UNIVERSITY OF PIREAUS & NCSR "DEMOKRITOS"
MSC PROGRAMME ARTIFICIAL INTELLIGENCE (AI)

Σημασιολογική ανάκτηση νομικών κειμένων

από

Κωνσταντάκος Σωτήριος

Υποβάλλεται για την εκπλήρωση των προϋποθέσεων λήψης
Μεταπτυχιακού Διπλώματος στην «Τεχνητή Νοημοσύνη» στο
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Κύριος επιβλέπων: Γιαννακόπουλος Γεώργιος
Ερευνητής

Μέλη εξεταστικής επιτροπής: Πετάσης Γεώργιος, Ερευνητής
Ρεντούμη Βασιλική, Ερευνήτρια

Αθήνα, 06 2021

Σημασιολογική ανάκτηση νομικών κειμένων

Κωνσταντάκος Σωτήριος

MSc. Thesis, MSc. Programme in Artificial intelligence (AI)

University of Pireaus & NCSR "Demokritos", 06 2021

UNIVERSITY OF PIREAUS & NCSR "DEMOKRITOS"
MSC PROGRAMME ARTIFICIAL INTELLIGENCE (AI)

Σημασιολογική ανάκτηση νομικών κειμένων

από

Κωνσταντάκος Σωτήριος

Υποβάλλεται για την εκπλήρωση των προϋποθέσεων λήψης

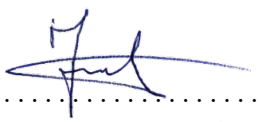
Μεταπτυχιακού Διπλώματος στην «Τεχνητή Νοημοσύνη» στο

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Κύριος επιβλέπων: Γιαννακόπουλος Γεώργιος
Ερευνητής

Μέλη εξεταστικής επιτροπής: Πετάσης Γεώργιος, Ερευνητής
Ρεντούμη Βασιλική, Ερευνήτρια

Έγινε αποδεκτή από την επιτροπή στις 06, 2021.

(Υπογραφή)	(Υπογραφή)	(Υπογραφή)
		
Γεώργιος Γιαννακόπουλος Ερευνητής	Γεώργιος Πετάσης Ερευνητής	Βασιλική Ρεντούμη Ερευνητής

Αθήνα, 06 2021

Ευχαριστίες

Πρώτα απ΄όλα, θέλω να εκφράσω την ευγνωμοσύνη μου προς τον Δρ. Γιώργο Γιαννακόπουλο, που δέχτηκε να είναι επιβλέπων μου για αυτήν την εργασία, που αφιέρωσε ένα σημαντικό μέρος του χρόνου και της προσπάθειάς του στην καθοδήγηση αυτής της διατριβής και απαντούσε άμεσα σε όποιο αίτημα είχα.

Θέλω επίσης να ευχαριστήσω βαθιά τον Δρ. Άρη Κοσμόπουλο, ο οποίος έθεσε το έδαφος για αυτήν τη διατριβή, ήταν πάντα χαρούμενος να με βοηθήσει στις προσπάθειές μου και με στήριξε σε όλες τις καταστάσεις αυτής της εργασίας.

Τέλος, είναι αδύνατο να μην ευχαριστήσω τους ανθρώπους που βρίσκονται κοντά μου, την οικογένεια μου, οι οποίοι υπομονετικά με στήριζαν στην εκπόνηση της διατριβής από την αρχή μέχρι το τέλος.

Στην οικογένεια μου.

Περίληψη

Σε μία πληθώρα νομικών εφαρμογών, υπάρχει η ανάγκη να χρησιμοποιηθεί ένα κείμενο ως ερώτημα σε μία βάση γνώσης, ώστε να ανακληθούν τα πιο σχετικά κείμενα. Στο εθιμικό δίκαιο, κάθε περίπτωση έχει σχετιζόμενες προϋπάρχουσες περιπτώσεις, στις οποίες ανατρέχουν δικηγόροι και δικαστικοί. Η επεξεργασία φυσικής γλώσσας μπορεί να υπολογίσει σημασιολογική ομοιότητα κειμένων, ενώ η ανάκληση πληροφορίας επιφορτίζεται με την αποδοτική ευρετηρίαση και εύρεση σχετικών εγγράφων. Αυτή η εργασία, θα λειτουργήσει στο όριο των δύο τομέων, προσπαθώντας να κτίσει γνωστές αναπαραστάσεις, μεθόδους υπολογισμού ομοιότητας και ευρετήρια για να επιτρέψει ανάκληση νομικών εγγράφων, με αξιοποίηση της σημασιολογικής εγγύτητας. Θα γίνει περιγραφή του τι είναι η σημασιολογική ομοιότητα, θα χρησιμοποιηθούν εργαλεία ανάλυσης φυσικής γλώσσας, βαθιά μάθηση για αναπαράσταση κειμένων και τεχνικές ανάκλησης για την αναζήτηση παρόμοιων κειμένων. Χρησιμοποιώντας τις παραπάνω γνωστές τεχνικές στο νομικό τομέα, εξετάζουμε ποια είναι πιο κατάλληλη στον τομέα αυτόν.

Σε αυτήν την εργασία δίνεται μεγαλύτερη έμφαση στο κομμάτι της αξιολόγησης. Για να γίνει η αξιολόγηση των τεχνικών που υλοποιήσαμε, προτείναμε μία μέθοδο ανθρώπινης αξιολόγησης. Στα πλαίσια της μεθόδου της αυτής δημιουργήσαμε ένα εργαλείο επισημείωσης, ώστε να μπορέσει να υλοποιηθεί η διαδικασία που προτείναμε. Σκοπός ήταν να γίνουν πειράματα με αξιολογητές, ώστε να δούμε πόσο κατάλληλες ήταν τελικά οι γνωστές τεχνικές που χρησιμοποιήσαμε για τον τομέα αυτόν αλλά και ποια κρίθηκε η πιο κατάλληλη.

Abstract

Ill a variety of legal settings, there is a clear need to use a text as a query, in order to retrieve related documents . For example, in customary law, each case has related cases in the past that the lawyers and judges need to consult. In other applications, a complaint or lawsuit is related to specific laws or decisions. Natural Language Processing can support semantic text similarity, while Information Retrieval can help in retrieving the related documentss. This project will touch the intersection of the two domains, trying to build efficient representations, comparison methods and indexes to facilitate semantic-relevance-based document retrieval in the legal domain. Using the above known techniques in the legal field, we examine which is more appropriate in this field.

In this work, more emphasis is placed on the evaluation part. To evaluate the techniques we have implemented, we proposed a method of human evaluation. As part of this method, we created a annotation tool, so that the process we proposed could be implemented. The purpose was to do experiments with annotators, to see how appropriate the techniques we used for this field were and which method is the most suitable.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Κεφάλαιο 1

Εισαγωγή

1.1 Περιγραφή προβλήματος

Μέχρι στιγμής, οι δικηγόροι αλλά και το ευρύ κοινό όταν αναζητούν το κοινό δίκαιο (προηγούμενες αποφάσεις που επηρεάζουν τις τρέχουσες αποφάσεις) για ομοιότητα υποθέσεων, πρέπει να αναζητήσουν ερωτήματα σε βάσεις δεδομένων, παίρνοντας συχνά είτε λανθασμένα είτε όχι τα αναμενόμενα αποτελέσματα. Η παρούσα διπλωματική έχει ως στόχο να συνεισφέρει προτείνοντας μία μέθοδο αξιολόγησης, ώστε να δούμε ποια τεχνική δείχνει να είναι πιο κατάλληλη στο νομικό τομέα για εύρεση παρόμοιων υποθέσεων.

Πρόσφατα, ο αριθμός των ψηφιακά διαθέσιμων νομικών εγγράφων αυξήθηκε ραγδαία λόγω των εξελίξεων στην τεχνολογία των πληροφοριών. Με αυτό το αυξημένο μέγεθος του τομέα, καθίσταται δύσκολο για τους νομικούς να βρουν χειροκίνητα σχετικές προηγούμενες υποθέσεις που θα βοηθούσαν μια υπό εξέλιξη υπόθεση υπέρ τους. Ως εκ τούτου, είναι απαραίτητο για τους νομικούς να διαθέτουν συστήματα τα οποία να επαναλαμβάνουν αυτόματα προηγούμενες περιπτώσεις. Αυτό γίνεται για να διασφαλιστεί ότι οι ίδιες καταστάσεις αντιμετωπίζονται με παρόμοιο τρόπο σε κάθε δικαστική υπόθεση. Το πρόβλημα της δημιουργίας ενός αυτόματου συστήματος ανάκτησης υποθέσεων είναι στο χώρο της Ανάκτησης Πληροφορίας. Τα παραδοσιακά συστήματα ανάκτησης πληροφορίας αποστέλλονται με ερωτήματα που βρίσκονται εντός ενός εύρους μερικών λέξεων. Στην περίπτωση νομικών συστημάτων ανάκτησης πληροφορίας, ένας

νομικός επαγγελματίας θα πρέπει να είναι σε θέση να παρέχει στο σύστημα μια περιγραφή της υπό εξέλιξη υπόθεσης. Ένα σύστημα αυτόματης ανάκτησης προτεραιότητας από νομικά έγγραφα βοηθά τους νομικούς επαγγελματίες να αναφέρονται εύκολα στις υποθέσεις που έχουν σχέση με την τρέχουσα υπόθεση.

Η μέτρηση της ομοιότητας μεταξύ δύο εγγράφων δικαστικής υπόθεσης αποτελεί κρίσιμη πρόκληση που πρέπει να αντιμετωπιστεί για την εξεύρεση συνάφειας μεταξύ της τρέχουσας περιγραφής της υπόθεσης και μιας προφανώς αποφασισμένης υπόθεσης. Αυτό το έργο είναι ιδιαίτερα δύσκολο, λαμβάνοντας υπόψη το γεγονός ότι τα νομικά έγγραφα είναι γενικά μακρά και περίπλοκα στη δομή [1]. Επιπλέον, ένα έγγραφο νομικής υπόθεσης μπορεί να περιλαμβάνει συζητήσεις για πολλά διαφορετικά νομικά ζητήματα. Ως εκ τούτου, η αυτόματη διαδικασία μέτρησης ομοιότητας μεταξύ δύο εγγράφων δικαστηρίου έχει ιδιαίτερη σημασία.

Θα γίνουν πειράματα με προσεγγίσεις από την Τεχνητή Νοημοσύνη και την Επεξεργασία Φυσικής γλώσσας για να μελετήσουμε και να συγκρίνουμε γνωστές μεθόδους και να δούμε ποια έχει την καλύτερη επίδοση. Χρησιμοποιήθηκαν δύο κατάλληλα σύνολα δεδομένων, από διαφορετικές χώρες, τα οποία υποβλήθηκαν σε επεξεργασία και αποθηκεύτηκαν σε κατάλληλη μορφή ώστε να μπορούν να χρησιμοποιηθούν σε πειράματα. Υλοποιήσαμε τις προσεγγίσεις TF-IDF, Doc2vec, BERT για αναπαράσταση των νομικών κειμένων σε διανυσματικό χώρο και υπολογίσαμε την ομοιότητα συνημιτόνου για σύγκριση διανυσμάτων. Ο στόχος ήταν να συγκρίνουμε όλα τα ζεύγη εγγράφων, να τα ταξινομήσουμε με βάση την ομοιότητά τους, για να δούμε αν σχετίζονται οι κατατάξεις των ζευγαριών με βάση την ομοιότητα μεταξύ των τεχνικών, μέσω της ανάλυσης ευαισθησίας.

Σε αυτήν την εργασία, δόθηκε περισσότερο έμφαση να βρεθεί μία μέθοδος αξιολόγησης ώστε να μπορέσει να γίνει η ανθρώπινη αξιολόγηση και να δούμε του κατά πόσο σχετίζονται οι αυτόματες μετρικές ομοιότητάς με την ομοιότητα σύμφωνα με τους ανθρώπους. Η ανθρώπινη αξιολόγηση όμως για όλα τα ζευγάρια του σώματος κειμένου είναι πολύ ακριβή σε χρόνο και ανέφικτη. Για το λόγο αυτό, προτείναμε μια διαδικασία που θα διευκόλυνε αυτό το έργο, ώστε να μπορέσει να γίνει η αξιολόγηση.

Στα πλαίσια της μεθόδου αξιολόγησης που προτείναμε, αναπτύχθηκε ένα εργαλείο επισημείωσης (annotation tool) με σκοπό να το το χρησιμοποιήσουν οι αξιολογητές για

να σχολιάσουν την ομοιότητα μεταξύ νομικών κειμένων. Σκοπός μας ήταν να γίνουν πειράματα με τους αξιολογητές, ώστε στη συνέχεια να συγκρίνουμε τις επιλογές τους με τα αποτελέσματα που δίνουν οι τεχνικές αναπαράστασης που χρησιμοποιήσαμε. Έτσι, θα μπορέσουμε να καταλήξουμε στο συμπέρασμα αν οι γνωστές αυτόματες μετρικές ομοιότητας που χρησιμοποιήσαμε οδηγούν, όντως, σε όμοιες νομικές υποθέσεις και ποια από αυτές τις τεχνικές είναι καλύτερη σε σχέση με την πλευρά του ανθρώπου.

Έτσι, συνοπτικά, οι συνεισφορές αυτής της εργασίας είναι:

- Μετασχηματισμός ενός συνόλου δεδομένων σε ένα υποσύνολο, ώστε να γίνει σε αυτό η ανθρώπινη αξιολόγηση
- Μελέτη 3 αναπαραστάσεων και η επίδραση τους στη σημασιολογική ομοιότητα νομικών κειμένων
- Δημιουργία εργαλείου επισημείωσης (annotation tool), ώστε να γίνουν πειράματα με αξιολογητές
- Έλεγχος του κατά πόσο υπάρχει συμφωνία μεταξύ των τεχνικών και των ανθρώπων-αξιολογητών

1.2 Δομή εργασίας

Στις ακόλουθες ενότητες, οι παραπάνω πληροφορίες εξηγούνται λεπτομερώς. Στο δεύτερο κεφάλαιο παρουσιάζονται κάποιες άλλες σχετικές εργασίες μαζί με το τεχνικό υπόβαθρο, στο κεφάλαιο 3 περιγράφεται η προτεινόμενη μέθοδος που χρησιμοποιήθηκε για την αναπαράσταση των νομικών εγγράφων καθώς και τον υπολογισμό ομοιότητας και πως επιτεύχθηκε. Στο κεφάλαιο 4 εξηγούμε σε βάθος τα σύνολα δεδομένων, τα προβλήματα που συναντώνται σε αυτά, τη διαδικασία που προτείναμε για να γίνει αξιολόγηση καθώς και αναλύουμε το εργαλείο επισημείωσης που αναπτύξαμε. Στο κεφάλαιο 5 αναλύονται τα αποτελέσματα και τα ευρήματά της αξιολόγησης καθώς και οι παρατηρήσεις που έγιναν σε αυτά. Τέλος, στο κεφάλαιο 6 γίνεται συζήτηση σχετικά με αυτήν την εργασία και τα συμπεράσματα που βγήκαν από το πείραμα και επιπλέον δίνονται κατευθύνσεις για μελλοντικές επεκτάσεις.

Όλα αυτά δίνουν μια εικόνα της δουλειάς που πραγματοποιήθηκε στο πλαίσιο της διπλωματικής, καθώς και διαφορετικά στοιχεία που θα μπορούσαν να ξαναχρησιμοποιηθούν σε άλλα πειράματα.

Κεφάλαιο 2

Τεχνικό υπόβαθρο και σχετική εργασία

Στο κεφάλαιο αυτό περιγράφουμε τα μοντέλα που χρησιμοποιήσαμε για αυτό το πρόβλημα αλλά και αφιερώνουμε μία ενότητα για να αναφέρουμε σχετικές εργασίες που βρίσκουμε στη βιβλιογραφία, σχετικά με το πρόβλημα μας και το πως προσεγγίζεται.

