Δράση: ΕΡΕΥΝΩ – ΔΗΜΙΟΥΡΓΩ - ΚΑΙΝΟΤΟΜΩ

MediaPot [ΤΑΕΔΚ-06196]: Πλατφόρμα συλλογής, ανάλυσης και σύνθεσης πολυμεσικού περιεχομένου από κοινωνικά δίκτυα στην υπηρεσία των Ψηφιακών Μέσων

Π3.1 Μεθοδολογία Εξόρυξης Οντοτήτων

Ενότητα Εργασίας	ΕΕ3: Ανάλυση Φυσικής Γλώσσας και Κοινωνικών Δικτύων
Ημερομηνία	31/05/2024
Τύπος εγγράφου	Final
Υπεύθυνος Φορέας	ΕΚΕΦΕ «Δημόκριτος»
Συμμετέχοντες Φορείς	
Επιμελητές	Σωτήρης Λέγκας, Ελισάβετ Παλογιαννίδη, Σωτήρης Νικολέτος, Βογιατζής Δημήτριος, Αναστασία Κριθαρά
Συνοπτική περιγραφή	Το παραδοτέο περιγράφει το *χρονικό γράφο* που δημιουργείται μετά από την εξόρυξη οντοτήτων από ειδησεογραφικά κείμενα. Αναφέρεται η μορφή των κειμένων, συμπεριλαμβανομένης της μορφής κειμένων από κοινωνικά δίκτυα. Επίσης εντοπίζονται θέματα σχετικά με την ποιότητα δεδομένων. Στη συνέχεια αναφέρεται το API που αναπτύχθηκε προκειμένου να διευκολύνει την πρόσβαση στο γράφο.

Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Περιεχόμενα

Εικόνες

Πίνακες

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Υπόμνημα

Η εργασία υλοποιήθηκε στο πλαίσιο της Δράσης ΕΡΕΥΝΩ – ΔΗΜΙΟΥΡΓΩ – ΚΑΙΝΟΤΟΜΩ συγχρηματοδοτήθηκε από το Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης (ΕΤΠΑ) της Ευρωπαϊκής Ένωσης και εθνικούς πόρους μέσω του Ε.Π. Ανταγωνιστικότητα, Επιχειρηματικότητα & Καινοτομία (ΕΠΑνΕΚ) (κωδικός έργου MediaPot: ΤΑΕΔΚ-06196)

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

1 Εισαγωγή

Ο σκοπός του παραδοτέου είναι να περιγράψει διάφορες μεθοδολογίες εξόρυξης οντοτήτων που εφαρμόστηκαν (ή αναμένεται να εφαρμοστούν) σε κείμενα. Στη συνέχεια αναφερόμαστε στη μορφή των κειμένων (άρθρων εν προκειμένω) όπου εφαρμόστηκαν οι εν λόγω μεθοδολογίες.

Επίσης παρήχθησαν και σχέσεις μεταξύ των οντοτήτων και των άρθρων. Το τελικό αποτέλεσμα είναι ένας χρονικός γράφος (temporal graph) οντοτήτων, σχέσεων, και άρθρων. Παρέχουμε επίσης κάποιες στατιστικές πληροφορίες για τον χρονικό γράφο. Τέλος αναφερόμαστε στην διεπαφή χρήστη (Web API), που σχεδιάστηκε και υλοποιήθηκε προκειμένου να διευκολυνθεί η πρόσβαση στον γράφο των οντοτήτων και των σχέσεων.

2 Μορφή Δεδομένων

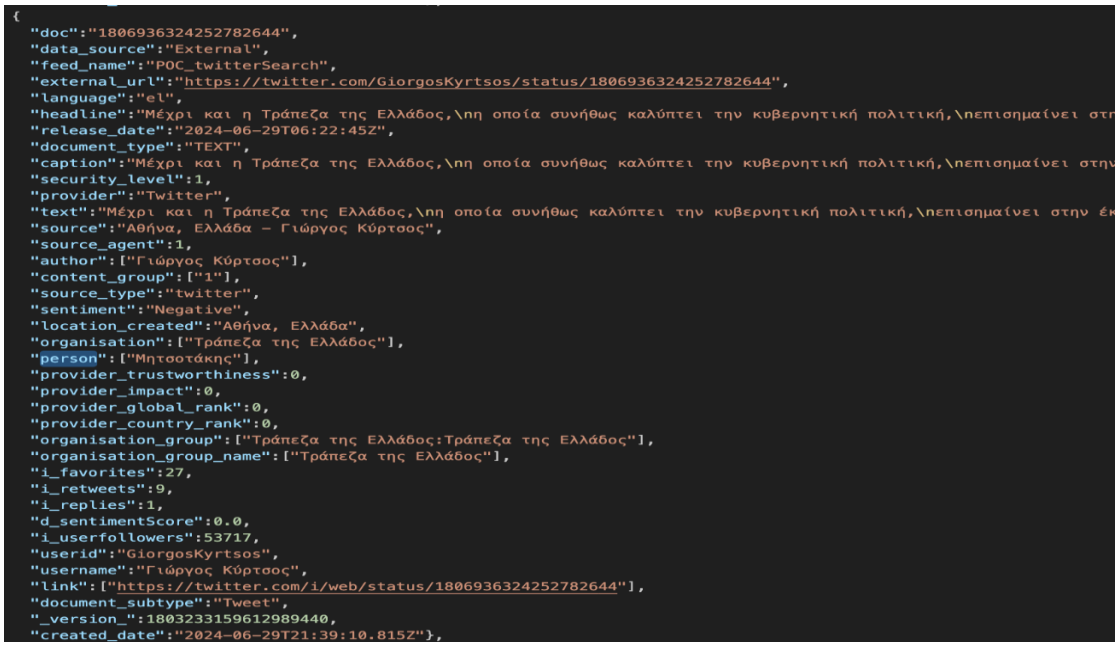
Σε αυτό το σημείο θα αναφερθούμε στην μορφή των δεδομένων επί των οποίων εφαρμόστηκε η *Εξόρυξη Οντοτήτων (ΕΟ).* Τα δεδομένα είναι μία συλλογή κειμένων, και κάθε κείμενο αναπαρίσταται με τη μορφή JSON. Η προέλευση των δεδομένων είναι από *ιστοσελίδες ελληνικών μέσων ενημέρωσης*. Πέρα από το ίδιο το κείμενο (πεδίο content) υπάρχουν και διάφορα μεταδεδομένα όπως η ημερομηνία δημοσίευσης, το URL του αρχικής δημοσίευσης, κάποιες επισημειώσεις (tags) κ.α.

Η πρώτη συλλογή δεδομένων που ελήφθη από την ATC αποτελείται από 20 χιλιάδες άρθρα από ειδησεογραφικές ιστοσελίδες στα ελληνικά. Επίσης υπάρχουν και μεταδεδομένα όπως:

- η ημερομηνία
- το URL της ιστοσελίδας
- εικόνες
- “λέξεις κλειδιά”
- οντότητες που αφορούν τις κατηγορίες *πρόσωπο (person), οργανισμός, (organization), τοποθεσία (location)*

Υπάρχει επίσης και μία συλλογή *9,014 Tweets*, η οποία ελήφθη επίσης από την ATC. Κάθε Tweet περιέχει πληθώρα μεταδεδομένων όπως φαίνεται και στο παρακάτω παράδειγμα. Από αυτά επιλέγουμε να κρατήσουμε και να χρησιμοποιήσουμε τα προαναφερθέντα (*ημερομηνία, το URL της ιστοσελίδας, εικόνες*, *λέξεις κλειδιά*, που αφορούν τις κατηγορίες *πρόσωπο*, *οργανισμός*, *τοποθεσία*).

Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης



```
{
  "doc":"1806936324252782644",
  "data_source":"External",
  "feed_name":"POC_twitterSearch",
  "external_url":"https://twitter.com/GiorgosKyrtsos/status/1806936324252782644",
  "language":"el",
  "headline":"Μέχρι και η Τράπεζα της Ελλάδος,\nη οποία συνήθως καλύπτει την κυβερνητική πολιτική,\nεπισημαίνει στη
  "release_date":"2024-06-29T06:22:45Z",
  "document_type":"TEXT",
  "caption":"Μέχρι και η Τράπεζα της Ελλάδος,\nη οποία συνήθως καλύπτει την κυβερνητική πολιτική,\nεπισημαίνει στην
  "security_level":1,
  "provider":"Twitter",
  "text":"Μέχρι και η Τράπεζα της Ελλάδος,\nη οποία συνήθως καλύπτει την κυβερνητική πολιτική,\nεπισημαίνει στην έ
  "source":"Αθήνα, Ελλάδα – Γιώργος Κύρτσος",
  "source_agent":1,
  "author":["Γιώργος Κύρτσος"],
  "content_group":["1"],
  "source_type":"twitter",
  "sentiment":"Negative",
  "location_created":"Αθήνα, Ελλάδα",
  "organisation":["Τράπεζα της Ελλάδος"],
  "person":["Μητσοτάκης"],
  "provider_trustworthiness":0,
  "provider_impact":0,
  "provider_global_rank":0,
  "provider_country_rank":0,
  "organisation_group":["Τράπεζα της Ελλάδος:Τράπεζα της Ελλάδος"],
  "organisation_group_name":["Τράπεζα της Ελλάδος"],
  "i_favorites":27,
  "i_retweets":9,
  "i_replies":1,
  "d_sentimentScore":0.0,
  "i_userfollowers":53717,
  "userid":"GiorgosKyrtsos",
  "username":"Γιώργος Κύρτσος",
  "link":["https://twitter.com/i/web/status/1806936324252782644"],
  "document_subtype":"Tweet",
  "_version_":1803233159612989440,
  "created_date":"2024-06-29T21:39:10.815Z"},
```


Εικόνα 1: Παράδειγμα ενός Tweet σε μορφή JSON

3 Μέθοδοι Εξόρυξης Οντοτήτων

3.1 Διάφορες προσεγγίσεις

Το πρόβλημα που θα προσεγγίσουμε αφορά στην ΕΟ στα ελληνικά, με μεθόδους που είναι διαθέσιμες μέσω ελεύθερου λογισμικού. Στη συνέχεια θα παρουσιάσουμε κάποιες από τις μεθόδους που έχουν αναπτυχθεί για τα Ελληνικά. Εδώ πρέπει να τονίσουμε ότι γενικότερα δεν υπάρχουν πολλά γλωσσικά εργαλεία για την Ελληνική γλώσσα σε σχέση με άλλες ευρέως ομιλούμενες γλώσσες.[1]

Γενικά θα μπορούσαμε να κατατάξουμε τις προσεγγίσεις για ΕΟ στις παρακάτω κατηγορίες.

- ΕΟ με προκατασκευασμένους κανόνες (rule-based) και λίστες ονομάτων (gazetteers)
- Συστήματα με βασίζονται σε τεχνικές μηχανικής μάθησης
- Συνδυασμοί των παραπάνω
- Μεγάλα Γλωσσικά Μοντέλα (LLMs)

Μία από τις παλαιότερες προσεγγίσεις εκτίθεται στο [1]. Χρησιμοποιήθηκαν συστήματα κανόνων, καθώς και μηχανική μάθησης. Στο [2] χρησιμοποιήθηκαν επίσης ένα σύστημα 110 κανόνων που εφαρμόστηκε σε ένα σύνολο 12,000,000 λέξεων. Ένα προϋπάρχον σύνολο κανόνων ενημερώνεται με τεχνικές μηχανικής μάθησης στο [3].

Μία νεότερη προσέγγιση για τα Ελληνικά βασίζεται στο μοντέλο βαθιάς μάθησης BERT, το οποίο εκπαιδεύτηκε πάνω σε μία μεγάλη συλλογή ελληνικών κειμένων προκειμένου το παραχθεί το greek-

[1] http://nlp.cs.aueb.gr/software.html, https://www.clarin.gr/el/content/nlpel-clarin-knowledge-centre-natural-language-processing-greece

Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

BERT [4]. Το greek-BERT μπορεί να χρησιμοποιηθεί, μεταξύ των άλλων, και για εξόρυξη οντοτήτων και είναι δημόσια διαθέσιμο. [2]

Μία δημοφιλής βιβλιοθήκη για την επεξεργασία φυσικής γλώσσας (ΕΦΓ) είναι η spaCY.[3] Η spaCY υποστηρίζει πολλές γλώσσες συμπεριλαμβανομένων τον ελληνικών.[4] Είναι δε δημόσια διαθέσιμη και μπορεί να χρησιμοποιηθεί για ΕΟ.

Η NLTK είναι επίσης μία δημοφιλής βιβλιοθήκη για ΕΦΓ.[5] Δεν παρέχει υποστήριξη για τα ελληνικά, αλλά έχει τις κατάλληλες διεπαφές ώστε να ενσωματωθούν μονάδες ελληνικών από διάφορες άλλες πηγές.

Η FLAIR είναι μία βιβλιοθήκη ΕΦΚ που αναπτύχθηκε από τη Zalando Research .[6] Προσφέρει πολλές δυνατότητες, μεταξύ των άλλων και ΕΟ. Μπορεί να χρησιμοποιηθεί μέσω της Python, ενώ βασίζεται στο PyTorch. Η FLAIR είναι πολυγλωσσική, και μέσα στις γλώσσες που υποστηρίζει είναι και τα ελληνικά. Αλλά αναμένεται η ποιότητα διαφόρων εργασιών ΦΓ να είναι χαμηλότερη στα ελληνικά απ' ότι σε άλλες ευρέως ομιλούμενες γλώσσες.

Τέλος, τα Μεγάλα Γλωσσικά Μοντέλα (LLMs) βελτιώνουν τη διαδικασία εξαγωγής οντοτήτων, αξιοποιώντας την εκτενή εκπαίδευσή τους σε ποικίλα δεδομένα, επιτρέποντάς τους να κατανοούν το πλαίσιο και τις ιδιαιτερότητες της γλώσσας πιο αποτελεσματικά από τις παραδοσιακές μεθόδους (Zhao et al. 2023), (Minae et al. 2024). Τα μοντέλα αυτά μπορούν να εκπαιδευτούν περαιτέρω με επιβλεπόμενη μάθηση ή ακόμα και να χρησιμοποιηθούν απλώς με τη ρύθμιση των "οδηγιών" του μοντέλου (prompt engineering), ώστε να επιτελέσει το συγκεκριμένο έργο. Επίσης βιβλιοθήκες όπως το LangChain, και LlamaIndex μπορούν να διευκολύνουν την ένταξη ενός LLM σε ένα ευρύτερο σύστημα λογισμικού (Rostam et al. 2024).

4 Δημιουργία γράφου οντοτήτων και σχέσεων

Η πρώτη συλλογή δεδομένων που προαναφέρθηκε στην ενότητα 2 προέρχεται από την ΑΤΟ. Σε αυτά τα δεδομένα η εξόρυξη των οντοτήτων έγινε με την χρήση της βιβλιοθήκης SpaCy που αναφέρθηκε στην ενότητα 3. Οι τύποι των οντοτήτων είναι *τοποθεσία*, *οργανισμός* και *πρόσωπο*. Ουσιαστικά τα άρθρα (κόμβοι) συνδέονται με τις οντότητες (κόμβους) στον γράφο.

Προσθέτοντας σε μία βάση δεδομένων για γράφους (Neo4j[7]), τα άρθρα μαζί με τις οντότητες, καταλήξαμε σε έναν γράφο (Εικόνα 2) με κόμβους 20,000 τίτλους άρθρων, με 26,042 κόμβους για τους οργανισμούς, 33,326 κόμβους για τα πρόσωπα και 14,857 κόμβους για τις τοποθεσίες. Τέλος, οι σχέσεις μεταξύ των κόμβων ανήλθαν σε 200,317.

[2] https://github.com/nlpaueb/greek-bert

[3] https://spacy.io/

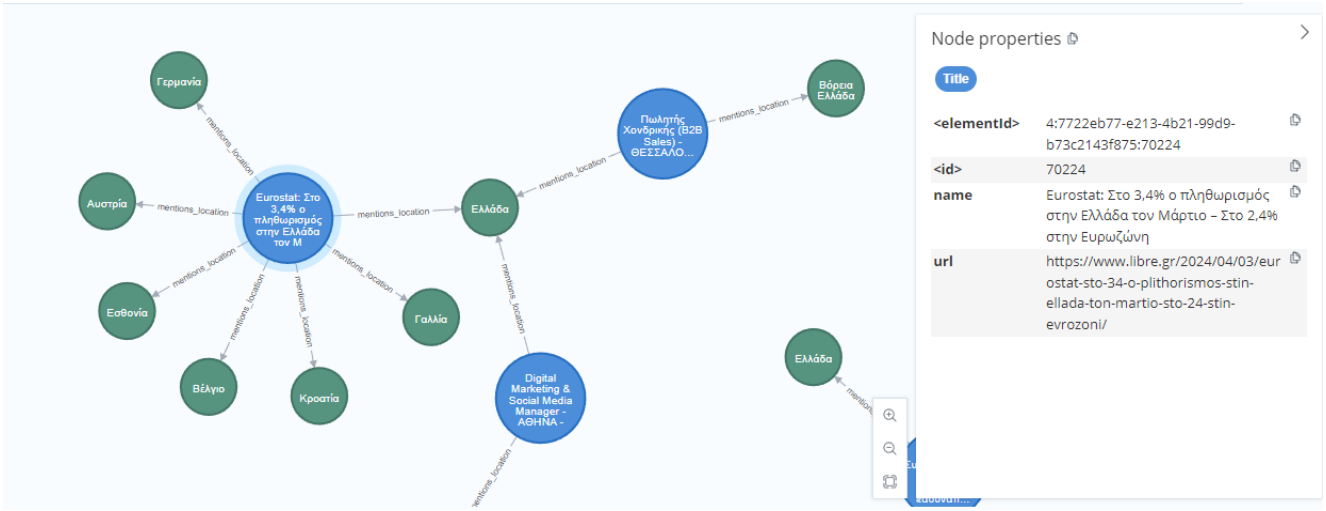
[4] https://spacy.io/models/el

[5] https://www.nltk.org/

[6] https://flairnlp.github.io/

[7] https://neo4j.com/

Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης



Εικόνα 2: Στιγμιότυπο ενός μέρους του γράφου από την βάση δεδομένων Neo4j. Με μπλε απεικονίζονται οι κόμβοι με τον τίτλο των άρθρων, και με πράσινο οι κόμβοι των οντοτήτων.

Υπολογίζοντας το PageRank (Zafarani, R. 2014), που μετρά την σημαντικότητα κάθε κόμβου μέσα στον γράφο με βάση τον αριθμό των εισερχόμενων σχέσεων και τη σημασία των συγκεκριμένων κόμβων[8], παρατηρήθηκε ότι ο μέσος όρος του PageRank των κόμβων ανέρχεται σε 0.94 και η διάμεσος σε 0.47. Συμπεραίνεται λοιπόν, ότι οι πολλοί κόμβοι έχουν κατά μέσο όρο χαμηλή σημασία, εφόσον οι περισσότεροι κόμβοι έχουν χαμηλότερο PageRank από το μέσο όρο. Οι σημαντικότεροι κόμβοι με το μεγαλύτερο PageRank αναφέρονται στον παρακάτω πίνακα:

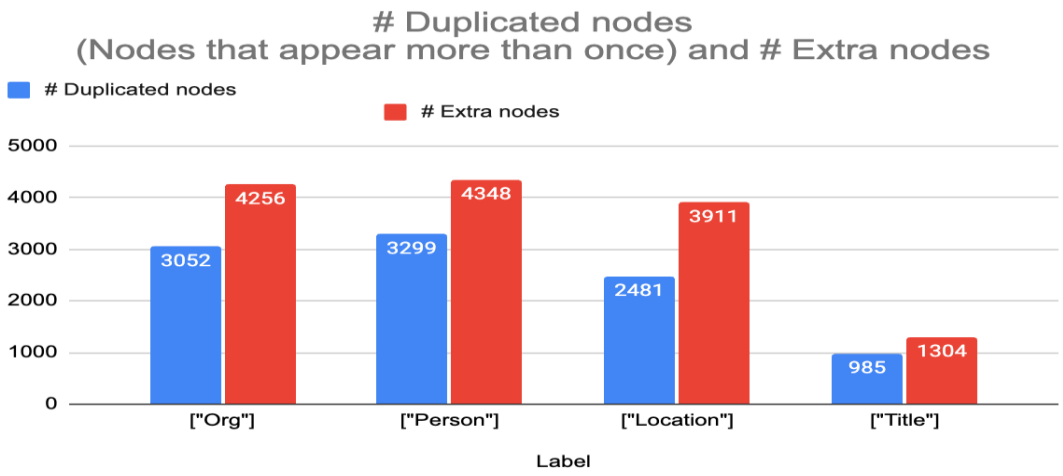
Node	PageRank
"Ελλάδα"	221.04
"Ελλάδα"	114.15
"Google News"	114.11
"Τι θα δούμε στην Επίδαυρο"	105.95
"Κόσμο"	101.94
"La tormenta más fría del año azota el sur de California con nieve y lluvia"	88.24
"Ελλάδα"	77.72
"Αστυνομία"	69.46

