive Recurrent Neural Networks. *Proceedings 2016 Conference of the North American Chapter of the Association for Computational Linguistics,* (σσ. 93–98). California: Human Language Technologies.

Evdaimon, I., Abdine, H., Xypolopoulos, C., Outsios, S., Vazirgiannis, M., & Stamou, G. (2023). *GreekBART: The First Pretrained Greek Sequence-to-Sequence Model.* Ανάκτηση από arxiv.org: https://arxiv.org/abs/2304.00869

Fabbri, A., Li, I., She, T., Li, S., & Radev, D. (2019). Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (σσ. 19-1102). Florence: Association for Computational Linguistics. doi:10.18653/v1/P19-1102

Gahman, G., & Vinayak, E. (2023). A Comparison of Document Similarity Algorithms. *International Journal of Artificial Intelligence and Applications, 14*.

Harinatha, Sreeya, Tasara, Beauty, & Qomariyah, Nunung Nurul. (2021). Evaluating Extractive Summarization Techniques on News Articles. *International Seminar on Intelligent Technology and Its Applications (ISITIA).* doi:DOI: 10.1109/ISITIA52817.2021.9502230

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. . *Advances in Neural Information Processing Systems*, 1693-1701.

Hong, K., Marcus, M., & Nenkova, A. (2015). System Combination for Multi-document Summarization. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (σσ. 107–117). Lisbon: 2015. doi:10.18653/v1/D15-1011

Koutsikakis, J., Chalkidis, I., Malakasiotis, P., & Androutsopoulos, I. (2020). *GREEK-BERT: The Greeks visiting Sesame Street.* Ανάκτηση από arxiv.org: https://arxiv.org/abs/2008.12014

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., . . . Zettlemoyer, L. (2019). *Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.* Ανάκτηση από arxiv.org: https://arxiv.org/abs/1910.13461

Lin, Chin-Yew, & Hovy, E. (2000). The Automated Acquisition of Topic Signatures for Text Summarization. *In COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics.*

Liu, Xin, & Gong, Yihong. (2001). Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. *SIGIR '01: Proceedings of the 24th annual*

international ACM SIGIR conference on Research and development in information retrieval.

Liu, Y. (2019, September 5). *Fine-tune BERT for Extractive Summarization.* Ανάκτηση από arxiv.org: https://arxiv.org/abs/1903.10318v2

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Text. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (σσ. 404–411). Barcelona, Spain: Association for Computational Linguistics. Ανάκτηση από https://aclanthology.org/W04-3252

Miller, D. (2019). *Leveraging BERT for Extractive Text Summarization on Lectures.* Ανάκτηση από arxiv.org: https://arxiv.org/abs/1906.04165

Mishra, R., & Gayen, T. (2018). Automatic Lossless-Summarization of News Articles with Abstract Meaning Representation. *Procedia Computer Science, 135*, 178-185.

Nenkova, A. &. (2005). *The impact of frequency on summarization.*

Nenkova, A., & McKeown, K. (2011). Automatic Summarization. *Fountations and Trends in Information Retrieval, 5*(2-3), 103-233. doi:http://dx.doi.org/10.1561/1500000015

Rush, A. e. (2015). A Neural Attention Model for Abstractive Sentence Summarization. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (σσ. 379-389). Lisbon, Portugal: Association for Computational Linguistics. doi:10.18653/v1/D15-1044

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Networks, 61*, 85-117.

See, A., Liu, P., & Manning, C. (2017). *Get To The Point: Summarization with Pointer-Generator Networks.* doi:https://doi.org/10.48550/arXiv.1704.04368

Sharkar, D. (2019). *Text analytics with Python* (2nd εκδ.). Bangalore, Karnataka, India: Apress. doi:https://doi.org/10.1007/978-1-4842-4354-1

Shearing, S., Gertner, A., & Wellner, B. M. (2020). *Automated Text Summarization: A Review and Recommendations.* MITRE. Ανάκτηση από https://www.mitre.org/news-insights/publication/automated-text-summarization-review-and-recommendations

Steinberger, J., & Ježek, K. (2012). Evaluation Measures for Text Summarization. *COMPUTING AND INFORMATICS, 28*, 251–275. Ανάκτηση από https://www.cai.sk/ojs/index.php/cai/article/view/37

style, A. (2010). *δαφσδφασδ φασδφ ασδφ φ σδαα*. Ανάκτηση April 23, 2012, από http://asdfs.gg.gsf

Tan, P., Steinbach, M., Karpatne, A., & Kumar, V. (2020). *Εισαγωγή στην εξόρυξη δεδομένων.* (Β. Βερύκιος, Επιμ., & Σ. Σουραβλάς, Μεταφρ.) Θεσσαλονίκη: Εκδόσεις Τζιόλα.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., . . . Polosukhin, I. (2017). *Attention Is All You Need.* Ανάκτηση από arxiv.org: https://arxiv.org/abs/1706.03762

Wolhalander, R., Cattan, A., Ernst, O., & Dagan, I. (2022). *How "Multi" is Multi-Document Summarization*. Ανάκτηση από arxiv.com: https://arxiv.org/abs/2210.12688

Zangh, M., Gang, Z., Wanting, Y., Ningbo, H., & Wenfen, L. (2022). A Comprehensive Survey of Abstractive Text Summarization Based on Deep Learning. *Computational Intelligence and Neuroscience*, 21. doi:https://doi.org/10.1155/2022/7132226

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2016). *PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization.* doi:https://doi.org/10.48550/arXiv.1912.08777