2.1 Τεχνικό υπόβαθρο

Τα Word embeddings είναι λέξεις που αναπαριστώνται ως διανύσματα πραγματικών αριθμών, ενώ η απόσταση των δύο διανυσμάτων των λέξεων δείχνει τη σημασιολογική ομοιότητα μεταξύ τους.

Στις εργασίες [2][3] προτείνεται ένας αλγόριθμος μη επιβλεπόμενης μηχανικής μάθησης, που ονομάζεται Word2Vec, που χρησιμοποιεί ένα νευρωνικό δίκτυο για να εκπαιδεύσει αποτελεσματικά δισεκατομμύρια αναπαραστάσεις λέξεων σε έναν διανυσματικό χώρο.

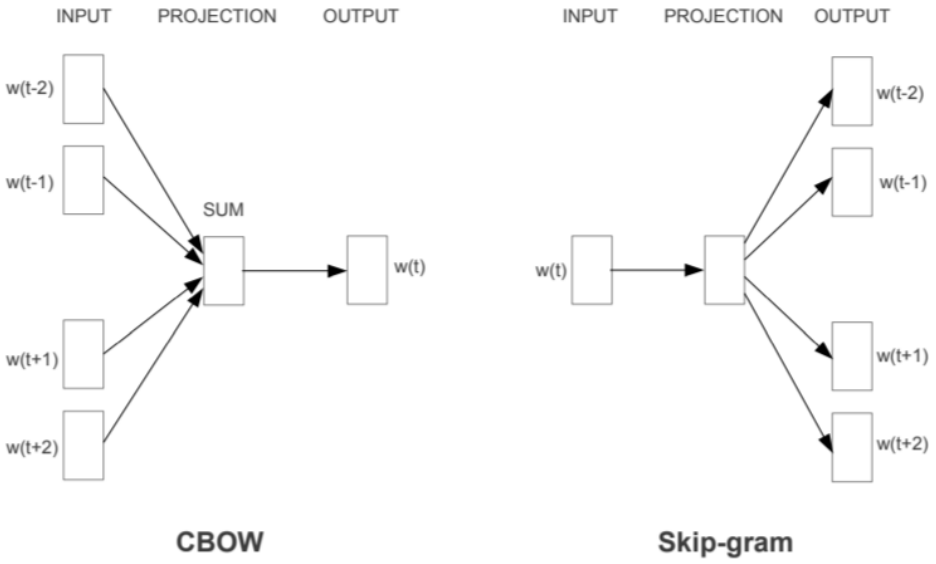
Το word2vec παίρνει ως είσοδο μία συλλογή κειμένων και παράγει ως έξοδο ενσωματώσεις λέξεων. Έχει δύο διαφορετικές αρχιτεκτονικές όπως δείχνει στο σχήμα 2.1: το συνεχές μοντέλο bag-of-words που προβλέπει μία λέξη βασισμένη στο περιεχόμενο και το συνεχές μοντέλο skip-gram που προβλέπει το περιεχόμενο δοθέντος μιας λέξης.

Το skip-gram μοντέλο, αποδίδει καλύτερα από το CBOW σύμφωνα με τον Micolov[2]. Το Word2vec έχει πολλές παραμέτρους για την βελτιστοποίηση των διανυσμάτων.

Δύο από αυτές που εστιάζουμε στο δικό μας μοντέλο είναι:

- Το μέγεθος του διανύσματος: το μήκος του διανύσματος λέξης, μεγάλο μέγεθος σημαίνει υψηλότερη χρήση μνήμης αλλά όμως περισσότερο χώρο για να περιγράψουμε τη λέξη αριθμητικά.
- Εποχές: ο αριθμός των επαναλήψεων για την εκπαίδευση ολόκληρου του σώματος κειμένου

Σχήμα 2.1: CBOW και Skip-gram μοντέλο



2.1.1 TF-IDF

Ο όρρος συχνότητα - η αντίστροφη συχνότητα εγγράφων TF-IDF, είναι ένα σύστημα στάθμισης όρων [4] που εκφράζει τη συνάφεια ενός όρου για ένα έγγραφο D σε μια συλλογή κειμένων. Είναι το προϊόν της συχνότητας όρου (tf) και της αντίστροφης συχνότητας εγγράφου (idf).

Το TF-IDF είναι μια αριθμητική στατιστική χρήση που αντικατοπτρίζει πόσο σημαντική είναι μια λέξη για το έγγραφο. Η συχνότητα του όρου TF είναι ο λόγος του αριθμού των φορών που εμφανίζεται μια λέξη σε ένα έγγραφο σε σύγκριση με τον συνολικό αριθμό λέξεων στο έγγραφο. Δίνεται από τον τύπο:

TF(t) = (Αριθμός των φορών που εμφανίζεται ο όρος t σε ένα έγγραφο) / Συνολικός αριθμός όρων του εγγράφου.)

Η αντίστροφη συχνότητα εγγράφου χρησιμοποιείται για τον υπολογισμό του βάρους των μοναδικών λέξεων σε όλα τα έγγραφα. Οι σπάνιες λέξεις έχουν υψηλή βαθμολογία IDF.

IDF(t) = log (Συνολικός αριθμός των εγγράφων / Αριθμός των εγγράφων που εμφανίζεται ο όρος t.)

Το TF-IDF εστιάζει σε λέξεις που είναι συχνές σε ένα κείμενο αλλά σπάνιες στα υπόλοιπα του σώματος κειμένου. Για να υπολογίσουμε το TF-IDF score συνδυάζουμε τους δύο προηγούμενους τύπους όπως φαίνεται στον τύπο 2.1

$$TF - IDF_{t,d} = TF_{t,d} \cdot IDFt \tag{2.1}$$

2.1.2 Dov2vec

Οι ενσωματώσεις εγγράφων είναι έγγραφα (αποσπάσματα κειμένου) που αντιπροσωπεύονται ως διανύσματα πραγματικών αριθμών, ενώ η απόσταση δύο διανυσμάτων εγγράφων υποδηλώνει σημασιολογική ομοιότητα. Lee και ο Mikolov [5] προτείνουν τα διανύσματα παραγράφων που μπορούν να δημιουργήσουν διανυσματικές αναπαραστάσεις από κείμενο οποιουδήποτε μήκους, όπως προτάσεις ή ολόκληρα έγγραφα. Ο μη εποπτευόμενος αλγόριθμος, που ονομάζεται Doc2Vec, είναι μια επέκταση του Word2Vec και οι αλγεβρικές λειτουργίες μπορούν επίσης να καταλήξουν σε σημασιολογικά ουσιαστικά αποτελέσματα [6].

Όπως και το word2vec το Doc2vec έχει δύο διαφορετικά μοντέλα, το κατανεμημένο μοντέλο μνήμης (PV-DM) και το κατανεμημένο σύνολο λέξεων (DBOW).

Το κατανεμημένο μοντέλο μνήμης (PV-DM) είναι παρόμοιο με το CBOW με τη διαφορά ότι ένα πρόσθετο διακριτικό παραγράφου (P-ID) προστίθεται στο περιεχόμενο. Αυτό το διακριτικό λειτουργεί ως μνήμη ή θέμα του τρέχοντος περιεχομένου. Μαζί προβλέπουν την επόμενη λέξη. Σημειώνεται ότι αυτό το μοντέλο λαμβάνει υπόψη τη σειρά λέξεων. Το κατανεμημένο σύνολο λέξεων (DBOW) είναι παρόμοιο με το Skip-gram με τη διαφορά ότι ένα διακριτικό παραγράφου αντί για μια λέξη προβλέπει το

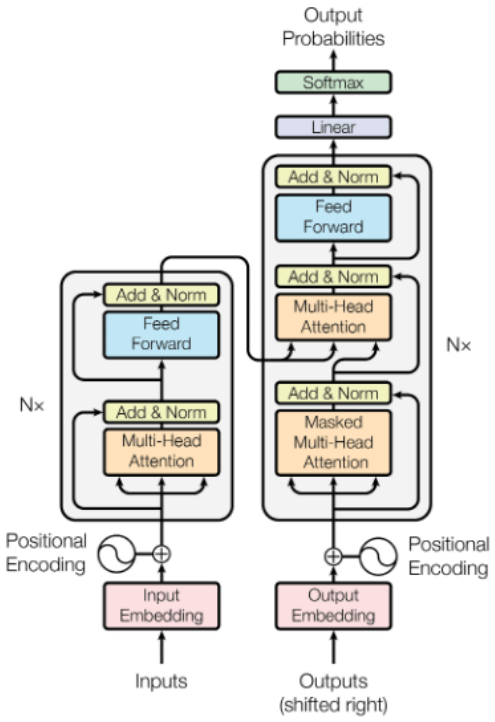
περιεχόμενο του. Η σειρά των λέξεων δεν λαμβάνεται υπόψη.

Παρόλο που οι Le and Mikolov [5] δηλώνουν μια καλύτερη απόδοση για PV-DM, το σύστημά μας εφαρμόζει το PV-DBOW καθώς οι Lau and Baldwin [7], δείχνουν αντίθετα αποτελέσματα.

2.1.3 BERT

Το BERT είναι ένα μοντέλο βαθιάς μάθησης που χρησιμοποιείται σε αρκετές εργασίες για επεξεργασία φυσικής γλώσσα και εισήχθη από τον Devlin[8]. Το μοντέλο αυτό έχει προ εκπαιδευτεί σε μεγάλα σώματα κειμένου, αλλά χρειάζεται fine tunning για κάθε πρόβλημα που χρησιμοποιείται. Το BERT βασίζεται στην αρχιτεκτονική Transformer [9]. Αυτή η αρχιτεκτονική βασίζεται σε μεγάλο βαθμό σε έναν μηχανισμό που αποκαλείται attention, και πιο συγκεκριμένα, self-attention και multi-head attention. Σε υψηλό επίπεδο, το self-attention καθορίζει το αντίκτυπο που έχει κάθε λέξη στην υπόλοιπη πρόταση, ενώ το multi-head attention υπολογίζει το attention σε διαφορετικούς χρόνους, με διαφορετικούς πίνακες βάρους και στη συνέχεια συνδυάζει τα αποτελέσματα μαζί. Το multi-head attention διευρύνει την ικανότητα του μοντέλου να επικεντρώνεται σε διαφορετικές θέσεις και δίνει επίσης στο attention layer πολλαπλά υποδιαστήματα αναπαράστασης [10]. Η αρχιτεκτονική του BERT είναι multi layer bidirectional Transformer βασισμένο στην υλοποίηση που περιγράφεται την εργασία [9].

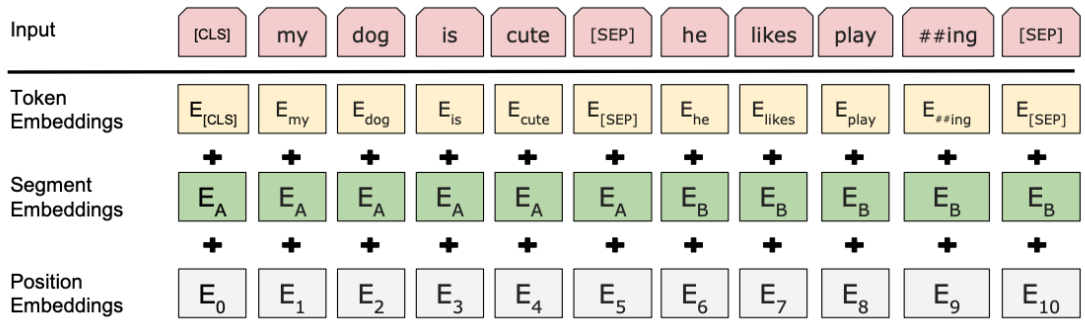
Στο στάδιο της προ-εκπαίδευσης, το BERT εκπαιδεύεται σε δύο μη εποπτευόμενες εργασίες ταυτόχρονα, δηλαδή το Μοντέλο Κρυμμένης Γλώσσας και την Πρόβλεψη Επόμενης Πρότασης. Στο πρώτο μοντέλο, το BERT εμφανίζεται με μια ακολουθία λέξεων, όπου μερικές από αυτές τυχαία καλύπτονται ως διακριτικά και πρέπει να προβλέψει τη σωστή λέξη πίσω από τα καλυμμένα διακριτικά. Αυτή η εργασία διαφέρει από τα συνηθισμένα autoregressive, όπου ζητείται από το μοντέλο να προβλέψει την επόμενη λέξη σε μια ακολουθία. Συγκεκριμένα, αυτό επιτρέπει στο μοντέλο να εξερευνήσει το περιεχόμενο με αμφίδρομο τρόπο. Το δεύτερο μοντέλο είναι μια εργασία δυαδικής ταξινόμησης στην οποία, δεδομένου ενός ζεύγους προτάσεων, το μοντέλο καλείται να προβλέψει εάν η δεύτερη πρόταση είναι η πραγματική επόμενη πρόταση της πρώτης



Σχήμα 2.2: Αρχιτεκτονική Transformer

πρότασης.

Το tokenization που χρησιμοποιείται από τον BERT ονομάζεται WordPiece tokenization [11]. Αυτό σημαίνει ότι το λεξιλόγιο αρχικοποιείται με όλους τους μεμονωμένους χαρακτήρες στη γλώσσα και, στη συνέχεια, οι πιο συχνές συνδυασμοί των υπαρχόντων χαρακτήρων στο λεξιλόγιο προστίθενται επαναληπτικά. Αυτή η τεχνική επιτρέπει στο μοντέλο να χειρίζεται τις λέξεις εκτός λεξιλογίου, χωρίζοντάς τις σε γνωστές υπό λέξεις.



Σχήμα 2.3: Είσοδος BERT για ζεύγη προτάσεων

Η αναπαράσταση εισαγωγής του BERT είναι καθολική, τόσο για εργασίες πρόβλε-

ψης μεμονωμένων προτάσεων όσο και για ζεύγη προτάσεων. Το πρώτο διακριτικό κάθε ακολουθίας εισόδου είναι το ειδικό διακριτικό ταξινόμησης, που ονομάζεται [CLS]. Αυτό το διακριτικό χρησιμοποιείται σε εργασίες ταξινόμησης και συνοψίζει ολόκληρη την αναπαράσταση ακολουθίας. Στο τέλος κάθε φράσης, είτε για εργασίες μεμονωμένων ή ζευγών, το διακριτικό [SEP] ακολουθεί. Η τελική αναπαράσταση της εισόδου για ζεύγη προτάσεων φαίνεται στο σχήμα 2.3.

Όπως βλέπουμε και στο σχήμα, εφαρμόζονται δύο ακόμη δείκτες που απαιτούνται εκ νέου από το μοντέλο. Πρώτα, μια πρόταση ενσωμάτωσης segment embeddings προστίθεται σε κάθε διακριτικό και δείχνει αν είναι στην πρόταση A ή B. Τέλος, προστίθεται επίσης η λεγόμενη τοποθέτηση θέσης position embendings σε κάθε σύμβολο για να δείξει τη θέση της στην ακολουθία.

Το BERT χρησιμοποιείται σε πολλές εφαρμογές NLP και ο κύριος περιορισμός του είναι ότι δεν μπορεί να χειριστεί συμβολοσειρές άνω των 512 λέξεων.

2.2 Βιβλιογραφική επισκόπηση

Το δίκαιο (common law), επίσης γνωστός ως δικαστικό προηγούμενο ή νομολογία, είναι το σώμα του δικαίου που προέρχεται από δικαστικές αποφάσεις και όχι από καταστατικά ή συντάγματα [12]. Αυτό επιτρέπει στο δικαστήριο, εάν υπάρχει διαφωνία σχετικά με τον ακριβή ορισμό του νόμου, να εξετάζει τις προγενέστερες αποφάσεις των σχετικών δικαστηρίων,και να σημειώσουμε τις αρχές που ισχύουν για την παρούσα υπόθεση. Εάν η διαφορά είναι παρόμοια, το δικαστήριο είναι υποχρεωμένο να ακολουθήσει την προηγούμενη απόφαση, γνωστή ως κοίταξε την κρίση. Ωστόσο, εάν το δικαστήριο διαπιστώσει μια υπόθεση διαφορετική από τις παρούσες περιστάσεις και οι νομοθετικοί νόμοι δεν υπάρχουν ή είναι διφορούμενοι στη διατύπωση τους, οι δικαστές έχουν την δυνατότητα να επιλύσουν το ζήτημα. Αυτό θέτει ένα προηγούμενο, ισοδύναμο με αυτό που αναφέρθηκε παραπάνω, που επηρεάζει τις αποφάσεις των μελλοντικών δικαστών σε παρόμοιες υποθέσεις.

Το δίκαιο εμφανίζεται σε πολλές χώρες, οι περισσότερες από τις οποίες βασίζονται στο αγγλικό δίκαιο. Σε αυτήν την εργασία η συζήτηση αφορά δύο από αυτές, το κοινό δίκαιο των Ηνωμένων Πολιτειών, καθώς και το κοινό δίκαιο της Αυστραλίας,

ανεξάρτητα από τα εμπλεκόμενα δικαστήρια. Οι ίδιες οι υποθέσεις, ως κείμενο, ήταν το κύριο επίκεντρο.

Ένα σύστημα αυτόματης ανάκτησης προτεραιότητας από νομικά έγγραφα βοηθά τους νομικούς επαγγελματίες να αναφέρονται εύκολα στις υποθέσεις που σχετίζονται με την τρέχουσα υπόθεση. Ένα τέτοιο σύστημα προκαταρκτικής ανάκτησης έχει πολλές εφαρμογές, όπως εκτίμηση βάσει υποθέσεων [13][14], νομικές παραπομπές και ανάκτηση νομικών πληροφοριών [15]. Αρκετές προσεγγίσεις, όπως η εξαγωγή πληροφοριών με βάση την επεξεργασία φυσικής γλώσσας [16], βάσει κανόνα Προσέγγιση [17] και τεχνικές μηχανικής μάθησης [18], χρησιμοποιούνται για την ανάκτηση των προηγούμενων περιπτώσεων με εκ νέου εμφάνιση στην τρέχουσα υπόθεση.

Όταν σκοπεύουμε να αυτοματοποιήσουμε την κατανόηση ή, σε αυτήν την περίπτωση, σύγκριση κειμένου, μπαίνουμε σε ένα πεδίο γνωστό ως Επεξεργασία Φυσικής Γλώσσας (NLP). Το NLP είναι μια μαθηματική-στατιστική προσέγγιση για την ανάλυση κειμένου και την αναπαράσταση φυσικού κειμένου σε ένα ή περισσότερα επίπεδα γλωσσικής ανάλυσης, με σκοπό την επίτευξη επεξεργασίας γλωσσών που μοιάζει με ανθρώπους για μια σειρά από εργασίες ή εφαρμογές [19].

Η πλειονότητα των υπαρχουσών μεθόδων αυτόματης ευρετηρίασης επιλέγουν όρους φυσικής γλώσσας από τα κείμενα εγγράφων [20][21]. Το πιο ουσιαστικό μέρος είναι η ανάλυση του κειμένου σε λέξεις, όπου το κείμενο αναλύεται σε μεμονωμένες λέξεις, μετά από αυτό, οι επιλογές είναι διαφορετικές αλλά προαιρετικές. Μια πολύ ευεργετική μέθοδος είναι η αφαίρεση λέξεων-κλειδιών, λέξεων που χρησιμοποιούνται συνήθως αλλά δεν συμβάλλουν τίποτα στην ουσία του κειμένου, όπως άρθρα, προθέσεις και σύνδεσμοι. Μια άλλη μέθοδος είναι το stemming, που μειώνει τις λέξεις στη ριζική τους μορφή, αγνοώντας τις καταλήξεις οι οποίες είναι περιττές για την ερμηνεία της λέξης. Επιπλέον, χρήσιμη είναι η χρήση φράσεων ως όρων ευρετηρίου, καθώς και η αντικατάσταση όρων ευρετηρίου, με τον ομοιόμορφο όρο τους. Τελευταίο, αλλά εξίσου σημαντικό, είναι η χρήση του βάρους κάθε όρου δείκτη [21].

Ο όρος Frequency-Inverse Document Frequency (TF-IDF), είναι η πιο δημοφιλής προσέγγιση στάθμισης [22]). Ουσιαστικά ο αλγόριθμος λειτουργεί καθορίζοντας τη σχετική συχνότητα των όρων σε ένα συγκεκριμένο έγγραφο σε σύγκριση με την αντίστροφη αναλογία αυτής της λέξης σε ολόκληρο το σώμα. Αυτό είναι σημαντικό,

καθώς οι λέξεις που είναι κοινές σε μια μεμονωμένη ή μικρή ομάδα εγγράφων τείνουν να έχουν υψηλότερους αριθμούς TF-IDF από τις συνηθισμένες λέξεις όπως άρθρα και προθέσεις [23]). Επίσης, επηρεάζει κοινούς όρους που εμφανίζονται συχνά σε ένα σώμα, όπως συμβαίνει σε αυτό το πείραμα, σε νομικά έγγραφα.

Στην εργασία τους οι Kumar et al. [24] πρότεινε δύο μετρικές ομοιότητας βασισμένες στο κείμενο,την ομοιότητα συνημιτόνου όλων των όρων και την ομοιότητα συνημιτόνου των νομικών όρων. Και οι δύο αυτές μετρικές είναι βασισμένες στις βαθμολογίες TF-IDF των όρων που υπάρχουν στα νομικά έγγραφα. Τα έγγραφα αναπαριστώνται ως διανύσματα όπου το μέγεθος του διανύσματος είναι ο αριθμός των διακριτών λέξεων του λεξιλογίου σε όλη τη συλλογή των εγγράφων. Έτσι κάθε στοιχείο του διανύσματος αντιστοιχεί σε έναν όρο του λεξιλογίου και περιέχει τη βαθμολογία TF-IDF του αντίστοιχου όρου. Μόλις σχηματιστούν τα διανύσματα και των δύο εγγράφων, η ομοιότητα των συνημίτονων υπολογίζεται μεταξύ των δύο διανυσμάτων για να ληφθεί η ομοιότητα μεταξύ των εγγράφων.

Κεφάλαιο 3

Προτεινόμενη προσέγγιση και μεθοδολογία

3.1 Προτεινόμενη προσέγγιση

Για αυτήν την εργασία, έχει εφαρμοστεί μια προσέγγιση σημασιολογικής ομοιότητας εγγράφων για τη σημασιολογική ανάκτηση νομικών κειμένων. Έχουν υλοποιηθεί τρεις παραλλαγές, κάθε μία με διαφορετική τεχνική αναπαράστασής του κειμένου, χρησιμοποιώντας τόσο εργαλεία επεξεργασίας φυσικής γλώσσας, όσο και μοντέλα νευρωνικών δικτύων-βαθιάς μάθησης. Πιο συγκεκριμένα χρησιμοποιήθηκαν οι εξής τεχνικές Bag of words με TF-IDF, Word2Vec-Doc2vec και BERT. Τα βήματα που ακολουθήθηκαν στην προσέγγιση μας δίνονται παρακάτω.

- Συλλογή σωμάτων κειμένου(datasets)
- Προ-επεξεργασία νομικών κειμένων
- Αναπαράσταση των νομικών κειμένων σε διανυσματική μορφή μέσω των μεθόδων Bag of words με TF-IDF, word2vec-Dov2vec, BERT
- Για όλα τα δυνατά μοναδικά ζεύγη ανά δύο των κειμένων του συνόλου δεδομένων, υπολογίστηκε η ομοιότητα συνημιτόνου(cosine similarity score)
- Ανθρώπινη αξιολόγηση των αποτελεσμάτων μέσω επί σημείωσης(annotation) των κειμένων από εθελοντές, για να φτάσουμε στο συμπέρασμα κατά πόσο τελι-

κά οι αυτόματες μετρικές ομοιότητας στις επιλεγμένες αναπαραστάσεις μπορούν να οδηγήσουν, όντως, στην ανακάλυψη ομοίων ζευγαριών νομικών κειμένων.

Τα βήματα που χρησιμοποιούνται και στις τρεις μεθόδους εξηγούνται λεπτομερώς στη συνέχεια.

3.2 Μεθοδολογία

Όπως αναφερθήκαμε και νωρίτερα, σε αυτήν την εργασία, εξετάζουμε μεθοδολογίες και αναπαραστάσεις βασισμένες σε κείμενο(text-based) για τον υπολογισμό της ομοιότητας δύο νομικών κειμένων.
Η μέτρηση της ομοιότητας μεταξύ δύο νομικών εγγράφων περιλαμβάνει δύο υπόεργασίες:

- Εύρεση κατάλληλης αναπαράστασης ενός εγγράφου:
  Δηλαδή, του ποια κομμάτια του εγγράφου θα χρησιμοποιηθούν για την διανυσματική αναπαράσταση του νομικού κειμένου.Η πιο απλή προσέγγιση, είναι να χρησιμοποιηθεί όλο το κείμενο στην τεχνική αναπαράστασης. Ωστόσο, ένα νομικό κείμενο είναι πιθανό να συζητάει διαφορετικά νομικά ζητήματα. Ως εκ τούτου, μια εναλλακτική μέθοδος θα ήταν να χρησιμοποιηθούν υποσύνολα του εγγράφουέτσι ώστε κάθε υποσύνολο να καταγράφει ένα μόνο νομικό ζήτημα- στην τεχνική αναπαράστασης και στη συνέχεια να υπολογιστεί η ομοιότητα μεταξύ των υποσυνόλων των δύο εγγράφων. Εφόσον αποφασιστεί ποια κομμάτια του κειμένου θα συμμετέχουν στην αναπαράσταση, μέσω των τεχνικών που θα αναλύσουμε, παρακάτω παίρνουμε τη διανυσματική αναπαράσταση των εγγράφων.

- Μέτρηση της ομοιότητας μεταξύ των αναπαραστάσεων δύο δοθέντων εγγράφων:
  Μόλις επιλεγούν οι αναπαραστάσεις των δύο εγγράφων, πρέπει να χρησιμοποιηθεί μια κατάλληλη μαθηματική μέθοδος η οποία να είναι συγκρίσιμη, ώστε να υπολογιστεί η ομοιότητα μεταξύ των αναπαραστάσεων και ο υπολογιστής να μπορεί να μας πει ποια έγγραφα είναι παρόμοια και ποια άσχετα μεταξύ τους. Επίσης, οι αναπαραστάσεις των κειμένων πρέπει να είναι σε ποσοτικοποιημένη μορφή, ή ένα μαθηματικό αντικείμενο, το οποίο είναι μια διανυσματική μορφή, έτσι ώστε να κάνουμε τους υπολογισμούς ομοιότητας πάνω σε αυτό.

Σε αυτήν την ενότητα, περιγράφουμε διάφορες μεθοδολογίες για τις δύο υπό εργασίες που αναφέρονται παραπάνω. Διαφορετικοί τρόποι επιλογής της αναπαράστασης ενός νομικού εγγράφου περιγράφονται στην υπό ενότητα 4.2.1, ενώ η Ενότητα 4.2.2 συζητά διάφορους τρόπους μέτρησης της ομοιότητας μεταξύ των αναπαραστάσεων.

3.2.1 Επιλογή αναπαραστάσεων του νομικού κειμένου

Ένα νομικό κείμενο ή μία νομική υπόθεση ενδέχεται να συζητά πολλά νομικά θέματα και όχι μόνο ένα, καθώς κάποιος κατηγορούμενος μπορεί να κατηγορείται για πολλά αδικήματα και όχι μόνο με ένα έτσι κάποιο από όλα αυτά θα έχει νομικά όμοια υπόθεση.

Έτσι, αποφασίσαμε να χρησιμοποιήσουμε ολόκληρο το κείμενο στην αναπαράσταση. Και αυτό γιατί θα μπορέσουμε να πιάσουμε ενδεχομένως όλες τις υποθέσεις που συζητιούνται σε μία υπόθεση. Σε μελλοντική εργασία βέβαια, θα πρέπει να δοκιμαστούν και να χρησιμοποιηθούν άλλα μέρη στην αναπαράσταση.

Εξαίρεση αποτελεί όπως θα δούμε και παρακάτω, αποτελεί η αναπαράσταση BERT για το λόγο του ότι δεν μπορεί να χειριστεί συμβολοσειρές άνω του μεγέθους των 512, οπότε αποφασίσαμε να χρησιμοποιήσουμε σε εκείνη την τεχνική μόνο τους 512 πρώτους χαρακτήρες, θεωρώντας κιόλας ότι είναι αρκετοί αυτοί οι χαρακτήρες για να κρατήσουν τι πραγματεύεται η κάθε υπόθεση. Αυτό βέβαια θα φανεί στα αποτελέσματα του κατά πόσο ήταν αποδοτικό αυτό ή αν σε μελλοντική εργασία θα χρειαστεί να γίνει ιεραρχική αναπαράσταση με BERT ώστε να χρησιμοποιηθεί ολόκληρο το νομικό κείμενο στην αναπαράσταση.

3.2.2 Μέτρηση ομοιότητας μεταξύ των αναπαραστάσεων

Μόλις επιλέξουμε τα σημεία του εγγράφου που θα χρησιμοποιηθούν στην αναπαράσταση του νομικού κειμένου, τότε πρέπει να μετρήσουμε την ομοιότητα μεταξύ των αυτών των αναπαραστάσεων. Το πρόβλημα της μέτρησης της ομοιότητας μεταξύ δύο εγγράφων κειμένου έχει λάβει μεγάλη προσοχή από τις ερευνητικές κοινότητες Ανάκτησης και Ανάκτησης Δεδομένων, και υπάρχουν πολλές γνωστές τεχνικές στη βιβλιογραφία.[25] Οι πιο δημοφιλείς τεχνικές σχηματίζουν μια διανυσματική αναπαράσταση των δύο εγγράφων, όπου οι διαστάσεις της αναπαράστασης του διανύσματος πχ θα μπορούσαν να είναι οι όροι που περιέχονται στα έγγραφα.

Η ομοιότητα συνημιτόνου (cosine similarity) μπορεί να χρησιμοποιηθεί για να μετρηθεί η ομοιότητα δύο διανυσμάτων. Δοθέντων 2 διανυσμάτων A και B διαστάσεως n το καθένα η ομοιότητα συνημιτόνου (cosine similarity) δίνεται από τον τύπο που φαίνεται στο σχήμα 3.1.

Σχήμα 3.1: Ομοιότητα συνημιτόνου (Cosine Similarity)

$$\frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Η Cosine Similarity είναι ένα μέτρο ομοιότητας που χρησιμοποιείται για να υπολογίζει πόσο όμοια είναι τα έγγραφα ανεξάρτητα από το μέγεθός τους. Μαθηματικά, μετρά το συνημίτονο της γωνίας μεταξύ δύο διανυσμάτων που προβάλλονται σε έναν πολυδιάστατο χώρο.

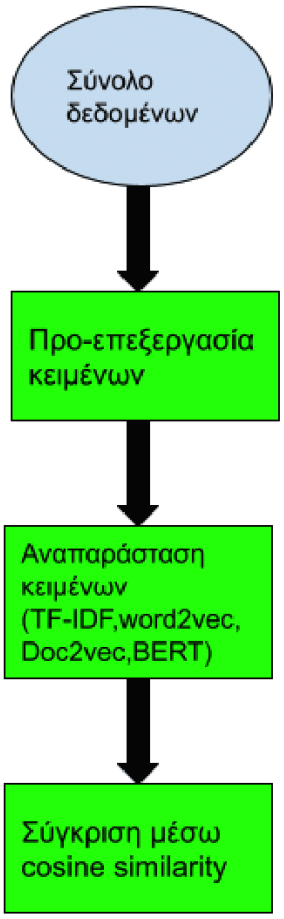
Χρησιμοποιήθηκε αυτή η προσέγγιση για τη μέτρηση της ομοιότητας και όχι κάποια άλλη, όπως για παράδειγμα η ευκλείδεια απόσταση και αυτό γιατί έχει το πλεονέκτημα ότι μπορούμε να μετρήσουμε πόσο παρόμοια είναι τα έγγραφα ανεξάρτητα από το μέγεθός τους.

