Εφαρμοσμένη Επιστήμη Δεδομένων

Ατομική 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: 25 Μαΐου 2025

Μέρος Β: Μηχανική Μάθηση σε Ελληνικά Νομικά Κείμενα

Στο Β' μέρος της εργασίας, θα εφαρμόσετε μεθόδους επιβλεπόμενης και μη επιβλεπόμενης μηχανικής μάθησης σε ελληνικά νομικά κείμενα. Η εργασία επικεντρώνεται στην ταξινόμηση νομικών εγγράφων και στην εξαγωγή και ανάλυση θεμάτων από αυτά, χρησιμοποιώντας διαφορετικές προσεγγίσεις.

Β1. Ταξινόμηση Νομικών Εγγράφων με Επιβλεπόμενη Μηχανική Μάθηση

Στο πρώτο ερώτημα θα υλοποιήσετε μοντέλα ταξινόμησης για την κατηγοριοποίηση νόμων. Συγκεκριμένα, θα χρησιμοποιήσετε το dataset Greek Legal Code του πανεπιστημίου ΕΚΠΑ (θα το βρείτε στο σύνδεσμο: AI-team-UoA/greek_legal_code), το οποίο περιλαμβάνει 47 χιλιάδες έγγραφα νόμων από την συλλογή Ραπτάρχης. Για κάθε έγγραφο περιλαμβάνονται και τρεις ετικέτες: συλλογή, κεφάλαιο, θέμα. Στόχος σας είναι να προβλέψετε τις ετικέτες του εγγράφου από το κείμενο του. Υλοποιήστε και συγκρίνετε τα ακόλουθα μοντέλα ταξινόμησης:

i) Support Vector Machines (SVM) με Bag-of-Words (BoW) και TF-IDF αναπαράσταση.

ii) Logistic Regression με dense embeddings που προέρχονται από μία υλοποίηση εκ των: Word2Vec, fastText, Glove.

iii) Ένα τρίτο μοντέλο ταξινόμησης από τα εξής: K-Nearest Neighbors, Naive Bayes, Random Forest, MLP, XGBoost με όποια embeddings θέλετε από τα παραπάνω (TF-IDF ή κάποιο από τα dense embeddings).

Για την αξιολόγηση να χρησιμοποιήσετε τις μετρικές: Accuracy, Precision, Recall και F1-score και να παρουσιάσετε σε πίνακα τα αποτελέσματα των τριών μοντέλων ανά κατηγορία. Η εκπαίδευση των μοντέλων να γίνει στο train set, η επιλογή των υπερπαραμέτρων στο validation set και η αξιολόγηση στο test set.

Β2. Ανάλυση Θεμάτων Νομικών Αποφάσεων Αρείου Πάγου

Στο δεύτερο ερώτημα θα κάνετε ανάλυση των θεμάτων των νομικών αποφάσεων του Αρείου Πάγου που είδαμε και στο Ά μέρος της εργασίας. Αυτή τη φορά θα χρησιμοποιήσετε το έτοιμο dataset Greek Legal Sum και δεν θα χρειαστεί να κάνετε κάποιο crawling. Το dataset περιέχει το κείμενο της απόφασης (text), την περίληψη της απόφασης (summary), την κατηγορία (case category) και ετικέτες (case tags). Το subset δεν θα το χρειαστούμε.

i) Εφαρμόστε μια διερεύνηση (exploratory data analysis) των παρεχόμενων θεματικών ετικετών (case_category, case_tags**)** χρησιμοποιώντας κατάλληλα διαγράμματα (π.χ., με βάση τη συχνότητα). Σχολιάστε τα αποτελέσματα.

ii) Εφαρμόστε την μέθοδο ομαδοποίησης K-μέσων (K-means). Θα πρέπει να αναπαραστήσετε τα κείμενα των αποφάσεων (ή των περιλήψεων τους) με κάποια αναπαράσταση, όπως αυτές που χρησιμοποιήσατε στο Β1 ερώτημα (TF-IDF ή dense embeddings). Για την επιλογή του βέλτιστου K θα χρησιμοποιήσετε τον συντελεστή Silhouette (με micro και macro aggregation) και το NMI (για τις ετικέτες και τις κατηγορίες των αποφάσεων). Οπτικοποιήστε και σχολιάστε τα αποτελέσματα που σας βοήθησαν στην επιλογή του K.

iii) Χρησιμοποιήστε ένα Μεγάλο Γλωσσικό Μοντέλο (LLM) για την εξαγωγή τίτλου για κάθε συστάδα, δίνοντάς του τρεις κατάλληλα επιλεγμένες αποφάσεις ανά συστάδα (3-shot learning). Συγκρίνετε τους τίτλους που παράγονται όταν δίνετε τυχαία επιλεγμένες αποφάσεις της συστάδας και αποφάσεις που βρίσκονται κοντά στο κεντροειδές. Επιλέξτε με επιχειρήματα τον καλύτερο τρόπο. Προτείνεται το μοντέλο Llama-Krikri-8B-Instruct με τη βιβλιοθήκη Unsloth και χρήση δωρεάν GPUs από το Google Colab (εναλλακτικά, αν έχετε δυσκολίες με το colab, μπορείτε να χρησιμοποιήσετε τα μικρότερα μοντέλα: gemma-3-4b-it, gemma-3-1b-it, ή κάποιο άλλο της επιλογής σας).

- Είναι πιθανόν το πλήρες κείμενο να μην μπορεί να δοθεί ολόκληρο, οπότε αν χρειαστεί δώστε του κάποιο απόσπασμα του κειμένου ή την περίληψη.
- Μην χρησιμοποιήσετε στο prompt τα **case_category** και **case_tags** στα οποία έχετε πρόσβαση από το dataset.

Παράδειγμα προτροπής (prompt) για εξαγωγή θέματος από νομική απόφαση:

“Σου δίνεται ένα κείμενο νομικής απόφασης. Ποιο είναι το κεντρικό θέμα της απόφασης; Απάντησε στην μορφή: ‘Θέμα:\n…’. Το κείμενο είναι αυτό:\n{ΚΕΙΜΕΝΟ}”

Την απάντηση του LLM θα πρέπει να την εξάγετε με κατάλληλο κώδικα με κανονικές εκφράσεις (regular expressions).

Προαιρετικά, για κάθε συστάδα, μελετήστε (και παρουσιάστε με κατάλληλη οπτικοποίηση) τη κατανομή των ετικετών και των κατηγοριών. Αναλύστε τη σύνδεση μεταξύ τίτλων και ετικετών/κατηγοριών για συστάδες της επιλογής σας.

Παραδοτέο: ένα συμπιεσμένο αρχείο (.zip) με όνομα τον αριθμό μητρώου σας, το οποίο θα περιέχει τα εξής: (α) Ένα PDF αρχείο (έως 2 σελίδες συνολικά, Arial font size 11) που θα περιγράφει συνοπτικά τι κάνατε στο Μέρος Β, με ξεχωριστές ενότητες για τα υποερωτήματα Β1 και Β2, (β) Δύο Jupyter notebooks (.ipynb) με τον πλήρη κώδικά σας για τα ερωτήματα Β1 και Β2. Ο κώδικας θα πρέπει να είναι εκτελέσιμος και να περιλαμβάνει τα κελιά εξόδου (outputs) με τα αποτελέσματα και τον σχολιασμό σας.

Υποβάλετε το συμπιεσμένο αρχείο (.zip) στο φάκελο "Εργασίες" στο e-class έως τις **25 Μαΐου 2025 και ώρα 23:55**.