Ανάκτηση Πληροφορίας

*Απαλλακτικό Project για το Σεπτέμβριο*

**Διδάσκων**: Μανώλης Μαραγκουδάκης

**Ακαδημαϊκό Εξάμηνο**: Εαρινό 2021

**Τίτλος**: Σημασιολογική Ανάκτηση Πληροφορίας των Ερευνητικών και Αναπτυξιακών Έργων Elasticsearch και SentenceTransformers

**Τύπος**: Ατομική

**Παράδοση**: 5 Σεπτεμβρίου 2021, 23.00 στο open courses.

**Περιγραφή**: Την εποχή των άμεσων καναλιών επικοινωνίας των φορέων με τους πολίτες, οι μηχανές αναζήτησης έχουν το καθιερωμένο τους καθήκον να επιστρέφουν στο χρήστη ανακτημένα έγγραφα και πληροφορίες που ταιριάζουν καλύτερα με την πληροφοριακή του ανάγκη, εκφραζόμενη ως ένα ερώτημα σε φυσική γλώσσα ή ως ακολουθία λέξεων. Παράλληλα με την ανάπτυξη της Μηχανικής Μάθησης, η σημασιολογική πληροφορία για το περιεχόμενο μιας πρότασης ή ενός σύντομου κειμένου είναι πλέον εφικτή να αποτυπωθεί με τη χρήση της τεχνολογίας BERT embedding. Ο σκοπός της παρούσας εργασίας είναι η ανάπτυξη ενός συστήματος που θα επιτρέπει στο χρήστη να ανακτά με σημασιολογική ομοιότητα ερευνητικά και αναπτυξιακά έργα που έχουν χρηματοδοτηθεί από την ΕΕ.

**Τεχνολογίες**: Elastic search, Python ή/και KNIME, SentenceTransformers

**Δεδομένα**: Στο παρακάτω site θα δείτε μια ιστοσελίδα από την οποία ο κάθε πολίτης μπορεί να κατεβάσει τη λίστα όλων των χρηματοδοτούμενων έργων του Horizon 2020 από την ΕΕ.

<https://data.europa.eu/data/datasets/cordish2020projects?locale=en>

Μπορείτε να κατεβάσετε τα έργα σε μορφή Excel ή CSV, όπως σας εξυπηρετεί καλύτερα

Graphical user interface

Description automatically generated with medium confidence

Τα δεδομένα αυτά έχουν ένα πεδίο με όνομα Objective*,* το οποίο περιγράφει τους σκοπούς του έργου. Αλλα ενφιαφέροντα πεδία είναι τα: acronym, title, startDate, endDate, ecMaxContribution, coordinator και coordinatorCountry.

**Εργασίες**: Ο χρήστης θα πρέπει είτε μέσα από ένα Jupyter Notebook ή μέσω του KNIME να γράφει ελεύθερο κείμενο και το σύστημα να του επιστρέφει πίσω τα Ν (20) πιο σχετικά έργα ταιριάζουν με αυτό που ρώτησε, με βάση το πεδίο Objective. Θα πρέπει να υπάρχουν 2 τρόποι λειτουργίας, (1) με το απλό elastic search index και (2) με SentenceTransformers.

Στην απάντηση του συστήματος θα πρέπει να φαίνεται η πρόταση η οποία παρουσίασε την μεγαλύτερη σχετικότητα ως προς το ερώτημα του χρήστη, το ακρωνύμιο του έργου, τον τίτλο του και όλα τα υπόλοιπα πεδία. ***Θα δοθεί 1 μονάδα bonus εάν η υλοποίηση γίνει σε web εφαρμογή και στα αποτελέσματα εμφανίζεται και χάρτης με τις χώρες των coordinator των έργων***.

Για να γίνει αυτό θα πρέπει τα δεδομένα από το παραπάνω βήμα (αφού χωριστούν σε προτάσεις) να γίνουν και αυτά SentenceTransformers embeddings και όλα μαζί (τα πεδία που προαναφέρθησαν) μαζί με το embedding να μπουν σε ένα Elastic Search index και με χρήση του cosine similarity να πραγματοποιηθεί η ανάκτηση (όπως έχει περιγραφεί στο συνοδευτικό υλικό και στον έτοιμο κώδικα που υπάρχει εκεί).

**SentenceTransformers:** ένα πλαίσιο με ακόμη καλύτερες και αποδοτικότερες μεθόδους BERT embeddings.

Θα βρείτε εδώ οδηγίες για τη χρήση του:

* <https://www.sbert.net/index.html>
* <https://www.sbert.net/docs/quickstart.html>

Θα χρησιμοποιήσετε το μοντέλο:  **paraphrase-mpnet-base-v2** ή το **paraphrase-MiniLM-L6-v2** από τη λίστα που υπάρχει εδώ: <https://www.sbert.net/docs/pretrained_models.html>

Καλή επιτυχία,

Μανώλης Μαραγκουδάκης

***Σημείωση***: για οποιαδήποτε απορία, στέλνετε email στο [mmarag@ionio.gr](mailto:mmarag@ionio.gr).