Εργασία Ανάκτησης Πληροφορίας

Μαρία Νεφέλη Τυχάλα, 4152, mtychala@csd.auth.gr
Άγγελος Μανουσέλης, 3520, manousela@csd.auth.gr

Η εργασία αυτή στοχεύει στην εφαρμογή τεχνικών Ανάκτησης Πληροφορίας στις ομιλίες της Βουλής των Ελλήνων χρησιμοποιούνται τεχνικές εξαγωγής λέξεων-κλειδιών, υπολογισμού ομοιότητας μεταξύ βουλευτών και Latent Semantic Indexing για θεματική ανάλυση.

Τεχνολογίες που χρησιμοποιήσαμε

Η εφαρμογή υλοποιήθηκε σε Python 3.13.9 (σε 3.14 δεν λειτουργεί), με τοπικό server Flask και front-end βασισμένο σε Jinja και Bootstrap. Οι ομιλίες αποθηκεύονται σε βάση δεδομένων SQLite και για την επεξεργασία κειμένου χρησιμοποιούνται το εργαλείο greek_stemmer_plus για stemming. Το LSI γίνεται μέσω SVD του scikit-learn.

Επεξεργασία κειμένου

Για αρχή γίνεται ένας καθαρισμός του κειμένου με regular expressions και μετά stemming με την χρήση της βιβλιοθήκης greek_stemmer_plus. Σε αυτήν την διαδικασία φτιάχνεται και ένα map με τα stems και κάποια ενδεικτική λέξη ώστε να είναι πιο κατανοητά τα αποτελέσματα. Η υλοποίηση γίνεται στο αρχείο text_processor.py.

TF-IDF

To term frequency υπολογίζεται με
$tf(t,d)=1+\log(f_{td})$

To inverse document frequency υπολογίζεται με
$idf(t)=\log(\frac{N+1}{df(t)})$

Και τα βάρη με
$w_{td}=tf(t,d)*idf(t)$

Οι τύποι είναι ίδιοι και για το query. Για την αναζήτηση βρίσκουμε τις μεγαλύτερες τιμές από την πράξη.
$w*q$
δεν γίνεται κανονικοποίηση γιατί παρατηρήσαμε καλύτερα αποτελέσματα στο απλό dot product από το cosine similarity.
Τα βάρη αποθηκεύονται σε έναν αραιό πίνακα της βιβλιοθήκης scipy. Η υλοποίηση βρίσκεται στο αρχείο tfidf.py

Keywords
Για την εξαγωγή των keywords βρίσκουμε τις λέξεις με τα μεγαλύτερα βάρη σε κάθε ομιλία (ανά βουλευτή ή κόμμα). Η υλοποίηση βρίσκεται στο αρχείο keywords.py

LSI
Για να εντοπιστούν θεματικές περιοχές χρησιμοποιήθηκε SVD από το scikit-learn, κάθε ομιλία εκφράζεται ως διάνυσμα σε πολυδιάστατο χώρο και στην συνέχεια υπολογίζεται μια ‘μέση ομιλία’ για κάθε βουλευτή που λειτουργεί σαν ένα διάνυσμα χαρακτηριστικών. Στην συνέχεια γίνεται σύγκριση ανά ζεύγη. Η υλοποίηση βρίσκεται στο αρχείο lsi.py

Οδηγίες
Για να τρέξετε την εφαρμογή βεβαιωθείτε ότι έχετε εγκατεστημένες τις βιβλιοθήκες που χρειάζεται, υπάρχει αρχείο requirements.txt, άρα μπορείτε να εγκαταστήσετε με

pip install -r requirements.txt

Επίσης βεβαιωθείτε ότι το αρχείο του dataset με όνομα Greek_Parliament_Proceedings_1989_2020.csv βρίσκεται στον φάκελο του project. Μέτα από αυτά τα βήματα μπορείτε να τρέξετε το αρχείο app.py και θα ανοίξει αυτόματα καρτέλα στον browser σας. Αν δεν ανοίξει για κάποιον λόγο θα βρείτε την εφαρμογή στο localhost:5000/.

Η πρώτη φορά που θα τρέξει η εφαρμογή μπορεί να αργήσει λόγο των υπολογισμών, αλλά γίνεται caching ώστε να τρέχει γρήγορα μετά την πρώτη εκκίνηση.