Παράρτημα Α: Πηγαίος κώδικάς

Ακολουθεί, σε νέα σελίδα, το παράρτημα της Εργασίας.

```
import pandas as pd
from ArticleRetriever import NewsFinder
from ModeSelector import ModeSelector


if __name__ == '__main__':
    # περιοχή ορισμού παραμέτρων
    # παράμετροι αναζήτησης googlenews: γλώσσα, χρονικό διάστημα, θέμα
    lang = 'en'
    period = '7d'
    topic = 'greek elections'
    # αριθμός άρθρων που θέλουμε να συλλεχθούν
    icount = 5
    # επιλογή τρόπου λειτουργίας
    # 5: περίληψη κάθε άρθρου ξεχωριστά, συνένωση περιλήψεων σε ένα
κείμενο και περίληψη, abstractive summarization μόνο, model:
google/pegasus-cnn_dailymail
    # 15: περίληψη κάθε άρθρου ξεχωριστά extractive summy.TextRank,
συνένωση περιλήψεων σε ένα κείμενο και περίληψη abstractive
summarization model: google/pegasus-cnn_dailymail


    mode = 5

    # αν θέλω να χρησιμοποιήσω μόνο summarizer και όχι να κάνω νέα
συλλογή άρθρων (πρέπει να υπάρχει articles.csv)
    UseExistinDF = True
    if UseExistinDF is False:
    #δημιουργώ instance της κλάσης NewsFinder με τις παραμέτρους
αναζήτησης και τον αριθμό των αποτελεσμάτων
        nf = NewsFinder(lang, period, topic, icount)
    #η μέθοδος iterate δημιουργεί τη συλλογή των άρθρων
        news_df = nf.iterate()
    # για οπτικό έλεγχο του dataframe κατά την εκτέλεση. Μπορεί να
αφαιρεθεί
        print(news_df)
    else:
    # αν χρησιμοποιώ υφιστάμενο dataframe, τότε το διαβάζω
        news_df = pd.read_csv('articles.csv')
    # δημιουργία ModeSelector. Ακολουθεί διαφορετική διαδρομή για κάθε
mode
    ms = ModeSelector(news_df, mode)
    ms.selectmode()

from Processing import TextProcessor
from Helpers import Helpers
from PegasusSummarizer import PegasusSummarizer
from ExtractiveSummarizer import ExtrSummarizer


class ModeSelector:
    def __init__(self, df, mode):
        self.df = df
        self.mode = mode
        # δημιουργούμε τους summarizers που πρόκειται να
```

```
χρησιμοποιήσουμε
        # 1 pegasus/cnn_dailymail, 2: facebook/bart-large-cnn
        if mode in (1, 5):
            self.summarizer = PegasusSummarizer(1)
        if mode in (11, 15):
            self.summarizer_ex = ExtrSummarizer()
            self.summarizer = PegasusSummarizer(1)



    def mode_5(self, df, mode):
        Helpers.writetotxt("mode: ", mode)
        # περίληψη κάθε άρθρου ξεχωριστά
        text = ''
        new_rows = []
        for ind in df.index:
            Helpers.writetotxt("article: ", ind)
            # αφαίρεση κενών γραμμών
            article = TextProcessor.RemoveEmptyLines(df['Article'][ind])
            print(ind)
            # ορισμός αυθαίρετου ορίου κειμένου περίληψης
            # αν ξεπεραστεί, τότε ξαναγίνεται περίληψη με τον ίδιο
μηχανισμό
            countchar = 4000
            summary = self.summarizer.summarize(article, countchar)
            # αντικατάσταση χαρακτήρων που παράγει ο summarizer μεταξύ
των λέξεων ή των προτάσεων
            # <n>, Ġ, Âł)
            summary = TextProcessor.Replace(summary)
            # δημιουργία εγγραφής για το dataframe και τοποθέτηση της σε
λίστα
            new_row = {'mode': mode, 'num': ind, 'title':
df['Title'][ind], 'summary': summary}
            new_rows.append(new_row)
            print('Article summary')
            print(summary)
            # συνένωση περιλήψεων
            text = TextProcessor.concatstrings(text, summary)
        print('Summaries')
        print(text)
        # εγγραφή των αποτελεσμάτων στο dataset
        iret = Helpers.writetodataset_mul(new_rows)
        if iret == -1:
            print("Write to dataframe failed")
        countchar = 6000
        # περίληψη των περιλήψεων με όριο χαρακτήρων της παραγόμενης
        summary = self.summarizer.summarize(text, countchar)
        # αντικατάσταση χαρακτήρων που παράγει ο summarizer μεταξύ των
λέξεων ή των προτάσεων
        summary = TextProcessor.Replace(summary)
        # εγγραφή της τελικής περίληψης στο dataframe
        iret = Helpers.writetodataset(mode, 1000, "SummaryofSummaries",
summary)
        if iret == -1:
            print("Write to dataframe failed")
        print("Final summary")
        print(summary)

    def mode_15(self, df, mode):
        # περίληψη άρθρων extractive και μετά abstractive
        text = ''
```

```
        new_rows = []
        Helpers.writetotxt("mode: ", mode)
        for ind in df.index:
            Helpers.writetotxt("article: ", ind)
            # αφαίρεση κενών γραμμών
            article = TextProcessor.RemoveEmptyLines(df['Article'][ind])
            print(ind)
            # εδώ γίνεται πρώτα extractive. πρέπει να αφαιρεθούν τα url
            text = TextProcessor.removeurlsections(text)
            summary =
self.summarizer_ex.extractive_summarization_one(article)
            # αφαίρεση χαρακτήρων
            summary = TextProcessor.Replace(summary)
            Helpers.writetotxt("summary of article: ", summary)
            # δημιουργία εγγραφής για το dataframe και τοποθέτηση της σε
λίστα
            new_row = {'mode': mode, 'num': ind, 'title':
df['Title'][ind], 'summary': summary}
            new_rows.append(new_row)
            print('Article summary')
            print(summary)
            # συνένωση των παραγόμενων περιλήψεων
            text = TextProcessor.concatstrings(text, summary)
        print('Summaries')
        print(text)
        # εγγραφή αποτελέσματος στο dataframe
        iret = Helpers.writetodataset_mul(new_rows)
        if iret == -1:
            print("Write to dataframe failed")
        countchar = 6000
        # περίληψη των περιλήψεων με όριο χαρακτήρων της παραγόμενης
        summary = self.summarizer.summarize(text, countchar)
        # αντικατάσταση χαρακτήρων που παράγει ο summarizer μεταξύ των
λέξεων ή των προτάσεων
        summary = TextProcessor.Replace(summary)
        # εγγραφή της τελικής περίληψης στο dataframe
        iret = Helpers.writetodataset(mode, 1000, "SummaryofSummaries",
summary)
        if iret == -1:
            print("Write to dataframe failed")
        print("Final summary")
        print(summary)

    def selectmode(self):
        mode = self.mode
        df = self.df


        # 5: περίληψη κάθε άρθρου ξεχωριστά, συνένωση περιλήψεων σε ένα
κείμενο και περίληψη, abstractive summarization μόνο, model:
google/pegasus-cnn_dailymail
        if mode == 5:
            self.mode_5(df, mode)

        # 15: περίληψη κάθε άρθρου ξεχωριστά extractive summy.TextRank,
συνένωση περιλήψεων σε ένα κείμενο και περίληψη abstractive
summarization model: google/pegasus-cnn_dailymail
        if mode == 15:
            self.mode_15(df, mode)
```

```
import requests
from GoogleNews import GoogleNews
import pandas as pd
from newspaper import Article, Config
from SimilarityChecker import SimilarityChecker


class NewsFinder:
    # η κλάση που δημιουργεί τη συλλογή άρθρων. τα instances έχουν ως
attributes τα κριτήρια αναζήτησης
    def __init__(self, lang, period, topic, icount):
        self.lang = lang
        self.period = period
        self.topic = topic
        self.icount = icount

    def checkforvalid(self, df):
        # η συνάρτηση ελέγχει αν ένα url είναι έγκυρο, δηλαδή αν
επιστρέφει status 200, με timeout limit
        # ως είσοδο δέχεται το dataframe που παράγεται από την αναζήτηση
googlenews
        for url in df.loc[:, "link"]:
            try:
                response = requests.get(url, timeout=10)
                if response.status_code == 200:
                    pass
                else:
                    # αν το url δεν είναι valid εκτυπώνει μήνυμα με το
url και το status
                    # στη συνέχεια αφαιρεί την εγγραφή από το dataframe
                    print(f"{url} is invalid (status code:
{response.status_code})")
                    df.drop(df.index[df['link'] == url], inplace=True)
            except requests.exceptions.Timeout:
                # αν έχουμε timeout επιστέφει κατάλληλο μήνυμα
                # στη συνέχεια αφαιρεί την εγγραφή από το dataframe
                print(f"{url} is invalid (The request timed out)")
                df.drop(df.index[df['link'] == url], inplace=True)
            except:
                print(f"{url} is invalid")

        return df

    def getgooglenews(self, gncount, bexclude=False, df_exclude=None):
        # η μέθοδος κάνει την αναζήτηση στο Googlenews
        googlenews = GoogleNews()
        # οι τιμές των παραμέτρων αναζήτησης
        googlenews.set_lang(self.lang)
        googlenews.set_period(self.period)
        googlenews.set_encode('utf-8')
        googlenews.search(self.topic)
        # επιστοφή αποτελεσμάτων
        result = googlenews.results()
        # μετατροπή σε dataframe
        df = pd.DataFrame(result)
        df = self.checkforvalid(df)

        # το googlenews αρχικά επιστρέφει τα αποτελέσματα της πρώτης
σελίδας.
        # αν ο αριθμός τους δεν επαρκεί, θα πρέπει να πάρουμε και
επόμενων
```

```
        ipage = 1

        # το bexclude είναι μια συνθήκη, για τα επόμενα iterations
        # δείχνει ότι έχω ήδη προχωρήσει στην εξαγωγή άρθρων
        if bexclude:
            # τα αποτελέσματα που έχω ήδη συγκεντρώσει συγχωνεύονται με
τα νέα
            df = pd.concat([df, df_exclude], ignore_index=True)
            # αφαιρούνται τα διπλά αποτελέσματα
            df.drop_duplicates(inplace=True)
        # εδω γίνεται ουσιαστικά η αναζήτηση.
        # Επαναλαμβάνεται μέχρι να συμπληρωθεί ο απαιτούμενος αριθμός
αποτελεσμάτων
        while df.shape[0] < gncount:
            # ανάκτηση αποτελεσμάτων από την επόμενη σελίδα
            ipage += 1
            googlenews.getpage(ipage)
            result = googlenews.results()
            # προσωρινό dataframe για την αποθήκευση των αποτελεσμάτων
της δεύτερης σελίδας
            df_temp = pd.DataFrame(result)
            df_temp = self.checkforvalid(df_temp)
            # ενοποίηση των dataframes και αφαίρεση διπλότυπων
            df = pd.concat([df, df_temp], ignore_index=True)
            df.drop_duplicates(inplace=True)
            df = df.drop_duplicates(subset='title', keep="first")
            df = df.drop_duplicates(subset='desc', keep="first")
        df.reset_index(drop=True, inplace=True)
        print(df.loc[:, "link"])
        print(df.shape[0])
        # αποθήκευση σε csv
        df.to_csv("dataframe_initial.csv")
        return df

    def getarticlesdf(self, df):
        # εδώ γίνεται η ανάκτηση των κειμένων από τις ιστοσελίδες
        # ρυθμίσεις browser user agent
        config = Config()
        config.browser_user_agent = "Mozilla/5.0 (Windows NT 10.0;
Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/113.0.0.0
Safari/537.36"
        config.request_timeout = 20
        news_list = []
        excluded = []
        # ανάκτηση άρθρων. από κάθε link του dataframe ανακτώ το άρθρο
        for ind in df.index:
            dict = {}
            try:
                a = Article(df['link'][ind], language='en',
config=config)
                a.url.strip()
                a.download()
                a.parse()
                # δημιουργία λεξικού για εισαγωγή στο νέο dataframe
                dict['Link'] = df['link'][ind]
                dict['Date'] = df['date'][ind]
                dict['Media'] = df['media'][ind]
                dict['Title'] = a.title
                dict['Article'] = a.text
                # έλεγχος μεγέθους άρθρου. αν δεν έχει περιεχόμενο θα
εξαιρεθεί
```

```
                if len(dict['Article']) > 400:
                    news_list.append(dict)
                else:
                    excluded.append(ind)

            except:
                # αν η διαδικασία δώσει σφάλμα, η εγγραφή τηρείται σε
λίστα exclude για να αφαιρεθεί
                print(f"{df['link'][ind]} is invalid")
                excluded.append(ind)
                continue
        #δημιουργία dataframe άρθρων news_df
        news_df = pd.DataFrame(news_list)
        # ότι δε γύρισε αποτέλεσμα αφαιρείται από το dataframe με τα
αποτελέσματα απο το googlenews
        for i in excluded:
            df.drop(i, inplace=True)
        df.drop_duplicates(inplace=True)
        df.reset_index(drop=True)


        #αφαίρεση διπλοτύπων. αφαιρούνται και από τα δύο dataframes
        news_df = news_df.drop_duplicates(subset='Title', keep="first")
        df = df.drop_duplicates(subset='title', keep="first")
        if news_df.shape[0] > 0:


            news_df.reset_index(drop=True, inplace=True)
            if news_df.shape[0] > 1:
                sc = SimilarityChecker()
                exclude = sc.compare(news_df)
                for i in exclude:
                    link = news_df['Link'][i]
                    df.drop(df[df['link'] == link].index, inplace=True)
                    news_df.drop(i, inplace=True)
            df.reset_index(drop=True)
            news_df.reset_index(drop=True)
            # τα τελικά αποτελέσματα αποθηκεύονται σε csv
            df.to_csv("dataframe_new.csv")
            news_df.to_csv("articles.csv")
        return news_df, df

    def iterate(self):
        # πρώτη αναζήτηση και ανάκτηση άρθρων
        df = self.getgooglenews(self.icount)
        news_df, df_new = self.getarticlesdf(df)

        while df_new.shape[0] < self.icount:
        # αν τα άρθρα είναι λιγότερα, η διαδικασία επαναλαμβάνεται
            print("Count is")
            print(df_new.shape[0])
            # ο νέος αριθμός άρθρων που θέλουμε να ανακτήσουμε
            now_count = self.icount-df_new.shape[0]
            df_new = self.getgooglenews(now_count, True, df_new)
            news_df, df_new = self.getarticlesdf(df_new)

        return news_df
from sumy.parsers.plaintext import PlaintextParser
from sumy.nlp.tokenizers import Tokenizer
from sumy.summarizers.text_rank import TextRankSummarizer
import re
```

```
class ExtrSummarizer:
    #extractive summarization με TextRamk
    def __init__(self):
        self.summarizer = TextRankSummarizer()

    def extractive_summarization_one(self, text):
    # tokenization
        parser = PlaintextParser.from_string(text, Tokenizer('english'))

        # μέτρηση προτάσεων αρχικού κειμένου.
        # θέλουμε η περίληψη να περιέχει το 20% του αρχικού
        summary = ""
        sentences = re.split(r'(?<=[.!?])\s+', text)
        or_sentence_count = len(sentences)
        sentence_count = int(or_sentence_count * 0.2)
        # για να μην πάρουμε πολύ μικρή περίληψη
        if sentence_count < 5:
            sentence_count = 5
        # περίληψη κειμένου
        summary_sum = self.summarizer(parser.document, sentence_count)
        # συνένωση των προτάσεων σε κείμενο
        summary = ' '.join(map(str, summary_sum))

        return summary



import pandas as pd
import os.path


class Helpers:

    @staticmethod
    def writetodataset(mode, num, title, summary):
        # ανάκτηση dataframe αποτελεσμάτων και προσθήκη νέας εγγραφής
(για μία εγγραφή)
        # αποθήκευση του αποτελέσματος σε csv
        iret = 0
        try:
            if os.path.exists('Results.csv'):
                final_df = pd.read_csv('Results.csv')
            else:
                final_df = pd.DataFrame(columns=['mode', 'num', 'title',
'summary'])
            new_row = {'mode': mode, 'num': num, 'title': title,
'summary': summary}
            final_df = pd.concat([final_df, pd.DataFrame([new_row])],
ignore_index=True)
            final_df.to_csv('Results.csv')
            iret = 1
        except:
            print("Could not write to dataframe")
            iret = -1
        return iret

    @staticmethod
    def writetodataset_mul(rows):
        # ανάκτηση dataframe αποτελεσμάτων και προσθήκη λίστας εγγραφών
```

```
        # αποθήκευση του αποτελέσματος σε csv
        iret = 0
        try:
            if os.path.exists('Results.csv'):
                final_df = pd.read_csv('Results.csv')
            else:
                final_df = pd.DataFrame(columns=['mode', 'num', 'title',
'summary'])
            for new_row in rows:
                final_df = pd.concat([final_df,
pd.DataFrame([new_row])], ignore_index=True)
            final_df.to_csv('Results.csv')

            iret = 1
        except:
            print("Could not write to dataframe")
            iret = -1
        return iret
    @staticmethod
    def writetotxt (labeltxt, text):
        with open('segments.txt', 'a', encoding='utf-8') as f:
            f.write('\n')
            f.write(f" {labeltxt}: {text}")
            f.write('\n')
from transformers import PegasusForConditionalGeneration,
PegasusTokenizer
import re
from Processing import TextProcessor
from Helpers import Helpers


class PegasusSummarizer:
    def __init__(self, model):
        self.model = model
        self.tokenizer =
PegasusTokenizer.from_pretrained("google/pegasus-cnn_dailymail")
        self.summarizer =
PegasusForConditionalGeneration.from_pretrained("google/pegasus-
cnn_dailymail")


    def SplitText(self, text, max_token_limit):
        # τμηματοποίηση κειμένου. Το μέγεθος του κειμένου δεν ξεπερνά τα
1024 tokens

        # χωρισμός παραγράφων (νέα γραμμή)
        paragraphs = text.split("/n")

        segments = []
        current_segment = ""
        for paragraph in paragraphs:
            # αν η παράγραφος είναι πολύ μικρή δεν κρατιέται
            if len(paragraph) < 50:
                continue
            # χωρισμός της παραγράφου σε προτάσεις
            sentences = re.split(r'(?<=[.!?])\s+', paragraph)
            # tokenization με τον tokenizer του μοντέλου
            tokens = self.tokenizer.tokenize(paragraph)
            paragraph_len = 0
            # μήκος παραγράφου σε tokens
            for token in tokens:
```

```
                paragraph_len += len(token)
                # αν το μήκος της παραγράφου ξεπερνάει το όριο
                #τότε χωρίζουμε σε προτάσεις
            if paragraph_len > max_token_limit:
                for sentence in sentences:
                    # προτάσεις με πολύ μικρό μήκος απορρίπτονται
                    if len(sentence) < 40:
                        continue
                    # tokenization πρότασης
                    tokens = self.tokenizer.tokenize(sentence)
                    sentence_len = 0
                    # προσθέτουμε το μήκος των tokens μέχρι το όριο
                    for token in tokens:
                        sentence_len += len(token) + 1

                    if sentence_len + len(current_segment) <
max_token_limit:
                    # αν το μήκος των προτάσεων του segment επαρκεί,
προσθέτω την πρόταση στο segment
                        for token in tokens:
                            current_segment += token
                    else:
                        # αν όχι προσθέτω το segment σε λίστα και
δημιουργώ νέο
                        segments.append(current_segment.strip())
                        current_segment = ""
            else:

                if paragraph_len + len(current_segment) <
max_token_limit:
                    # αν το μήκος δεν ξεπερνάει το όριο
                    # τότε προστίθεται στο segment
                    for token in tokens:
                        current_segment += token
                else:
                    segments.append(current_segment.strip())
                    current_segment = ""

        # αν έχει περισσέψει segment, το προσαρτώ στη λίστα
        if current_segment:
            segments.append(current_segment.strip())

        print(len(segments))
        with open('segments.txt', 'a',  encoding='utf-8') as f:
            f.write('\n')
            for i, segment in enumerate(segments):
                strsegment =
self.tokenizer.convert_tokens_to_string(segment)
                # εγγραφή segments σε αρχείο
                f.write(f"Segment {i + 1}: {strsegment}")
                f.write('\n')
                print(f"Segment {i + 1}:
{self.tokenizer.convert_tokens_to_string(segment)}")

        # επιστρέφεται η λίστα με τα segments

        return segments

    def Summarize_Chunks(self, text):

        # καλείται η μέθοδος για την τμηματοποίηση και επιστρέφει λίστα
```

```
με τα τμήματα
        chunks = self.SplitText(text, 1024)
        print(len(chunks))
        summaries = []
        summarized = ''
        for chunk in chunks:
            print(len(chunk))
            # κάθε τμήμα γίνεται περίληψη
            summary_text = self.Summarize_Peg(chunk)
            Helpers.writetotxt("segmentsummary: ", summary_text)
            # οι περιλήψεις μπαίνουν σε λίστα
            summaries.append(summary_text)

        print(len(summaries))
        print(summaries)
        for summary in summaries:
            # συνένωση σε ένα κείμενο
            summarized = TextProcessor.concatstrings(summarized,summary)

        final = summarized

        return final, len(chunks)

    def summarize(self, text, countchar):
        # περίληψη με τμηματοποίηση
        summary_text, icount = self.Summarize_Chunks(text)
        print(summary_text)
        # αφαίρεση <n> ή άλλων χαρακτήρων
        summary_text = TextProcessor.Replace(summary_text)
        # αν το μήκος της περίληψης > από το όριο και προέρχεται από
πολλά κομμάτια
        # τότε ξανά περίληψη για ομογενοποίηση
        if len(summary_text) > countchar and icount > 1:
            summary_text, icount = self.Summarize_Chunks(summary_text)

        return summary_text

    def Summarize_Peg(self, text):
        tokenizer = self.tokenizer
        summarizer = self.summarizer
        # ορισμός μέγιστου μήκους περίληψης
        max_length = int(0.2 * len(text))
        # ορισμός μέγιστων ορίων για κάθε μοντέλο,
        # ώστε να μην είναι μικρότερα από τα min
        if self.model == 1:
            if max_length < 35:
                max_length = 35
        elif self.model == 2:
            if max_length < 60:
                max_length = 60
        # ξαναγίνεται tokenization
        tokens = tokenizer(text, truncation=False, padding="longest",
return_tensors="pt")
        # παραγωγή περίληψης
        summary = summarizer.generate(**tokens, length_penalty=2.0,
max_length=int(max_length),
                                       early_stopping=True, num_beams=5,
no_repeat_ngram_size=2)
        summary_text = tokenizer.decode(summary[0],
skip_special_tokens=True)
```

```
        return summary_text
import os
import re


class TextProcessor:
    @staticmethod
    def RemoveEmptyLines(text):
        #αφαίρεση κενών γραμμών
        text = os.linesep.join(
            [line for line in text.splitlines() if line]
        )
        return text

    @staticmethod
    def CharCount(text):
        text_length = len(text)
        return text_length

    @staticmethod
    # συνένωση string με νέα γραμμή
    def concatstrings(text1, text2):
        text = text1 + "\n" + text2
        return text

    @staticmethod
    # αντικατάσταση ειδικών χαρακτήρων
    def Replace(text):
        text = text.replace("<n>", " ")
        text = text.replace("Ġ", " ")
        text = text.replace("Âł", " ")
        return text

    @staticmethod
    # αφαίρεση τμημάτων που δεν έχουν σχέση με το άρθρο
    def remove_unwanted(article):
        unwanted_phrases =[]
        unwanted_phrases.append("RELATED TOPICS:")
        unwanted_phrases.append("Advertisement")

        # βρίσκουμε την αρχή του τμήματος που θα αφαιρεθεί
        for start_of_section in unwanted_phrases:
            unwanted_start = article.find(start_of_section)

        # όταν το βρούμε, το αφαιρούμε
        if unwanted_start != -1:
            # Only keep the part of the article before the unwanted
section
            article = article[:unwanted_start]

        return article

    @staticmethod
    #αφαίρεση url από το κείμενο
    def removeurlsections(text):
        url_pattern = re.compile(r'\S*https?://\S*\s?|www\.\S*\s?')
        sentences = text.split('. ')
        clean_sentences = [sentence for sentence in sentences if not
url_pattern.search(sentence)]
```

```
        clean_text = '. '.join(clean_sentences)
        return clean_text
import nltk
import string
from sklearn.feature_extraction.text import TfidfVectorizer
from nltk.corpus import stopwords


class SimilarityChecker:
    def __init__(self):
        # nltk.download('punkt')
        self.stemmer = nltk.stem.porter.PorterStemmer()
        self.remove_punctuation_map = dict((ord(char), None) for char in
string.punctuation)
        self.stop_words = set(stopwords.words('english'))
        self.vectorizer = TfidfVectorizer(tokenizer=self.tokenize,
stop_words=self.stopwords)


    # tokenization, stemming και αφαίρεση stopwords
    def tokenize(self, text):
        tokens =
nltk.word_tokenize(text.lower().translate(self.remove_punctuation_map))
        stemmed_tokens = [self.stemmer.stem(token) for token in tokens
if token not in self.stop_words]
        return stemmed_tokens

    # υπολογισμός ομοιότητας συνημιτόνου
    def cosine_sim(self, text1, text2):
        #μετατροπή σε διανύσματα
        tfidf = self.vectorizer.fit_transform([text1, text2])
        return ((tfidf * tfidf.T).A)[0, 1]

    def compare(self, df):
        imax = df[df.columns[0]].count()
        exclude = []
        # σύγκριση κειμένων ανά δύο και υπολογισμός ομοιότητας
συνημιτόνου
        for i in df.index:
            for j in range(i + 1, imax):
                # υπολογισμός ομοιότητας
                similarity = self.cosine_sim(df['Article'][i],
df['Article'][j])
                print(similarity)
                # όριο ομοιότητας για την απόρριψη ενός κειμένου
                if similarity > 0.95:
                    exclude.append(j)
                    exclude = list(set(exclude))

        return exclude
```