Πίνακας 1 Κόμβοι και βαθμολόγηση με PageRank

[8] Υποθέσαμε ότι κάθε κόμβος είναι εξίσου σημαντικός με τους κόμβους που συνδέονται με αυτόν.

Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Η ύπαρξη διπλότυπων κόμβων και η εμφάνιση παραπλήσιων οντοτήτων που αναφέρονται ουσιαστικά στην ίδια οντότητα (Εικόνα 3), οδήγησε στην εφαρμογή τεχνικών εκκαθάρισης του γράφου.



Εικόνα 3: Στο γράφημα εμφανίζεται ο αριθμός των διπλότυπων κόμβων (μπλε) και ο αριθμός των παραπλήσιων κόμβων (κόκκινο).

Ο καθαρισμ'ος του γράφου υλοποιήθηκε με τις παρακάτω τεχνικές:

- Διαγραφή των ορφανών κόμβων
- Αφαίρεση των κόμβων που οι τιμές τους δεν είναι ούτε ελληνικές ούτε αγγλικές
- Κανονικοποίηση του κειμένου, δηλαδή μετατροπή σε κεφαλαία, αφαίρεση των τόνων και των μη γραμματικών χαρακτήρων
- Εφαρμογή ανίχνευσης τοποθεσίας
- Εφαρμογή αφαίρεσης κατάληξης
- Εφαρμογή φιλτραρίσματος ομοιότητας
- Αφαίρεση διπλότυπων κόμβων
- Αφαίρεση των σχέσεων που περιλαμβάνουν κόμβους που έχουν διαγραφεί

Ο γράφος που δημιουργήθηκε μετά την εφαρμογή της εκκαθάρισης περιέχει 16,315 άρθρα, καθώς και 11,540 κόμβους για τους "οργανισμούς", 17,001 κόμβους για τα "πρόσωπα", 3,535 κόμβους για τις "τοποθεσίες" και 3,603 κόμβους για την επιπρόσθετη οντότητα "μέρος". Η συγκεκριμένη οντότητα δημιουργήθηκε με σκοπό τον διαχωρισμό της "τοποθεσίας" ώστε να περιέχει πόλεις, χώρες, ηπείρους και το "μέρος" ώστε να περιέχει οτιδήποτε επιπρόσθετο υπήρχε στην "τοποθεσία" (π.χ. ποδοσφαιρικό γήπεδο, εστιατόριο κ.α.). Τέλος, οι σχέσεις μεταξύ των κόμβων που παραμένουν μετά τον καθαρισμό, ανέρχονται σε 110,922.

5 Δημιουργία API

Η διαδικασία δημιουργίας του API περιλαμβάνει:

1. Τη δημιουργία μιας βάσης δεδομένων γράφων Neo4j, εισάγοντας τα δεδομένα που αναλύθηκαν στην ενότητα 2.

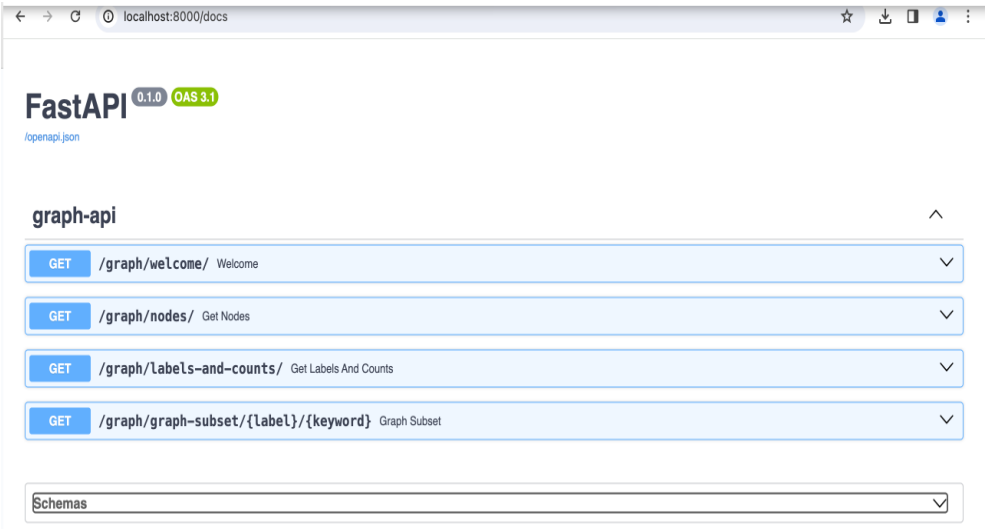
Ευρωπαϊκή Ένωση Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020 ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

2. Την ανάπτυξη ενός FastAPI backend[9], που θα είναι υπεύθυνο για την σύνδεση στην βάση δεδομένων και την εκτέλεση διαφόρων queries, ώστε να ανακτηθούν τα απαραίτητα δεδομένα (Εικόνα 4).
3. Την ανάπτυξη ενός ReactJS[10] frontend, που αρέχει στον τελικό χρήστη μια διεπαφή για την εισαγωγή των λέξεων-κλειδιών αναζήτησης και την ανάκτηση των πληροφοριών (Εικόνα 5).

Μερικές από τις λειτουργικότητες που ο χρήστης θα μπορεί να επιτελέσει μέσω του API είναι:

- Με την εισαγωγή μιας λέξης-κλειδί για την “τοποθεσία”, εύρεση άρθρων που αναφέρονται σε αυτή με χρονολογική σειρά.
- Με την εισαγωγή μιας λέξης-κλειδί για το “πρόσωπο”, εύρεση άρθρων που αναφέρονται σε αυτή με χρονολογική σειρά.
- Με την εισαγωγή μιας λέξης-κλειδί για την “οργάνωση”, εύρεση άρθρων που αναφέρονται σε αυτή με χρονολογική σειρά.
- Εμφάνιση επιπλέον λεπτομερειών για ένα άρθρο.
- Κύλιση στην κορυφή της σελίδας όταν ο χρήστης βρίσκεται στο τέλος της.
- Ενημέρωση ότι δεν βρέθηκαν άρθρα όταν ο συνδυασμός label και λέξης-κλειδιού δεν είναι έγκυρος.
- Δυνατότητα εμφάνισης του περιεχομένου του κειμένου.

Όσον αφορά τη διεπαφή χρήστη (ReactJS frontend), στην Εικόνα 5 περιέχεται η αναζήτηση με λέξη-κλειδί την τοποθεσία “Βόρεια Ελλάδα”.

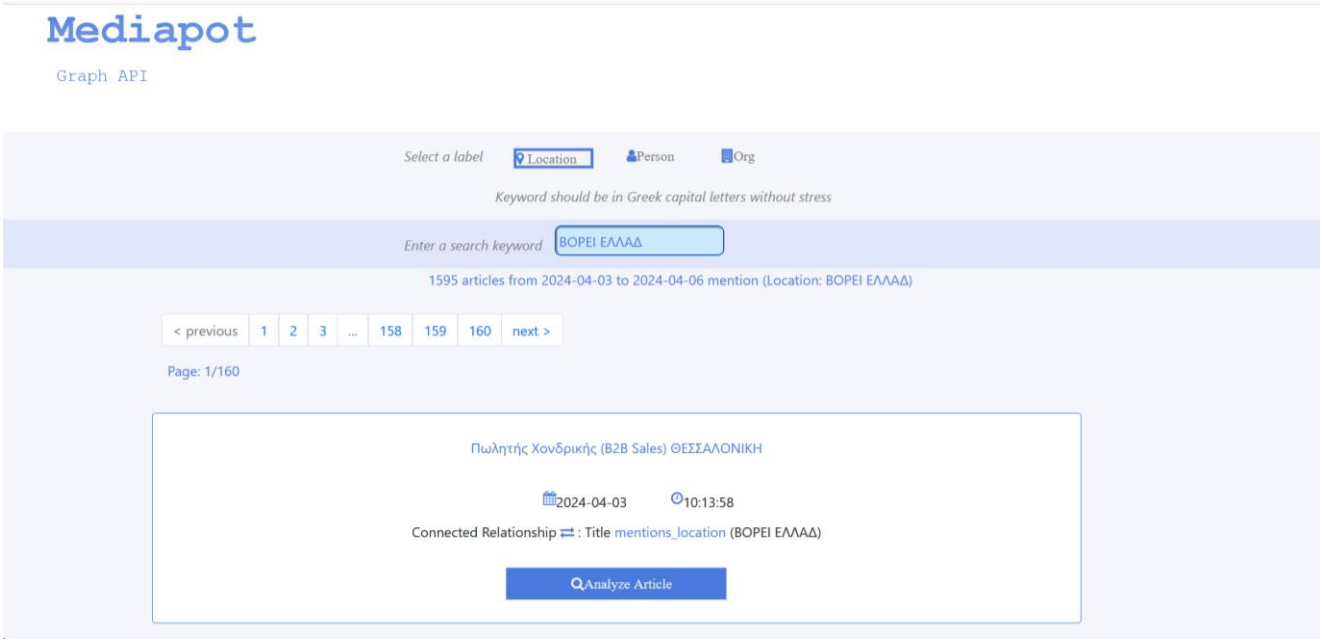


Εικόνα 4: Backend API

[9] https://fastapi.tiangolo.com/

[10] https://react.dev/

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης



Εικόνα 5: Η διεπαφή χρήστη (frontend)

6 Συμπεράσματα

Κατασκευάστηκε ένα χρονικός γράφος οντοτήτων-σχέσεων που αφορά σε ειδησεογραφικές πληροφορίες. Συγκεκριμένα τα άρθρα συνδέονται με 3 τύπους οντοτήτων όπως είναι οι γεωγραφικές τοποθεσίες, τα πρόσωπα και οργανισμοί. Ταυτόχρονα καταγράφεται και χρονική πληροφορία, δηλαδή η ημερομηνία δημοσίευσης του άρθρου. Επίσης εφαρμόστηκαν διάφορες τεχνικές καθαρισμού των δεδομένων του γράφου, προκειμένου να αποφευχθούν επαναλήψεις της ίδιας οντότητας με διαφορετική μορφή. Στη συνέχεια εξήχθησαν κάποιες γραφο-θεωρητικές πληροφορίες που μπορούν να δώσουν μία εκτίμηση της σπουδαιότητας της κάθε οντότητας.

Τέλος, κατασκευάστηκε ένα API και μία διεπαφή χρήστη προκειμένου να υπάρχει πρόσβαση στο γράφο, αλλά και να διευκολυνθεί μελλοντική ενσωμάτωσή του σε άλλα αρθρώματα

Η μελλοντικές εργασίες θα εστιάσουν σε ενσωμάτωση πληροφορίας από κοινωνικά δίκτυα (X), αλλά και σε θέματα ποιότητας δεδομένων. Επίσης, αναμένεται να εξαχθούν και άλλες πληροφορίες από τις ειδησεογραφικές πηγές με τη χρήση LLMs. Ειδικότερα, θα εστιάσουμε στην εξαγωγή μικρο-ιστοριών, που συνυπάρχουν σε ένα μεγάλο άρθρο. Αυτή η πληροφορία θα εμπλουτίσει τον γράφο (Rostam 2024). Επίσης αναμένονται και καινούργια δεδομένα από την ATC τα οποία θα ενταχθούν στον γράφο.

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

7 Αναφορές

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Ren, X., Tang, J., Yin, D., Chawla, N., & Huang, C. (2024). A survey of large language models for graphs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 6616-6626).

Rostam, Z. R. K., Szénási, S., & Kertész, G. (2024). Achieving Peak Performance for Large Language Models: A Systematic Review. *IEEE Access*.

Zafarani, R. (2014). *Social Media Mining: An Introduction*. Cambridge University Press

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Εθνικής Οικονομίας και Οικονομικών
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης