Εφαρμοσμένη Επιστήμη Δεδομένων

Εξαμηνιαία Εργασία 2025 - Οικονομικό Πανεπιστήμιο Αθηνών

Οδυσσέας Χλαπάνης, ΥΔ τμήμα Επιστήμης Υπολογιστών (odyhlapanis@aueb.gr)
Παρασκευή Πλατάνου, ΥΔ τμήμα Επιστήμης Υπολογιστών (platanou@aueb.gr)
Προθεσμία υποβολής: 1η Μαΐου 2025

Μέρος Α: Συλλογή Δεδομένων Αποφάσεων Αρείου Πάγου

Στο Α' μέρος της εργασίας, θα αναπτύξετε λογισμικό για web crawling για να συλλέξετε δεδομένα από αποφάσεις του Αρείου Πάγου (https://areiospagos.gr). Επικεντρωθείτε στις αποφάσεις του έτους 2024, συμπεριλαμβάνοντας τόσο ποινικές όσο και πολιτικές υποθέσεις. Ο τελικός στόχος είναι να δημιουργήσετε ένα dataframe που θα περιλαμβάνει όλες τις αποφάσεις. Θα πρέπει να συλλέξετε (τουλάχιστον) τα εξής δεδομένα για κάθε απόφαση: αριθμός και έτος απόφασης, τμήμα (Πολιτικό ή Ποινικό), αριθμός τμήματος (πχ Β2), δικαστές, ΤΟ ΔΙΚΑΣΤΗΡΙΟ ΤΟΥ ΑΡΕΙΟΥ ΠΑΓΟΥ (και το εισαγωγικό κείμενο που ακολουθεί), ΣΚΕΦΘΗΚΕ ΣΥΜΦΩΝΑ ΜΕ ΤΟ ΝΟΜΟ, ΓΙΑ ΤΟΥΣ ΛΟΓΟΥΣ ΑΥΤΟΥΣ, άρθρα ΠΚ (Ποινικού Κώδικα) και άρθρα ΚΠΔ (Κώδικα Ποινικής Δικονομίας) αν πρόκειται για Ποινικό Τμήμα, άρθρα ΑΚ (Αστικού Κώδικα) και άρθρα ΚΠολΔ (Κώδικα Πολιτικής Δικονομίας) αν πρόκειται για Πολιτικό τμήμα.

Α1. Εξάγετε τις προδιαγραφές του dataframe. Χρησιμοποιήστε το dataframe που θα βρείτε στο σχετικό dataset του Huggingface (DominusTea/GreekLegalSum · Datasets at Hugging Face). Προσέξτε ότι μπορεί να απουσιάζουν οι περιλήψεις στο έτος 2024.

Α2. Ανάπτυξη κώδικα και εκτέλεσή του για scraping και crawling.

Α3. Αναλύστε τα δεδομένα που συλλέξατε σε σχέση με αυτά που θα περιμένατε (π.χ., σε σχέση με αυτά είχε το GreekLegalSum).

Α4. Παρουσιάστε οπτικά τα δεδομένα με σωστό τρόπο, αιτιολογώντας τις επιλογές σας. Φτιάξτε κατάλληλα διαγράμματα για κάθε είδους δεδομένα που συλλέξατε.

Παραδοτέο: ένα συμπιεσμένο αρχείο με τα εξής: (α) ένα CSV αρχείο (DataFrame) το οποίο θα περιέχει τα δεδομένα που έχετε συλλέξει στη μορφή που σας έχει ζητηθεί, (β) ένα PDF μιας σελίδας (Arial font, size 11) που θα αναφέρετε τι κάνατε στις εξής ενότητες (1. Περιγραφή της προσέγγισής σας, 2. Προδιαγραφές, 3. Κώδικας και Διαδικασία Crawling, 4. Ανάλυση Αποτελεσμάτων), (γ) ένα notebook με τον κώδικά σας (θα πρέπει να περιλαμβάνει τα output κελιά).