Παράρτημα Β: Άρθρα, ενδιάμεσα τμήματα και τελικό αποτέλεσμα

Εφαρμογή 2 σταδίων αφαιρετικής περίληψης

Είδος	**Κείμενο**
Αρχικό κείμενο	ATHENS, June 29 (Reuters) - With backing from a politician jailed for leading a party declared a criminal gang, Greek far-right groups swept up over 12% of the vote in Sunday's election, mirroring the rise of populist and ultra-nationalist politicians across Europe. The surge of three parties with their ultra-nationalist views - including 'Spartans' which barely registered in polls until Ilias Kasidiaris from the banned Golden Dawn party endorsed it from his prison cell -- could swing public debate at home and prove a springboard in European elections. While the centre-right New Democracy party of Kyriakos Mitsotakis stormed to victory in the June 25 Greek vote, winning 158 seats in the 300-seat parliament, the Spartans emerged as the fifth largest group. Spartans and two more parties, Greek Solution and Niki - which together have 34 seats - view migration as a threat to Greece's national identity, believe LGBTQ+ issues undermine the sanctity of family and deeply resent authorities for forcing people to get vaccinated during the COVID-19 pandemic. The deaths of hundreds of migrants when their boat sank off Greece while being tracked by the Greek coastguard has sharpened the debate over immigration. Despite mourning the tragedy, many Greeks want to halt the stream of migrants.

Είδος	Κείμενο
	"The re-emergence of the far right is a byproduct of a political strategy of the New Democracy government which tried to appeal to the centre with an agenda of economic liberalism and, at the same time to the far right with an agenda of law and order and anti-immigrant discourse," said Akritas Kaidatzis, assistant professor of constitutional law at the Aristotle University in Thessaloniki. EUROPE'S FAR-RIGHT RESURGENCE Gains by the far right in Greece mirror a trend in several other European countries. Support for Germany's anti-immigrant AfD party is at its highest since the wake of Europe’s migrant crisis in 2018, in Italy a former far-right activist is prime minister and in Sweden a far-right party has joined the government. Some analysts said it was unlikely the three parties would be able to define government policy. But they are expected to pressure Mitsotakis on LGBTQ+ issues, migration and relations with Greece's historic rival Turkey. All three are expected to run in European Parliament elections next year. First elected in 2019, Mitsotakis has been resolute in promoting LGBTQ+ rights but has himself taken a hard line on migration, prompting criticism from rights groups. Some analysts said Mitsotakis's hard line on migration gave licence to a more xenophobic narrative from the far right. However, some said ultra-nationalists had claimed the turf on the far right because Mitsotakis was shifting to the centre.

Είδος	Κείμενο
	"I sense that it's less likely they will impose their agenda on him, and a lot more likely it will propel him more to the political centre," said Akis Georgakellos, political advisor and managing director of Athens-based communications firm Stratego. "It's a given they will compete among themselves on who is the most extreme," he said. TROJAN HORSE FROM SPARTA? The Spartans party are led by businessman Vassilios Stigas. Despite a far-right agenda, it lacks the militancy of Golden Dawn, which was known for its torch-lit marches through Athens, vitriolic speeches and Nazi-like salutes. In February, Greece's parliament passed a law banning parties whose leaders are convicted of crimes. Yet Kasidiaris, Golden Dawn's former frontman, has shown he is still a force to reckon with, even from behind bars. "I will vote, and support with all my strength the Spartans," Kasidiaris wrote on Twitter from his cell in the high-security Domokos prison after Greece's top court disqualified his group from contesting the elections. Kasidiaris is serving a 13-year jail term for his leadership role in Golden Dawn, once Greece's third largest party. It was declared a criminal gang linked to hate crimes in a 2020 court ruling. Both convictions are being appealed. Reporting By Michele Kambas and Renee Maltezou, writing by Michele Kambas, editing by Edmund Blair and Christina Fincher Our Standards: The Thomson Reuters Trust Principles.