Όταν σχεδιάζεται σε έναν πολυδιάστατο χώρο, όπου κάθε διάσταση αντιστοιχεί σε μια λέξη στο έγγραφο, η ομοιότητα συνημίτονου καταγράφει τον προσανατολισμό (τη γωνία) των εγγράφων και όχι το μέγεθος. Αν είχε σημασία το μέγεθος, θα υπολογίζαμε την ευκλείδεια απόσταση.

Η ομοιότητα συνημιτόνου Cosine Similarity είναι χρήσιμη επειδή ακόμα και αν τα δύο παρόμοια έγγραφα είναι πολύ μακριά με βάση την ευκλείδεια απόσταση λόγω του μεγέθους, θα μπορούσαν να έχουν ακόμα μικρότερη γωνία μεταξύ τους. Μικρότερη γωνία σημαίνει μεγαλύτερη ομοιότητα.

Μια εικόνα για τη διαδικασία της μεθοδολογίας δίνεται στο σχήμα 3.2.

Σχήμα 3.2: Διαδικασία μεθοδολογίας



3.2.2.1 Προ-επεξεργασία κειμένων

Προτού αναπαραστήσουμε το κάθε κείμενο σε διανυσματική μορφή, εφαρμόζουμε κάποιες βασικές ενέργειες που αφορούν την προ-επεξεργασία, ώστε να έχουμε μία κανονικοποιημένη αναπαράσταση των εγγράφων.

Αυτές οι ενέργειες είναι οι εξής:

1. Χωρισμός του κειμένου σε λέξεις (Tokenization). Χωρίσαμε τα στοιχεία του κειμένου που ακολουθούνται από κενό. Αυτό μας βοηθάει και στο να προχωρήσουμε στις ενέργειες που περιγράφουμε παρακάτω, αλλά και στο να μπορέσει να γίνει η ενσωμάτωση (embeddings) των λέξεων στο μοντέλο (doc2vec) που θα χρησιμοποιήσουμε. Το μοντέλο doc2vec δέχεται λέξεις και όχι προτάσεις οπότε μας είναι χρήσιμο να χωρίσουμε το κείμενο σε λέξεις.

2. Μετατροπή σε πεζά γράμματα. Μετατρέπουμε ολόκληρο το σώμα κειμένων σε πεζά, αφού στη μέτρηση της ομοιότητας αυτό δεν παίζει ρόλο. Π.χ. "Legal" και "legal", δεν έχει καμία απολύτως διαφορά, θεωρείται η ίδια ακριβώς λέξη, οπότε με αυτήν την ενέργεια στη μηχανή να καταλάβει ότι πρόκειται για την ίδια ακριβώς λέξη.

3. Αφαίρεση stopwords, σημείων στίξης και αριθμητικών χαρακτήρων. Τα Stopwords είναι οι λέξεις που εμφανίζονται πολυ συχνά σε μία γλώσσα, στην περίπτωση μας στα αγγλικά. Οι λέξεις αυτές δεν προσθέτουν πολύ νόημα σε μια πρόταση. Εκτός του ότι μπορεί να εμφανίζονται συχνά σε ένα κείμενο, δεν θα θέλαμε αυτές τις λέξεις να συμμετέχουν στις αναπαραστάσεις ή να καταναλώνουν χρόνο στην επεξεργασία. Έτσι, μπορούν ασφαλώς να αγνοηθούν χωρίς να θυσιάζεται η έννοια μιας πρότασης. Χρησιμοποιήσαμε τη λίστα για stopwords που δίνει το εργαλείο nltk [1]. Για τον τομέα που μελετάμε, δηλαδή το νομικο, οι αριθμοί και τα σημεία στίξης μπορούν επίσης να αγνοηθούν χωρίς να αλλοιώνεται το νόημα της νομικής πρότασης.

4. Εφαρμογή stemming porter σε όλες τις λέξεις. Το Stemming είναι η διαδικασία που μετατρέπει μια λέξη στο στέλεχος της αφαιρώντας το επίθεμα της. Στόχος

[1] http://www.nltk.org/

του stemming είναι να μειώσει τις διαστάσεις, προβάλλοντας ουσιαστικά όλες τις λέξεις που έχουν κοινή ρίζα σε μία διάσταση του τελικού διανυσματικού χώρου.

Για τις αναπαραστάσεις BERT και Doc2vec, στο κομμάτι της προ επεξεργασίας ακολουθήθηκαν οι ενέργειες 1-3 ενώ στο TF-IDF οι 1-4.

3.2.3 Διανυσματικές αναπαράστεις

3.2.3.1 Bag of words με TF-IDF

Αυτή είναι μια απλή, αλλά αποτελεσματική προσέγγιση για τη μετατροπή ενός κειμένου σε ένα μόνο διάνυσμα. Κάθε ξεχωριστός όρος-λέξη από το λεξιλόγιο ολόκληρου του σώματος (συλλογή κειμένων) θεωρείται ως διάσταση για τον διανυσματικό χώρο. Ένα έγγραφο d αναπαρίσταται ως διάνυσμα μήκους ίσο με τον αριθμό των μοναδικών λέξεων στο λεξιλόγιο της συλλογής κειμένων.[2] Το ν-ιοστό στοιχείο του διανύσματος αντιστοιχεί στην ν-ιοστή λέξη στο λεξιλόγιο και περιέχει το TF-IDF score της αντίστοιχης λέξης, όπου το TF (όρος συχνότητα) μετρά τον αριθμό των φορών που υπάρχει η λέξη στο έγγραφο d και το IDF μετρά το αντίστροφο της συχνότητας της λέξης σε ολόκληρη τη συλλογή κειμένων. Στην τεχνική αυτή και μόνο εφαρμόστηκε όπως είπαμε πριν το stemming με σκοπό να μειωθούν οι διαστάσεις του τελικού διανύσματος.

Για τον υπολογισμό της ομοιότητας μεταξύ ενός ζεύγους κειμένων, υπολογίζουμε απλώς την ομοιότητα συνημιτόνου (σχήμα 3.1) μεταξύ των δύο διανυσμάτων TF-IDF που αντιστοιχούν στα δύο κείμενα.

3.2.3.2 Doc2vec

Αυτή η τεχνική ενσωμάτωσης αποτελεί επέκταση του μοντέλου που χρησιμοποιείται στο Word2vec. Όπως υποδηλώνει το όνομα, αυτό το μοντέλο μετατρέπει άμεσα ένα δεδομένο κείμενο σε ένα διάνυσμα. Χρησιμοποιήσαμε ολόκληρη τη συλλογή δεδομένων

[2] `https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html`

μας για να εκπαιδεύσουμε το Doc2vec. [3]

Καθώς το Doc2vec απαιτεί να παρέχουμε το μέγεθος της ενσωμάτωσης ως παράμετρο εισαγωγής, δοκιμάσαμε την απόδοσή του με πολλά μεγέθη και βρήκαμε ότι τα μοντέλα Doc2vec με μέγεθος 100 έχουν καλύτερη απόδοση στην βαθμολογία της ομοιότητας συνημιτόνου από τα υπόλοιπα. Εκπαιδεύτηκε για 10 εποχές εκπαίδευσης. Έτσι, όλα τα αποτελέσματα για το Doc2vec που αναφέρονται σε αυτό το έγγραφο θεωρούν ένα μοντέλο με μέγεθος 100.

Ο υπολογισμός της ομοιότητας είναι απλός εφόσον έχουμε απευθείας τις διανυσματικές αναπαραστάσεις Doc2vec. Για ένα ζεύγος εγγράφων, υπολογίζουμε απλά την ομοιότητα του συνημίτονου μεταξύ αυτών των δύο διανυσμάτων.

3.2.3.3 BERT

Με τον περιορισμό που έχει το BERT που δεν μπορεί να χειριστεί συμβολοσειρές με πάνω από 512 λέξεις, χρησιμοποιήσαμε στην αναπαράσταση του νομικού κειμένου τις πρώτες 512 λέξεις κάθε εγγράφου. Χρησιμοποιήσαμε προ εκπαιδευμένο μοντέλο BERT [4] για να ενσωματώσουμε τη συλλογή κειμένων μας. Φορτώσαμε το βασικό μοντέλο BERT, το οποίο έχει 12 επίπεδα, 12 attention heads, 110 εκατομμύρια παραμέτρους και κρυφό επίπεδο μεγέθους των 768. Το συγκεκριμένο μοντέλο που χρησιμοποιήσαμε δίνει απευθείας την ενσωμάτωση (embeddings) μια πρότασης σε διάνυσμα. Έτσι θεωρήσαμε το κάθε νομικό κείμενο ως μία πρόταση των 512 λέξεων και πήραμε την αναπαράσταση του σε διανυσματικό χώρο μέσω του μοντέλου BERT που χρησιμοποιήσαμε.

Έτσι, αφού και εδώ έχουμε την αναπαράσταση του κειμένου σε διανυσματική μορφή υπολογίζουμε την ομοιότητα συνημιτόνου μεταξύ των δύο αυτών διανυσμάτων για να δούμε πόσο όμοια είναι τα δύο νομικά έγγραφα.

[3] Για την υλοποίηση του Doc2vec χρησιμοποιήσαμε `https://radimrehurek.com/gensim/models/doc2vec.html`

[4] `https://huggingface.co/sentence-transformers/bert-base-nli-mean-tokens`

Κεφάλαιο 4

Πειραματικό μέρος και μέθοδος αξιολόγησης

4.1 Πειραματικό μέρος

Το ίδιο το πείραμα είχε πολλά επίπεδα, είτε κατά την προετοιμασία και την πρόβλεψη μελλοντικών αναγκών, είτε απλά στην ανάλυση χρήσιμων βιβλιοθηκών, εργαλείων και μεθόδων για τα αναμενόμενα αποτελέσματα. Ο στόχος στην αρχή του πειράματος ήταν η ερώτηση, ¨Πώς να βοηθήσετε ένα άτομο να αναζητήσει νομικά έγγραφα με σημαντική συσχέτιση με μια παρεχόμενη υπόθεση¨.

Καθώς αυτό ήταν ένα φιλόδοξο εγχείρημα, ο πρώτος στόχος ήταν να βρεθούν ζευγάρια για ένα σύνολο νομικών εγγράφων που φαινόταν να σχετίζονται μεταξύ τους, χρησιμοποιώντας εργαλεία εργαλείων Τεχνητής Νοημοσύνης και Επεξεργασίας Φυσικής Γλώσσας. Αυτό ήταν το επίκεντρο αυτής της εργασίας, καθώς και το να αξιολογήσουμε του κατά πόσο οι τεχνικές που χρησιμοποιήσαμε, οδηγούν όντως, στην ανακάλυψη όμοιων ζευγαριών νομικών υποθέσεων.

Επίσης, μέσω στατιστικής ανάλυσης της ευαισθησίας της ομοιότητας με βάση την αναπαράσταση, μελετήθηκε του κατά πόσο διαφορετικές αναπαραστάσεις οδηγούν σε διαφορετικά κριτήρια ομοιότητας κειμένων, που μπορούν να επηρεάσουν την ανάκληση νομικών κειμένων.

Για την αξιολόγηση, συγκρίνουμε τις βαθμολογίες που υπολογίσαμε μέσω της ο-

μοιότητας συνημιτόνου(cosine similarity), με εκείνες που παρέχονται από ανθρώπινη αξιολόγηση, ώστε να συμπεράνουμε αν υπάρχει συνοχή ανάμεσα στα αποτελέσματα των αυτόματων μετρικών ομοιότητας και σε αυτά της ανθρώπινης οπτικής. Η αξιολόγηση αποτελεί τη βασική συνεισφορά αυτής της εργασίας και θα αναλυθεί λεπτομερώς, σε αυτό το κεφάλαιο.

Τέλος, καταλήγουμε στο συμπέρασμα ποια τεχνική αναπαράστασης είναι καλύτερη, με βάση του ποια τεχνική είναι πιο ευθυγραμμισμένη με την ανθρώπινη αντίληψη σημασιολογικής εγγύτητας.

4.1.1 Σύνολα Δεδομένων

Το Australian Case Law dataset, από το αποθετήριο μηχανικής μάθησης (UCI Machine Learning Repository) [1], ήταν το πρώτο σύνολο δεδομένων που χρησιμοποιήθηκε σε αυτό το πείραμα. Το συγκεκριμένο σύνολο δεδομένων περιέχει 3890 νομικές υποθέσεις από τα έτη 2006-2009, οι οποίες χρησιμοποιήθηκαν όλες για αυτό το πείραμα. Επίσης, το σύνολο δεδομένων χωρίζεται σε τρεις φακέλους, με όλα του τα αρχεία να είναι σε μορφή XML. Ο πρώτος φάκελος αφορά παραπομπές για κάθε υπόθεση του συνόλου δεδομένων σε άλλες νομικές υποθέσεις, ο δεύτερος περιέχει τις περιλήψεις των νομικών υποθέσεων και ο τρίτος περιλαμβάνει ολόκληρα τα κείμενα των νομικών υποθέσεων.

Τα αρχεία που βρίσκονται στον τρίτο φάκελο, ο οποίος περιέχει ολόκληρα τα νομικά κείμενα, έχουν ετικέτες οι οποίες αφορούν:

- Το όνομα της υπόθεσης
- Τη σελίδα AUSTLII (`http://www.austlii.edu.au`)
- Φράσεις (catchphrases) οι οποίες δίνουν μια μικρή περίληψη της υπόθεσης.
- Προτάσεις οι οποίες αποτελούν το πλήρες κείμενο της υπόθεσης διαχωρισμένο με τελείες. Ουσιαστικά, οι προτάσεις αυτές, αφορούν, ότι συζητήθηκε στο δικαστήριο για τη συγκεκριμένη νομική υπόθεση.

[1] `https://archive.ics.uci.edu/ml/datasets/Legal+Case+Reports`

Το επίκεντρο του προγράμματος ήταν ο τελευταίος φάκελος και είναι το σύνολο δεδομένων που περιγράφεται στις ακόλουθες εξηγήσεις οι οποίες αφορούν την προ επεξεργασία των στοιχείων του, ώστε να περιέχουν όλες οι εγγραφές καθαρό κείμενο (raw text) και να μπορέσουμε να εφαρμόσουμε τις μεθόδους και τις τεχνικές αναπαράστασής των κειμένων που αναλύσαμε στο προηγούμενο κεφάλαιο.

Το ίδιο το σύνολο δεδομένων είχε κάποια προβλήματα που απαιτούσαν επίλυση στη φάση προ επεξεργασίας. Το πρώτο, και ίσως το πιο σημαντικό, ήταν η ακανόνιστη προσθήκη ετικετών στη μορφή XML, ειδικά στην ετικέτα id, λόγω της χρήσης του "id= 0" (0 χρησιμοποιήθηκε ως παράδειγμα αριθμού) αντί του id = "0", η οποία προκάλεσε την ανάγνωση της ετικέτας XML. Η λύση ήταν μια απλή αντικατάσταση των εισαγωγικών.

Ένα άλλο πρόβλημα, αν και λιγότερο κοινό στο σύνολο δεδομένων σε σύγκριση με το προηγούμενο, ήταν ότι μη αγγλικοί χαρακτήρες όπως φωνήεντα με διακριτικά σημάδια που χρησιμοποιήθηκαν σε ονόματα ή ονόματα εταιρειών, συντάχθηκαν ως οντότητες HTML. Αυτό έπρεπε να επιδιορθωθεί, λόγω των προβλημάτων που δημιουργήθηκαν από τον αναγνώστη XML, που δεν τα αναγνώριζαν. Για την επίλυση του προβλήματος, το κείμενο μεταβιβάστηκε από έναν αναγνώστη HTML ως παράγραφος, έτσι ώστε οι ετικέτες XML να μην δημιουργούν προβλήματα. Στη συνέχεια επιστράφηκε ως συμβολοσειρά unicode για να διατηρηθούν τυχόν αλλαγές μορφοποίησης στις οντότητες, οι ετικέτες παραγράφων αφαιρέθηκαν, έτσι το κείμενο καθαρίστηκε από τυχόν σφάλματα και ήταν έτοιμο για τον αναγνώστη XML.

Τέλος, οι ετικέτες αφαιρέθηκαν κρατώντας μόνο το καθαρό κείμενο από κάθε έγγραφο XML και κάθε αρχείο κειμένου αποθηκεύτηκε ως ακολουθία προτάσεων.

Το US [2] σύνολο δεδομένων, δημιουργήθηκε χρησιμοποιώντας ένα API που περιέχει σαρωμένες (scanned) περιπτώσεις από τη βιβλιοθήκη Harvard Law School, το CaseLaw Access Project. Το CAP επιτρέπει την ελεύθερη πρόσβαση σε ορισμένες δικαιοδοσίες, από τις οποίες επιλέχθηκε το κράτος του Ιλινόις. Το API δημιούργησε ένα αρχείο JSON από τις δεδομένες παραμέτρους, παρέχοντας βασικές πληροφορίες, όπως το όνομα της υπόθεσης, το αναγνωριστικό (id), παραπομπές σε άλλες υποθέσεις,

[2] `https://case.law/docs/site_features/api`

το δικαστήριο που έγινε η δίκη καθώς και την αντίστοιχη δικαιοδοσία.

Χρησιμοποιώντας την παράμετρο full = true καθώς και τις παραμέτρους που απαιτούνται για το επιλεγμένο ερώτημα, επιστρέφονται πρόσθετες πληροφορίες, μαζί με την πιο ουσιαστική, που είναι ολόκληρο το κείμενο της αντίστοιχης νομικής υπόθεσης, όπως φαίνεται στο παρακάτω σχήμα.

Σχήμα 4.1: Παράδειγμα δεύτερου κειμένου


```
"data": {
      "head_matter": "Fifth District\n(No. 70-17;\nThe People of the State of Illinois
...",
      "opinions": [
          {
              "author": "Mr. PRESIDING JUSTICE EBERSPACHER",
              "text": "Mr. PRESIDING JUSTICE EBERSPACHER\ndelivered the opinion of the
court: ...",
              "type": "majority"
          }
      ],
      "judges": [],
      "attorneys": [
          "John D. Shulleriberger, Morton Zwick, ...",
          "Robert H. Rice, State's Attorney, of Belleville, for the Peop ..."
      ]
  }
}
```


Σε αυτά τα δύο dataset εφαρμόστηκαν οι γνωστές τεχνικές αναπαράστασης που αναφέραμε στο προηγούμενο κεφάλαιο. Λόγω του φόρτου όμως, η μέθοδος αξιολόγησης που προτείναμε έγινε μόνο για το πρώτο σύνολο δεδομένων. (Australian Case Law dataset)

4.1.2 Ανάλυση ευαισθησίας

Για όλα τα δυνατά ζευγάρια κειμένων του συνόλου δεδομένων υπολογίστηκε η ομοιότητα συνημιτόνου. Χρησιμοποιώντας αυτή τη διαδικασία για κάθε μέθοδο, δημιουργο-

ύνται διαφορετικές βαθμολογίες και κατατάξεις των ζευγαριών με βάση την ομοιότητα για κάθε μέθοδο. Για να μετρήσουμε τη συσχέτιση μεταξύ αυτών των κατατάξεων, χρησιμοποιήσαμε τον συσχετισμό κατάταξης του Kendall [26].

Το Kendall's Tau είναι ένας μη παραμετρικός συντελεστής που αντιπροσωπεύει το βαθμό συνάφειας μεταξύ δύο κατατάξεων (rankings). Όσο μεγαλύτερος ο αριθμός των αντίστροφων κατατάξεων, τόσο μικρότερος είναι ο συντελεστής.

Δίνεται από τον παρακάτω τύπο 4.1:

$$k = \frac{C - D}{C + D} \tag{4.1}$$

Όπου C είναι ο αριθμός των αντιστοιχούντων ζευγών ανάμεσα στις κατατάξεις και D είναι ο αριθμός των ασύμφωνων ζευγών ανάμεσα στις κατατάξεις.

Έχει εύρος τιμών από 0 έως 1 όπου:

- 0: δεν υπάρχει συσχέτιση
- 1: υπάρχει τέλεια συσχέτιση

Οι συντελεστές συσχέτισης Tau-A και Tau-B χρησιμοποιούνται για ίδιο αριθμό στηλών ενώ το Tau-B μπορεί να χειριστεί και ισοπαλίες. Οι περισσοτέρες υλοποιήσεις στα στατιστικά πακέτα υλοποιούν το Tau-B και αυτόν χρησιμοποιήσαμε και εμείς στο πείραμα μας. Το επιλέξαμε και υπολογίσαμε αυτόν το συντελεστή μεταξύ στις κατατάξεις των αναπαραστάσεων, ώστε να δούμε του αν οι τεχνικές αναπαράστασης που χρησιμοποιήσαμε οδηγούν σε διαφορετικές ή σε παρόμοιες κατατάξεις με βάση την ομοιότητα συνημιτόνου.

Πίνακας 4.1: Συντελεστής συσχέτισης Kendall στις κατατάξεις των μεθόδων

	TF-IDF	Doc2vec	BERT
TF-IDF	1		
Doc2vec	0.34	1	
BERT	0.26	0.19	1

Βλέπουμε στον πίνακα 4.1 πως ο συντελεστής συσχέτισης ανάμεσα στις κατατάξεις των TF-IDF και Doc2vec βγήκε 0.34, το οποίο σημαίνει ότι δεν υπάρχει αρκετή συνάφεια μεταξύ των 2 αυτών κατατάξεων. Το ίδιο ισχύει και για το TF-IDF

και BERT όπου ο συντελεστής βγήκε 0.26, ενώ για τις κατατάξεις των BERT και Doc2vec βλέπουμε ότι ο συντελεστής βγαίνει 0.19 που σημαίνει ότι παράγουν εντελώς διαφορετικές κατατάξεις.

Επομένως, καταλήγουμε στο συμπέρασμα πως η επιλογή αναπαράστασης παίζει μεγάλο ρόλο και καθώς παράγουν αρκετά διαφορετικά αποτελέσματα.

Άρα, αξίζει να μελετηθεί μέσω ανθρώπινης αξιολόγησης ποια από τις 3 τεχνικές είναι πιο ευθυγραμμισμένη με την ανθρώπινη αντίληψη.

4.2 Αξιολόγηση και αποτελέσματα

Μία από τις συνεισφορές (contribution) αυτής της εργασίας αφορά την αξιολόγηση των μεθόδων που χρησιμοποιήσαμε για την αναπαράσταση των νομικών κειμένων και τις αυτόματες μετρικές ομοιότητας. Θέλουμε να αξιολογήσουμε του κατά πόσο τελικά, συμφωνούν υπολογιστής και ανθρώπινη αντίληψη σχετικά με την ομοιότητα νομικών κειμένων, δηλαδή του αν δύο νομικά κείμενα τα οποία σύμφωνα με τον υπολογιστή θεωρηθούν σχετικά, θεωρούνται και από τους ανθρώπους ως παρόμοιες νομικές υποθέσεις.

Για να μπορέσει να γίνει αυτή η σύγκριση μεταξύ υπολογιστή και ανθρώπου, χρειάστηκε να ακολουθηθεί μια διαδικασία, η οποία σχετίζεται με το κομμάτι της επί σημείωσης (annotation). Σκοπός αυτής της διαδικασίας ήταν να μπορέσει να γίνει ανθρώπινη αξιολόγηση ζευγαριών νομικών υποθέσεων, μέσω εθελοντών, ώστε να δούμε τελικά αν υπάρχει συνοχή μεταξύ των επιλογών, των σχολιαστών και των αλγορίθμων.

Σε αυτήν την ενότητα θα αναλυθεί εκτενώς τα βήματα που χρειάστηκαν για να γίνει η αξιολόγηση, η διαδικασία της επί σημείωσης και τέλος τα αποτελέσματα μαζί τα ευρήματα της αξιολόγησης.

4.2.1 Βήματα της αξιολόγησης

Λόγω του μεγάλου όγκου του συνόλου δεδομένων και της δυσκολίας του να αξιολογηθούν όλα τα ζευγάρια, μετασχηματίσαμε το σύνολο δεδομένων σε ένα μικρότερο ώστε να μπορέσει να γίνει η αξιολόγηση. Σε αυτήν την υπό ενότητα περιγράφουμε τα βήματα που ακολουθήσαμε για την ανθρώπινη αξιολόγηση. Τα βήματα για την αξιολόγηση

ήταν δύο:

1. Επιλογή καλών υποψηφίων για ταίριασμα (matching)
2. Κατάταξη κειμένων με βάση τη σχετικότητα

Περιληπτικά, η διαδικασία του πρώτου βήματος της αξιολόγησης αναλύεται στις παρακάτω ενέργειες:

- Επιλογή 10 κειμένων από το σύνολο δεδομένων, που τα ονομάζουμε **κείμενα αναφοράς**.
- Για κάθε κείμενο αναφοράς
  - Επιλογή 12 καλύτερων άλλων κειμένων με βάση την ομοιότητά τους με το κείμενο αναφοράς από τον υπολογιστή. Αυτά τα κείμενα τα ονομάζουμε **υποψήφια κείμενα**
  - Προβολή των 12 υποψηφίων κειμένων στο σχολιαστή-αξιολογητή
  - Ο αξιολογητής για κάθε υποψήφιο κείμενο επιλέγει αν το θεωρεί σχετικό ή μη σχετικό με το αντίστοιχο κείμενο αναφοράς.
  - Η διαδικασία για κάθε κείμενο αναοφράςσταματάει όταν βρεθούν 5 σχετικά κείμενα από τα υποψήφια κείμενα.

Έτσι, μετά το πέρας της πρώτης φάσης καταλήγουμε να έχουμε για κάθε κείμενο αναφοράς i Ki σχετικά κείμενα (το πολύ μέχρι 5) και 12-Ki άσχετα κείμενα. Διαλέξαμε να επιλέγουν το πολύ μέχρι 5 λόγω του φόρτου που απαιτείται για τη διαδικασία, έτσι με αυτήν την επιλογή μειώνουμε κάπως το φόρτο.

Έχοντας αυτό ως δεδομένο, συνεχίζουμε παρουσιάζοντας συνοπτικά και τις ενέργειες της δεύτερης φάσης της αξιολόγησης.
Για το δεύτερο βήμα της αξιολόγησης:

- Δημιουργούμε όλα τα δυνατά ζευγάρια των K σχετικών κειμένων ανά δύο
- Δείχνουμε τα παραπάνω ζευγάρια κειμένων στον αξιολογητή, μαζί με το κείμενο αναφοράς με το οποίο θεωρήθηκαν σχετικά τα κείμενα με τα οποία δημιουργήσαμε τα ζευγάρια

- Ο αξιολογητής επιλέγει ανάμεσα σε:
  - Αν κάποιο από τα δύο κείμενα προτιμά ως καλύτερο ταίριασμα match με το θέμα-κείμενο αναφοράς και ποιο είναι αυτό
  - Θεωρεί και τα δύο το ίδιο σχετικά με το κείμενο αναφοράς και δεν μπορεί να ξεχωρίσει κανένα από τα δύο πιο σχετικό από το άλλο σε σχέση με το θέμα

4.2.1.1 Πρώτη φάση αξιολόγησης

Αναφέραμε προηγουμένως επί γραμματικά τις ενέργειες που γίνονται στο πρώτο βήμα της αξιολόγησης. Τώρα, θα αναλύσουμε τις επιλογές που έγιναν σε κάθε ενέργεια ξεχωριστά. Σε αυτήν την ενότητα αναλύουμε τη μεθοδολογία που ακολουθήσαμε για να δημιουργήσουμε το υποσύνολο του συνόλου δεδομένων που θα χρησιμοποιηθεί στην αξιολόγηση.

Λόγω του ότι η ανθρώπινη αξιολόγηση είναι ακριβή και πολύ χρονοβόρα, ήταν ανέφικτο να συγκριθούν όλα τα νομικά ζευγάρια του σώματος κειμένου. Για αυτό και επιλέξαμε για κείμενα αναφοράς 10 ώστε να μειώσουμε το χρόνο αλλά και να καταστήσουμε εφικτή την ανθρώπινη αξιολόγηση.

Αρχικά, η επιλογή των 10 κειμένων αναφοράς έγινε εφαρμόζοντας αλγόριθμο συσταδοποίησης στο σύνολο δεδομένων. Σκοπός πίσω από αυτό το σκεπτικό, ήταν να διασφαλίσουμε ότι τα 10 κείμενα αναφοράς θα έχουν διαφορετική θεματολογία μεταξύ τους. Αυτό που θέλουμε να πετύχουμε, είναι να πάρουμε το κείμενο, με βάση την αναπαράσταση του σε διανυσματικό χώρο, που είναι πιο κοντινό στο κέντρο (centroid) της κάθε συστάδας (cluster) και να το ορίσουμε ως κείμενο αναφοράς.

Ένας αλγόριθμος που το κάνει απευθείας αυτό, είναι ο k-εσωτερικών αντιπροσώπων (k-medoids).[27] Η ιδέα πίσω από αυτόν τον αλγόριθμο είναι ότι μετά το τέλος της συσταδοποίησης κάθε κέντρο της συστάδας αποτελεί πραγματικό σημείο στο σύνολο δεδομένων. Έτσι, τα κέντρα των ομάδων είναι ερμηνεύσιμα και έχουμε αντιπρόσωπο για κάθε συστάδα από τη συλλογή δεδομένων. Αυτό κάνει τον k-medoids πιο χρήσιμο για το πρόβλημα μας και μας διευκολύνει, σε σχέση με τους υπόλοιπους κλασσικούς αλγορίθμους ομαδοποίησης.

Προτού εφαρμοστεί ο αλγόριθμος συσταδοποίησης πρέπει να έχουμε πρώτα τα κείμενα σε διανυσματική μορφή. Άρα, πρωτίστως υπολογίζουμε τα βάρη TF-IDF για κάθε κείμενο από τη συλλογή, αφαιρώντας πιο πριν τα σημεία στίξης από κάθε έγγραφο και μετατρέποντας όλους τους χαρακτήρες σε πεζά γράμματα. Στη συνέχεια, εφαρμόζουμε τον k-medoids για k=10 συστάδες, και στο τέλος κάθε κέντρο της συστάδας αποτελεί κείμενο αναφοράς. Έτσι, έχουμε εν τέλει τα 10 κείμενα αναφοράς.

Η δεύτερη ενέργεια αφορά του πως θα γίνει η επιλογή των 12 υποψήφιων κείμενων για κάθε κείμενο αναφοράς.

Αυτό που αποφασίστηκε για τα υποψηφία κείμενα για κάθε κείμενο αναφοράς, να είναι από όλες τις αναπαραστάσεις και με επιλεγμένες τιμές έχοντας ως κριτήριο την βαθμολογία της ομοιότητας συνημιτόνου που υπολογίσαμε.

Πιο αναλυτικά η επιλογή στην δεύτερη ενέργεια της πρώτης φάσης της αξιολόγησης έγινε ως εξής:

- Για κάθε **κείμενο αναφοράς** τα 12 υποψηφία κείμενα επιλέχθηκαν ως εξής:
  - Τα 6 πιο κοντινά. Δηλαδή, τα 6 εκείνα κείμενα που υπολογίσαμε ότι είχαν τη μεγαλύτερη τιμή στην ομοιότητα συνημιτόνου με το κείμενο αναφοράς. Ένα άλλο ζήτημα που προέκυψε εδώ ήταν βάση ποιας αναπαράστασης τη βαθμολογία στην ομοιότητα θα γίνει η επιλογή των 6. Αυτό που αποφασίστηκε, ήταν να πάρουμε 2 κείμενα με την πιο υψηλή ομοιότητα σε κάθε αναπαράστασή που υλοποιήσαμε. Άρα επιλέξαμε τα 2 κείμενα που είχαν την πιο υψηλή ομοιότητα με αναπαράσταση TF-IDF και ακολούθως με την ίδια λογική, 2 από την τεχνική Doc2vec και 2 από τη μέθοδο BERT.
  - 3 μεσαία. Τα 3 εκείνα κείμενα που είχαν τιμή ομοιότητας κοντά στη μεσαία τιμή του διαστήματος της χαμηλότερης ομοιότητας και της μέγιστης για το αντίστοιχο κείμενο αναφοράς. Επιλέχθηκαν 1 κείμενο με την μεσαία ομοιότητα από κάθε μέθοδο (1 TF-IDF 1 doc2vec και 1 BERT)
  - 3 μακρινά. Αυτά τα 3 νομικά κείμενα που υπολογίσαμε ότι είχαν τη χαμηλότερη ομοιότητα βάση cosine similarity score με το αντίστοιχο κείμενο αναφοράς. Εδώ ξανά, επιλέξαμε ένα νομικό κείμενο από κάθε αναπαράσταση που είχε τη χαμηλότερη ομοιότητα (1 TF-IDF, 1 doc2vec και 1 BERT)

Έτσι, με το παραπάνω πετύχαμε να συμμετέχουν, τόσο όλες οι αναπαραστάσεις το ίδιο στην επιλογή των υποψήφιων κειμένων όσο και να υπάρχουν τιμές ομοιότητας από όλες τις ακραίες τιμές -εύρη για να καλύψουμε όλες τις περιπτώσεις. Για παράδειγμα, όπως αναφέραμε, μπήκαν και κείμενα με χαμηλή ομοιότητα στο σύνολο των υποψήφιων κειμένων, για να εξετάσουμε και αν υπάρχει κείμενο που ενώ υπολογίστηκε από τη μηχανή ότι είχε χαμηλή ομοιότητα, αντίθετα όμως, σύμφωνα με την ανθρώπινη αντίληψη θεωρήθηκε ως σχετικό.

Οι τελευταίες δύο ενέργειες αφορούν τις επιλογές του αξιολογητή. Αρχικά επιλέγει για κάθε υποψήφιο κείμενο αν το θεωρεί σχετικό ή μη σχετικό με το αντίστοιχο κείμενο αναφοράς. Σκοπός πίσω από αυτό, ήταν να υπάρξουν σχετικά κείμενα για τη δεύτερη και σημαντικότερη φάση της αξιολόγησης. Ορίσαμε ως μέγιστο όριο να διαλέξει 5 σχετικά από τα 12 υποψήφια κείμενα για κάθε θέμα, γιατί θεωρούμε ότι είναι ένα αρκετά καλό δείγμα για την αξιολόγηση που θα γίνει στο δεύτερο βήμα, αλλά και για να μειωθεί ο φόρτος εργασίας για τους εθελοντές-αξιολογητές στη δεύτερη φάση της αξιολόγησης.

Στην υπό ενότητα που αφορά τη διαδικασία της επισημείωσης (annotation) θα επανέλθουμε και πιο αναλυτικά για τις 2 τελευταίες ενέργειες της πρώτης φάσης.

4.2.1.2 Δεύτερη φάση αξιολόγησης

Σε αυτή τη δεύτερη φάση της αξιολόγησης, αρχικά έχουμε ως δεδομένο από την προηγούμενη φάση ποια από τα 12 κείμενα -για κάθε κείμενο αναφοράς- τα οποία ορίσαμε πριν ως υποψήφια, ο αξιολογητής τα έκρινε ως σχετικά.

Έστω Ki ο αριθμός των σχετικών κειμένων που επέλεξε ο εθελοντής για το κείμενο αναφοράς i. Για κάθε κείμενο αναφοράς i δημιουργούμε όλα τα δυνατά ζευγάρια των Ki κειμένων ανά δύο. Έτσι, έχουμε για κάθε ένα από τα 10 κείμενα αναφοράς, ορισμένα ζευγάρια νομικών κειμένων.

Στη συνέχεια, δείχνοντας στο σχολιαστή για κάθε κείμενο αναφοράς, τα ζευγάρια κειμένων που δημιουργήθηκαν για αυτό, αυτός έχει να επιλέξει ποιο από τα δύο κείμενα του κάθε ζευγαριού θεωρεί πιο σχετικό σε σχέση με το κείμενο αναφοράς. Όταν όμως κάποιος έχει αν διαλέξει ανάμεσα σε 2 κείμενα, είναι καλό να εισάγουμε μία

τρίτη επιλογή η οποία είναι: δεν μπορώ να ξεχωρίσω κανένα από τα 2 ως σχετικό με το αντίστοιχο κείμενο αναφοράς. Αυτό είναι σημαντικό ώστε να μπορέσουμε να το συγκρίνουμε και με τις αντίστοιχες ισοπαλίες που μπορεί να βγάζει στην ομοιότητα 2 κειμένων μία τεχνική.

Τέλος, εφόσον έχουμε πλέον και τις επιλογές των αξιολογητών, εξετάζουμε τη συνοχή μεταξύ των επιλογών αυτών από την πλευρά των ανθρώπων και των επιλογών από τις αυτόματες μετρικές ομοιότητας, μέσω στατιστικών μεθόδων όπως θα δούμε εκτενώς στο κομμάτι των αποτελεσμάτων.

4.2.2 Δημογραφικά στοιχεία

Για αυτήν την έρευνα, όπως αναφέραμε χρειάστηκε να επιστρατευθούν εθελοντές οι οποίοι θα έχουν το ρόλο του αξιολογητή για τη σύγκριση ομοιότητας νομικών κειμένων. Σε αυτήν την υπό ενότητα αναφέρουμε ορισμένα στοιχεία για τους εθελοντές μας.

Στην πρώτη φάση συμμετείχαν 4 άνθρωποι με αντικείμενα εξειδίκευσής την φιλολογία, μαθηματικό και την πληροφορική. Στη δεύτερη φάση συμμετείχαν άνθρωποι από όλες τις βαθμίδες της τριτοβάθμιας εκπαίδευσής, προ πτυχιακοί φοιτητές, μεταπτυχιακοί καθώς και υποψήφιοί διδάκτορες. Οι περισσότεροι προερχόμενοι από τον κλάδο της Πληροφορικής υπήρχαν όμως και εθελοντές από τον κλάδο της ψυχολογίας, του μαθηματικού και της φιλολογίας.

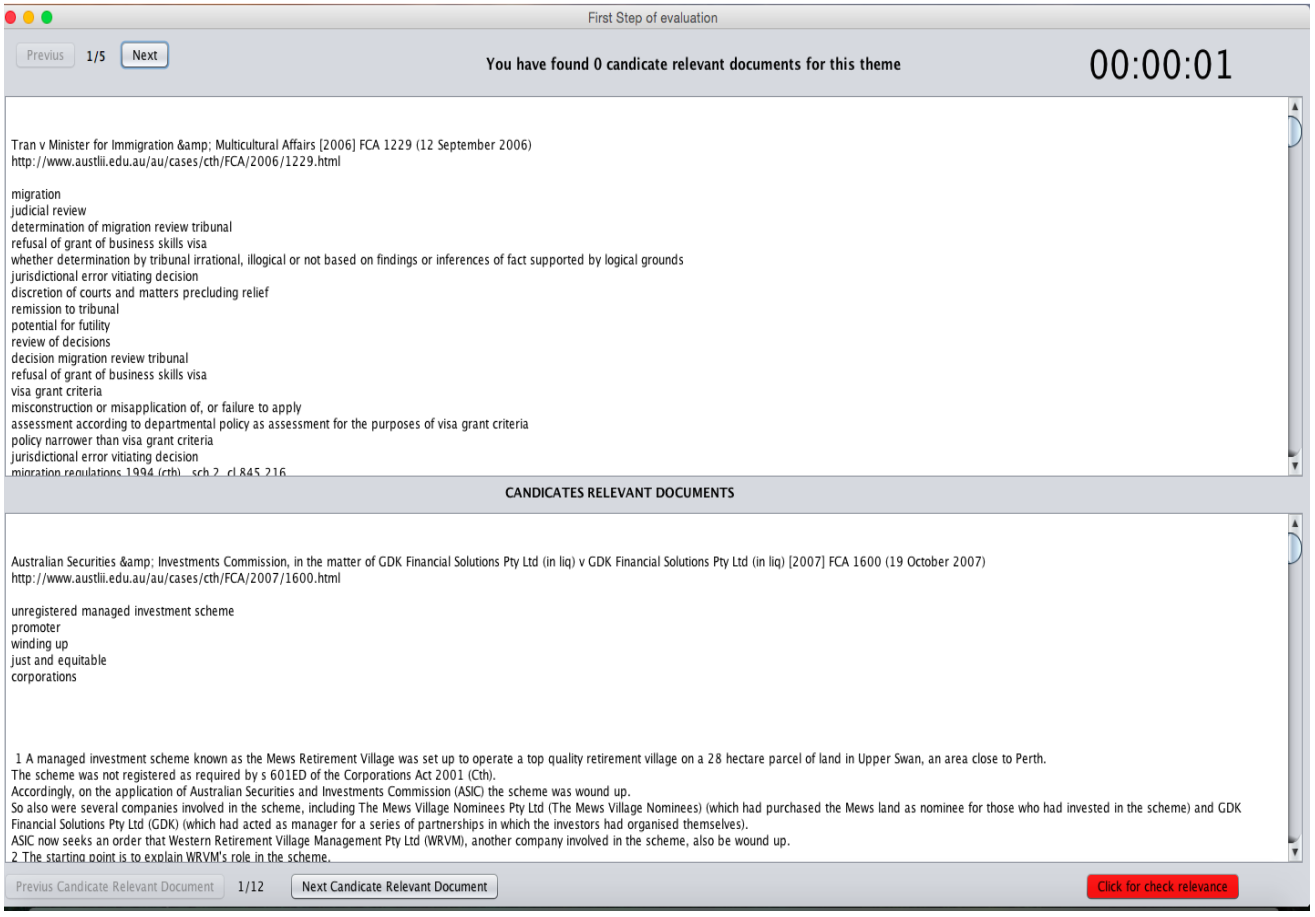
4.2.3 Υλοποίηση annotation tool

Για τις ανάγκες του να γίνει διαδικασία της επί σημειώσης εύκολα, ήταν απαραίτητο να υλοποιηθεί ένα εύχρηστο εργαλείο, το οποίο θα χρησιμοποιούσαν οι εθελοντές ώστε να εφαρμόσουν όλα όσα αναλύσαμε για τα βήματα της αξιολόγησης. Το εργαλείο αυτό γράφτηκε στη γλώσσα προγραμματισμού Java και για το GUI χρησιμοποιήθηκε η βιβλιοθήκη swing.

Στο σχήμα 4.2 φαίνεται ένα παράδειγμα εκτέλεσης του εργαλείου για την πρώτη φάση που αναλύσαμε σε προηγούμενη ενότητα. Όπως βλέπουμε στο πάνω μέρος της γραφικής επιφάνειας είναι το κείμενο αναφοράς και στο κάτω μέρος τα υποψήφια σχετικά κείμενα. Με το πάτημα των κουμπιών Next και Previus που βρίσκονται αριστερά

πάνω στην γραφική επιφάνεια, ο χρήστης αλλάζει τα κείμενα αναφοράς και ταυτόχρονα αλλάζουν και το σύνολο υποψήφιων κειμένων στο κάτω μέρος. Επίσης με τα κουμπιά Next Candicate Relevant Document και Next Candicate Relevant Document μπορεί να αλλάζει τα υποψήφια σχετικά κείμενα για το αντίστοιχο κείμενο αναφοράς που βρίσκεται στο πάνω μέρος.

Σχήμα 4.2: Εργαλείο της πρώτης φάσης της αξιολόγησης. Επισημείωση συναφών κειμένων για κάθε κείμενο αναφοράς

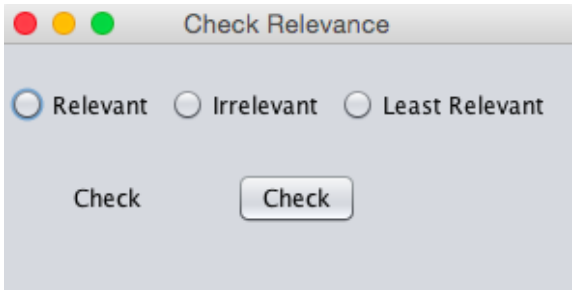


Ο χρήστης-αξιολογητής λοιπόν, διαβάζει πρώτα το πάνω κείμενο, στη συνέχεια τα 12 κάτω κείμενα και για κάθε ένα από αυτά πατάει το κόκκινο κουμπί (Click for check relevance) που βρίσκεται στο κάτω δεξί μέρος της επιφάνειας και επιλέγει ανάμεσα στις επιλογές που φαίνονται στο σχήμα 4.3. Relevant αν το θεωρεί σχετικό με το πάνω κείμενο, Irrelevant αν το θεωρεί άσχετο, ενώ η τελευταία επιλογή αφορά την σπάνια ακραία περίπτωση που για κάποιο κείμενο αναφοράς δεν έκρινε κανένα από τα υποψήφια να είναι σχετικό, τότε επιλέγει τα δύο λιγότερα άσχετα ως least irrelevant (Μόνο τότε

να γίνετε αυτή η επιλογή).

Στο πάνω δεξιά μέρος φαίνεται ο χρόνος που χρησιμοποιεί ο annotator το εργαλείο, γιατί θέλουμε να κρατήσουμε σαν πληροφορία το πόσο χρόνο έκανε γι΄ αυτό το task. Επιπλέον στη μέση στο πάνω μέρος εμφανίζεται ένα μήνυμα του πόσα σχετικά κείμενα έχει βρει για το κείμενο αναφοράς στο οποίο βρίσκεται. Τέλος κάθε φορά που κλείνει κάποιος το εργαλείο, αποθηκεύονται οι επιλογές του και έτσι δεν χρειάζεται να τρέξει κάποιος την επισημείωση σε ένα πέρασμα.

Σχήμα 4.3: Επιλογή σχετικότητας



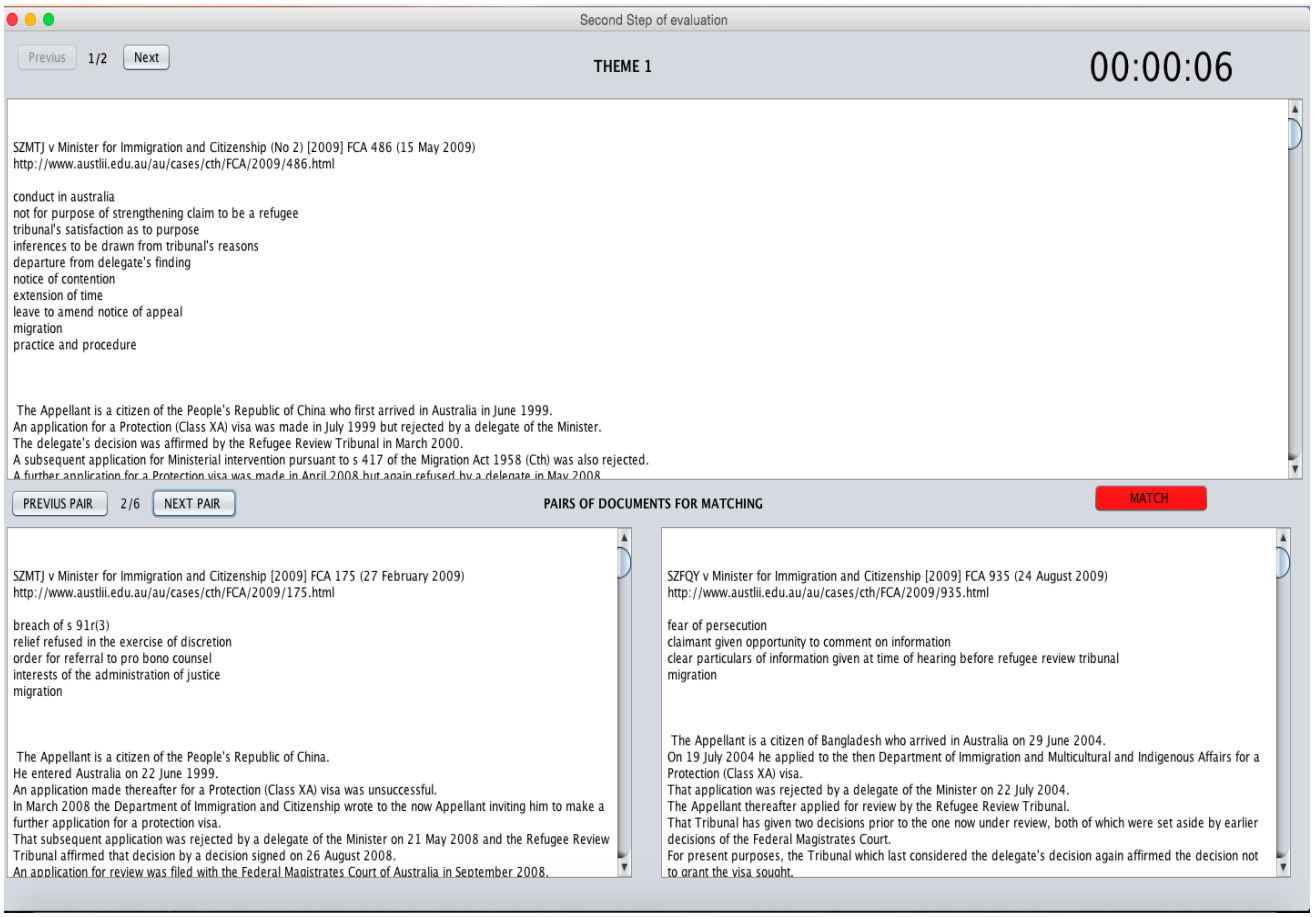
Με το που τελείωνε κάποιος την επισημείωση μας έστελνε πίσω ένα αρχείο JSON που περιείχε τις επιλογές του, δηλαδή πόσα σχετικά κείμενα βρήκε για κάθε κείμενο αναφοράς και ποια είναι αυτά, καθώς και την πληροφορία του πόσο χρόνο αφιέρωσε συνολικά για την διαδικασία.

Με την ίδια λογική, υλοποιήθηκε και το εργαλείο που θα χρειαστεί για τη δεύτερη φάση. Όπως βλέπουμε και στο σχήμα 4.4, τα πράγματα είναι παρόμοια με προηγουμένως. Πάνω, έχουμε τα κείμενα αναφοράς ξανά, καθώς και το χρόνο, ενώ στο κάτω μέρος της γραφικής επιφάνειας αυτή τη φορά βρίσκονται ζευγάρια κειμένων.

Ο αξιολογητής μελετά το κείμενα αναφοράς, διαβάζει το ζευγάρι στο κάτω μέρος και κάνει κλικ στο κουμπί MATCH ώστε να επιλέξει ποιο από τα δύο κείμενα του ζευγαριού θεωρεί πιο σχετικό με το πάνω κείμενο. Οι επιλογές του όπως φαίνονται και στο σχήμα είναι : First Doc αν θεωρεί πιο όμοιο το αριστερά κάτω κείμενο, Second Doc αν κρίνει το κάτω δεξιά κείμενο πιο σχετικό και τέλος Same αν δεν μπορεί να διακρίνει κανένα από τα δύο και τα θεωρεί το ίδιο σχετικά. Το κουμπί MATCH από κόκκινο γίνεται πράσινο ώστε να θυμάται ο χρήστης για ποια ζευγάρια έχει κάνει τις επιλογές του.

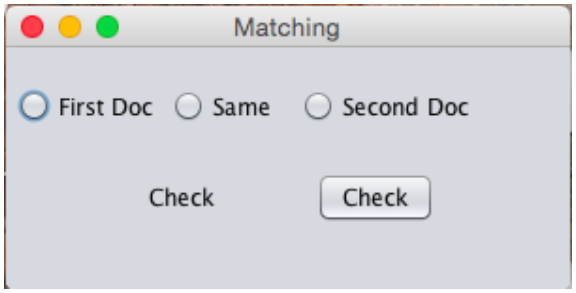
Τέλος, μετά το πέρας και αυτής της διαδικασίας ο χρήστης-αξιολογητής μας επι-

Σχήμα 4.4: Εργαλείο της δεύτερης φάσης της αξιολόγησης. Επισημείωση για ταίριασμα κειμένου ζευγαριού σε σχέση με το κείμενο αναφοράς.



στρέφει το αρχείο JSON με τις απαντήσεις του, που περιέχουν τις επιλογές του για όλα τα ζευγάρια αλλά και το χρόνο που διήρκεσε η διαδικασία.

Σχήμα 4.5: Ταίριασμα



Πίνακας 4.2: Κλίμακες για το Cohen's kappa και ερμηνείες

0.01-0.20	ελάχιστη συμφωνία
0.21-0.40	μικρή συμφωνία
0.41-0.60	μέτρια συμφωνία
0.61-0.80	ουσιώδη συμφωνία
0.81-1	σχεδόν τέλεια συμφωνία

4.2.3.1 Συμφωνία μεταξύ των των σχολιαστών

Η συμφωνία μεταξύ των σχολιαστών inter-annotator agreement [28] είναι ένα μέτρο του πόσο καλά μπορούν πολλοί σχολιαστές να λάβουν την ίδια απόφαση σχολιασμού για μια συγκεκριμένη κατηγορία. Το IAA μας δείχνει πόσο σαφείς είναι οι οδηγίες σχολιασμού που δώσαμε, πόσο ομοιόμορφα κατάλαβαν το πρόβλημα οι σχολιαστές μας και πόσο επαναχρησιμοποιήσιμο είναι το έργο σχολιασμού.

Στη δική μας περίπτωσή, για παράδειγμα στην πρώτη φάση θέλουμε για τα ίδια κείμενα αναφοράς και τα υποψήφια κείμενα τους που αξιολόγησαν δύο σχολιαστές πόσο ίδιες ήταν οι επιλογές τους και πόσο συμφωνούσαν μεταξύ τους. Εφόσον θέλουμε να υπολογίσουμε τη συνοχή μεταξύ δύο σχολιαστών, εφαρμόζουμε το Cohen kappa το οποίο υπολογίζεται μεταξύ ενός ζεύγους σχολιαστών.

Η στατιστική kappa του Cohen είναι η συμφωνία μεταξύ δύο βαθμολογητών λαμβάνοντας παράλληλα υπόψη την πιθανότητα συμφωνίας, όπου το Po είναι η σχετική παρατηρούμενη συμφωνία μεταξύ των βαθμολογητών και το Pe είναι η αναμενόμενη συμφωνία. Δίνεται από τον παρακάτω τύπο 4.2:

$$k = \frac{P_o - P_e}{1 - P_e} \tag{4.2}$$

Το k ερμηνεύεται στις κλίμακες [29] που βλέπουμε στον πίνακα 4.2 .

Πίνακας 4.3: Συμφωνία μεταξύ σχολιαστών για τη φάση 1

	Αξιολογητής 3	Αξιολογητής 4
Αξιολογητής 1	0.82	-
Αξιολογητής 2	-	0.75

Για την πρώτη φάση της αξιολόγησης, η διαδικασία θέλαμε να γίνει από δύο σχο-

λιαστές για κάθε κείμενο αναφοράς. Οι εθελοντές στο πρώτο βήμα που συμμετείχαν ήταν 4, και με δεδομένο ότι θέλαμε να μειώσουμε τον απαιτούμενο φόρτο, χωρίσαμε τα 10 θέματα σε κομμάτια των 5 ώστε να μπορέσει να γίνει η διαδικασία της επισημείωσης από δύο ανθρώπους. Ο αξιολογητής 1 με τον αξιολογητής 3 επισημείωσαν τα ίδια κείμενα οπότε υπολογίζουμε το k για αυτό το ζεύγος και αντίστοιχα για τον αξιολογητής 2 με τον αξιολογητής 4.

Όπως βλέπουμε στον πίνακα 4.3 και σύμφωνα με τις ερμηνείες που δώσαμε πριν για το k υπάρχει σχεδόν τέλεια συμφωνία μεταξύ του σχολιαστή 1 και 3, ενώ ανάμεσα στον 2 και στον 4 υπάρχει ουσιώδης συμφωνία.

Στον πίνακα 4.4 βλέπουμε τους χρόνους που χρειάστηκαν για να ολοκληρώσουν τη διαδικασία της πρώτης φάσης οι αξιολογητές. Παρατηρούμε ότι για να βρεθούν συναφή κείμενα για 5 κείμενα αναφοράς χρειάστηκε περίπου μία ώρα. Μπορούμε να συμπεράνουμε ότι η διαδικασία της εύρεσης συναφών νομικών κειμένων είναι όπως αναφέραμε και πριν χρονοβόρα για ανθρώπινη αξιολόγηση και ίσως οφείλεται στην δυσκολία κατανόησης του συγκεκριμένου τομέα. Βλέπουμε ότι οι αξιολογητές 1 και 3 που επισημειώσαν τα ίδια κείμενα, η διαδικασία τους πήρε τον ίδιο περίπου χρόνο όπως και αντίστοιχα για τον 2 και 4.

Πίνακας 4.4: Χρόνοι για την πρώτη φάση αξιολόγησης

	Χρόνος
Αξιολογητής 1	62′
Αξιολογητής 2	53′
Αξιολογητής 3	51′
Αξιολογητής 4	64′

Κεφάλαιο 5

Αποτελέσματα

Σε αυτό το κεφάλαιο αναλύουμε τα αποτελέσματα της δεύτερης φάσης της αξιολόγησης. Αρχικά περιγράφουμε τη συμφωνία μεταξύ των αξιολογητών στη δεύτερη φάση και στη συνέχεια αναλύουμε τα ευρήματα σχετικά με τη σύγκριση των μεθόδων και των αξιολογητών.

5.1 Συμφωνία μεταξύ των αξιολογητών

Τα 10 κείμενα αναφοράς μοιράστηκαν ανά δύο στον κάθε αξιολογητή μαζί με τα αντίστοιχα ζευγάρια για το κάθε κείμενο αναφοράς που χρειαζόταν για τη δεύτερη φάση της αξιολόγησης. Έτσι από εδώ και κάτω, όταν θα αναφερόμαστε στο Αξιολογητής 1 και 6 θα εννοούμε τον αξιολογητή εκείνον, που επισημείωσε τα 2 πρώτα κείμενα αναφοράς, Αξιολογητής 2 και 7 εκείνος που επισημείωσε τα κείμενα αναφοράς 3 και 4 και αντίστοιχα το ίδιο και για τους υπόλοιπους.

Η ίδια λογική ακολουθείται και στην ορολογία αξιολόγηση. Με τον όρο αξιολόγηση 1 αναφερόμαστε στην αξιολόγηση κατά μέσο όρο των κειμένων 1 και 2 που έγιναν από τους αξιολογητές που επισημείωσαν τα εκείνα τα κείμενα αναφοράς. Το ίδιο και για την αξιολόγηση 2 που αφορούσε τα κείμενα 3 και 4 και ου το καθεξής και για τις υπόλοιπες αξιολογήσεις.

Πίνακας 5.1: Συμφωνία μεταξύ σχολιαστών για τη φάση 2

	Αξιολογητής 6
Αξιολογητής 1	0.76

	Αξιολογητής 7
Αξιολογητής 2	0.69

	Αξιολογητής 8
Αξιολογητής 3	0.63

	Αξιολογητής 9
Αξιολογητής 4	0.84

Στον πίνακα 5.1 βλέπουμε του κατά πόσο συμφωνούσαν οι αξιολογητές που έκριναν τα ίδια κείμενα μεταξύ τους σύμφωνα με την κλίμακα του πίνακα 4.2. Παρατηρούμε ότι οι αξιολογητές 1 και 6 έχουν ουσιώδη συμφωνία, όπως και ο 2 με τον 7 αλλά και ο 3 με τον 8. Τέλος ο αξιολογητής 4 με τον 9 έχουν σχεδόν τέλεια συμφωνία ενώ αντίθετα ο 5 με τον 10 έχουν μικρή συμφωνία. Σε γενικές γραμμές πάντως βλέπουμε ότι υπάρχει ικανοποιητική συμφωνία ανάμεσα στους αξιολογητές εκτός από το τελευταίο ζευγάρι, το οποίο ίσως και να οφείλεται στη δυσκολία της διαδικασίας.

Επίσης στη συνέχεια αυτού που αναφέραμε παραθέτουμε τον πίνακα 5.2 τους χρόνους που χρειάστηκαν για τη διαδικασία οι αξιολογητές για τη δεύτερη φάση και δείχνει τη δυσκολία και το πόσο χρονοβόρα ήταν η διαδικασία. Σε αυτή τη δεύτερη φάση βλέπουμε μεγαλύτερους χρόνους από την πρώτη φάση της αξιολόγησης. Αυτό γίνεται γιατί σε αυτό το βήμα, οι αξιολογητές έχουν να διαβάσουν ζευγάρια κειμένων και έχουν να πάρουν πιο δύσκολη απόφαση από την προηγούμενη.

5.2 Σύγκριση μεθόδων με την ανθρώπινη αξιολόγηση

Σε αυτήν την ενότητα αναλύουμε τη σύγκριση των μεθόδων και της αξιολόγησής που ακολουθήσαμε. Για τα αποτελέσματα που θα περιγράψουμε παρακάτω ακολουθήσαμε την εξής μεθοδολογία: Θεωρήσαμε την κάθε μέθοδο που χρησιμοποιήσαμε στην αναπαράσταση ως εικονικό αξιολογητή.

Τρέξαμε δηλαδή τη διαδικασία της δεύτερης φάσης αξιολόγησης που αναλύσαμε πριν και για τις 3 τεχνικές, κρατήσαμε τις επιλογές τους και τις συγκρίναμε με τις

	Αξιολογητής 10
Αξιολογητής 5	0.37

Πίνακας 5.2: Χρόνοι για την δεύτερη φάση σε λεπτά

	Χρόνος
Αξιολογητής 1	70′
Αξιολογητής 2	85′
Αξιολογητής 3	92′
Αξιολογητής 4	120’
Αξιολογητής 5	63′
Αξιολογητής 6	81′
Αξιολογητής 7	97′
Αξιολογητής 8	110′
Αξιολογητής 9	108′
Αξιολογητής 10	114′

αντίστοιχες επιλογές των αξιολογητών ώστε να δούμε κατά πόσο είναι αποδοτικές στο συγκεκριμένο τομέα που εξετάζουμε.

Οι μέθοδοι επέλεγαν ανάλογα με την τιμή της ομοιότητας που υπολόγιζαν και στην περίπτωση της επιλογής του ότι δεν ξεχώρισαν κάποιο ως σχετικό με το κείμενο αναφοράς, δηλαδή ότι τα θεωρούν το ίδιο όμοια δώσαμε μια χαλάρωση του ότι αν δύο κείμενα έχουν διαφορά ομοιότητας στο εύρος 0.5 τότε οι μέθοδοι επιλέγουν την επιλογή θεωρώ και τα δύο το ίδιο όμοια με το αντίστοιχο κείμενο αναφοράς.

Έτσι, έχοντας αυτά ως δεδομένα και θεωρώντας την κάθε μέθοδο αξιολογητή στη συνέχεια μπορούμε να υπολογίσουμε τη συμφωνία μεταξύ τους μέσω του Cohen’s kappa που αναλύσαμε προηγουμένως και μπορέσουμε να βγάλουμε κάποια συμπεράσματα για την απόδοση των μεθόδων και στο κατά πόσο συμφωνούν με τους ανθρώπους μέσω του inter annotator agreement.

Αρχικά, βλέπουμε στον πίνακα 5.3 τον μέσο όρο απόδοσης. Δηλαδή τον μέσο όρο του Cohen’s kappa των αξιολογητών που επισημείωσαν τα ίδια κείμενα συγκρίνοντάς το με τα επιλογές των μεθόδων για τα αντίστοιχα ζευγάρια κειμένων. Ως όρος αξιολόγηση που εμφανίζεται στον πίνακα εννοούμε τα κομμάτια των κειμένων που δόθηκαν για επισημείωση. Δεδομένου του ότι είχαμε 10 αξιολογητές και θέλαμε να γίνουν 2 επισημείωσεις, έχοντας 10 κείμενα αναφοράς που αναλύσαμε νωρίτερα μοιράστηκαν 2

κείμενα στον καθένα από τους 10, μαζί με τα αντίστοιχα ζευγάρια για κάθε κείμενο αναφοράς.

Πίνακας 5.3: Μέσος όρος απόδοσης

	TF-IDF	Doc2vec	BERT
Αξιολόγηση 1	0.55	0.76	0.31
Αξιολόγηση 2	0.58	0.82	0.51
Αξιολόγηση 3	0.46	0.69	0.60
Αξιολόγηση 4	0.73	0.89	0.55
Αξιολόγηση 5	0.37	0.41	0.38

Όλες οι ερμηνείες που δίνονται ξανά είναι σύμφωνα με την κλίμακα του πίνακα 4.2. Παρατηρούμε ότι η μέθοδος Doc2vec βγάζει τα καλύτερα αποτελέσματα σύμφωνα με την ανθρώπινη αντίληψη καθώς δύο φορές συναντάμε σχεδόν τέλεια συμφωνία αλλά και γενικά δείχνει να τα πάει καλύτερα σε σχέση με τα ζευγάρια αξιολογητών. Ενώ και το TF-IDF δείχνει εμπειρικά να τα πάει λίγο καλύτερα συγκριτικά με τον BERT. Πιο αναλυτικά στο πρώτο ζευγάρι αξιολογητών βλέπουμε ότι υπάρχει μέτρια συμφωνία για τον TF-IDF ουσιώδη με τον Doc2vec ενώ υπάρχει μικρή για τον BERT. Για το δεύτερο ζευγάρι υπάρχει σχεδόν τέλεια συμφωνία μεταξύ Doc2vec και μέτρια για BERT και TF-IDF. Στο τρίτο ζευγάρι υπάρχει ουσιώδη για τον Doc2vec και μέτρια για TF-IDF BERT. Το τέταρτο ζευγάρι έχει ουσιώδη συμφωνία με TF-IDF μέτρια με BERT και σχεδόν τέλεια με Doc2vec. Τέλος το ζευγάρι 5 των αξιολογητών έχει μέτρια συμφωνία και με τις 3 μεθόδους και αυτό οφείλεται και στην μέτρια συμφωνία που είχαν οι αξιολογητές 5 και 10 μεταξύ τους.

Στον πίνακα 5.4 βλέπουμε την κορυφαία απόδοση. Δηλαδή τα αποτελέσματα του αξιολογητή εκείνου που τα πήγαινε καλύτερα με τους αλγορίθμους. Ως όρο αξιολογητής 1 2 κλπ στον πίνακα εννούμε τον αξιολογητή εκείνον που επισημείωσε τα αντίστοιχα κομμάτια κειμένων αναφοράς που μοιράστηκαν. Βλέπουμε και εδώ ότι το Doc2vec τα πάει παντού καλύτερα έχοντας 3 σχεδόν τέλειες συμφωνίες. Επίσης ξανά εδώ παρατηρούμε ότι το TF-IDF τα πάει λίγο καλύτερα σε σχέση με τον BERT έχοντας μία ουσιώδη συμφωνία, ενώ ο BERT φτάνει μέχρι μέτρια συμφωνία ακόμα και στην κορυφαία απόδοση. Αυτό ίσως αν οφείλεται στο ότι χρησιμοποιήσαμε μόνο τους 512 πρώτους χαρακτήρες και όχι ολόκληρο το κείμενο μιας και το BERT δεν μπορεί να χειριστεί συμβολοσειρές με μέγεθος άνω των 512.

Πίνακας 5.4: Κορυφαία απόδοση

	TF-IDF	Doc2vec	BERT
Αξιολογητής 1	0.59	0.82	0.43
Αξιολογητής 2	0.65	0.86	0.54
Αξιολογητής 3	0.52	0.76	0.65
Αξιολογητής 4	0.75	0.91	0.69
Αξιολογητής 5	0.45	0.64	0.54

Τέλος, στον πίνακα 5.5 βλέπουμε τη λιγότερο καλή απόδοση, δηλαδή τους αξιολογητές εκείνους που συμφώνησαν λιγότερο με τους αλγορίθμους, για τα αντίστοιχα κομμάτια της αξιολόγησης. Βλέπουμε και εδώ ότι παντού τα πάει καλύτερα ο Doc2vec. Ενώ, και εδώ βλέπουμε από τον πίνακα ότι και το TF-IDF τα πάει καλύτερα από το BERT. Επίσης παρατηρούμε πόσο διαφέρουν οι τιμές του Cohen's kappa για το 5 κομμάτι της αξιολόγησης ανάμεσα στους σχολιαστές κάτι που εξηγεί και πίνακας 5.1 έχοντας μικρή συμφωνία μεταξύ τους.

Πίνακας 5.5: Λιγότερο καλή απόδοση

	TF-IDF	Doc2vec	BERT
Αξιολογητής 1	0.51	0.70	0.19
Αξιολογητής 2	0.51	0.78	0.48
Αξιολογητής 3	0.40	0.62	0.55
Αξιολογητής 4	0.71	0.87	0.41
Αξιολογητής 5	0.29	0.18	0.22

Για να μπορέσουμε να υποστηρίξουμε και στατιστικά αυτό που αναφέραμε προηγουμένως ότι παρατηρήσαμε από τον πίνακα, ότι ο TF-IDF τα πάει καλύτερα συγκριτικά με το BERT, θα πρέπει σε μελλοντική εργασία θα πρέπει μέσω στατιστικού test να καταρρίψουμε την υπόθεση ότι το BERT δείχνει να συμφωνεί περισσότερο με τους ανθρώπους σε σχέση με το TF-IDF.

Κεφάλαιο 6

Συμπεράσματα και μελλοντικές επεκτάσεις

6.1 Συμπεράσματα

Σε αυτήν την εργασία αναλύσαμε το θέμα της εύρεσης όμοιων νομικών υποθέσεων μέσω μεθόδων βασισμένων στο κείμενο. Έγιναν πειράματα σε σύνολα δεδομένων νομικών κειμένων, στα οποία αρχικά έγινε προ επεξεργασία. Στη συνέχεια δημιουργήσαμε τις αναπαραστάσεις των νομικών κειμένων στο διανυσματικό χώρο για να μπορέσουμε να υπολογίσουμε την ομοιότητα συνημιτόνου.

Έπειτα, επιβεβαιώσαμε μέσω ανάλυσης ευαισθησίας της ομοιότητας με βάση την αναπαράσταση ότι οι 3 γνωστές αναπαραστάσεις που χρησιμοποιήσαμε (TF-IDF, Doc2vec BERT) παράγουν διαφορετικές κατατάξεις. Αυτό δημιούργησε την ανάγκη να γίνει ανθρώπινη αξιολόγηση ώστε να δούμε ποια μέθοδος είναι καλύτερη σε σχέση με τη γνώμη του ανθρώπου.

Λόγω του μεγάλου όγκου του συνόλου δεδομένων αλλά και της δυσκολίας να αξιολογηθούν όλα τα ζευγάρια, μετασχηματίσαμε το αρχικό σύνολο δεδομένων σε ένα υποσύνολο ώστε να μπορέσει να γίνει η αξιολόγηση. Έτσι προτείναμε μία διαδικασία για ανθρώπινη αξιολόγηση των αποτελεσμάτων. Στο πλαίσιο αυτής της διαδικασίας δημιουργήθηκε ένα εργαλείο ώστε η αξιολόγηση της ομοιότητας νομικών κειμένων από ανθρώπους να γίνει πιο εύκολα.

Οι αξιολογητές μέσω του annotation tool επισημείωσαν την ομοιότητα νομικών κειμένων. Έτσι, συγκρίναμε τις επιλογές των αξιολογητών με αυτές των μεθόδων, θεωρώντας την κάθε μέθοδο ως εικονικό αξιολογητή.

Μέσω της αξιολόγησής παρατηρήσαμε ότι οι αυτόματες μετρικές ομοιότητας οδηγούν σε αρκετά καλό βαθμό σε συναφή ζευγάρια νομικών κειμένων και καταλήξαμε στο συμπέρασμα ότι η μέθοδος Doc2vec είναι πιο ευθυγραμμισμένη με την ανθρώπινη αντίληψη της σημασιολογικής ομοιότητας.

Οι πόροι του κοινού δικαίου για τους δικηγόρους, τους δικαστικούς, και ακόμη και το ευρύ κοινό, είναι κυρίως προσβάσιμοι μέσω ερωτημάτων σε βάσεις δεδομένων, ένας ξεπερασμένος και μη βέλτιστος τρόπος αναζήτησης λόγω των πολλών περίπλοκων που μπορεί να έχει μια υπόθεση. Αυτή και παρόμοια έρευνα, παρέχει την προσπάθεια για συνεισφορά στο κομμάτι της αξιολόγησης για να βελτιωθεί η αναζήτηση προηγούμενων υποθέσεων για τρέχουσες. Η εφαρμογή των γνώσεων που αποκτήθηκαν από αυτό το έργο, θα μπορούσε να κάνει μια πραγματική αλλαγή στον τρόπο χρήσης των πόρων του κοινού δικαίου. Στο μέλλον, το κοινό δίκαιο θα είναι πιο προσιτό στο ευρύ κοινό, καθώς και στους επαγγελματίες, και η απόκτηση σχετικών υποθέσεων ευκολότερη, καθιστώντας την υποστήριξη των υποθέσεων πιο απλή από ποτέ.

6.2 Μελλοντικές επεκτάσεις

Ως μελλοντική εργασία, είναι πολύ σημαντικό να χρησιμοποιηθούν νομικοί για αξιολογητές στη μέθοδο της αξιολόγησης. Επειδή είναι ειδικοί στο συγκεκριμένο τομέα, θα μπορούσαν να αξιολογήσουν με αποτελεσματικότερο τρόπο την ομοιότητα νομικών υποθέσεων. Έτσι θα μπορέσουμε να βγάλουμε πιο ασφαλή συμπεράσματα για τις αυτόματες μετρικές ομοιότητας.

Επίσης, θα ήταν χρήσιμο να δοκιμαστούν και άλλες τεχνικές για αναπαράσταση νομικών κειμένων όπως Latent Dirichlet Allocation(LDA) ή ιεραρχικοί τρόποι αναπαράστασης μέσω BERT. Το BERT όπως είδαμε δείχνει να μην τα πάει καλά με την ομοιότητα χρησιμοποιώντας στην αναπαράσταση μόνο τους 512 πρώτους χαρακτήρες. Οπότε, θα πρέπει να δοκιμαστεί η αναπαράσταση ολόκληρου του κειμένου μέσω BERT. Για το LDA θα ήταν θα ήταν καλή ιδέα το να δούμε του πόσο αποδοτικά θα τα πάει

συγκριτικά με τις υπόλοιπες τεχνικές καθώς τα δεδομένα (νομικές υποθέσεις) μπορούν να μοντελοποιηθούν ως μια κατανομή Θεμάτων Συζήτησης (topics), και κάθε θέμα με τη σειρά του ως κατανομή Λέξεων(Words). Θα πρέπει επίσης, μέσω στατιστικού τεστ να δούμε ποια από τις μεθόδους TF-IDF και BERT τα πάει καλύτερα όσον αφορά τη συμφωνία με τον άνθρωπο.

Επιπλέον μία ακόμα πρόκληση θα ήταν να χρησιμοποιηθούν άλλα μέρη στην αναπαράσταση των νομικών κειμένων μέσω μεθόδων για συνόψιση summarization και όχι να χρησιμοποιηθεί ολόκληρο το κείμενο.

Άλλοι μέθοδοι για μέτρηση ομοιότητας, όπως η ευκλείδεια απόσταση, θα είχε ενδιαφέρον να δοκιμαστούν ώστε να δούμε πόσο απόκλιση υπάρχει σε σχέση με την ομοιότητα συνημιτόνου που εμείς χρησιμοποιήσαμε.

Τέλος, να γίνει η αξιολόγηση που προτείναμε και στο δεύτερο σύνολο δεδομένων που χρησιμοποιήσαμε για τις αναπαραστάσεις και να υλοποιηθεί το εργαλείο σε εφαρμογή διαδικτύου ώστε να είναι εύκολα προσβάσιμο από τους νομικούς που θα κληθούν να κάνουν την αξιολόγηση.

References

[1] Stefanie Brüninghaus and Kevin D Ashley. Improving the representation of legal case texts with information extraction methods. In *Proceedings of the 8th international conference on Artificial Intelligence and Law*, pages 42–51, 2001.

[2] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*, 2013.

[4] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[5] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.

[6] Andrew M Dai, Christopher Olah, and Quoc V Le. Document embedding with paragraph vectors. *arXiv preprint arXiv:1507.07998*, 2015.

[7] Jey Han Lau and Timothy Baldwin. An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:1607.05368*, 2016.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[10] Alammar Jay. The illustrated transformer, 2018.

[11] Mike Schuster and Kaisuke Nakajima. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE, 2012.

[12] A Bryan Garner. Black's law dictionary. 10th available at westlaw blacks, p. 334., 2014.

[13] Ramon Lopez De Mantaras, David McSherry, Derek Bridge, David Leake, Barry Smyth, Susan Craw, Boi Faltings, Mary Lou Maher, MICHAEL T COX, Kenneth Forbus, et al. Retrieval, reuse, revision and retention in case-based reasoning. *The Knowledge Engineering Review*, 20(3):215–240, 2005.

[14] Chieh-Yuan Tsai and Chuang-Cheng Chiu. Developing a significant nearest neighbor search method for effective case retrieval in a cbr system. In *2009 International Association of Computer Science and Information Technology-Spring Conference*, pages 262–266. IEEE, 2009.

[15] Marc Van Opijnen and Cristiana Santos. On the concept of relevance in legal information retrieval. *Artificial Intelligence and Law*, 25(1):65–87, 2017.

[16] Peter Jackson, Khalid Al-Kofahi, Alex Tyrrell, and Arun Vachher. Information extraction from case law and retrieval of prior cases. *Artificial Intelligence*, 150(1-2):239–290, 2003.

[17] Filippo Galgani, Paul Compton, and Achim Hoffmann. Lexa: Building knowledge bases for automatic legal citation classification. *Expert Systems with Applications*, 42(17-18):6391–6407, 2015.

[18] Khalid Al-Kofahi, Alex Tyrrell, Arun Vachher, and Peter Jackson. A machine learning approach to prior case retrieval. In *Proceedings of the 8th international conference on Artificial intelligence and law*, pages 88–93, 2001.

[19] Elizabeth D Liddy. Natural language processing. *mdm*, 2001.

[20] P Hípola. G. salton, automatic text processing: The transformation analysis and retrieval of information by computer. *Procesamiento del Lenguaje Natural*, 10, 1991.

[21] Marie-Francine Moens. Innovative techniques for legal text retrieval. *Artificial Intelligence and Law*, 9(1):35, 2001.

[22] Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. Paper recommender systems: a literature survey. *International Journal on Digital Libraries*, 17(4):305–338, 2016.

[23] Juan Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142. Citeseer, 2003.

[24] Sushanta Kumar, P Krishna Reddy, V Balakista Reddy, and Aditya Singh. Similarity analysis of legal judgments. In *Proceedings of the Fourth Annual ACM Bangalore Conference*, pages 1–4, 2011.

[25] Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Aritsugi. Semantic cosine similarity. In *The 7th International Student Conference on Advanced Science and Technology ICAST*, page 1, 2012.

[26] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[27] Xin Jin and Jiawei Han. *K-Medoids Clustering*, pages 564–565. Springer US, Boston, MA, 2010.

[28] Ron Artstein. Inter-annotator agreement. In *Handbook of linguistic annotation*, pages 297–313. Springer, 2017.

[29] Anthony J Viera, Joanne M Garrett, et al. Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5):360–363, 2005.