Δράση: ΕΡΕΥΝΩ – ΔΗΜΙΟΥΡΓΩ - ΚΑΙΝΟΤΟΜΩ

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

Π1.2: Αρχιτεκτονική συστήματος και προδιαγραφές

Ενότητα Εργασίας	ΕΕ1: Ανάλυση απαιτήσεων, σχεδιασμός πλατφόρμας και αξιολόγηση
Ημερομηνία	30/04/2024
Τύπος εγγράφου	Final v1.0
Υπεύθυνος Φορέας	ATC
Συμμετέχοντες Φορείς	ΕΚΕΦΕ «Δ», ΕΚΕΤΑ
Επιμελητές	Γεώργιος Ζήσης
Συνοπτική περιγραφή	Το παρόν έγγραφο περιγράφει την τεχνική δομή και τις βασικές λειτουργίες της πλατφόρμας MediaPot, εστιάζοντας στην γενική αρχιτεκτονική του συστήματος και των εργαλείων ανάλυσης πολυμεσικού περιεχομένου και ειδησεογραφικού περιεχομένου από το διαδίκτυο και τα κοινωνικά δίκτυα. Θα χρησιμεύσει ως οδηγός για την ανάπτυξη της πλατφόρμας και θα ενημερώνεται ανάλογα με τις ανάγκες κατά την ανάπτυξη και την πιλοτική χρήση.

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

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

Στην εποχή της ψηφιακής πληροφορίας, ο όγκος των ειδήσεων που διακινείται μέσω διαδικτύου και κοινωνικών δικτύων αυξάνεται εκθετικά. Η διαχείριση, ανάλυση και αξιοποίηση αυτού του περιεχομένου αποτελεί κρίσιμη ανάγκη για δημοσιογραφικούς οργανισμούς, εταιρείες και επαγγελματίες που επιδιώκουν να κατανοήσουν τάσεις και να δημιουργήσουν περιεχόμενο με απήχηση. Η πλατφόρμα MediaPot απαντά σε αυτή την ανάγκη, προσφέροντας ένα ισχυρό εργαλείο συλλογής, ανάλυσης και διαχείρισης δεδομένων από ποικίλες διαδικτυακές πηγές, με σκοπό την παραγωγή ιστοριών που βασίζονται σε τεκμηριωμένη πληροφορία και προηγμένα αναλυτικά στοιχεία.

Το παρόν έγγραφο καλύπτει λεπτομερώς την αρχιτεκτονική και τις βασικές λειτουργίες της πλατφόρμας. Ξεκινά με τις προδιαγραφές του συστήματος, περιγράφοντας τις δύο βασικές ροές εργασιών: την άντληση και ανάλυση δεδομένων και τη διαχείριση υλικού. Στη συνέχεια, περιγράφεται η αρχιτεκτονική και η τοπολογία του συστήματος, που περιλαμβάνει τη διεπαφή χρήστη, το backend, τις βάσεις δεδομένων, τα εργαλεία ανάλυσης και το υποσύστημα ειδοποιήσεων. Ακολουθεί η ενότητα που εστιάζει στα συστήματα ανάλυσης περιεχομένου, τα οποία καλύπτουν λειτουργίες όπως η κατάτμηση και επισημείωση βίντεο και εικόνων, η αντίστροφη αναζήτηση και η περίληψη πολυμέσων. Τέλος, γίνεται ανάλυση των διαδικασιών ανάκτησης και συσχέτισης δεδομένων από ειδησεογραφικές πηγές και κοινωνικά δίκτυα με τη βοήθεια του Γράφου Γνώσης.

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 Προδιαγραφές Συστήματος – Ροές εργασιών

Η συνολική λειτουργία του συστήματος μπορεί να διακριθεί σε δύο βασικές ροές εργασιών.

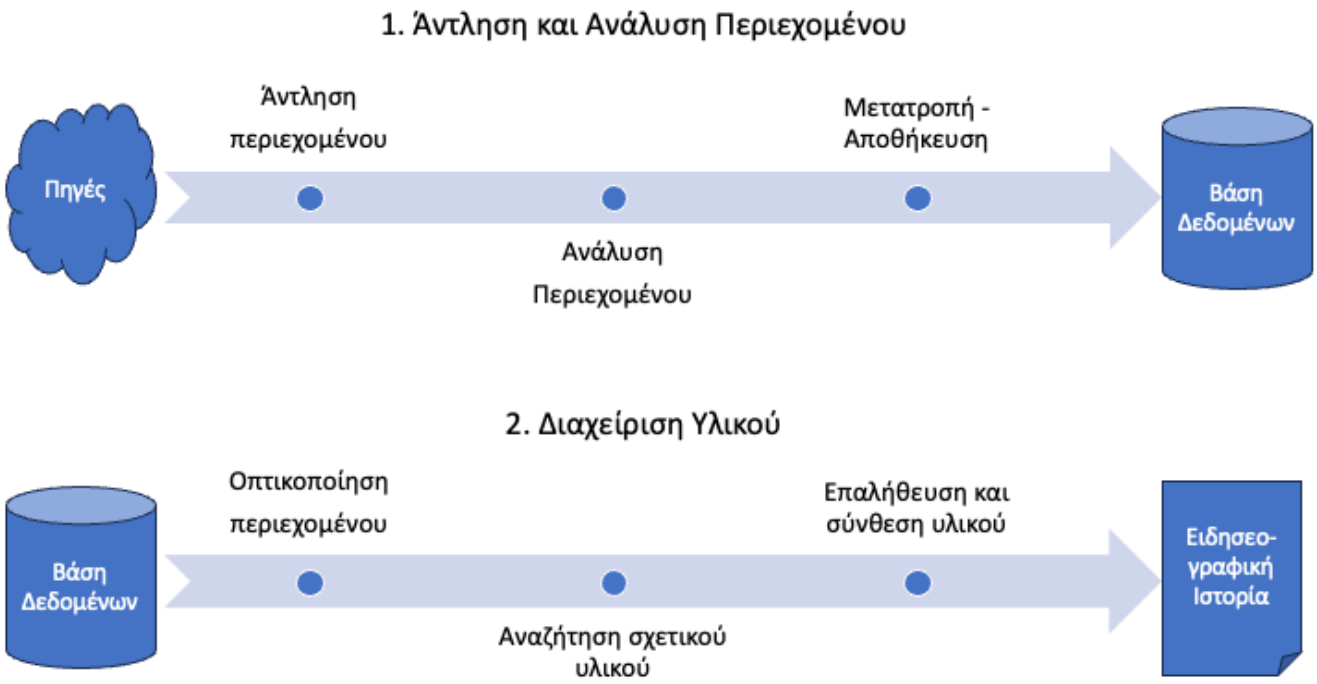
1. Άντληση και ανάλυση δεδομένων:

    i. Συλλογή περιεχομένου από διαδικτυακές πηγές και μέσα κοινωνικής δικτύωσης.
    ii. Ανάλυση περιεχομένου (κείμενο, εικόνα βίντεο) και δημιουργία συσχετίσεων.
    iii. Αποθήκευση περιεχομένου, μορφοποίηση και δημιουργία ευρετηρίου.

2. Διαχείριση υλικού:

    i. Οπτικοποίηση θεμάτων και περιεχομένου
    ii. Αναζήτηση σχετικού πολυμεσικού υλικού
    iii. Επαλήθευση περιεχομένου
    iv. Σύνθεση περιεχομένου και δημιουργία ιστοριών.

Οι δύο ροές εργασιών περιγράφονται στο παρακάτω διάγραμμα:



Εικόνα 1: Διάγραμμα ροών εργασιών

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

3 Αρχιτεκτονική και Τοπολογία Συστήματος

3.1 Αρχιτεκτονική Συστήματος

Στο παρακάτω διάγραμμα παρουσιάζεται η γενική αρχιτεκτονική του συστήματος καθώς και η τοπολογία εγκατάστασης.

Διεπαφή Χρήστη (user interface): Η κεντρική πλατφόρμα αποτελείται από τη διεπαφή χρήστη (user interface), η οποία βασίζεται σε τεχνολογία Angular και φιλοξενείται σε ένα Content Delivery Network μέσω του οποίου κάθε χρήστης μπορεί να έχει πρόσβαση σε αυτή του Web Browser και τη χρήση ενός μοναδικού URL.

Backend Πλατφόρμας: Η διασύνδεση της διεπαφής χρήστη με τη βάση δεδομένων, υλοποιείται μέσω του backend, το οποίο είναι υλοποιημένο με τεχνολογία Java Spring boot και παρέχει RESTful API ώστε να μπορεί να επικοινωνούν τα επιμέρους υποσυστήματα με τη χρήση μηνυμάτων HTTP.

Υποσύστημα Συλλογής: Η άντληση του υλικού από τις διάφορες πηγές και τα δίκτυα κοινωνικής δικτύωσης πραγματοποιείται από το Υποσύστημα Συλλογής, το οποίο αποτελεί ανεξάρτητο υποσύστημα υλοποιημένο σε τεχνολογία .net. Διαθέτει δική του διαχειριστική διεπαφή χρήστη μέσω της οποίας ο χρήστης ρυθμίζει τις πηγές από τις οποίες αντλεί υλικό, ορίζει τον τρόπο εκτέλεσης και τις δομές δεδομένων και παρακολουθεί την κατάσταση λειτουργίας του. Ο ρόλος του είναι να αντλεί υλικό από τις επιμέρους πηγές και να εμπλουτίζει τα μετα-δεδομένα των συλλεγόμενων άρθρων/αναρτήσεων με το να καλεί σειριακά όσα από τα εργαλεία διαθέτουν προγραμματιστική διεπαφή API και έχουν χαμηλό χρόνο απόκρισης. Τέλος, είναι υπεύθυνο για την αποθήκευση του περιεχομένου στις επιμέρους βάσεις δεδομένων του συστήματος. Συνεπώς, το Υποσύστημα Συλλογής, εκτελεί και ένα ρόλο συντονιστή (orchestrator) των εργαλείων ανάλυσης κατά τη συλλογή (fetching time).

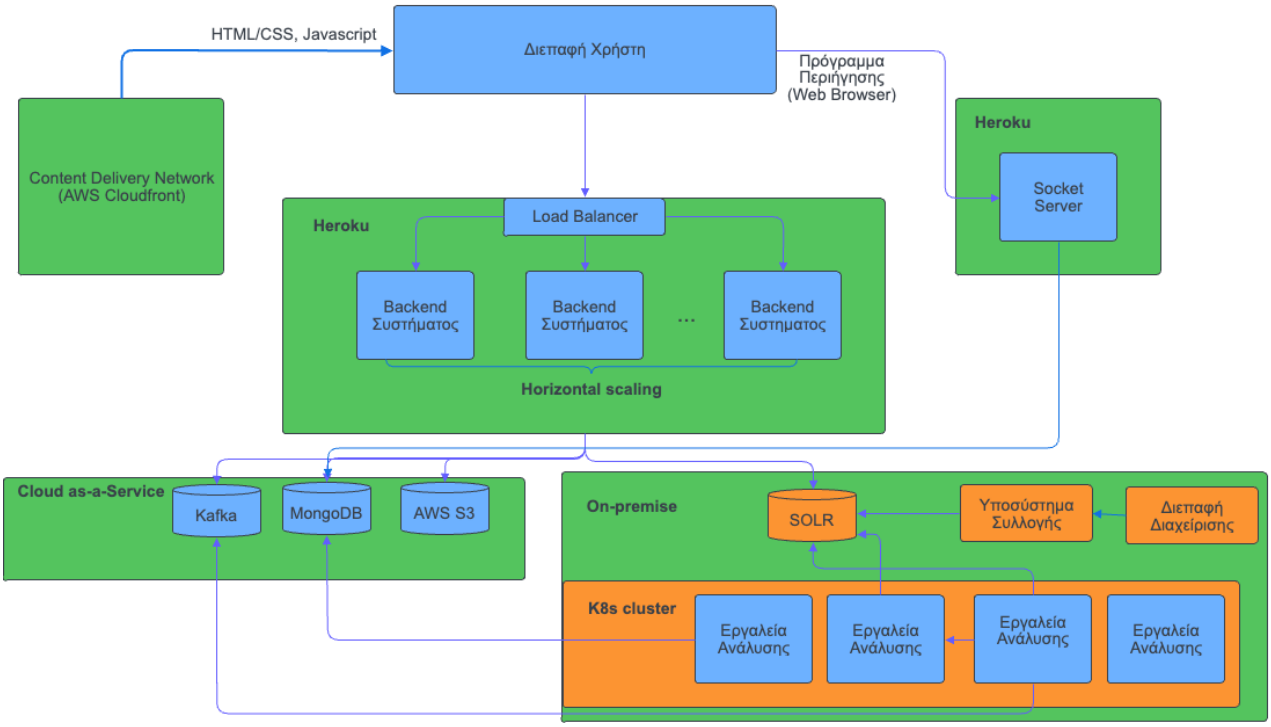
Εργαλεία Ανάλυσης: Τα επιμέρους εργαλεία ανάλυσης είναι υλοποιημένα ως microservices και καλούνται είτε με σύγχρονο είτε ασύγχρονο τρόπο μέσω μηνυμάτων HTTP. Αναλόγως με την ταχύτητα απόκρισης κάθε εργαλείου και τον τρόπο επεξεργασίας των δεδομένων υπάρχουν διάφοροι τρόποι να καλείται και να επικοινωνεί καθένα από αυτά με το κεντρικό σύστημα. Συνεπώς, η επικοινωνία μπορεί να γίνει είτε με τη σειριακή κλήση του από το backend ή το Υποσύστημα Συλλογής.

Βάσεις Δεδομένων: Σε αυτές συγκαταλέγονται η βάση αναζήτησης κειμένου (full-text search) όπως η Apache Solr, η βάση πολυμεσικών αντικειμένων (S3 object storage), η βάση μετα-δεδομένων (metadata) καθώς και ο Kafka, που αποτελεί ένα σύστημα διαχείρισης ροής δεδομένων με τη μορφή μηνυμάτων. Η τελευταία μπορεί να χρησιμοποιηθεί για την επικοινωνία μεταξύ αυτόνομων εργαλείων που λειτουργούν με ασύγχρονο τρόπο.

Σύστημα ειδοποιήσεων πραγματικού χρόνου: Το σύστημα ειδοποιήσεων υλοποιείται με τεχνολογία web sockets κρατώντας ανοιχτή μία σύνδεση με το frontend έτσι ώστε να μπορεί να στείλει ανά πάσα στιγμή ειδοποιήσεις πραγματικού χρόνου στον χρήστη. Το ίδιο υποσύστημα

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

χρησιμοποιείται και για την επικοινωνία μεταξύ χρηστών όταν απαιτείται η ταυτόχρονη απεικόνιση των ενεργειών τους. Έτσι, η πλατφόρμα μπορεί να αποκτήσει και δυνατότητες παράλληλης συνεργασίας δύο ή περισσοτέρων χρηστών καθώς ο ένας μπορεί να παρακολουθεί τις ενέργειες του άλλου σε πραγματικό χρόνο και έτσι να δουλέψουν στο ίδιο workspace.



Εικόνα 2: Γενική αρχιτεκτονική της πλατφόρμας MediaPot

3.2 Τοπολογία Συστήματος

Η τοπολογία εγκατάστασης του συστήματος φαίνεται στο προηγούμενο διάγραμμα. Τα επιμέρους μέρη του συστήματος ακολουθούν μια modular αρχιτεκτονική που επιτρέπει την αποσύνδεσή τους και την εγκατάσταση του καθενός σε ανεξάρτητο περιβάλλον. Η συνολική τοπολογία βασίζεται σε ένα συνδυασμό on-premise και cloud λύσεων όπως οι πλατφόρμες Heroku και AWS καθώς και η χρήση βάσεων as-a-service. Οι συγκεκριμένοι πάροχοι cloud infrastructure, αποτελούν αξιόπιστες λύσεις, καθώς δίνουν τη δυνατότητα προσθήκης εργαλείων σχετικά με την ασφάλεια την ταχύτητα και την επεκτασιμότητα του συστήματος.

Τέλος, ορισμένα εργαλεία του συστήματος μπορούν να υλοποιηθούν ως Docker containers και να τρέχουν σε ένα on-premise Kubernetes, ένα σύστημα συντονισμού και διαχείρισης containers. Αυτό μας δίνει την ευελιξία να αναπτύξουμε ανεξάρτητα τις επιμέρους εφαρμογές και τα εργαλεία του συστήματος, χωρίς να επηρεάζει η μία την άλλη.

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

4 Ολοκλήρωση και Εγκατάσταση Συστήματος

4.1 Ολοκλήρωση Συστήματος (Integration)

4.1.1 Ενσωμάτωση Υποσυστήματος Άντλησης Περιεχομένου

Το συγκεκριμένο υποσύστημα μπορεί να εγκατασταθεί ανεξάρτητα. Επίσης, διαθέτει δικό του διαχειριστικό interface. Η διασύνδεση με το υπόλοιπο σύστημα γίνεται μέσω της χρήσης των κοινών βάσεων και ειδικά της Apache Solr, στην οποία αποθηκεύονται τα δεδομένα που συλλέγονται. Επίσης, το υποσύστημα αυτό έχει ρόλο συντονιστή (orchestrator) καθώς ευθύνεται για τη σειριακή εκτέλεση των εργαλείων ανάλυσης και την ενσωμάτωση των αποτελεσμάτων της ανάλυσης σε κοινό αρχείο μετα-δεδομένων.

4.1.2 Ενσωμάτωση Εργαλείων Ανάλυσης

Όπως προαναφέρθηκε, η ευελιξία που παρέχει η συγκεκριμένη αρχιτεκτονική μας δίνει τη δυνατότητα να αναπτύξουμε τα επιμέρους εργαλεία ανάλυσης με όποια τεχνολογία επιθυμούμε και να τα εγκαταστήσουμε ανεξάρτητα το ένα από το άλλο, σε διαφορετικά περιβάλλοντα. Η σύγκλιση τους, θα γίνει μέσω διεπαφών, με τις οποίες θα επικοινωνούν με μηνύματα HTTP. Υπάρχουν δύο τρόποι επικοινωνίας:

i. Ο **σύγχρονος** τρόπος εκτέλεσης επιβάλλει να δοθεί η απάντηση σε μια κλήση άμεσα, με ελάχιστη αναμονή στον αποστολέα του αρχικού μηνύματος. Η επικοινωνία αυτή επιτυγχάνεται σχεδόν αποκλειστικά με μήνυμα HTTP, το οποίο δέχεται άμεσα την απάντηση στη ίδια σύνδεση (connection).
ii. Ο **ασύγχρονος** τρόπος επικοινωνίας περιλαμβάνει μια κλήση/μήνυμα ώστε να ξεκινήσει η εκτέλεση της ανάλυσης. Ο αποδέκτης του μηνύματος δεν υποχρεούται να απαντήσει άμεσα. Συνήθως αυτό αφορά εργαλεία ανάλυσης που χρειάζονται αρκετό χρόνο εκτέλεσης και δεν είναι εφικτή η διατήρηση της ίδιας σύνδεσης (connection timeout). Η ασύγχρονη επικοινωνία σε δεύτερο χρόνο ώστε να ληφθούν τα αποτελέσματα της ανάλυσης γίνεται είτε με τεχνική polling, είτε μέσω ροών μηνυμάτων (Kafka) και πρωτοκόλλων pub/sub. Σε αυτή τη λειτουργία, τα επιμέρους στοιχεία του συστήματος εγγράφονται σε «ουρές» μηνυμάτων και ειδοποιούνται για κάποιο αποτέλεσμα μέσω αυτών.

Η επιλογή του τρόπου επικοινωνίας μεταξύ των στοιχείων δε χρειάζεται να προαποφασιστεί και μπορεί να αλλάξει κατά τη φάση της υλοποίησης ή και αργότερα, αναλόγως με τις απαιτήσεις του χρόνου απόκρισης και τον όγκο των δεδομένων που θα κληθούν να αναλύσουν. Όσον αφορά την ανταλλαγή των δεδομένων (multimedia, metadata κλπ), αυτή μπορεί να γίνει πάντοτε μέσω της διεπαφής του backend, το οποίο θα είναι υπεύθυνο για την επικοινωνία με τις κεντρικές κοινές βάσεις δεδομένων (Apache Solr, AWS S3, MongoDB).

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

4.1.3 Ενσωμάτωση εργαλείων οπτικοποίησης και διαχείρισης περιεχομένου.

Τα εργαλεία αυτά θα ενσωματωθούν ως μέρος του κεντρικού συστήματος, μέσα στο λογισμικό της διεπαφής χρήστη σε γλώσσα Angular. Επίσης, μέρος της λειτουργία τους θα εκτελείται στο backend, στο οποίο αντίστοιχα θα υλοποιηθεί σε γλώσσα Java spring boot.

4.2 Εγκατάσταση Συστήματος (Deployment)

Η αποθήκευση του κώδικα λογισμικού ανάπτυξης του συστήματος απαιτεί ένα versioning tool. Η ATC παρέχει τη λύση του Gilab on-premise, το οποίο μπορεί να διατεθεί είτε ως code repository είτε ως container registry. Οι ομάδες υλοποίησης μπορούν να αποφασίσουν αν θα διαθέσουν τον κώδικα απευθείας στο repository ή αν θα παραδώσουν τα εργαλεία τους ως containers μέσα στο container registry.

Η εγκατάσταση του συστήματος γίνεται τμηματικά. Αναλόγως με το περιβάλλον εγκατάστασης που απαιτείται, κάθε υποσύστημα μπορεί να εγκατασταθεί με διαφορετικό τρόπο. Κάθε επιμέρους περιβάλλον θα πρέπει να τροποποιηθεί ώστε να είναι συμβατό με τις απαιτήσεις του έργου. Αυτή η εργασία απαιτεί εμπειρία σε DevOps και πρέπει να γίνει χειροκίνητα από την ομάδα integration. Στη συνέχεια θα υλοποιηθούν συγκεκριμένες ροές εργασιών (CI/CD pipelines) που θα εκτελούν αυτόματα όλες τις ενέργειες (compilation, deployment) που απαιτούνται για την εγκατάσταση ενός εργαλείου/υποσυστήματος.

Για τη υλοποίηση και τη ρύθμιση των αυτοματοποιημένων διεργασιών χρησιμοποιούμε την πλατφόρμα Jenkins, με την οποία εκτελείται η ροή εγκατάστασης για το κεντρικό σύστημα (backend, frontend). Για το υπόλοιπο σύστημα έχουμε την εμπειρία και την ευελιξία να χρησιμοποιήσουμε τόσο το Jenkins, όσο και το Gitlab CI/CD, εφόσον τα επιμέρους εργαλεία είναι αποθηκευμένα στο περιβάλλον του Gitlab της ATC που διατίθεται για τις ανάγκες του έργου.

Ειδικότερα για τα containerized εργαλεία συστήματος, η χρήση της πλατφόρμας Kubernetes που έχει στηθεί στην ATC μας παρέχει ένα σύνολο δυνατοτήτων όσον αφορά την εκτέλεση και την εγκατάσταση. Ωστόσο και εδώ χρειάζονται γνώσεις και υλοποίηση κατάλληλων ροών για τη σωστή εγκατάσταση των εργαλείων ανάλυσης (helm, docker compose).

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

5 Συστήματα ανάλυσης πολυμεσικού περιεχομένου

5.1 Σύστημα κατάτμησης βίντεο

Το σύστημα κατάτμησης βίντεο τμηματοποιεί ένα βίντεο στα πλάνα (shots) που το συνθέτουν. Κάθε πλάνο αποτελείται από μια ακολουθία καρέ (frames) που έχουν καταγραφεί χωρίς διακοπή από μια βίντεο κάμερα. Για τη δημιουργία του βίντεο, τα πλάνα συρράπτονται είτε χωρίς είτε με τη χρήση κάποιου εφέ βαθμιαίας μετάβασης, όπως για παράδειγμα την περίπτωση που το οπτικό περιεχόμενο ενός πλάνου σταδιακά αντικαθίσταται από το οπτικό περιεχόμενο του αμέσως επόμενου πλάνου. Με βάση το παραπάνω, ο στόχος του συστήματος κατάτμησης βίντεο είναι να εντοπίσει τα σημεία που αντιστοιχούν στις μεταβάσεις μεταξύ διαδοχικών πλάνων του βίντεο, είτε αυτές γίνονται άμεσα είτε βαθμιαία, και έτσι να οριοθετήσει χρονικά την έναρξη και το τέλος κάθε πλάνου. Στην περίπτωση που το βίντεο αποτελείται από ένα και μόνο πλάνο, όπως για παράδειγμα τα βίντεο που καταγράφονται με τη χρήση ενός κινητού τηλεφώνου, καθώς και στην περίπτωση που μια πιο λεπτομερής κατάτμηση είναι απαραίτητη, το σύστημα τμηματοποιεί το βίντεο σε ύπο-πλάνα (sub-shots). Για αυτού του είδους την κατάτμηση, το σύστημα εντοπίζει ακολουθίες καρέ που έχουν καταγραφεί κατόπιν διαφόρων συγκεκριμένων ενεργειών της κάμερας, όπως οριζόντια κίνηση προς τα δεξιά/αριστερά, κατακόρυφη κίνηση προς τα πάνω/κάτω, αλλαγή του επιπέδου εστίασης, και άλλων. Το σύστημα εκτελείται σε ένα Ubuntu processing server, ο οποίος επιτρέπει την χρήση γνωστών μορφών αρχείων μοντέλων (PyTorch, TensorFlow) και την παράλληλη επεξεργασία με χρήση GPU για μεγιστοποίηση της απόδοσης. Η χρήση του συστήματος γίνεται μέσω REST API.

5.2 Σύστημα επισημείωσης εικόνων και βίντεο

Ο στόχος του συστήματος επισημείωσης εικόνων και βίντεο είναι να παρέχει με αυτόματο τρόπο μεταδεδομένα για αρχεία πολυμέσων. Τα μεταδεδομένα αυτά αφορούν διαφορετικές πλευρές του σημασιολογικού περιεχομένου των πολυμέσων και παράγονται ύστερα από επεξεργασία μέσω μιας σειράς αλγορίθμων βαθιάς μάθησης που βρίσκονται στην αιχμή των τεχνολογικών εξελίξεων. Τα πολυμέσα που υποστηρίζονται στην τρέχουσα έκδοση του είναι εικόνες, βίντεο αλλά και αντικείμενα 3D.

Πιο συγκεκριμένα, υποστηρίζεται η αυτόματη παραγωγή κειμένου φυσικής γλώσσας για την περιγραφή του περιεχομένου αλλά και ειδικά η αναγνώριση 16000 διεθνών διασημοτήτων (π.χ. αθλητές, καλλιτέχνες, πολιτικοί), 400 ειδών δραστηριότητας (π.χ. γυμναστική, οργανοπαιξία, περπάτημα, μπάσκετ), 6500 αντικειμένων (π.χ. αυτοκίνητο, τραπέζι, άνθρωπος, τρένο) και ειδικών ετικετών (π.χ. Χριστουγεννιάτικο δέντρο, χρώματα, είδος σκηνής). Επίσης, υποστηρίζεται η αναγνώριση ακατάλληλου και σκληρού περιεχομένου, όπως και η ειδική επισημείωση εικόνων τύπου meme. Τέλος, παρέχεται και μία διανυσματική αναπαράσταση του πολυμέσου η οποία μπορεί να χρησιμοποιηθεί για την ανάκτηση σημασιολογικά όμοιων στοιχείων περιεχομένου.

Για την εκτέλεση των μοντέλων χρησιμοποιείται ένας Nvidia Triton Inference Server[1], ο οποίος επιτρέπει την χρήση γνωστών μορφών αρχείων μοντέλων (TensorFlow, PyTorch, ONNX), την παράλληλη επεξεργασία μέσα από κάρτες γραφικών για μεγιστοποίηση της απόδοσης καθώς επίσης παρέχει και μία σειρά από προηγμένα χαρακτηριστικά όπως σύνολα μοντέλων (model ensemble) και

[1] https://developer.nvidia.com/nvidia-triton-inference-server

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

ροή εκτέλεσης (streaming inferencing). Πρόκειται για έναν gRPC server που εκτελείται σε Python και ο οποίος μπορεί να δέχεται αιτήματα μέσω ενός gRPC AsyncIO API[2]. Τα μέρη του συστήματος εκτελούνται σε περιβάλλον docker container.

5.3 Σύστημα αντίστροφης αναζήτησης εικόνας και βίντεο

Το σύστημα αντίστροφης αναζήτησης εικόνας και βίντεο παρέχει τη δυνατότητα αποδοτικής ανάκτησης πολυμέσων (εικόνων και βίντεο) βάση του οπτικού και ακουστικού περιεχομένου τους σε συλλογές μεγάλης κλίμακας. Συγκεκριμένα, παρέχει τη δυνατότητα αναζήτησης σχεδόν όμοιων (near-duplicate) εικόνων και βίντεο, χρησιμοποιώντας άλλες εικόνες ή βίντεο ως ερωτήματα (queries). Η ομοιότητα μπορεί να εκτείνεται από τα αντίγραφα ενός πολυμέσου, έως οπτικοακουστικό περιεχόμενο που αναφέρεται στο ίδιο γεγονός, παρέχοντας τη δυνατότητα ρύθμισής της βάση των αναγκών της εκάστοτε εφαρμογής. Ταυτόχρονα, υποστηρίζει την αυτόματη κατάτμηση των βίντεο, προσφέροντας τη δυνατότητα αναζήτησης σκηνής προς σκηνή (shot-to-shot), ενώ δίνει τη δυνατότητα επιλογής πραγματοποίησης της αναζήτησης είτε βάση της εικόνας, είτε του ήχου. Το πολυμεσικό περιεχόμενο οργανώνεται σε συλλογές, το μέγεθος των οποίων μπορεί να εκτείνεται ακόμη και σε εκατομμύρια στοιχεία, δίχως να επηρεάζεται δραστικά η απόδοση της αναζήτησης. Ακόμη, παρέχεται ευρεία υποστήριξη διαδικτυακών πηγών για τη λήψη περιεχομένου. Οι δυνατότητες αναζήτησης δύναται να ενσωματωθούν σε τρίτες εφαρμογές μέσω του παρεχόμενου REST API.

Η τεχνολογία αναζήτησης βασίζεται σε μία αρθρωτή (modular) αρχιτεκτονική βαθιάς μηχανικής μάθησης δύο σταδίων, σχεδιασμένη για την αποδοτική ανάκτηση πολυμέσων σε συλλογές μεγάλης κλίμακας. Επιμέρους μοντέλα βαθιάς μηχανικής μάθησης εστιάζουν στις ιδιαιτερότητες του κάθε τύπου πολυμεσικού περιεχομένου (εικόνα, ήχος), μετατρέποντάς το σε ιεραρχικές διανυσματικές απεικονίσεις που χρησιμοποιούνται στα επιμέρους στάδια της αναζήτησης, καταλαμβάνοντας μικρό αποθηκευτικό χώρο. Η αρχιτεκτονική του συστήματος στηρίζεται στο κατανεμημένο (distributed) μοντέλο των microservices, παρέχοντας τη δυνατότητα για κλιμάκωση (scaling) δίχως περαιτέρω αλλαγές στη δομή του συστήματος. Υλοποιείται με χρήση Docker[3] containers, RabbitMQ[4] και Protocol Buffers[5] για την απομόνωση και επικοινωνία των microservices, MongoDB[6] και FAISS[7] για την αποθήκευση των δεδομένων, PyTorch[8] για την εκτέλεση των μοντέλων βαθιάς μηχανικής μάθησης και FastAPI[9] για την παροχή του REST API, ενώ το documentation του συστήματος ακολουθεί το πρότυπο OpenAPI[10].

[2] https://grpc.github.io/grpc/python/grpc_asyncio.html
[3] https://www.docker.com/
[4] https://rabbitmq.com/
[5] https://protobuf.dev/
[6] https://www.mongodb.com/
[7] https://github.com/facebookresearch/faiss
[8] https://pytorch.org/
[9] https://fastapi.tiangolo.com/
[10] https://swagger.io/specification/

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

5.4 Σύστημα περίληψης βίντεο

Το σύστημα περίληψης βίντεο δημιουργεί μια σύντομη σύνοψη του περιεχομένου του βίντεο, επιλέγοντας τα πιο σημαντικά και ενδιαφέροντα τμήματά του, καθώς και έναν αριθμό από αντιπροσωπευτικά καρέ. Ο στόχος του συστήματος είναι, αξιολογήσει τη σημαντικότητα κάθε καρέ του βίντεο με βάση τη γνώση που έχει αποκτήσει μετά από διαδικασία εκπαίδευσης η οποία περιλαμβάνει διάφορα κριτήρια. Ενδεικτικά, ένα πρώτο κριτήριο αφορά την οπτική και σημασιολογική συσχέτιση του περιεχομένου ενός καρέ του βίντεο με το υπόλοιπο περιεχόμενο, ώστε να εντοπιστούν οι βασικές οπτικές και σημασιολογικές οντότητες του βίντεο. Άλλα κριτήρια σχετίζονται με την ποικιλία και τη μοναδικότητα/διακριτότητα του οπτικού περιεχομένου κάθε καρέ του βίντεο, υπό το σκεπτικό ότι η επιλογή τέτοιων καρέ αυξάνει την περιγραφικότητα της περίληψης, και την καλαισθησία του οπτικού περιεχομένου κάθε καρέ του βίντεο. Μετά το πέρας της διαδικασίας αξιολόγησης σε επίπεδο καρέ, το σύστημα περίληψης βίντεο εξάγει έναν, προκαθορισμένο από το χρήστη, αριθμό από αντιπροσωπευτικά καρέ επιλέγοντας τα πιο υψηλά βαθμολογημένα από αυτά. Επίσης, υπολογίζει τη σημαντικότητα των διαφορετικών τμημάτων του βίντεο, λαμβάνοντας υπόψη το αποτέλεσμα του συστήματος κατάτμησης βίντεο. Κατόπιν, η περίληψη διαμορφώνεται επιλέγοντας τα τμήματα του βίντεο που μεγιστοποιούν τη συνολική σημαντικότητά της, υπό την προϋπόθεση οτι η διάρκεια της περίληψης δεν υπερβαίνει την επιλεγμένη από το χρήστη διάρκειά της (π.χ. το 15% της διάρκειας του αρχικού βίντεο). Το σύστημα εκτελείται σε ένα Ubuntu processing server, ο οποίος επιτρέπει την χρήση γνωστών μορφών αρχείων μοντέλων (PyTorch, TensorFlow) και την παράλληλη επεξεργασία με χρήση GPU για μεγιστοποίηση της απόδοσης. Η χρήση του συστήματος γίνεται μέσω REST API.

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

6 Σύστημα ανάλυσης ειδησεογραφικών κειμένων και κοινωνικών δικτύων

Το ΕΚΕΦΕ «Δημόκριτος» θα εστιάσει στην ανάλυση κειμένων από ειδησεογραφικές πηγές προκειμένου να ορυχθεί χρήσιμη πληροφορία από αυτά. Συγκεκριμένα θα εξαχθεί ένα Γράφος Γνώσης (ΓΓ) (δηλαδή ένα δίκτυο) οντοτήτων και σχέσεων. Τα στοιχεία του δικτύου (δηλαδή οι οντότητες και οι σχέσεις) θα επισημειώνονται με διάφορα μεταδεδομένα, όπως χρονική πληροφορία, προέλευση, κ.α. Επίσης θα συγκεντρώνεται πληροφορία από κοινωνικά μέσα όπως το Twitter προκειμένου να εμπλουτιστεί περαιτέρω ό ΓΓ.

Στο παρακάτω σχήμα απεικονίζεται το σύστημα διασωλήνωσης (pipeline) της ανάλυσης κειμένων. Τα αρχικά δημοσιογραφικά άρθρα θα είναι αποθηκευμένα σε μία βάση τύπου Elastic Search.[11] Στη συνέχεια ανακτώνται κάποια άρθρα (1: Ανάκτηση ειδησεογραφικών Δεδομένων) και αποθηκεύονται σε μία βάση (Επιλεγμένα Δεδομένα). Μετά τα άρθρα αναλύονται (3: Εξόρυξη Δομής Κειμένου) προκειμένου να δημιουργηθεί ο Γράφος Γνώσης και να αποθηκευτεί σε μία βάση γράφων (Neo4j[12]).

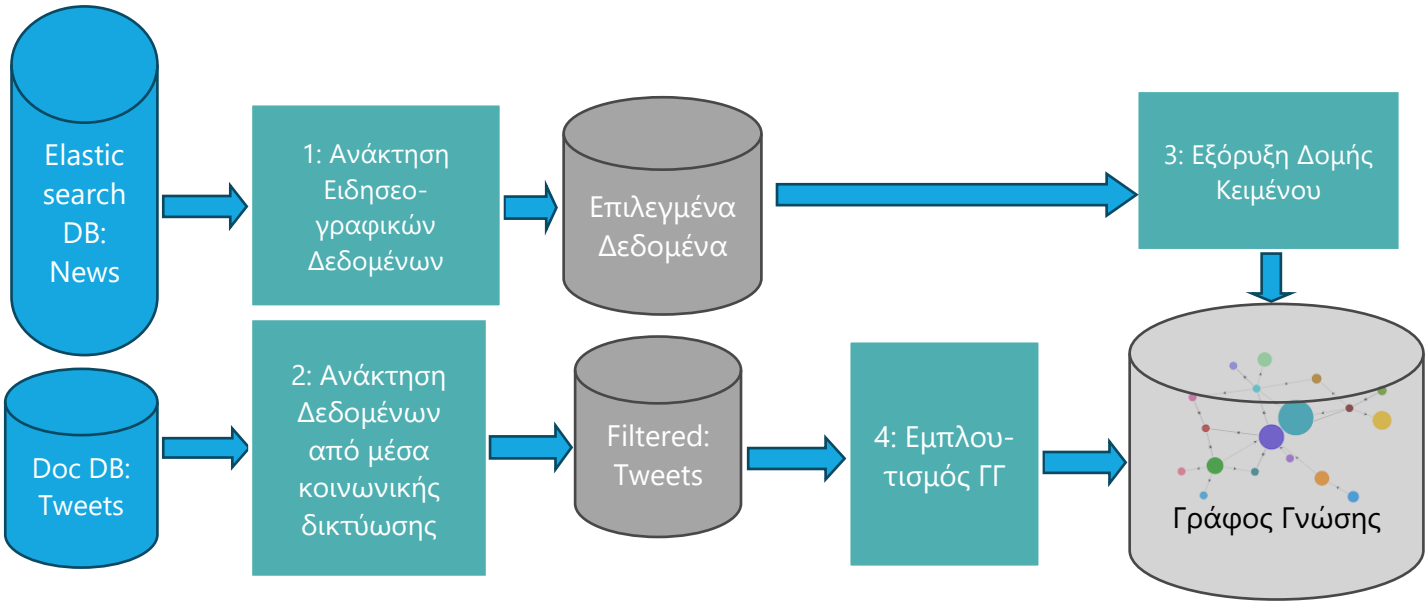
Παράλληλα ανακτώνται δημοσιεύματα σε κοινωνικά δίκτυα (π.χ. από το Twitter), τα οποία αποθηκεύονται σε μία βάση (Filtered Tweets). Στη συνέχεια τα Tweets θα χρησιμοποιηθούν για τον εμπλουτισμό του Γράφου Γνώσης (ΓΓ). Τέλος ο ΓΓ θα αποθηκεύεται σε μία βάση τύπου Neo4j.

Ακολούθως αναφερόμαστε στις εισόδους και εξόδους των προαναφερθέντων αρθρωμάτων, παρέχοντας ταυτόχρονα μία σύντομη περιγραφή. Οι αλγόριθμοι των αρθρωμάτων θα παρουσιαστούν στα παραδοτέα: Π3.1: Μεθοδολογία εξόρυξης οντοτήτων, Π3.2: Μεθοδολογία όρυξης πληροφορίας από κοινωνικά δίκτυα, Π3.3: Εξαγωγή σχέσεων οντοτήτων, και Π3.4: Συγχώνευση πληροφορίας από ανάλυση κειμένων και κοινωνικών δικτύων.

[11] https://www.elastic.co/

[12] 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 3: Σύστημα διασωλήνωσης (pipeline) της ανάλυσης κειμένων

Τίτλος Αρθρώματος: Ανάκτηση Ειδησεογραφικών Δεδομένων
Περιγραφή: Ανάκτηση άρθρων από δημοσιογραφικές ιστοσελίδες ελληνικού ενδιαφέροντος. Τα άρθρα μπορεί να είναι στα Ελληνικά ή τα Αγγλικά. Το κείμενα θα συλλέγονται για περαιτέρω ανάλυση και συγκριμένα για την εξαγωγή δομής από αυτά.
Είσοδος:
Μία βάση δεδομένων (ΒΔ) τύπου *Elastic search* «φιλοξενεί» κείμενα σε μορφή JSON. Ένα ερώτημα (query) θα εκτελείτε περιοδικά (π.χ. κάθε δύο μήνες), και θα ανακτά κάποια κείμενα βάσει συγκεκριμένων κριτηρίων. Για παράδειγμα τα κριτήρια μπορεί να είναι θεματικά, ή να αναφέρονται σε κάποια χρονική περίοδο.
Έξοδος:
Ένα επιλεγμένο σώμα κειμένων σε μορφή JSON. Τα αποτελέσματα αποθηκεύονται σε μία βάση κειμένων (π.χ. MongoDB)

Τίτλος Αρθρώματος: Ανάκτηση Δεδομένων από μέσα κοινωνικής δικτύωσης
Περιγραφή: Ανάκτηση δεδομένων που προέρχονται από κοινωνικά μέσα (Twitter) προκειμένου να χρησιμοποιηθούν περαιτέρω και συγκεκριμένα στον εμπλουτισμό του Γράφου Γνώσης.
Είσοδος:
Μία βάση κειμένων (όπως είναι η MongoDB) περιέχει Tweets σε μορφή JSON. Ένα ερώτημα (query) θα εκτελείται προκειμένου να ανακτηθούν Tweets με συγκεκριμένα κριτήρια. Για παράδειγμα τα κριτήρια μπορεί να αναφέρονται σε κάποιο θέμα, ή σε κάποια χρονική περίοδο.
Έξοδος:
Μία ομάδα από Tweets σε μορφή σε μορφή JSON, Τα αποτελέσματα αποθηκεύονται σε μία άλλη βάση κειμένων (π.χ. MongoDB)

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

Τίτλος Αρθρώματος: Εξόρυξη Δομής Κειμένου
Περιγραφή: Δημιουργία ενός Γράφου Γνώσης (ΓΓ) από ένα σώμα κειμένων. Ο ΓΓ είναι συνίσταται σε ένα δίκτυο *οντοτήτων* και *σχέσεων* και είναι ένας πολυσχεσιακός γράφος. Για παράδειγμα οι οντότητες μπορεί να είναι ονόματα ανθρώπων & οργανισμών, γεωγραφικές τοποθεσίες, κ.α. Οι σχέσεις αποτελούν τις συνδέσεις μεταξύ των οντοτήτων, για παράδειγμα: «*οντότητα 1:* ο πρωθυπουργός της Βρετανίας, *σχέση:* επισκέφθηκε, *οντότητα* 2: τη Γερμανία». (Δες και [7] για τη χρήση γράφων στην αναπαράσταση ειδησεογραφικών δεδομένων). Αυτό θα επιτευχθεί με μεθόδους επεξεργασίας φυσικής γλώσσας, και αναμένεται να χρησιμοποιηθούν μεγάλα γλωσσικά μοντέλα (ΜΓΜ) (Large Language Models-LLMs). Επίσης μπορούν να χρησιμοποιηθούν και παλαιότερες τεχνικές που βασίζονται σε απλούστερα γλωσσικά μοντέλα, αλλά και τεχνικές που βασίζονται σε τεχνικές όπως οι κανονικές εκφράσεις [1], [2], [3], [4], [5]. Το παρόν άρθρωμα θα χρησιμοποιείται μέσω ενός REST-API.
Είσοδος:
Ένα σώμα (δημοσιογραφικών) κειμένων στα Ελληνικά ή Αγγλικά. Κάθε κείμενο αναπαρίσταται με τη μορφή JSON. Επίσης κάθε κείμενο θα συνοδεύεται και από μετα-δεδομένα. Όλα το κείμενα είναι αποθηκευμένα σε μία βάση του είδους της MongoDB. Τα κείμενα θα ανακτώνται από τη MongoDB και στη συνέχεια θα ορύσσεται δομή με τη μορφή πολυσχεσιακού γράφου.
Έξοδος:
Ένας γράφος γνώσης (knowledge graph) με τη μορφή τριπλετών. Η αναπαράσταση θα είναι με τη μορφή JSON. π.χ. {entity-1: entityName, relation: relationName, entity-2: entityName}. Η έξοδος αποθηκεύεται σε μία βάση γράφων (π.χ. Neo4j).

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

6.1 Συσχέτιση κοινωνικών μέσων με τον Γράφο Γνώσης

Τίτλος Αρθρώματος: Εμπλουτισμός του Γράφου Γνώσης
Περιγραφή: Ο Γράφος Γνώσης που έχει δημιουργηθεί από το άρθρωμα **εξόρυξη δομής κειμένου** θα εμπλουτίζεται με πληροφορίες από κοινωνικά δίκτυα (τυπικά Tweets). Το παρόν άρθρωμα θα χρησιμοποιείται μέσω ενός REST-API.
Είσοδος:
Ένα σύνολο από επιλεγμένα Tweets στα Ελληνικά ή Αγγλικά. Κάθε Tweet αναπαρίσταται με τη μορφή JSON. Τα Tweets είναι αποθηκευμένα σε μία βάση του είδους της MongoDB. Αυτό το άρθρωμα θα ανακτά Tweets και θα τα συσχετίζει με το Γράφο Γνώσης. Θα συσχετίζει οντότητες ή σχέσεις με Tweets βάσει θεματικής συνάφειας. Για παράδειγμα, ο τίτλος ή τα hashtags του tweet μπορούν να χρησιμοποιηθούν για την πραγματοποίηση των συσχετίσεις [7]. Το παρόν άρθρωμα θα χρησιμοποιείται μέσω ενός REST-API.
Έξοδος:

Ένας εμπλουτισμένος Γράφος Γνώσης (knowledge graph) με τη μορφή τριπλετών. Η αναπαράσταση ακολουθεί τη μορφή JSON. Για παράδειγμα στα παρακάτω δεδομένα το πεδίο properties αναπαριστά πληροφορίες από σχετιζόμενα Tweet. Τα prop1, prop2 κτλ. Αποτελούν πεδία του Tweet. Το entity-1, entity-2 και το relation προέρχονται από την ανάλυση ειδησεογραφικής πηγής.
 	
{
{entity-1: entityVal,
{properties: {prop1: val1, prop2: val2}}
},
{relation1: relationName,
{properties: { prop1: val4, prop2: val5}
},
}
{entity-2: entityVal,
{properties: {prop1: val1, prop2: val2}
}
}
}
 	
Η έξοδος αποθηκεύεται σε μία βάση γράφων (π.χ. Neo4j).

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 Συμπεράσματα

Το έγγραφο "Αρχιτεκτονική Συστήματος και Προδιαγραφές" καταγράφει την τεχνική δομή, τις βασικές ροές εργασιών και τα κύρια υποσυστήματα της πλατφόρμας MediaPot. Οι αναλυτικές προδιαγραφές και οι αρχιτεκτονικές/τεχνολογικές επιλογές που περιλαμβάνονται εδώ θα χρησιμεύσουν ως οδηγός για την ανάπτυξη της πλατφόρμας, διασφαλίζοντας τη συστηματική υλοποίηση των λειτουργιών συλλογής, ανάλυσης και διαχείρισης δεδομένων από διαδικτυακές πηγές.

Με την πρόοδο της ανάπτυξης και κατά τη διάρκεια της πιλοτικής φάσης, το παρόν έγγραφο θα ενημερώνεται ώστε να ανταποκρίνεται σε νέες απαιτήσεις και λειτουργικές ανάγκες που ενδέχεται να προκύψουν. Αυτή η δυναμική προσαρμογή είναι απαραίτητη για την επιτυχή ενσωμάτωση πρόσθετων λειτουργιών και τη συνεχή βελτιστοποίηση των υποσυστημάτων, συμβάλλοντας στην επίτευξη των στόχων και στη συνολική απόδοση της πλατφόρμας.

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

8 Βιβλιογραφία - Αναφορές

[1] Karkaletsis, V., Paliouras, G., Petasis, G., Manousopoulou, N., & Spyropoulos, C. D. (1999). Named-entity recognition from Greek and English texts. *Journal of Intelligent and Robotic Systems*, *26*(2), 123-135.
[2] Boutsis, S., Demiros, I., Giouli, V., Liakata, M., Papageorgiou, H., & Piperidis, S. (2000). A system for recognition of named entities in Greek. In *Natural Language Processing—NLP 2000: Second International Conference Patras, Greece, June 2–4, 2000 Proceedings 2* (pp. 424-435). Springer Berlin Heidelberg.
[3] Petasis, G., Vichot, F., Wolinski, F., Paliouras, G., Karkaletsis, V., & Spyropoulos, C. D. (2001, July). Using machine learning to maintain rule-based named-entity recognition and classification systems. In *proceedings of the 39th annual meeting of the association for computational linguistics* (pp. 426-433).
[4] Koutsikakis, J., Chalkidis, I., Malakasiotis, P., & Androutsopoulos, I. (2020, September). Greek-bert: The greeks visiting sesame street. In *11th Hellenic conference on artificial intelligence* (pp. 110-117).
[5] 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*.
[6] Opdahl, A. L., Al-Moslmi, T., Dang-Nguyen, D. T., Gallofré Ocaña, M., Tessem, B., & Veres, C. (2022). Semantic knowledge graphs for the news: A review. *ACM Computing Surveys*, *55*(7), 1-38.
[7] Harandizadeh, B., & Singh, S. (2020, November). Tweeki: Linking named entities on Twitter to a knowledge graph. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)* (pp. 222-231).

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