Είδος	Κείμενο
Segment 1	ATHENS, June 29 (Reuters) - With backing from a politician jailed for leading a party declared a criminal gang, Greek far-right groups swept up over 12% of the vote in Sunday's election, mirroring the rise of populist and ultra-nationalist politicians across Europe. The surge of three parties with their ultra-nationalist views - including 'Spartans' which barely registered in polls until Ilias Kasidiaris from the banned Golden Dawn party endorsed it from his prison cell -- could swing public debate at home and prove a springboard in European elections. While the centre-right New Democracy party of Kyriakos Mitsotakis stormed to victory in the June 25 Greek vote, winning 158 seats in the 300-seat parliament, the Spartans emerged as the fifth largest group.
Summary 1	Greek far-right groups swept up over 12% of the vote in Sunday's election.<n>The surge of three parties with their ultra-nationalist views could swing public debate at home and prove a springboard in European elections,<n>Ilias Kasidiaris from the banned Golden Dawn party endorsed it from his prison cell, mirroring the rise of populist politicians across Europe.
Segment 2	The deaths of hundreds of migrants when their boat sank off Greece while being tracked by the Greek coastguard has sharpened the debate over immigration. Despite mourning the tragedy, many Greeks want to halt the stream of migrants. "The re-emergence of the far right is a byproduct of a political strategy of the New Democracy government which tried to appeal to the centre with an agenda of economic liberalism and, at the same time to the far right with an agenda of law and order and anti-immigrant discourse," said Akritas Kaidatzis, assistant professor of constitutional law at the Aristotle University in Thessaloniki. EUROPE'S FAR-RIGHT RESURGENCE Gains by the far right in Greece mirror a trend in several other European countries. Support for Germany's anti-immigrant AfD party is at its highest since the wake of Europe’s migrant crisis in 2018, in Italy a former far-right activist is prime minister and in Sweden a far-right party has joined the government.
Summary 2	Far-right gains in Greece mirror a trend in several other European countries.<n>Support for Germany's anti-immigrant AfD party is at its highest since the wake of Europe’s migrant crisis in 2018.<n>In

Είδος	Κείμενο
	Italy a former far-Right activist is prime minister and in Sweden a far right party has joined the government.
Segment 3	But they are expected to pressure Mitsotakis on LGBTQ+ issues, migration and relations with Greece's historic rival Turkey. All three are expected to run in European Parliament elections next year. First elected in 2019, Mitsotakis has been resolute in promoting LGBTQ+ rights but has himself taken a hard line on migration, prompting criticism from rights groups. Some analysts said Mitsotakis's hard line on migration gave licence to a more xenophobic narrative from the far right. However, some said ultra-nationalists had claimed the turf on the far right because Mitsotakis was shifting to the centre. "I sense that it's less likely they will impose their agenda on him, and a lot more likely it will propel him more to the political centre," said Akis Georgakellos, political advisor and managing director of Athens-based communications firm Stratego. "It's a given they will compete among themselves on who is the most extreme," he said. The Spartans party are led by businessman Vassilios Stigas.
Summary 3	All three are expected to run in European Parliament elections next year.<n>Mitsotakis has been resolute in promoting LGBTQ+ rights but has himself taken a hard line on migration, prompting criticism from rights groups.
Segment 4	In February, Greece's parliament passed a law banning parties whose leaders are convicted of crimes. Yet Kasidiaris, Golden Dawn's former frontman, has shown he is still a force to reckon with, even from behind bars. "I will vote, and support with all my strength the Spartans," Kasidiaris wrote on Twitter from his cell in the high-security Domokos prison after Greece's top court disqualified his group from contesting the elections. Kasidiaris is serving a 13-year jail term for his leadership role in Golden Dawn, once Greece's third largest party. It was declared a criminal gang linked to hate crimes in a 2020 court ruling. Reporting By Michele Kambas and Renee Maltezou, writing by Michele Kambas, editing by Edmund Blair and Christina Fincher Our Standards: The Thomson Reuters Trust Principles.
Summary 4	Kasidiaris is serving a 13-year jail term for his leadership role in Golden Dawn.<n>Golden Dawn was declared a criminal gang linked to hate crimes in a 2020 court ruling, and was banned in

Είδος	Κείμενο
	Greece in 2011 <n>It was once Greece's third largest party, but has been on the wane since the financial crisis of 2008
Summary of Segment 1	Greek far-right groups swept up over 12% of the vote in Sunday's election.<n>The surge of three parties with their ultra-nationalist views could swing public debate at home and prove a springboard in European elections,<n>Ilias Kasidiaris from the banned Golden Dawn party endorsed it from his prison cell, mirroring the rise of populist politicians across Europe.
Summary of Segment 2	Far-right gains in Greece mirror a trend in several other European countries.<n>Support for Germany's anti-immigrant AfD party is at its highest since the wake of Europe’s migrant crisis in 2018.<n>In Italy a former far-Right activist is prime minister and in Sweden a far right party has joined the government.
Summary of Segment 3	All three are expected to run in European Parliament elections next year.<n>Mitsotakis has been resolute in promoting LGBTQ+ rights but has himself taken a hard line on migration, prompting criticism from rights groups.
Summary of Segment 4	Kasidiaris is serving a 13-year jail term for his leadership role in Golden Dawn.<n>Golden Dawn was declared a criminal gang linked to hate crimes in a 2020 court ruling, and was banned in Greece in 2011 <n>It was once Greece's third largest party, but has been on the wane since the financial crisis of 2008
Article Summary 0	Greek far-right groups swept up over 12% of the vote in Sunday's election. The surge of three parties with their ultra-nationalist views could swing public debate at home and prove a springboard in European elections, Ilias Kasidiaris from the banned Golden Dawn party endorsed it from his prison cell, mirroring the rise of populist politicians across Europe. Far-right gains in Greece mirror a trend in several other European countries. Support for Germany's anti-immigrant AfD party is at its highest since the wake of Europe’s migrant crisis in 2018. In Italy a former far-Right activist is prime minister and in Sweden a far right party has joined the government. All three are expected to run in European Parliament elections next year. Mitsotakis has been resolute in promoting LGBTQ+ rights but has himself taken a hard line on migration, prompting criticism from rights groups.

Είδος	Κείμενο
	Kasidiaris is serving a 13-year jail term for his leadership role in Golden Dawn. Golden Dawn was declared a criminal gang linked to hate crimes in a 2020 court ruling, and was banned in Greece in 2011 It was once Greece's third largest party, but has been on the wane since the financial crisis of 2008
Αρχικό κείμενο	ATHENS, Greece -- Greece's firebrand opposition leader, Alexis Tsipras, announced his decision Thursday to step down as leader of the left-wing Syriza party, days after a crushing general election defeat. Tsipras, 48, served as Greece's prime minister from 2015 to 2019 during politically tumultuous years as the country struggled to remain in the euro zone and end a series of international bailouts. "I have therefore decided to propose the election of a new leadership by the members of the party ... Of course I will not be a candidate," Tsipras said in a televised address. "I make no secret of the fact that this is a painful decision ... I don't take hasty decisions. I put them under my pillow and torture myself with them first," he added. In Sunday's general election, Tsipras' left-wing Syriza party received just under 18% of the vote — losing almost half its support over the past four years — while Prime Minister Kyriakos Mitsotakis' winning New Democracy party topped 40%. "The party must take difficult and courageous decisions, which are called upon to serve a new vision. This obviously concerns me too," Tsipras said. Heading the Syriza since 2012, Tsipras forged a more cohesive party, taking it from a small political group to general election victory in 2015 on a pledge to push back on harsh austerity measures demanded by bailout

Είδος	Κείμενο
	lenders from other euro zone members and the International Monetary Fund. The effort proved unsuccessful and Greece was given a third bailout later that year to avoid bankruptcy and an exit from the shared euro currency. He eased his combative stance toward the European Commission and eventually forged closed ties with European leaders, including former German Chancellor Angela Merkel. He was widely praised by Western allies for finalizing an agreement with North Macedonia that ushered Greece's neighbor into NATO and advanced its effort to join the European Union. Although regarded as a skillful politician, Tsipras' main opponent said he hoped the opposition leader's departure would improve the quality of political debate. "I think decision was to be expected," Greece's conservative Prime Minister Mitsotakis said while attending a meeting of EU leaders in Brussels. "Syriza, both in government and in opposition, was a party characterized by toxicity, divisive rhetoric and with striking inefficiency," Mitsotakis said. "Political parties must unite citizens and propose realistic, cost-effective and workable solutions to people's problems — a road Syriza has never taken. I sincerely hope it does now." Tsipras is expected to stay on as leader for several weeks until his successor is elected by the party's rank-an-file membership. No prominent members of the party has publicly called on Tsipras to step down after the election defeat, though Euclid Tsakalotos, a former Syriza finance minister, had urged him to reflect on the results and "take the necessary actions." Effie Achtsioglou, a 38-year-old former social security

Είδος	Κείμενο
	minister, has received support from a section of the party to seek a leadership role but has not publicly discussed her plans. Commentators blamed Syriza's poor election result on the party's largely-negative campaign, the resurgence of the traditionally-strong Socialist party Pasok, and the appearance of splinter parties headed by Tsipras's former allies, including former Finance Minister Yanis Varoufakis. Largely rooted in fierce political confrontations during the 2010-2018 international bailouts, Syriza and the Socialists have been unable reach any agreement on potential collaboration, despite support from some senior members in both parties.
Segment 1	ATHENS, Greece -- Greece's firebrand opposition leader, Alexis Tsipras, announced his decision Thursday to step down as leader of the left-wing Syriza party, days after a crushing general election defeat. Tsipras, 48, served as Greece's prime minister from 2015 to 2019 during politically tumultuous years as the country struggled to remain in the euro zone and end a series of international bailouts. "I have therefore decided to propose the election of a new leadership by the members of the party ... Of course I will not be a candidate," Tsipras said in a televised address. "I make no secret of the fact that this is a painful decision ... I put them under my pillow and torture myself with them first," he added. In Sunday's general election, Tsipras' left-wing Syriza party received just under 18% of the vote — losing almost half its support over the past four years — while Prime Minister Kyriakos Mitsotakis' winning New Democracy party topped 40%.
Summary of Segment 1	Alexis Tsipras announced his decision to step down as leader of the left-wing Syriza party, days after a crushing general election defeat.<n>Tsipras, 48, served as Greece's prime minister from 2015 to 2019 during politically tumultuous years as the country struggled to remain in the euro zone and end a series of international bailouts.
Segment 2	This obviously concerns me too," Tsipras said. Heading the Syriza since 2012, Tsipras forged a more cohesive party, taking it from a small political group to general election victory in 2015 on a pledge

Είδος	Κείμενο
	to push back on harsh austerity measures demanded by bailout lenders from other euro zone members and the International Monetary Fund. The effort proved unsuccessful and Greece was given a third bailout later that year to avoid bankruptcy and an exit from the shared euro currency. He eased his combative stance toward the European Commission and eventually forged closed ties with European leaders, including former German Chancellor Angela Merkel. He was widely praised by Western allies for finalizing an agreement with North Macedonia that ushered Greece's neighbor into NATO and advanced its effort to join the European Union.
Summary of Segment 2	Heading the Syriza since 2012, Tsipras forged a more cohesive party, taking it from a small political group to general election victory in 2015.<n>He was widely praised by Western allies for finalizing an agreement with North Macedonia that ushered Greece's neighbor into NATO and advanced its effort to join the European Union.
Segment 3	"I think decision was to be expected," Greece's conservative Prime Minister Mitsotakis said while attending a meeting of EU leaders in Brussels. "Syriza, both in government and in opposition, was a party characterized by toxicity, divisive rhetoric and with striking inefficiency," Mitsotakis said. "Political parties must unite citizens and propose realistic, cost-effective and workable solutions to people's problems — a road Syriza has never taken. I sincerely hope it does now." Tsipras is expected to stay on as leader for several weeks until his successor is elected by the party's rank-an-file membership.
Summary of Segment 3	Mitsotakis: "Syriza was a party characterized by toxicity, divisive rhetoric and with striking inefficiency"<n>Tsipras is expected to stay on as leader for several weeks until his successor is elected by the party's rank-an-file membership.
Segment 4	Commentators blamed Syriza's poor election result on the party's largely-negative campaign, the resurgence of the traditionally-strong Socialist party Pasok, and the appearance of splinter parties headed by Tsipras's former allies, including former Finance Minister Yanis Varoufakis. Largely rooted in fierce political confrontations during the 2010-2018 international bailouts, Syriza and the Socialists have been unable reach any agreement on potential collaboration, despite support from some senior members in both parties.

Είδος	Κείμενο
Summary of Segment 4	Syriza and the Socialists have been unable to reach any agreement on potential collaboration.<n>Syriza's poor election result was blamed on the largely-negative campaign of the party and its former allies, including Yanis Varoufakis, the finance minister during the 2010-2018 bailouts, and Pasok, a traditionally-strong Socialist party, which returned to power after a five-year hiatus.
Summary of Segment 1	Alexis Tsipras announced his decision to step down as leader of the left-wing Syriza party, days after a crushing general election defeat.<n>Tsipras, 48, served as Greece's prime minister from 2015 to 2019 during politically tumultuous years as the country struggled to remain in the euro zone and end a series of international bailouts.
Summary of Segment 2	Heading the Syriza since 2012, Tsipras forged a more cohesive party, taking it from a small political group to general election victory in 2015.<n>He was widely praised by Western allies for finalizing an agreement with North Macedonia that ushered Greece's neighbor into NATO and advanced its effort to join the European Union.
Summary of Segment 3	Mitsotakis: "Syriza was a party characterized by toxicity, divisive rhetoric and with striking inefficiency"<n>Tsipras is expected to stay on as leader for several weeks until his successor is elected by the party's rank-an-file membership.
Summary of Segment 4	Syriza and the Socialists have been unable to reach any agreement on potential collaboration.<n>Syriza's poor election result was blamed on the largely-negative campaign of the party and its former allies, including Yanis Varoufakis, the finance minister during the 2010-2018 bailouts, and Pasok, a traditionally-strong Socialist party, which returned to power after a five-year hiatus.
Summary of Article 1	Alexis Tsipras announced his decision to step down as leader of the left-wing Syriza party, days after a crushing general election defeat. Tsipras, 48, served as Greece's prime minister from 2015 to 2019 during politically tumultuous years as the country struggled to remain in the euro zone and end a series of international bailouts. Heading the Syriza since 2012, Tsipras forged a more cohesive party, taking it from a small political group to general election victory in 2015. He was widely praised by Western allies for finalizing an agreement with North Macedonia that ushered Greece's neighbor into NATO and advanced its effort to join the European Union.

Είδος	Κείμενο
	Mitsotakis: "Syriza was a party characterized by toxicity, divisive rhetoric and with striking inefficiency" Tsipras is expected to stay on as leader for several weeks until his successor is elected by the party's rank-an-file membership. Syriza and the Socialists have been unable to reach any agreement on potential collaboration. Syriza's poor election result was blamed on the largely-negative campaign of the party and its former allies, including Yanis Varoufakis, the finance minister during the 2010-2018 bailouts, and Pasok, a traditionally-strong Socialist party, which returned to power after a five-year hiatus.
Αρχικό άρθρο	Our new issue on conspiracy is out now. Subscribe today to get it in print at a special discounted rate! The May 2023 Greek elections surprised most people on the Left within Greece and across the world. After four years of major political scandals, controlling the media, and accusations of illegally spying on political opponents and journalists, New Democracy won the elections with a twenty-point margin against Syriza. Understanding how we ended up here is essential for progressive social movements and left parties to offer a convincing progressive policy agenda and return to power. Rejecting austerity policies altogether, reconnecting with social movements and unions, and offering reliable policy alternatives are arduous but necessary tasks. The return to power for the Left is likely to be a slow process, and avoiding the repetition of the fall of Syriza should be the main objective. The Rise and the Fall of Syriza Syriza’s 2015 rise to power was the most significant political shock in Europe in recent memory. After five years of irrelevant austerity policies imposed on Greece by the European Union and the International Monetary Fund as part of the economic adjustment programs, the Greek people voted in favor of a potential rupture with the EU. Syriza promised a realistic possibility for progressive economic policies within the Eurozone. Between 2012 and 2015, the right-wing New Democracy party and the formerly social democratic Pasok implemented austerity-oriented economic policy programs that led to mass impoverishment and social unrest. During this time, Syriza actively engaged with progressive social movements and openly critiqued the harsh and inefficient austerity policies. The story of the ultimate rise of Syriza

Είδος	Κείμενο
	to power is relatively well known. After forming a coalition government on January 2015, the Syriza-led administration attempted to convince EU institutions, powerful lobbies, and the political leaders of the EU North that a progressive alternative policy agenda is possible. But these aspirations were crushed after six months of negotiations. The last act of Syriza's attempt to change the direction of economic policies within the Eurozone was the 2015 Greek bailout referendum. While an overwhelming 61.3 percent of the popular vote was against new austerity policies and the support for a Grexit was higher than ever, Prime Minister Aléxis Tsípras decided to go against the result and sign a new bailout agreement. Leading members of the first Syriza-led administration, including the finance minister Yanis Varoufakis, resigned, and new elections were announced for September 2015. Syriza issued an implicit promise that the new government would do anything to fairly distribute the costs of the upcoming austerity measures. People accepted that this was the only alternative after previous failed attempts to change EU economic policies, and Tsípras's Syriza re-won the majority of the parliament and formed a new coalition government. In the upcoming years, some limited social policies that protected poorer households were implemented, but economic policy has been focused primarily on maintaining budget surpluses and servicing public debt payments. Naturally, the process of creating a surplus that eventually reached thirty-one billion euros included heavy taxation and the stagnation of social spending, which particularly hurt the poorer segments of Greek society. Syriza argued that this was an unavoidable but temporary program that would be reverted once the memorandum agreements ended. There Is No Alternative New elections took place in July 2019. Syriza's narrative was that after the completion of certain austerity-oriented agreements, it was time to implement their own, more progressive agenda and return the accumulated budget surplus to society. Unsurprisingly, the majority of the voters were not convinced, and New Democracy won the majority of seats in the parliament, forming the first single-party government since 2009. Under the leadership of the political heir Kyriakos Mitsotakis, son of a former prime minister and brother of a former minister, New Democracy absorbed various far-right politicians from smaller political parties. The party presented a xenophobic migration and

Είδος	Κείμενο
	foreign policy agenda. But in terms of economic policy, their narrative was that New Democracy was a party of well-educated technocrats that, unlike Syriza, knew how to efficiently implement trickle-down economics and create a more efficient, market-oriented public administration. A few months after the reelection of New Democracy, the pandemic started and the new government had to abandon its austerity-oriented agenda out of necessity. Social unrest caused by the party's authoritarian public health approach and disinvestment in public health care, which caused thousands of excess deaths, pushed the government to spend a big share of the surplus created by Syriza in wage subsidies and social benefits. Needless to say, this was not a deliberate policy choice, but a necessary measure to limit social unrest in extraordinary circumstances. In the meantime, the New Democracy administration also spent twenty million euros on COVID-19 public health media campaigns. But soon after, criticism emerged claiming that the distribution of these funds was unequal and based on political affiliation with the governing party. This partisan favoritism, along with clear restrictions on journalistic freedom (Greece ranks 107th in a recent global press freedom ranking), strengthened the public's impression of New Democracy's corruption. On top of that, recent evidence also suggests that the Greek intelligence services, which are under the direct control of the prime minister, used illegal surveillance systems to spy on journalists and political opponents. Last but not least, during the final months of the New Democracy administration, a major train crash killed fifty-seven people. This devastating event was largely the outcome of disinvestment in public infrastructure and the privatization of the railways, since no signaling system had been in place for years and the operators were not properly trained. And as a result of these black marks on New Democracy's record, many believed that the Greek left's chances to return to power have increased significantly. Dream Deferred, Again The outcome of the 2023 elections is the most unexpected result in Greek politics in years. While the reelection of New Democracy was not unanticipated, winning with a margin of 20 percentage points against Syriza after four years of authoritarianism and scandals was unforeseen. At the same time, three smaller far-right parties managed to enter the parliament, where they now control over 10 percent of the seats. How did we

Είδος	Κείμενο
	end up here? First, the temporary ease of fiscal constraints by the EU due to COVID allowed the New Democracy government to provide some benefits in cash as well as coupons in response to the cost-of-living crisis, using the budget surplus that Syriza generated. If the social unrest that COVID created had not pushed the government to provide such benefits, the actual economic policy agenda of New Democracy would look very different to what was actually implemented. Smartly, in the preelection period of 2023, New Democracy took advantage of this turn of events and argued that they, not Syriza, are the party that actually cares about working-class people. Additionally, Syriza's election strategy was to focus almost exclusively on New Democracy's political scandals and blatant authoritarianism, without offering a more concrete economic policy agenda of their own. This approach missed that a precarious and impoverished population primarily cares about its economic survival and can be easily deceived by the rhetoric of the far right. As for why Syriza toned down its economic promises, the party perhaps recognized that financial markets would have reacted very aggressively against the election of a left-wing government, leading to higher borrowing rates. As a result, a left-wing government would have had to implement austerity anyways, a lesson Syriza already learned the hard way — and which raises major questions about how financial markets constrain democracy, especially within an interconnected monetary system like the Eurozone.
Segment 1	Subscribe today to get it in print at a special discounted rate! The May 2023 Greek elections surprised most people on the Left within Greece and across the world. After four years of major political scandals, controlling the media, and accusations of illegally spying on political opponents and journalists, New Democracy won the elections with a twenty-point margin against Syriza. Understanding how we ended up here is essential for progressive social movements and left parties to offer a convincing progressive policy agenda and return to power. Rejecting austerity policies altogether, reconnecting with social movements and unions, and offering reliable policy alternatives are arduous but necessary tasks. The return to power for the Left is likely to be a slow process, and avoiding the repetition of the fall of Syriza should be the main objective. The Rise and the Fall of Syriza

Είδος	Κείμενο
	Syriza's 2015 rise to power was the most significant political shock in Europe in recent memory.
Summary of Segment 1	After four years of major political scandals, controlling the media, and accusations of illegally spying on political opponents and journalists, New Democracy won the elections with a twenty-point margin against Syriza.<n>The return to power for the Left is likely to be a slow process and avoiding the repetition of the fall of Syrizo should be the main objective, argues Pavlos Tsimas and his co-authors, in this special edition of The Left's Revolution: How to Win the Fight for a Just and Progressive Europe
Segment 2	Syriza promised a realistic possibility for progressive economic policies within the Eurozone. Between 2012 and 2015, the right-wing New Democracy party and the formerly social democratic Pasok implemented austerity-oriented economic policy programs that led to mass impoverishment and social unrest. During this time, Syriza actively engaged with progressive social movements and openly critiqued the harsh and inefficient austerity policies. The story of the ultimate rise of Syriza to power is relatively well known. After forming a coalition government on January 2015, the Syriza-led administration attempted to convince EU institutions, powerful lobbies, and the political leaders of the EU North that a progressive alternative policy agenda is possible. But these aspirations were crushed after six months of negotiations. The last act of Syriza's attempt to change the direction of economic policies within the Eurozone was the 2015 Greek bailout referendum.
Summary of Segment 2	Syriza promised a realistic possibility for progressive economic policies within the Eurozone.<n>Between 2012 and 2015, the right-wing New Democracy party and the formerly social democratic Pasok implemented austerity-oriented economic policy programs that led to mass impoverishment and social unrest, according to this study, from 2012 to 2015, and openly critiqued the harsh and inefficient austerity policies.
Segment 3	Leading members of the first Syriza-led administration, including the finance minister Yanis Varoufakis, resigned, and new elections were announced for September 2015. Syriza issued an implicit promise that the new government would do anything to fairly distribute the costs of the upcoming austerity measures. People accepted that this was the only alternative after previous failed

Είδος	Κείμενο
	attempts to change EU economic policies, and Tsípras's Syriza re-won the majority of the parliament and formed a new coalition government. In the upcoming years, some limited social policies that protected poorer households were implemented, but economic policy has been focused primarily on maintaining budget surpluses and servicing public debt payments. Naturally, the process of creating a surplus that eventually reached thirty-one billion euros included heavy taxation and the stagnation of social spending, which particularly hurt the poorer segments of Greek society.
Summary of Segment 3	The first Syriza-led administration, including the finance minister Yanis Varoufakis, resigned, and new elections were announced for September 2015.<n>The process of creating a surplus that eventually reached thirty-one billion euros included heavy taxation and the stagnation of social spending, which particularly hurt the poorer segments of Greek society.
Segment 4	There Is No Alternative New elections took place in July 2019. Syriza's narrative was that after the completion of certain austerity-oriented agreements, it was time to implement their own, more progressive agenda and return the accumulated budget surplus to society. Unsurprisingly, the majority of the voters were not convinced, and New Democracy won the majority of seats in the parliament, forming the first single-party government since 2009. Under the leadership of the political heir Kyriakos Mitsotakis, son of a former prime minister and brother of a former minister, New Democracy absorbed various far-right politicians from smaller political parties. The party presented a xenophobic migration and foreign policy agenda.
Summary of Segment 4	Syriza's narrative was that after the completion of certain austerity-oriented agreements, it was time to implement their own, more progressive agenda and return the accumulated budget surplus to society.<n> Unsurprisingly, the majority of the voters were not convinced, and New Democracy won themajority of seats in the parliament, forming the first single-party government since 2009.
Segment 5	A few months after the reelection of New Democracy, the pandemic started and the new government had to abandon its austerity-oriented agenda out of necessity. Social unrest caused by the party's authoritarian public health approach and disinvestment in public health care, which caused thousands of excess deaths, pushed the government to spend a big share of the surplus created by Syriza in

Είδος	Κείμενο
	wage subsidies and social benefits. Needless to say, this was not a deliberate policy choice, but a necessary measure to limit social unrest in extraordinary circumstances. In the meantime, the New Democracy administration also spent twenty million euros on COVID-19 public health media campaigns. But soon after, criticism emerged claiming that the distribution of these funds was unequal and based on political affiliation with the governing party.
Summary of Segment 5	Social unrest caused by the party's authoritarian public health approach pushed the government to spend a big share of the surplus created by Syriza in wage subsidies and social benefits.<n>The New Democracy administration also spent twenty million euros on COVID-19 public Health media campaigns. But soon after, criticism emerged claiming that the distribution of these funds was unequal based on political affiliation with the governing party.
Segment 6	On top of that, recent evidence also suggests that the Greek intelligence services, which are under the direct control of the prime minister, used illegal surveillance systems to spy on journalists and political opponents. Last but not least, during the final months of the New Democracy administration, a major train crash killed fifty-seven people. This devastating event was largely the outcome of disinvestment in public infrastructure and the privatization of the railways, since no signaling system had been in place for years and the operators were not properly trained. And as a result of these black marks on New Democracy's record, many believed that the Greek left's chances to return to power have increased significantly. Dream Deferred, Again The outcome of the 2023 elections is the most unexpected result in Greek politics in years.
Summary of Segment 6	Recent evidence suggests that the Greek intelligence services used illegal surveillance systems to spy on journalists and political opponents.<n>During the final months of the New Democracy administration, a major train crash killed fifty-seven people, and many believed the left's chances of returning to power have increased significantly.
Segment 7	At the same time, three smaller far-right parties managed to enter the parliament, where they now control over 10 percent of the seats. First, the temporary ease of fiscal constraints by the EU due to COVID allowed the New Democracy government to provide some benefits in cash as well as coupons in response to the cost-of-living

Είδος	Κείμενο
	crisis, using the budget surplus that Syriza generated. If the social unrest that COVID created had not pushed the government to provide such benefits, the actual economic policy agenda of New Democracy would look very different to what was actually implemented. Smartly, in the preelection period of 2023, New Democracy took advantage of this turn of events and argued that they, not Syriza, are the party that actually cares about working-class people.
Summary of Segment 7	Three far-right parties now control over 10 percent of the seats in the Greek parliament.<n>New Democracy took advantage of this turn of events and argued that they, not Syriza, are the party that actually cares about working-class people, says Yanis Varoufakis, author of 'Syriza: The New Democrats'
Segment 8	This approach missed that a precarious and impoverished population primarily cares about its economic survival and can be easily deceived by the rhetoric of the far right. As for why Syriza toned down its economic promises, the party perhaps recognized that financial markets would have reacted very aggressively against the election of a left-wing government, leading to higher borrowing rates. As a result, a left-wing government would have had to implement austerity anyways, a lesson Syriza already learned the hard way — and which raises major questions about how financial markets constrain democracy, especially within an interconnected monetary system like the Eurozone.
Summary of Segment 8	This approach missed that a precarious and impoverished population primarily cares about its economic survival.<n>The party perhaps recognized that financial markets would have reacted very aggressively against the election of a left-wing government, leading to higher borrowing rates.
Summary of Segment 1	After four years of major political scandals, controlling the media, and accusations of illegally spying on political opponents and journalists, New Democracy won the elections with a twenty-point margin against Syriza.<n>The return to power for the Left is likely to be a slow process and avoiding the repetition of the fall of Syrizo should be the main objective, argues Pavlos Tsimas and his

Είδος	Κείμενο
	co-authors, in this special edition of The Left's Revolution: How to Win the Fight for a Just and Progressive Europe
Summary of Segment 2	Syriza promised a realistic possibility for progressive economic policies within the Eurozone.<n>Between 2012 and 2015, the right-wing New Democracy party and the formerly social democratic Pasok implemented austerity-oriented economic policy programs that led to mass impoverishment and social unrest, according to this study, from 2012 to 2015, and openly critiqued the harsh and inefficient austerity policies.
Summary of Segment 3	The first Syriza-led administration, including the finance minister Yanis Varoufakis, resigned, and new elections were announced for September 2015.<n>The process of creating a surplus that eventually reached thirty-one billion euros included heavy taxation and the stagnation of social spending, which particularly hurt the poorer segments of Greek society.
Summary of Segment 4	Syriza's narrative was that after the completion of certain austerity-oriented agreements, it was time to implement their own, more progressive agenda and return the accumulated budget surplus to society.<n> Unsurprisingly, the majority of the voters were not convinced, and New Democracy won themajority of seats in the parliament, forming the first single-party government since 2009.
Summary of Segment 5	Social unrest caused by the party's authoritarian public health approach pushed the government to spend a big share of the surplus created by Syriza in wage subsidies and social benefits.<n>The New Democracy administration also spent twenty million euros on COVID-19 public Health media campaigns. But soon after, criticism emerged claiming that the distribution of these funds was unequal based on political affiliation with the governing party.
Summary of Segment 6	Recent evidence suggests that the Greek intelligence services used illegal surveillance systems to spy on journalists and political opponents.<n>During the final months of the New Democracy administration, a major train crash killed fifty-seven people, and many believed the left's chances of returning to power have increased significantly.
Summary of Segment 7	Three far-right parties now control over 10 percent of the seats in the Greek parliament.<n>New Democracy took advantage of this turn of events and argued that they, not Syriza, are the party that actually cares about working-class people, says Yanis Varoufakis, author of 'Syriza: The New Democrats'

Είδος	Κείμενο
Summary of Segment 8	This approach missed that a precarious and impoverished population primarily cares about its economic survival.<n>The party perhaps recognized that financial markets would have reacted very aggressively against the election of a left-wing government, leading to higher borrowing rates.
Summary of Article 2	After four years of major political scandals, controlling the media, and accusations of illegally spying on political opponents and journalists, New Democracy won the elections with a twenty-point margin against Syriza.The return to power for the Left is likely to be a slow process and avoiding the repetition of the fall of Syrizo should be the main objective, argues Pavlos Tsimas and his co-authors, in this special edition of The Left's Revolution: How to Win the Fight for a Just and Progressive Europe Syriza promised a realistic possibility for progressive economic policies within the Eurozone. Between 2012 and 2015, the right-wing New Democracy party and the formerly social democratic Pasok implemented austerity-oriented economic policy programs that led to mass impoverishment and social unrest, according to this study, from 2012 to 2015, and openly critiqued the harsh and inefficient austerity policies. The first Syriza-led administration, including the finance minister Yanis Varoufakis, resigned, and new elections were announced for September 2015. The process of creating a surplus that eventually reached thirty-one billion euros included heavy taxation and the stagnation of social spending, which particularly hurt the poorer segments of Greek society. Syriza's narrative was that after the completion of certain austerity-oriented agreements, it was time to implement their own, more progressive agenda and return the accumulated budget surplus to society. Unsurprisingly, the majority of the voters were not convinced, and New Democracy won themajority of seats in the parliament, forming the first single-party government since 2009. Social unrest caused by the party's authoritarian public health approach pushed the government to spend a big share of the surplus created by Syriza in wage subsidies and social benefits. The New Democracy administration also spent twenty million euros on COVID-19 public Health media campaigns. But soon after,

Είδος	Κείμενο
	criticism emerged claiming that the distribution of these funds was unequal based on political affiliation with the governing party. Recent evidence suggests that the Greek intelligence services used illegal surveillance systems to spy on journalists and political opponents. During the final months of the New Democracy administration, a major train crash killed fifty-seven people, and many believed the left's chances of returning to power have increased significantly. Three far-right parties now control over 10 percent of the seats in the Greek parliament. New Democracy took advantage of this turn of events and argued that they, not Syriza, are the party that actually cares about working-class people, says Yanis Varoufakis, author of 'Syriza: The New Democrats' This approach missed that a precarious and impoverished population primarily cares about its economic survival. The party perhaps recognized that financial markets would have reacted very aggressively against the election of a left-wing government, leading to higher borrowing rates.
Αρχικό κείμενο	Greece's Alexis Tsipras has stepped down from the helm of the leftist Syriza party following a heavy election defeat. "The time has come to start a new cycle," Tsipras said in a televised address, adding that reform of the party was needed. Led by Tsipras, Syriza stormed to power in 2015 at the height of Greece's deep economic crisis, riding a wave of anti-austerity and anti-bailout anger among Greeks. It lost to Prime Minister Kyriakos Mitsotakis's New Democracy in 2019. In Sunday's vote, Syriza won just 17.8 per cent of the votes against 40.5 per cent for New Democracy. "The negative result can - and must - become the beginning of this cycle," Tsipras said. He said he was stepping down to pave the way for elections for a new party leader, saying he would not be a candidate. "I am proud of everything that happened," he said. "This difficult journey had compromises, and difficult decisions, and injuries and attrition, but it was a journey that left a mark on history."
Segment 1	Greece's Alexis Tsipras has stepped down from the helm of the leftist Syriza party following a heavy election defeat. "The time

Είδος	Κείμενο
	has come to start a new cycle," Tsipras said in a televised address, adding that reform of the party was needed. Led by Tsipras, Syriza stormed to power in 2015 at the height of Greece's deep economic crisis, riding a wave of anti-austerity and anti-bailout anger among Greeks. It lost to Prime Minister Kyriakos Mitsotakis's New Democracy in 2019. In Sunday's vote, Syriza won just 17.8 per cent of the votes against 40.5 per cent for New Democracy. "The negative result can - and must - become the beginning of this cycle," Tsipras said. He said he was stepping down to pave the way for elections for a new party leader, saying he would not be a candidate. "I am proud of everything that happened," he said. "This difficult journey had compromises, and difficult decisions, and injuries and attrition, but it was a journey that left a mark on history."
Summary of segment 1	"The time has come to start a new cycle," Tsipras said in a televised address.<n>Led by Tsipras, Syriza stormed to power in 2015 at the height of Greece's deep economic crisis, riding a wave of anti-austerity and Anti-bailout anger among Greeks!<n>It lost to Prime Minister Kyriakos Mitsotakis's New Democracy in 2019.
Summary of article 3	"The time has come to start a new cycle," Tsipras said in a televised address. Led by Tsipras, Syriza stormed to power in 2015 at the height of Greece's deep economic crisis, riding a wave of anti-austerity and Anti-bailout anger among Greeks! It lost to Prime Minister Kyriakos Mitsotakis's New Democracy in 2019.
Αρχικό κείμενο	Greece's left-wing opposition leader, Alexis Tsipras, has announced his decision to step down after a crushing election defeat. Mr Tsipras, 48, served as Greece's prime minister from 2015 to 2019 during politically tumultuous years as the country struggled to remain in the euro zone and end a series of international bailouts. Advertisement In Sunday's general election, Mr Tsipras's left-wing Syriza party received just under 18% of the vote, while the winning New Democracy party topped 40%. Mr Tsipras is expected to stay on as leader until his successor is elected by the party membership. Advertisement
Segment 1	Greece's left-wing opposition leader, Alexis Tsipras, has announced his decision to step down after a crushing election defeat. Mr Tsipras, 48, served as Greece's prime minister from 2015 to 2019

Είδος	Κείμενο
	during politically tumultuous years as the country struggled to remain in the euro zone and end a series of international bailouts. Advertisement In Sunday's general election, Mr Tsipras's left-wing Syriza party received just under 18% of the vote, while the winning New Democracy party topped 40%. Mr Tsipras is expected to stay on as leader until his successor is elected by the party membership. Advertisement
Summary of segment 1	Alexis Tsipras, 48, served as Greece's prime minister from 2015 to 2019.<n>His left-wing Syriza party received just under 18% of the vote, while the winning New Democracy party topped 40% in Sunday's general election <n>Mr Tsipras is expected to stay on as leader until his successor is elected by the party membership.
Summary of article 4	Alexis Tsipras, 48, served as Greece's prime minister from 2015 to 2019. His left-wing Syriza party received just under 18% of the vote, while the winning New Democracy party topped 40% in Sunday's general election Mr Tsipras is expected to stay on as leader until his successor is elected by the party membership.
Summary of Article 0	Greek far-right groups swept up over 12% of the vote in Sunday's election. The surge of three parties with their ultra-nationalist views could swing public debate at home and prove a springboard in European elections, Ilias Kasidiaris from the banned Golden Dawn party endorsed it from his prison cell, mirroring the rise of populist politicians across Europe. Far-right gains in Greece mirror a trend in several other European countries. Support for Germany's anti-immigrant AfD party is at its highest since the wake of Europe's migrant crisis in 2018. In Italy a former far-Right activist is prime minister and in Sweden a far right party has joined the government. All three are expected to run in European Parliament elections next year. Mitsotakis has been resolute in promoting LGBTQ+ rights but has himself taken a hard line on migration, prompting criticism from rights groups. Kasidiaris is serving a 13-year jail term for his leadership role in Golden Dawn. Golden Dawn was declared a criminal gang linked to hate crimes in a 2020 court ruling, and was banned in Greece in 2011 It was once Greece's third largest party, but has been on the wane since the financial crisis of 2008
Summary of Article 1	Alexis Tsipras announced his decision to step down as leader of the left-wing Syriza party, days after a crushing general election defeat.

Είδος	Κείμενο
	Tsipras, 48, served as Greece's prime minister from 2015 to 2019 during politically tumultuous years as the country struggled to remain in the euro zone and end a series of international bailouts. Heading the Syriza since 2012, Tsipras forged a more cohesive party, taking it from a small political group to general election victory in 2015. He was widely praised by Western allies for finalizing an agreement with North Macedonia that ushered Greece's neighbor into NATO and advanced its effort to join the European Union. Mitsotakis: "Syriza was a party characterized by toxicity, divisive rhetoric and with striking inefficiency" Tsipras is expected to stay on as leader for several weeks until his successor is elected by the party's rank-an-file membership. Syriza and the Socialists have been unable to reach any agreement on potential collaboration. Syriza's poor election result was blamed on the largely-negative campaign of the party and its former allies, including Yanis Varoufakis, the finance minister during the 2010-2018 bailouts, and Pasok, a traditionally-strong Socialist party, which returned to power after a five-year hiatus.
Summary of Article 2	After four years of major political scandals, controlling the media, and accusations of illegally spying on political opponents and journalists, New Democracy won the elections with a twenty-point margin against Syriza. The return to power for the Left is likely to be a slow process and avoiding the repetition of the fall of Syrizo should be the main objective, argues Pavlos Tsimas and his co-authors, in this special edition of The Left's Revolution: How to Win the Fight for a Just and Progressive Europe Syriza promised a realistic possibility for progressive economic policies within the Eurozone. Between 2012 and 2015, the right-wing New Democracy party and the formerly social democratic Pasok implemented austerity-oriented economic policy programs that led to mass impoverishment and social unrest, according to this study, from 2012 to 2015, and openly critiqued the harsh and inefficient austerity policies. The first Syriza-led administration, including the finance minister Yanis Varoufakis, resigned, and new elections were announced for September 2015. The process of creating a surplus that eventually reached thirty-one billion euros included heavy taxation and the

Είδος	Κείμενο
	stagnation of social spending, which particularly hurt the poorer segments of Greek society. Syriza's narrative was that after the completion of certain austerity-oriented agreements, it was time to implement their own, more progressive agenda and return the accumulated budget surplus to society. Unsurprisingly, the majority of the voters were not convinced, and New Democracy won themajority of seats in the parliament, forming the first single-party government since 2009. Social unrest caused by the party's authoritarian public health approach pushed the government to spend a big share of the surplus created by Syriza in wage subsidies and social benefits. The New Democracy administration also spent twenty million euros on COVID-19 public Health media campaigns. But soon after, criticism emerged claiming that the distribution of these funds was unequal based on political affiliation with the governing party. Recent evidence suggests that the Greek intelligence services used illegal surveillance systems to spy on journalists and political opponents. During the final months of the New Democracy administration, a major train crash killed fifty-seven people, and many believed the left's chances of returning to power have increased significantly. Three far-right parties now control over 10 percent of the seats in the Greek parliament. New Democracy took advantage of this turn of events and argued that they, not Syriza, are the party that actually cares about working-class people, says Yanis Varoufakis, author of 'Syriza: The New Democrats' This approach missed that a precarious and impoverished population primarily cares about its economic survival. The party perhaps recognized that financial markets would have reacted very aggressively against the election of a left-wing government, leading to higher borrowing rates.
Summary of article 3	"The time has come to start a new cycle," Tsipras said in a televised address. Led by Tsipras, Syriza stormed to power in 2015 at the height of Greece's deep economic crisis, riding a wave of anti-austerity and Anti-bailout anger among Greeks! It lost to Prime Minister Kyriakos Mitsotakis's New Democracy in 2019.
Summary of article 4	Alexis Tsipras, 48, served as Greece's prime minister from 2015 to 2019. His left-wing Syriza party received just under 18% of the vote, while the winning New Democracy party topped 40% in Sunday's

Είδος	Κείμενο
	general election Mr Tsipras is expected to stay on as leader until his successor is elected by the party membership.
Summary of summaries	Greek far-right groups swept up over 12% of the vote in Sunday's election. The surge of three parties with their ultra-nationalist views could swing public debate at home and prove a springboard in European elections Ilias Kasidiaris from the banned Golden Dawn party endorsed it from his prison cell, mirroring the rise of populist politicians across Europe. Tsipras, 48, served as Greece's prime minister from 2015 to 2019. He forged a more cohesive party, taking it from a small political group to general election victory in 2015. Tsipras is expected to stay on as leader for several weeks until his successor is elected by his rank-an-file membership. After four years of major political scandals, New Democracy won the elections with a twenty-point margin against Syriza. The return to power for the Left is likely to be a slow process and avoiding the repetition of the fall of Syrizo should be the main objective, argues Pavlos Tsimas and his co-authors. Syriza's narrative was that after the completion of certain austerity-oriented agreements, it was time to implement their own, more progressive agenda and return the accumulated budget surplus to society. Unsurprisingly, the majority of the voters were not convinced, and New Democracy won themajority of seats in the parliament, forming the first single-party government since 2009. New Democracy also spent twenty million euros on COVID-19 public Health media campaigns. Evidence suggests Greek intelligence services spied on journalists and political opponents. Yanis Varoufakis: New Democracy took advantage of turn of events and argued that Syriza is not the party that cares about working-class people' Varoufikis: Financial markets would have reacted aggressively against the election of a left-wing government, leading to higher borrowing rates. Alexis Tsipras, 48, served as Greece's prime minister from 2015 to 2019. His left-wing Syriza party received just under 18% of the vote, while the winning New Democracy party topped 40% in Sunday's general election Mr Tsipras is expected to stay on as leader until his successor is elected by the party membership.

Αφαιρετική και στη συνέχεια αποσπασματική περίληψη (τα αρχικά άρθρα παραλείπονται καθώς έχουν παρατεθεί παραπάνω)

Είδος	**Κείμενο**
Article Summary 0	ATHENS, June 29 (Reuters) - With backing from a politician jailed for leading a party declared a criminal gang, Greek far-right groups swept up over 12% of the vote in Sunday's election, mirroring the rise of populist and ultra-nationalist politicians across Europe. The surge of three parties with their ultra-nationalist views - including 'Spartans' which barely registered in polls until Ilias Kasidiaris from the banned Golden Dawn party endorsed it from his prison cell -- could swing public debate at home and prove a springboard in European elections. While the centre-right New Democracy party of Kyriakos Mitsotakis stormed to victory in the June 25 Greek vote, winning 158 seats in the 300-seat parliament, the Spartans emerged as the fifth largest group. "The re-emergence of the far right is a byproduct of a political strategy of the New Democracy government which tried to appeal to the centre with an agenda of economic liberalism and, at the same time to the far right with an agenda of law and order and anti-immigrant discourse," said Akritas Kaidatzis, assistant professor of constitutional law at the Aristotle University in Thessaloniki. Support for Germany's anti-immigrant AfD party is at its highest since the wake of Europe’s migrant crisis in 2018, in Italy a former far-right activist is prime minister and in Sweden a far-right party has joined the government.
Summary of Article 1	Tsipras, 48, served as Greece's prime minister from 2015 to 2019 during politically tumultuous years as the country struggled to remain in the euro zone and end a series of international bailouts. “I have therefore decided to propose the election of a new leadership by the members of the party ... Of course I will not be a candidate,” Tsipras said in a televised address. Heading the Syriza since 2012, Tsipras forged a more cohesive party, taking it from a small political group to general election victory in 2015 on a pledge to push back on harsh austerity measures demanded by bailout lenders from other euro zone members and the International Monetary Fund. No prominent members of the party has publicly called on Tsipras to step down after the election defeat, though Euclid Tsakalotos, a former Syriza finance minister, had urged him to reflect on the results and “take the necessary actions.” Effie Achtsioglou, a 38-year-old former social security minister, has received support from

Είδος	Κείμενο
	a section of the party to seek a leadership role but has not publicly discussed her plans. Commentators blamed Syriza's poor election result on the party's largely-negative campaign, the resurgence of the traditionally-strong Socialist party Pasok, and the appearance of splinter parties headed by Tsipras's former allies, including former Finance Minister Yanis Varoufakis.
Summary of Article 2	After four years of major political scandals, controlling the media, and accusations of illegally spying on political opponents and journalists, New Democracy won the elections with a twenty-point margin against Syriza.The return to power for the Left is likely to be a slow process and avoiding the repetition of the fall of Syrizo should be the main objective, argues Pavlos Tsimas and his co-authors, in this special edition of The Left's Revolution: How to Win the Fight for a Just and Progressive Europe Syriza promised a realistic possibility for progressive economic policies within the Eurozone. Between 2012 and 2015, the right-wing New Democracy party and the formerly social democratic Pasok implemented austerity-oriented economic policy programs that led to mass impoverishment and social unrest, according to this study, from 2012 to 2015, and openly critiqued the harsh and inefficient austerity policies. The first Syriza-led administration, including the finance minister Yanis Varoufakis, resigned, and new elections were announced for September 2015. The process of creating a surplus that eventually reached thirty-one billion euros included heavy taxation and the stagnation of social spending, which particularly hurt the poorer segments of Greek society. Syriza's narrative was that after the completion of certain austerity-oriented agreements, it was time to implement their own, more progressive agenda and return the accumulated budget surplus to society. Unsurprisingly, the majority of the voters were not convinced, and New Democracy won themajority of seats in the parliament, forming the first single-party government since 2009. Social unrest caused by the party's authoritarian public health approach pushed the government to spend a big share of the surplus created by Syriza in wage subsidies and social benefits. The New Democracy administration also spent twenty million euros on COVID-19 public Health media campaigns. But soon after,

Είδος	Κείμενο
	criticism emerged claiming that the distribution of these funds was unequal based on political affiliation with the governing party. Recent evidence suggests that the Greek intelligence services used illegal surveillance systems to spy on journalists and political opponents. During the final months of the New Democracy administration, a major train crash killed fifty-seven people, and many believed the left's chances of returning to power have increased significantly. Three far-right parties now control over 10 percent of the seats in the Greek parliament. New Democracy took advantage of this turn of events and argued that they, not Syriza, are the party that actually cares about working-class people, says Yanis Varoufakis, author of 'Syriza: The New Democrats' This approach missed that a precarious and impoverished population primarily cares about its economic survival. The party perhaps recognized that financial markets would have reacted very aggressively against the election of a left-wing government, leading to higher borrowing rates.
Αρχικό κείμενο	Greece's Alexis Tsipras has stepped down from the helm of the leftist Syriza party following a heavy election defeat. "The time has come to start a new cycle," Tsipras said in a televised address, adding that reform of the party was needed. Led by Tsipras, Syriza stormed to power in 2015 at the height of Greece's deep economic crisis, riding a wave of anti-austerity and anti-bailout anger among Greeks. It lost to Prime Minister Kyriakos Mitsotakis's New Democracy in 2019. In Sunday's vote, Syriza won just 17.8 per cent of the votes against 40.5 per cent for New Democracy. "The negative result can - and must - become the beginning of this cycle," Tsipras said. He said he was stepping down to pave the way for elections for a new party leader, saying he would not be a candidate. "I am proud of everything that happened," he said. "This difficult journey had compromises, and difficult decisions, and injuries and attrition, but it was a journey that left a mark on history."
Segment 1	Greece's Alexis Tsipras has stepped down from the helm of the leftist Syriza party following a heavy election defeat. "The time has

Είδος	Κείμενο
	come to start a new cycle," Tsipras said in a televised address, adding that reform of the party was needed. Led by Tsipras, Syriza stormed to power in 2015 at the height of Greece's deep economic crisis, riding a wave of anti-austerity and anti-bailout anger among Greeks. It lost to Prime Minister Kyriakos Mitsotakis's New Democracy in 2019. In Sunday's vote, Syriza won just 17.8 per cent of the votes against 40.5 per cent for New Democracy. "The negative result can - and must - become the beginning of this cycle," Tsipras said. He said he was stepping down to pave the way for elections for a new party leader, saying he would not be a candidate. "I am proud of everything that happened," he said. "This difficult journey had compromises, and difficult decisions, and injuries and attrition, but it was a journey that left a mark on history."
Summary of segment 1	"The time has come to start a new cycle," Tsipras said in a televised address.<n>Led by Tsipras, Syriza stormed to power in 2015 at the height of Greece's deep economic crisis, riding a wave of anti-austerity and Anti-bailout anger among Greeks!<n>It lost to Prime Minister Kyriakos Mitsotakis's New Democracy in 2019.
Summary of article 3	"The time has come to start a new cycle," Tsipras said in a televised address. Led by Tsipras, Syriza stormed to power in 2015 at the height of Greece's deep economic crisis, riding a wave of anti-austerity and Anti-bailout anger among Greeks! It lost to Prime Minister Kyriakos Mitsotakis's New Democracy in 2019.
Αρχικό κείμενο	Greece's left-wing opposition leader, Alexis Tsipras, has announced his decision to step down after a crushing election defeat. Mr Tsipras, 48, served as Greece's prime minister from 2015 to 2019 during politically tumultuous years as the country struggled to remain in the euro zone and end a series of international bailouts. Advertisement In Sunday's general election, Mr Tsipras's left-wing Syriza party received just under 18% of the vote, while the winning New Democracy party topped 40%. Mr Tsipras is expected to stay on as leader until his successor is elected by the party membership. Advertisement
Segment 1	Greece's left-wing opposition leader, Alexis Tsipras, has announced his decision to step down after a crushing election defeat. Mr Tsipras, 48, served as Greece's prime minister from 2015 to 2019 during politically tumultuous years as the country

Είδος	Κείμενο
	struggled to remain in the euro zone and end a series of international bailouts. Advertisement In Sunday's general election, Mr Tsipras's left-wing Syriza party received just under 18% of the vote, while the winning New Democracy party topped 40%. Mr Tsipras is expected to stay on as leader until his successor is elected by the party membership. Advertisement
Summary of segment 1	Alexis Tsipras, 48, served as Greece's prime minister from 2015 to 2019.<n>His left-wing Syriza party received just under 18% of the vote, while the winning New Democracy party topped 40% in Sunday's general election <n>Mr Tsipras is expected to stay on as leader until his successor is elected by the party membership.
Summary of article 4	Alexis Tsipras, 48, served as Greece's prime minister from 2015 to 2019. His left-wing Syriza party received just under 18% of the vote, while the winning New Democracy party topped 40% in Sunday's general election Mr Tsipras is expected to stay on as leader until his successor is elected by the party membership.
Summary of Article 0	Greek far-right groups swept up over 12% of the vote in Sunday's election. The surge of three parties with their ultra-nationalist views could swing public debate at home and prove a springboard in European elections, Ilias Kasidiaris from the banned Golden Dawn party endorsed it from his prison cell, mirroring the rise of populist politicians across Europe. Far-right gains in Greece mirror a trend in several other European countries. Support for Germany's anti-immigrant AfD party is at its highest since the wake of Europe's migrant crisis in 2018. In Italy a former far-Right activist is prime minister and in Sweden a far right party has joined the government. All three are expected to run in European Parliament elections next year. Mitsotakis has been resolute in promoting LGBTQ+ rights but has himself taken a hard line on migration, prompting criticism from rights groups. Kasidiaris is serving a 13-year jail term for his leadership role in Golden Dawn. Golden Dawn was declared a criminal gang linked to hate crimes in a 2020 court ruling, and was banned in Greece in 2011 It was once Greece's third largest party, but has been on the wane since the financial crisis of 2008
Summary of Article 1	Alexis Tsipras announced his decision to step down as leader of the left-wing Syriza party, days after a crushing general election defeat. Tsipras, 48, served as Greece's prime minister from 2015 to 2019

Είδος	Κείμενο
	during politically tumultuous years as the country struggled to remain in the euro zone and end a series of international bailouts. Heading the Syriza since 2012, Tsipras forged a more cohesive party, taking it from a small political group to general election victory in 2015. He was widely praised by Western allies for finalizing an agreement with North Macedonia that ushered Greece's neighbor into NATO and advanced its effort to join the European Union. Mitsotakis: "Syriza was a party characterized by toxicity, divisive rhetoric and with striking inefficiency" Tsipras is expected to stay on as leader for several weeks until his successor is elected by the party's rank-an-file membership. Syriza and the Socialists have been unable to reach any agreement on potential collaboration. Syriza's poor election result was blamed on the largely-negative campaign of the party and its former allies, including Yanis Varoufakis, the finance minister during the 2010-2018 bailouts, and Pasok, a traditionally-strong Socialist party, which returned to power after a five-year hiatus.
Summary of Article 2	The return to power for the Left is likely to be a slow process, and avoiding the repetition of the fall of Syriza should be the main objective. After five years of irrelevant austerity policies imposed on Greece by the European Union and the International Monetary Fund as part of the economic adjustment programs, the Greek people voted in favor of a potential rupture with the EU. The last act of Syriza's attempt to change the direction of economic policies within the Eurozone was the 2015 Greek bailout referendum. While an overwhelming 61.3 percent of the popular vote was against new austerity policies and the support for a Grexit was higher than ever, Prime Minister Aléxis Tsípras decided to go against the result and sign a new bailout agreement. People accepted that this was the only alternative after previous failed attempts to change EU economic policies, and Tsípras's Syriza re-won the majority of the parliament and formed a new coalition government. Unsurprisingly, the majority of the voters were not convinced, and New Democracy won the majority of seats in the parliament, forming the first single-party government since 2009. A few months after the reelection of New Democracy, the pandemic started and the new government had to abandon its austerity-oriented agenda out of necessity. Social unrest

Είδος	Κείμενο
	caused by the party's authoritarian public health approach and disinvestment in public health care, which caused thousands of excess deaths, pushed the government to spend a big share of the surplus created by Syriza in wage subsidies and social benefits. This devastating event was largely the outcome of disinvestment in public infrastructure and the privatization of the railways, since no signaling system had been in place for years and the operators were not properly trained. First, the temporary ease of fiscal constraints by the EU due to COVID allowed the New Democracy government to provide some benefits in cash as well as coupons in response to the cost-of-living crisis, using the budget surplus that Syriza generated.
Summary of article 3	Greece's Alexis Tsipras has stepped down from the helm of the leftist Syriza party following a heavy election defeat. "The time has come to start a new cycle," Tsipras said in a televised address, adding that reform of the party was needed. Led by Tsipras, Syriza stormed to power in 2015 at the height of Greece's deep economic crisis, riding a wave of anti-austerity and anti-bailout anger among Greeks. He said he was stepping down to pave the way for elections for a new party leader, saying he would not be a candidate. "This difficult journey had compromises, and difficult decisions, and injuries and attrition, but it was a journey that left a mark on history."
Summary of article 4	Greece's left-wing opposition leader, Alexis Tsipras, has announced his decision to step down after a crushing election defeat. Mr Tsipras, 48, served as Greece's prime minister from 2015 to 2019 during politically tumultuous years as the country struggled to remain in the euro zone and end a series of international bailouts. Advertisement In Sunday's general election, Mr Tsipras's left-wing Syriza party received just under 18% of the vote, while the winning New Democracy party topped 40%. Mr Tsipras is expected to stay on as leader until his successor is elected by the party membership. Advertisement
Summary of summaries	Greek far-right groups swept up over 12% of the vote in Sunday's election. The surge of three parties with their ultra-nationalist views could swing public debate at home and prove a springboard in European elections, Ilias Kasidiaris from the banned Golden Dawn party endorsed it from his prison cell, mirroring the rise of populist politicians across Europe.

Είδος	Κείμενο
	Tsipras served as Greece's prime minister from 2015 to 2019. He forged a more cohesive party, taking it from a small political group to general election victory in 2015 on a pledge to push back on harsh austerity measures demanded by bailout lenders from other euro zone members and the International Monetary Fund. Commentators blamed Syriza's poor election result on the party's largely-negative campaign, the resurgence of the traditionally-strong Socialist party Pasok, and the appearance of splinter parties. After five years of irrelevant austerity policies, Greek people voted in favor of a potential rupture with the EU, writes Pavlos Tsimas, who argues that the return to power for the Left is likely to be a slow process, with a main objective of avoiding the repetition of this fall of Tsipras' party, is the main goal. Tspras's Syriza re-won the majority of the parliament and formed a new coalition government. A few months after the reelection of New Democracy, the pandemic started and the new government had to abandon its austerity-oriented agenda out of necessity Social unrest caused by the party's authoritarian public health approach pushed the government to spend a big share of its surplus on wage subsidies and social benefits. "The time has come to start a new cycle," Tsipras said in a televised address. Led by Tsipras, Syriza stormed to power in 2015 at the height of Greece's deep economic crisis, riding a wave of anti-austerity and Anti-bailout anger among Greeks. Mr Tsipras, 48, served as Greece's prime minister from 2015 to 2019. During that time, the country struggled to remain in the euro zone and end a series of international bailouts.

Υπεύθυνη Δήλωση Συγγραφέα:

Δηλώνω ρητά ότι, σύμφωνα με το άρθρο 8 του Ν.1599/1986, η παρούσα εργασία αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης.