Ανάκτηση Πληροφορίας
Χειμερινό εξάμηνο 2022-2023

Ομιλίες Ελληνικού Κοινοβουλίου 1989-2020

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ
Στην εργασία αυτή θα ασχοληθούμε με την εφαρμογή τεχνικών Ανάκτησης Πληροφορίας στις ομιλίες της Βουλής των Ελλήνων. Τα δεδομένα έχουν συγκεντρωθεί από τον ιστότοπο της Βουλής, καθώς οι ομιλίες της Βουλής είναι διαθέσιμες στον οποιονδήποτε. Τα δεδομένα έχουν συγκεντρωθεί από το σύνδεσμο

https://www.hellenicparliament.gr/Praktika/Synedriaseis-Olomeleias

Αυτό το σύνολο δεδομένων περιλαμβάνει 1.280.918 ομιλίες βουλευτών της Ελληνικής Βουλής με συνολικό όγκο 2,30 GB, που εξήχθησαν από 5.355 αρχεία καταγραφής συνεδριάσεων της Βουλής. Εκτείνονται χρονολογικά από τις αρχές του Ιουλίου 1989 έως τα τέλη Ιουλίου 2020. Το σύνολο των δεδομένων αποτελείται από ένα αρχείο .csv σε κωδικοποίηση UTF-8 και περιλαμβάνει πολλές στήλες που σχετίζονται με το μέλος, το κόμμα, την ημερομηνία ομιλίας κ.λπ. Περισσότερες πληροφορίες σχετικά με τα δεδομένα μπορούν να θα βρείτε στο ακόλουθο αποθετήριο Github:

https://github.com/iMEdD-Lab/Greek_Parliament_Proceedings

Στόχος σας είναι η οργάνωση και επεξεργασία των δεδομένων έτσι ώστε να μπορούμε να βρίσκουμε χρήσιμες πληροφορίες από τις ομιλίες αυτές. Αναλυτικότερα, θα πρέπει να εστιάσουμε στα ακόλουθα.

ΥΛΟΠΟΙΗΣΗ

1. Θα πρέπει να υλοποιηθεί μία web-based εφαρμογή η οποία θα δίνει τη δυνατότητα να αναζητούμε πληροφορίες από τα δεδομένα των ομιλιών. Η εφαρμογή αυτή θα πρέπει να προσφέρει τις δυνατότητες μιας μηχανής αναζήτησης για το συγκεκριμένο δύνολο δεδομένων.

2. Ανά ομιλία, ανά βουλευτή και ανά κόμμα, θα πρέπει να βρούμε τις σημαντικότερες λέξεις-κλειδιά (keywords) και πως αυτές αλλάζουν στο χρόνο.

3. Δεδομένων όλων των ομιλιών, πρέπει να ανιχνεύσουμε ομοιότητες ανά ζεύγη μεταξύ των μελών του κοινοβουλίου. Συγκεκριμένα, πρέπει να βρούμε έναν τρόπο να εξαγάγουμε ένα διάνυσμα χαρακτηριστικών για κάθε μέλος και στη συνέχεια να εκτελέσουμε ομοιότητες ανά ζεύγη για να μπορέσουμε να ανιχνεύσουμε τα top-*k* ζεύγη με τον υψηλότερο βαθμό ομοιότητας (όπου *k* είναι μια παράμετρος).

4. Λαμβάνοντας υπόψη όλες τις ομιλίες, θα πρέπει να χρησιμοποιήσουμε την τεχνική LSI, ώστε να βρούμε τις σημαντικότερες θεματικές περιοχές και να εκφράσουμε την κάθε ομιλία ως διάνυσμα σε κάποιον πολυδιάστατο χώρο.

5. Μπορούμε να χρησιμοποιήσουμε ομαδοποίηση στις ομιλίες έτσι ώστε να σχηματίσουμε ομάδες ομιλιών έτσι ώστε οι ομιλίες της ίδιας ομάδας να έχουν μεγάλη ομοιότητα ?

6. Εδώ θα πρέπει η κάθε ομάδα να προτείνει και να υλοποιήσει μία συγκεκριμένη εργασία που θα έχει ενδιαφέρον και θα δίνει στην έξοδο ενδιαφέροντα αποτελέσματα.

ΟΔΗΓΙΕΣ
Μπορείτε να χρησιμοποιήσετε όποιες τεχνολογίες και γλώσσες προγραμματισμού θέλετε. Σε περιπτώσεις όπου το αρχικό σύνολο δεδομένων φαίνεται πολύ μεγάλο, μπορείτε να χρησιμοποιήσετε ένα δείγμα των δεδομένων. Ωστόσο, το τελικό παραδοτέο θα πρέπει να

αναφέρεται σε όλα τα δεδομένα. Θα πρέπει να σχηματίσετε ομάδες των 2 ατόμων. Τα παραδοτέα του έργου είναι:

1) ο πηγαίος κώδικας με σχόλια,
2) η τεχνική έκθεση που θα συνοψίζει την εργασία σας.

Αυτή η εργασία λαμβάνει το **50%** του συνολικού βαθμού. Θα έχετε την ευκαιρία να εργαστείτε σε ένα ενδιαφέρον σύνολο δεδομένων και να επεξεργαστείτε τις τεχνικές που συζητάμε στην τάξη. Επίσης, η ομαδική εργασία είναι πολύ σημαντική για τη μελλοντική σας καριέρα.

Προσπαθήστε να ανακαλύψετε ενδιαφέροντα πράγματα!

Καλή επιτυχία,

α.π.