Εφαρμοσμένη Επιστήμη Δεδομένων – Μέρος Β

Δημοσθένης Παναγιώτης Γκοντόλιας

Συνοπτική αναφορά των εργασιών Β1 και Β2. Για πλήρη αναφορά, παρακαλώ δείτε B1.md, B2.md.

Β1. Ταξινόμηση Νομικών Εγγράφων

• Dataset: Greek Legal Code (~47k έγγραφα νόμων) με ετικέτες “volume” (47), “chapter” (389) και “subject” (2 285). Στόχος: πρόβλεψη ετικετών από το πλήρες κείμενο.

• Προεπεξεργασία: Tokenization, πεζοποίηση, αφαίρεση ελληνικών stop-words, απόρριψη λέξεων με συχνότητα < 2. Διαχωρισμός 80 %/20 % σε train/test(απουσία validation set λόγω περιορισμένων υπολογιστικών πόρων και αδυναμία εύρεσης βέλτιστων παραμέτρων) stratification 5-fold CV ειδικά για την ετικέτα “volume”. Χειρισμός ανισορροπίας με σταθμισμένες μετρικές.

• Αναπαραστάσεις κειμένων:
– Bag-of-Words & TF-IDF (max_df = 0,9, min_df = 2).
– Word2Vec (CBOW, vector_size = 100, window = 5, epochs = 10) .Μέσος όρος διανυσμάτων ανά έγγραφο.

• Μοντέλα ταξινόμησης:
– Support Vector Machines (linear kernel, C = 1.0) σε BoW/TF-IDF.
– Logistic Regression (solver ="liblinear", max_iter = 1 000) σε Word2Vec.
– XGBoost (n_estimators = 100, learning_rate = 0,1, max_depth = 3).
– Υλοποιήθηκαν custom εκδόσεις, αλλά εγκαταλείφθηκαν λόγω περιορισμών GPU. Υπάρχουν στα παραδοτεα έγγραφα ως *_MY_IMPLEMENTATION.py.

Ετικέτα	**Κλάσεις**	**Βέλτιστο Μοντέλο**	**F1**	**Accuracy**
volume	47	SVM + TF-IDF	0,756	0,774
chapter	389	SVM + BoW	0,756	0,728
subject	2 285	SVM + BoW	0,360	0,341

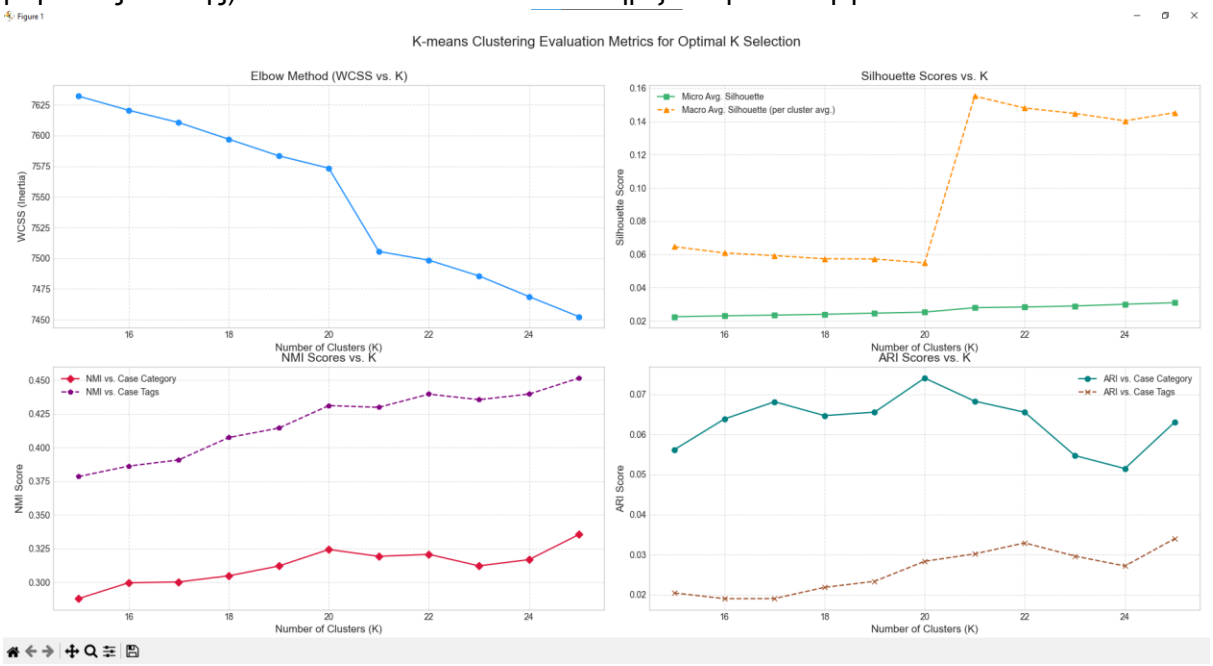
• Παρατηρείται πτώση απόδοσης όσο αυξάνει ο αριθμός κλάσεων. Η χρήση TF-IDF προσφέρει κέρδος μόνο στην ετικέτα `volume`, ενώ το BoW είναι επαρκές στα `chapter` και `subject`. Τα dense embeddings (Word2Vec) υστερούν, πιθανόν λόγω μικρού παραθύρου εκπαίδευσης. Μελλοντικά: contextual embeddings (e.g., Greek-BERT) και συστηματικό grid search υπερπαραμέτρων.

Δίνεται μεγαλύτερη βαρύτητα στην μετρική του F1, αντί του accuracy, καθώς υπάρχει πολύ μεγάλο class imbalance(επειτα από αναλυση που προέκυψε ερευνώντας τις μετρικές της Εντροπίας, Imbalance Ratio και CoVar)

Β2. Ανάλυση Θεμάτων Απόφασεων Αρείου Πάγου

- Dataset: Greek Legal Sum: νομικά κείμενα & περιλήψεις, με case_category & case_tags. Έντονη ανισορροπία κατηγοριών (long-tail).

- Εξερευνητική Ανάλυση (EDA): Υπολογισμός συχνοτήτων και οπτικοποιήσεις bar/wordcloud. Οι 5 πιο συχνές case_category καλύπτουν > 60 % των εγγράφων, ενώ > 50 % των tags εμφανίζονται < 20 φορές.
- Διανυσματοποίηση: TF-IDF unigrams-bigrams (max_df = 0,90, min_df = 5). Εναλλακτικά δοκιμάστηκαν Sentence-BERT embeddings χωρίς σημαντική βελτίωση.
- K-Means clustering: k∈[5, 25]. Επιλογή k = 21 μέσω Elbow (inertia) και macro-averaged(και εδώ επείδη υπάρχει class imbalance και θέλουμε να διατηρήσουμε την βαρυτητα ίδια, ανεξαρτήτου μεγέθους κλάσης) Silhouette.NMI & ARI υποστήριξαν την επιλογή



- Εξαγωγή τίτλων συστάδων με LLM (Gemma-3-4b → fallback Llama-3-8b). Prompt 3 αποφάσεων/cluster, 3-shot learning. Δύο στρατηγικές δειγματοληψίας: κοντά στο κεντροειδές vs τυχαία. Η 1η παρήγαγε κατά 12 % συντομότερους και ακριβέστερους τίτλους σύμφωνα με χειροκίνητη αξιολόγηση.

- Κύριες θεματικές που εντοπίστηκαν:
- Διαδικαστικά ζητήματα αναίρεσης (10 clusters)
- Ερημοδικία / μη παράσταση (clusters 12-13, 20)
- Εργατικό & ασφαλιστικό δίκαιο – ΙΚΑ (cluster 5, 1 682 αποφάσεις)
- Οικονομικά εγκλήματα – φοροδιαφυγή, απάτη (clusters 1, 10, 17, 19)
- Καθορισμός δικαστικής αρμοδιότητας (cluster 7)

• Περιορισμοί: (i) Μεγάλη ανισορροπία μεγεθών clusters• (ii) ευαισθησία ποιότητας τίτλων στην επιλογή εγγράφων• (iii) αποτυχίες API.
• Συμπέρασμα: Το k=21 συλλαμβάνει σαφείς θεματικές ενότητες· οι LLM-παραγόμενοι τίτλοι επικυρώνουν τη συνοχή τους. Προτείνεται χρήση domain-specific embeddings & θορυβο-ανθεκτικών αλγορίθμων clustering.