Newsletter 7

- Φεβρουάριος 2023 -

inPOINT

ΑΝΑΠΤΥΞΗ ΟΛΟΚΛΗΡΩΜΕΝΗΣ ΠΛΑΤΦΟΡΜΑΣ ΓΙΑ ΤΗΝ ΥΠΟΣΤΗΡΙΞΗ ΚΑΙ ΕΝΙΣΧΥΣΗ ΔΡΑΣΕΩΝ ΑΝΟΙΧΤΗΣ ΚΑΙΝΟΤΟΜΙΑΣ

inpoint-project.eu

Περιεχόμενα

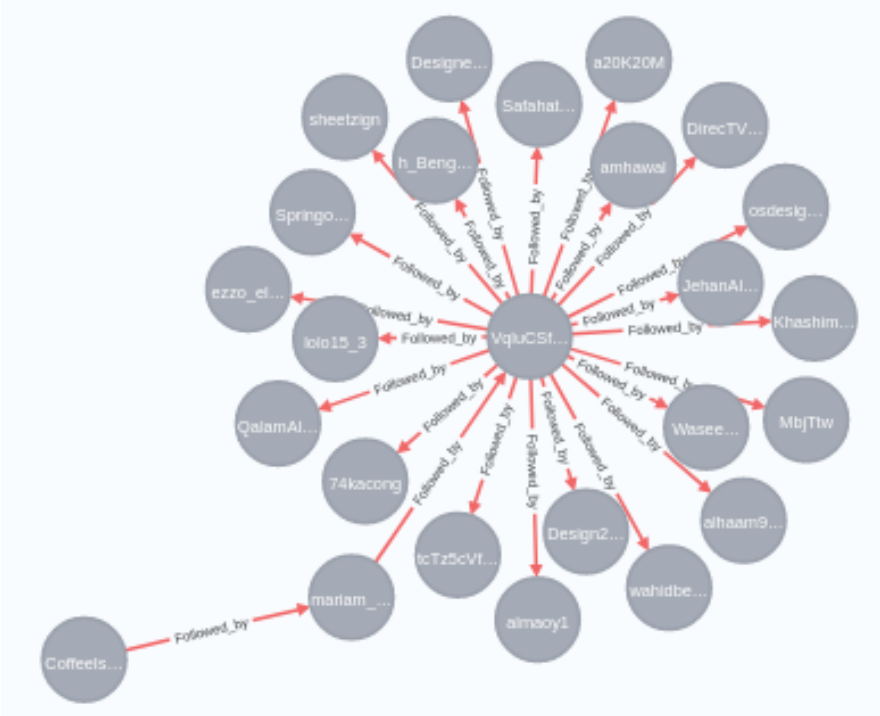
1. Υπηρεσίες Εξόρυξης και Ανάλυσης Δεδομένων

Καθώς το έργο οδεύει προς τους τελευταίους μήνες σχεδίασης και ανάπτυξης, η πλατφόρμα Ανοιχτής Καινοτομίας και οι ενσωματωμένες υπηρεσίες βελτιώνονται συνεχώς, χρησιμοποιώντας την ανατροφοδότηση και τα σχόλια που έχουν προκύψει από τους δύο κύκλους αξιολόγησης (Παραδοτέα **Π5.2** και **Π5.3**).

Στο παραπάνω πλαίσιο έχει ολοκληρωθεί η δεύτερη έκδοση των υπηρεσιών εξόρυξης και ανάλυσης δεδομένων (Παραδοτέο **Π2.2**) και επιλύονται σταδιακά θέματα σταθερότητας και απόδοσης που έχουν επισημανθεί από τους χρήστες. Συνοπτικά, οι υπηρεσίες που έχουν αναπτυχθεί με σκοπό την αξιοποίηση μεγάλου όγκου δεδομένων από τα κοινωνικά δίκτυα είναι οι εξής:

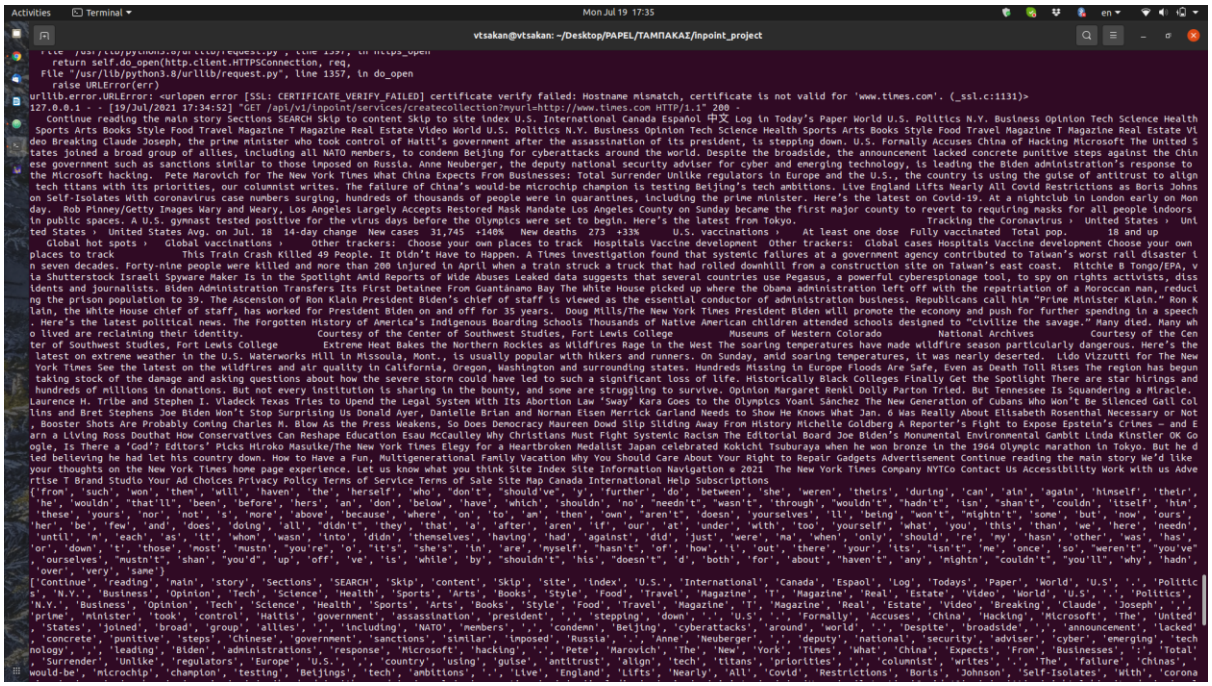
1. Ανάλυση συναισθήματος από δεδομένα κοινωνικών δικτύων
2. Δημιουργία γράφου «εγγύτητας» από δεδομένα κοινωνικών δικτύων
3. Ανάκτηση LinkedIn profiles μέσω web-crawling
4. Συλλογή κειμένων και εξόρυξη γνώσης

Η υπηρεσία ανάλυσης συναισθήματος από δεδομένα κοινωνικών δικτύων συλλέγει δεδομένα από το Twitter τα οποία αποθηκεύονται σε μια μη σχεσιακή βάση δεδομένων. Στη συνέχεια, αφού λαμβάνει χώρα η απαραίτητη προ-επεξεργασία (καθαρισμός των tweets από άσκοπους χαρακτήρες, mentions, URLs κ.λπ.) χρησιμοποιείται το μοντέλο VADER (Valence Aware Dictionary for Sentiment Reasoning για ανάλυση συναισθημάτων, το οποίο είναι ευαίσθητο τόσο στην πολικότητα (θετική/αρνητική/ουδέτερη) όσο και στην ένταση (δύναμη) του συναισθήματος. Η δεύτερη υπηρεσία δημιουργεί έναν γράφο εγγύτητας ενός χρήστη του Twitter. Ως γράφος εγγύτητας βάθους *d* ορίζεται ένας γράφος ο οποίος περιλαμβάνει ως κόμβους τους followers του κεντρικού χρήστη καθώς και τους followers αυτών (σε βάθος *d*) και ως ακμές τις σχετικές συνδέσεις φιλίας. Στην Εικόνα 1 απεικονίζεται ο γράφος εγγύτητας που παράγεται όταν ο χρήστης αναζητά τον γράφο εγγύτητας του χρήστη «Coffee_Island_GR» και τους φίλους αυτών σε βάθος 2.



Εικόνα 1: Αρχή δημιουργίας του γράφου εγγύτητας (βάθος 2).

Η Υπηρεσία ανάκτησης LinkedIn προφίλ μέσω web-crawling αποσκοπεί στην εξαγωγή τόσο εταιρικών όσο και προσωπικών profiles από το μέσο κοινωνικής δικτύωσης LinkedIn σύμφωνα με κάποιες λέξεις κλειδιά. Το αποτέλεσμα που επιστρέφει η υπηρεσία είναι ένα **δομημένο** σύνολο από πληροφορίες που περιγράφουν τους χρήστες/εταιρίες. Η υπηρεσία αυτή έρχεται να ενισχύσει τις προαναφερθείσες διαδικασίες ανάκτησης, αποθήκευσης και επεξεργασίας δεδομένων κοινωνικής δικτύων. Η υπηρεσία αυτή αξιοποιεί την βιβλιοθήκη BeautifulSoup για την διαδικασία του web-crawling ενώ ταυτόχρονα εντοπίζει το όνομα του χρήστη, την τοποθεσία του, την επαγγελματική του εμπειρία, τις σπουδές του και την τρέχουσα επαγγελματική του θέση. Τέλος, η υπηρεσία συλλογής κειμένων και εξόρυξης γνώσης συλλέγει με αυτόματο τρόπο κείμενα, τα αποθηκεύει σε κατάλληλες βάσεις δεδομένων και στη συνέχεια ανακτά γνώση από αυτά. Η μικρο-αρχιτεκτονική της υπηρεσίας περιλαμβάνει δύο συνιστώσες: (α) την διαδικασία web-crawling και εξόρυξης κειμένων, όπου συλλέγονται τα δεδομένα κειμένων και αποθηκεύονται σε μία βάση δεδομένων και (β) την διαδικασία συσταδοποίησης όπου δημιουργούνται ομάδες παρόμοιων κειμένων που μπορούν να προσπελάσουν ανά πάσα ώρα και στιγμή οι χρήστες της πλατφόρμας. Στην Εικόνα 2 μπορούμε να δούμε ένα παράδειγμα της εφαρμογής.



Εικόνα 2: Παράδειγμα κλήσης της υπηρεσίας συλλογής κειμένων και εξόρυξης γνώσης.

2. Δημοσιεύσεις

Το ερευνητικό έργο των φορέων υλοποίησης του έργου inPOINT συνεχίζεται. Στο πλαίσιο της Ερευνητικής Ενότητας 2 του έργου inPOINT, όπου αναπτύχθηκαν οι προαναφερθείσες υπηρεσίες, δημοσιεύτηκαν τα παρακάτω άρθρα, τα οποία έχουν γίνει δεκτά στο υψηλού κύρους πανελλαδικό συνέδριο πληροφορικής (Pan-Hellenic Conference on Informatics – PCI). Παραθέτουμε τις δημοσιεύσεις αυτές και τα βασικά ερευνητικά στοιχεία με τα οποία καταπιάνονται:

1. Ballas, I., Tsakanikas, V., Pefanis, E., & Tampakas, V. (2021, November). On Exploring the Optimum Configuration of Apache Spark Framework in Heterogeneous Clusters. In *25th Pan-Hellenic Conference on Informatics* (pp. 250-253).

Κατά την προηγούμενη δεκαετία, τόσο η βιομηχανία όσο και ο ακαδημαϊκός χώρος άρχισαν να εφαρμόζουν το παράδειγμα των μεγάλων δεδομένων. Καθώς αυξάνεται ο όγκος των δεδομένων που συλλέγονται, οι απαιτούμενες υπολογιστικές υποδομές πρέπει να αυξήσουν τις δυνατότητές τους προκειμένου να είναι σε θέση να επεξεργαστούν τα δεδομένα. Το άρθρο αυτό προτείνει ένα μοντέλο για την αξιολόγηση των βέλτιστων παραμέτρων διαμόρφωσης σε ένα ετερογενές Spark Cluster, το οποίο επικυρώνεται σε δύο διαφορετικές περιπτώσεις χρήσης. Τα πειράματα που πραγματοποιήθηκαν έδειξαν ότι το προτεινόμενο μοντέλο μπορεί να εκτιμήσει με επιτυχία τις βέλτιστες παραμέτρους διαμόρφωσης του Spark, για εφαρμογές που απαιτούν πολλούς υπολογιστικούς πόρους. Η Εικόνα 3 παρουσιάζει τα αποτελέσματα της ανάλυσης που αφορούν τον χρόνο εκτέλεσης 2 εργασιών σε συνάρτηση με τον αριθμό των υπολογιστικών πόρων (executors) που απαρτίζουν το Spark Cluster.

Εικόνα 3: Χρόνος εκτέλεσης 2 εργασιών σε συνάρτηση με τον αριθμό των υπολογιστικών πόρων (executors) που απαρτίζουν το Spark Cluster.

2. Kalogeras, G., Tsakanikas, V., Ballas, I., Aggelopoulos, V., & Tampakas, V. (2022, November). Community Detection at scale: A comparison study among Apache Spark and Neo4j. In *Proceedings of the 26th Pan-Hellenic Conference on Informatics* (pp. 21-26).

Η χρήση ολοένα και περισσότερων συσκευών παραγωγής δεδομένων, συμπεριλαμβανομένων συσκευών Internet of Things (IoT) και edge computing, έχει οδηγήσει στο πρότυπο των μεγάλων δεδομένων, το οποίο έχει ασκήσει σημαντική πίεση στις καθιερωμένες σχεσιακές βάσεις δεδομένων κατά την τελευταία δεκαετία. Οι ερευνητές έχουν προτείνει διάφορα εναλλακτικά μοντέλα βάσεων δεδομένων προκειμένου να μοντελοποιήσουν τα δεδομένα πιο αποτελεσματικά. Μεταξύ αυτών των προσεγγίσεων, οι βάσεις δεδομένων γράφων φαίνονται ως ο πιο υποσχόμενος υποψήφιος για τη συμπλήρωση των σχεσιακών προτύπων. Στο πλαίσιο της δημοσίευσης αυτής, πραγματοποιείται σύγκριση μεταξύ της Neo4j, μιας από τις κορυφαίες βάσεις δεδομένων γράφων, και της Apache Spark, μιας ενοποιημένης μηχανής για κατανεμημένο περιβάλλον επεξεργασίας δεδομένων μεγάλης κλίμακας, όσον αφορά τα όρια επεξεργασίας. Πιο συγκεκριμένα, τα δύο πλαίσια συγκρίνονται ως προς την ικανότητά τους να εκτελούν αλγορίθμους ανίχνευσης κοινοτήτων.

Φορείς Υλοποίησης

1. Τομέας Διοίκησης και Οργάνωσης, Τμήμα Μηχανολόγων & Αεροναυπηγών Μηχανικών, **Πανεπιστήμιο Πατρών**
2. Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Η/Υ, **Πανεπιστήμιο Πελοποννήσου**
3. **Coffee Island**
4. **Ergologic**

1

2

3

4

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΥΠΟΥΡΓΕΙΟ ΟΙΚΟΝΟΜΙΑΣ & ΑΝΑΠΤΥΞΗΣ
ΕΙΔΙΚΗ ΓΡΑΜΜΑΤΕΙΑ ΕΠΑ & ΤΕ
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ ΚΑΙΝΟΤΟΜΙΑ
ΕΣΠΑ 2014-2020
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης