

Ανάπτυξη Web εφαρμογής για patent classification search



ΔΙΕΘΝΕΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΗΣ ΕΛΛΑΔΟΣ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ
ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΕΥΦΥΕΙΣ ΤΕΧΝΟΛΟΓΙΕΣ ΔΙΑΔΙΚΤΥΟΥ

**Ανάπτυξη Web εφαρμογής για patent classification
search**

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΙΩΑΝΝΗ ΤΣΙΟΥΜΠΡΗ

Επιβλέπων : Μιχάλης Σαλαμπάσης

Θεσσαλονίκη, Ιούνιος 2024

Η σελίδα αυτή είναι σκόπιμα λευκή.



ΔΙΕΘΝΕΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΗΣ ΕΛΛΑΔΟΣ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ
ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΕΥΦΥΕΙΣ ΤΕΧΝΟΛΟΓΙΕΣ ΔΙΑΔΙΚΤΥΟΥ – WEB
INTELLIGENCE

Ανάπτυξη Web εφαρμογής για patent classification search

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΙΩΑΝΝΗ ΤΣΙΟΥΜΠΡΗ

Επιβλέπων : Μιχάλης Σαλαμπάσης
Καθηγητής ΔΙ.ΠΑ.Ε.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή στις 21 Ιουνίου 2024.

(Υπογραφή)

(Υπογραφή)

(Υπογραφή)

.....
Όνομα Επώνυμο
Choose an item. ΔΙ.ΠΑ.Ε.

.....
Όνομα Επώνυμο
Choose an item. ΔΙ.ΠΑ.Ε.

.....
Όνομα Επώνυμο
Choose an item. ΔΙ.ΠΑ.Ε.

Υπόδειγμα φύλλου τίτλου (μπροστινή σελίδα) του αντιτύπου ΜΔΕ

Θεσσαλονίκη, Ιούνιος 2024

(Υπογραφή)

.....

ΙΩΑΝΝΗΣ ΤΣΙΟΥΜΠΡΗΣ

ΠΛΗΡΟΦΟΡΙΚΗ ΕΑΠ

© 2024– All rights reserved

Περίληψη

Οι πατέντες, ένας από τους πυλώνες της ανθρώπινης εξέλιξης, είναι ιδέες υλοποιημένες ή περιγραφικές που προσδιορίζουν εξελίξεις στην τεχνολογία. Αυτές μπορεί να κυμαίνονται από συστήματα προώθησης διαστημικών σκαφών έως απλές τεχνικές προδιαγραφές σιδηροδρόμων. Παρόλο που οι πατέντες είναι σημαντικές για την πρόοδο, ο μεγάλος αριθμός τους δημιουργεί προβλήματα παραβίασης και αντιγραφής, καθιστώντας την διάκρισή τους ως νέες και μοναδικές, μια χρονοβόρα διαδικασία.

Λόγω της πολυπλοκότητας των τεχνικών περιγραφών, και του γεγονότος ότι οι πατέντες αποτελούν εθνικό κυριαρχικό δικαίωμα ενός κράτους, έχουν δημιουργηθεί γραφεία που εξετάζουν αν μια πατέντα είναι πραγματικά νέα και δεν «αντιγράφει» στοιχεία από άλλες ή επικαλύπτει άλλες. Αυτό απαιτεί χρόνο, με αποτέλεσμα πολλές πατέντες να περιμένουν για έλεγχο.

Σε αυτή την πρόκληση η Τεχνητή Νοημοσύνη δεν θα μπορούσε να είναι αμέτοχη και με χρήση αλγορίθμων Βαθιάς Μάθησης έχει κάνει σημαντική βήματα στην υποβοήθηση επίλυσης του συγκεκριμένου προβλήματος καθώς και άλλων που έχουν ανακύψει στο όλο οικοσύστημα των πατεντών.

Στόχος της διπλωματικής είναι η δημιουργία ενός εύχρηστου γραφικού περιβάλλοντος μέσω διαδικτυακής εφαρμογής, όπου οι χρήστες θα μπορούν να εισάγουν λέξεις-κλειδιά ή κείμενα αιτήσεων για νέες πατέντες και να βρίσκουν συναφείς κωδικούς ταξινόμησης που θα τους βοηθήσουν να αναζητήσουν παρόμοιες πατέντες. Το περιβάλλον θα επιτρέπει τη δυναμική επιλογή δικτύου, μοντέλου και άλλων παραμέτρων για την εξαγωγή συναφών πατεντών για περαιτέρω έλεγχο. Θα πρέπει να είναι λειτουργικό, προσβάσιμο και ευέλικτο, διευκολύνοντας την ανάλυση και ανεύρεση πατεντών με αυτοματοποιημένα εργαλεία τεχνητής νοημοσύνης.

Η ταξινόμηση ή αλλιώς η ανάθεση κωδικών σε λέξεις κλειδιά ή κείμενα που έχουν εξαχθεί από πατέντες στηρίζεται στο ερευνητικό έργο της υποψήφιας διδάκτορος Ελένης Καματέρη, η οποία χρησιμοποιεί βαθιά Νευρωνικά Δίκτυα και γλωσσικά μοντέλα για αυτόματη ανάλυση κειμένων πατεντών με βάση λέξεις-κλειδιά, εντοπίζοντας τους αντίστοιχους κωδικούς ταξινόμησης

Υπόδειγμα περίληψης (abstract) στα ελληνικά του αντιτύπου που υποβάλλεται στις βιβλιοθήκες

Λέξεις Κλειδιά: πατέντες, Βαθιά Μάθηση, γλωσσικά μοντέλα , γραφικό περιβάλλον ,δικτυακή εφαρμογή

Η σελίδα αυτή είναι σκόπιμα λευκή.

Abstract

Patents, one of the pillars of human advancement, are implemented or descriptive ideas that define technological developments. These can range from spacecraft propulsion systems to simple technical specifications for railways. Although patents are crucial for progress, their sheer number creates issues of infringement and duplication, making the distinction of new and unique patents a time-consuming process.

Due to the complexity of technical descriptions and the fact that patents are a sovereign right of a state, offices have been established to examine whether a patent is genuinely new and does not "copy" elements from others or overlap with others. This requires time, resulting in many patents waiting for review.

In facing this challenge, Artificial Intelligence, particularly using Deep Learning algorithms, has made significant strides in aiding the resolution of this specific problem as well as others arising within the patent ecosystem.

The aim of the thesis is to create a user-friendly graphical interface through a web application, where users can input keywords or texts from applications for new patents and find relevant classification codes that will help them search for similar patents. The interface will allow dynamic selection of the network, model, and other parameters for extracting relevant patents for further review. It should be functional, accessible, and flexible, facilitating the analysis and retrieval of patents with automated artificial intelligence tools.

The classification or assignment of codes to keywords or texts extracted from patents is based on the research work of PhD candidate Eleni Kamateri, who employs Deep Neural Networks and language models for the automatic analysis of patent texts based on keywords, identifying the corresponding classification codes.

Keywords: patents, deep learning, language models, user interface, web application

Η σελίδα αυτή είναι σκόπιμα λευκή.

Πίνακας περιεχομένων

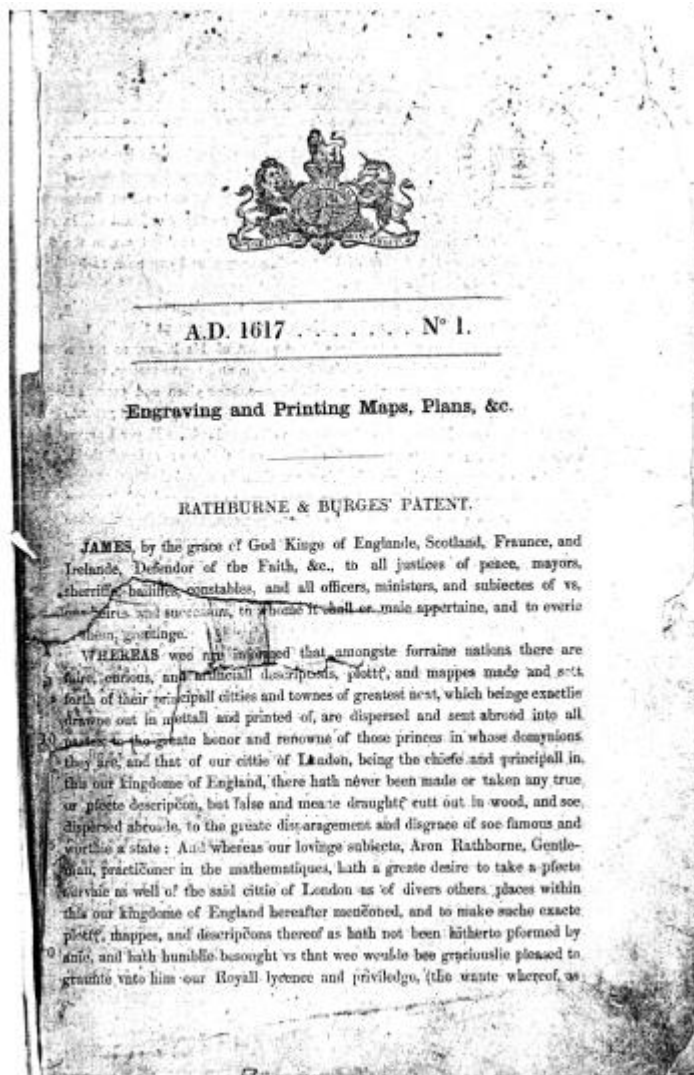
1	Εισαγωγή	1
1.1	Αντικείμενο διπλωματικής	3
1.2	Οργάνωση κειμένου	4
2	Πατέντες και ο ρόλος της Τεχνητής Νοημοσύνης	6
2.1	Η παρούσα κατάσταση σχετικά με τις πατέντες	6
2.2	Αυτοματοποιημένες διαδικασίες στην υποβολή μια νέας πατέντας	8
3	Γνωσιακό Υπόβαθρο.....	9
3.1	Εισαγωγή	9
3.1.1	<i>Νευρωνικά Δίκτυα</i>	<i>10</i>
3.1.2	<i>Σύντομη παρουσίαση των Νευρωνικών δικτύων που χρησιμοποιούνται στην εργασία που εξυπηρετεί η διπλωματική</i>	<i>13</i>
3.1.3	<i>Δίκτυα Μνήμης Μακράς Βραχείας Προθεσμίας (Long Short-Term Memory - LSTM)</i>	<i>13</i>
3.1.4	<i>Δίκτυα Επαναλαμβανόμενων Πυλών (Gated Recurrent Units - GRUs)....</i>	<i>14</i>
3.2	Γλωσσικά Μοντέλα	15
3.2.1	<i>Word Embeddings</i>	<i>15</i>
3.2.2	<i>Επισκόπηση Γλωσσικών μοντέλων</i>	<i>17</i>
3.2.3	<i>Word2Vec</i>	<i>21</i>
3.2.4	<i>FastText (300-D)</i>	<i>22</i>
3.2.5	<i>Risch and Krestel (300-D):.....</i>	<i>23</i>
3.2.6	<i>Συλλογές Πατεντών (Datasets).....</i>	<i>24</i>
4	Λειτουργικότητα Εφαρμογής	27
4.1	Εισαγωγή	27
4.2	Στοιχεία Λειτουργικότητας της εφαρμογής.....	27
4.3	Σχεδιασμός της εφαρμογής.....	29
5	Παρουσίαση της διπλωματικής.....	31
5.1	Το περιβάλλον ανάπτυξης	31
5.2	Δομή εφαρμογής	31

5.2.1	<i>Ανάπτυξη της εφαρμογής</i>	33
5.2.2	<i>Αρχική Παραμετροποίηση</i>	35
5.2.3	<i>Διακομιστής</i>	37
5.3	Σύστημα αυθεντικοποίησης.....	39
5.4	Η κύρια οθόνη.....	41
5.4.1	<i>Αποτελέσματα</i>	48
6	Λοιπές Αναφορές	54
6.1	Εισαγωγή	54
6.2	Στοιχεία Κώδικα	54
6.3	EspaceNet	57
7	Επίλογος	60
7.1	Σύνοψη και συμπεράσματα	60
7.2	Μελλοντικές επεκτάσεις.....	60
8	Βιβλιογραφία	62

1

Εισαγωγή

Οι πατέντες αποτελούν ένα θεμελιώδες στοιχείο για την προστασία της πνευματικής ιδιοκτησίας και την προώθηση της καινοτομίας. Μέσω των πατεντών, οι εφευρέτες εξασφαλίζουν αποκλειστικά δικαιώματα για την εκμετάλλευση των εφευρέσεών τους για ορισμένο χρονικό διάστημα, συνήθως 20 χρόνια, επιτρέποντάς τους να αποκομίσουν οικονομικά οφέλη και να έχουν ένα ανταγωνιστικό πλεονέκτημα στην αγορά. Ωστόσο, η διαδικασία ελέγχου και υποβολής πατεντών παρουσιάζει σημαντικές προκλήσεις.[1]



Εικ.1 Πατέντα εκδοθείσα στην Αγγλία το 1617 [1]

Η υποβολή μιας πατέντας απαιτεί λεπτομερή περιγραφή της εφεύρεσης, τεχνικές προδιαγραφές, καθώς και αποδείξεις ότι η εφεύρεση είναι καινοτόμος και δεν έχει προϋπάρξει. Η διαδικασία αυτή μπορεί να είναι χρονοβόρα και κοστοβόρα, καθώς περιλαμβάνει νομικές, τεχνικές και διοικητικές διαδικασίες. Επιπλέον, η επαλήθευση της καινοτομίας και η σύγκριση με προϋπάρχουσες πατέντες απαιτεί εξειδικευμένες γνώσεις και πολλές φορές εμπλέκει εκτενή αναζήτηση σε μεγάλες βάσεις δεδομένων πατεντών.

Σε επιστημονικό επίπεδο, υπάρχει έντονο ενδιαφέρον για την εφαρμογή τεχνολογιών βαθιάς μάθησης (deep learning) στην κατηγοριοποίηση και ανάλυση των πατεντών. Οι αλγόριθμοι βαθιάς μάθησης έχουν τη δυνατότητα να επεξεργάζονται μεγάλες ποσότητες δεδομένων, να αναγνωρίζουν πρότυπα και να ταξινομούν πληροφορίες με υψηλή ακρίβεια. Αυτό μπορεί να συμβάλει σημαντικά στην αυτοματοποίηση της διαδικασίας ελέγχου των πατεντών, καθιστώντας την πιο αποδοτική και αξιόπιστη.

Χρησιμοποιώντας βαθιά μάθηση[2][3], είναι δυνατόν να αναλυθούν τεχνικά κείμενα και να κατηγοριοποιηθούν οι πατέντες βάσει του περιεχομένου τους, διευκολύνοντας την ανεύρεση σχετικών πληροφοριών και την σύγκριση με προϋπάρχουσες πατέντες. Επιπλέον, μπορεί να συμβάλει στην αναγνώριση νέων τεχνολογικών τάσεων και στην πρόβλεψη του δυνητικού αντίκτυπου νέων εφευρέσεων στην αγορά.

Αυτό είναι και το αντικείμενο της υποψήφιας διδάκτορος Ελένη Καματέρη που χρησιμοποιώντας εξελιγμένα Νευρωνικά δίκτυα (LSTM, GRU) και γλωσσικά μοντέλα δέχεται σαν είσοδο λέξεις κλειδιά που θα μπορούσαν να είναι οι βασικές έννοιες για κάποια πατέντα και βρίσκει τις αντιστοιχίες μεταξύ αυτών και συλλογών από πατέντες.

1.1 Αντικείμενο διπλωματικής

Η προαναφερθείσα εργασία, αν και παράγει εξαιρετικά αποτελέσματα στην αναζήτηση πατεντών, χρειαζόταν ένα εύχρηστο περιβάλλον όπου ο τελικός χρήστης θα είχε δυνατότητα για εύκολη χρήση χωρίς να χρειάζεται να καταφύγει στον κώδικα.

Η παρούσα διπλωματική εργασία στοχεύει να καλύψει αυτό το «κενό», δημιουργώντας μια διαδικτυακή εφαρμογή που επιτρέπει στους χρήστες να εισάγουν λέξεις-κλειδιά και να αναζητούν σχετικές πατέντες. Στηριζόμενη στην προαναφερθείσα εφαρμογή αξιοποιεί τις δυνατότητες της, παρέχοντας όλες τις δυνατές παραμέτρους για εύρυθμη λειτουργία και έχει αναπτυχθεί με γνώμονα την πλήρη υποστήριξη της εφαρμογής όπως έχει αλλά παράλληλα βασιζόμενη στην δυναμικότητα της έχει κάνει πρόβλεψη για ευκολία σε προσθήκες και επεκτάσεις εφόσον αυτές γίνουν διαθέσιμες και στις περισσότερες περιπτώσεις χωρίς αλλαγές στον κώδικα.

Συγκεκριμένα η δομή της εφαρμογής επιτρέπει τη δυναμική προσθήκη και άλλων μοντέλων Βαθιάς Μάθησης, γλωσσικών μοντέλων και λοιπών παραμέτρων, εξασφαλίζοντας ευελιξία και προσαρμοστικότητα στην μελλοντική ανάπτυξη της εφαρμογής αλλά και στις συνεχώς εξελισσόμενες ανάγκες των χρηστών. Με αυτόν τον τρόπο, οι χρήστες θα έχουν τη δυνατότητα να πραγματοποιούν αποτελεσματικές και στοχευμένες αναζητήσεις, εξασφαλίζοντας ότι θα βρίσκουν τις πατέντες που σχετίζονται περισσότερο με τα ενδιαφέροντα και τις ανάγκες τους.

Η διαδικτυακή εφαρμογή υλοποιήθηκε σε Python, μια γλώσσα προγραμματισμού που είναι ευρέως χρησιμοποιούμενη για την ανάπτυξη εφαρμογών μηχανικής μάθησης και ανάλυσης δεδομένων λόγω της ευκολίας στη χρήση και της πλούσιας συλλογής βιβλιοθηκών.

Συγκεκριμένα, θα χρησιμοποιηθεί το framework Flask για την ανάπτυξη του περιβάλλοντος της εφαρμογής. Το Flask είναι ένα μικρό αλλά ισχυρό web framework που επιτρέπει την εύκολη ανάπτυξη και διαχείριση εφαρμογών web, παρέχοντας μεγάλη ευελιξία και επεκτασιμότητα.

Η βάση δεδομένων που θα χρησιμοποιηθεί είναι η SQLite, μια ελαφριά και εύκολη στη χρήση βάση δεδομένων, ιδανική για τη διαχείριση των δεδομένων που απαιτούνται από την εφαρμογή.

Επιπλέον, η εφαρμογή παρέχει διασύνδεση με το ESPACENET, μια πλατφόρμα που παρέχει πρόσβαση σε παγκόσμια δεδομένα πατεντών. Η διασύνδεση αυτή επιτρέπει στους χρήστες να έχουν πρόσβαση στην αναλυτική περιγραφή των πατεντών όπως εξαγονται από τα νευρωνικά δίκτυα ενώ παρέχεται και δυνατότητα σύγκρισης.

1.2 Οργάνωση κειμένου

Η δομή της παρούσας διπλωματικής ακολουθεί έναν οργανωμένο τρόπο προσέγγισης του θέματος, όπως περιγράφεται παρακάτω:

Το Κεφάλαιο 2 παρέχει μια επισκόπηση του οικοσυστήματος των πατεντών και της ανάγκης για τη δημιουργία ενός εύχρηστου μέσου που θα διευκολύνει την εύρεση και την αντιστοίχιση πατεντών. Αναλύονται οι προϋποθέσεις που πρέπει να πληροί μια πατέντα για να υποβληθεί και η σημασία της μη επικάλυψης με άλλες πατέντες.

Στο Κεφάλαιο 3 πραγματοποιείται μια ανάλυση των νευρωνικών δικτύων ως μέσο επεξεργασίας των πατεντών. Εξετάζεται η εφαρμογή τους στην επεξεργασία συλλογών πατεντών με σκοπό την εύρεση αντιστοιχιών μεταξύ αυτών. Επιπλέον, αναφέρονται τα datasets που χρησιμοποιούνται για την εκπαίδευση των μοντέλων και τα γλωσσικά μοντέλα που ενσωματώνονται για την βελτίωση της ανάλυσης του κειμένου.

Στο Κεφάλαιο 4, η επικέντρωση είναι στην παρουσίαση της δικτυακής εφαρμογής που αναπτύσσεται για την υποβολή λέξεων-κλειδιών και την αναζήτηση σχετικών πατεντών. Πρώτα, παρουσιάζεται ο τρόπος λειτουργίας της εφαρμογής, από τη στιγμή που ο χρήστης εισάγει τις λέξεις-κλειδιά μέχρι την εμφάνιση των αποτελεσμάτων. Έπειτα, αναλύονται τα βήματα της διαδικασίας αναζήτησης, συμπεριλαμβανομένης της ανάκτησης δεδομένων από τη βάση δεδομένων, την επεξεργασία των δεδομένων με χρήση των νευρωνικών μοντέλων και την παρουσίαση των αποτελεσμάτων στον χρήστη. Επιπλέον, αναφέρονται οι τεχνολογίες

και οι βιβλιοθήκες που χρησιμοποιούνται για την υλοποίηση της εφαρμογής, με έμφαση στην επιλογή του framework Flask για τη δημιουργία του server-side λογισμικού.

Στο Κεφάλαιο 5, παρουσιάζεται μια λεπτομερής ανάλυση του διαγράμματος ροής της εφαρμογής. Κάθε στάδιο της διαδικασίας αναζήτησης πατεντών αναλύεται αναλυτικά, περιγράφοντας τη λειτουργία του και τον τρόπο που συνδέεται με τα υπόλοιπα στάδια. Επιπλέον, παρέχεται μια πλήρης παρουσίαση της εφαρμογής σε κάθε σημείο, περιγράφοντας την αλληλεπίδραση του χρήστη με τη διεπαφή και τις λειτουργίες που είναι διαθέσιμες. Τέλος, γίνεται ανάλυση του τρόπου λειτουργίας της εφαρμογής σε πραγματικό χρόνο, παρουσιάζοντας παραδείγματα χρήσης και δίνοντας οδηγίες για τη βελτιστοποίηση της απόδοσης και της ακρίβειας. Τέλος, γίνεται συνοπτική αναφορά στις δυνατότητες επέκτασης της εφαρμογής και σε πιθανές μελλοντικές επεκτάσεις που μπορούν να προστεθούν για τη βελτίωση της λειτουργικότητάς της.

Στο κεφάλαιο 6, γίνονται αναφορές σε τμήματα κώδικα που παρουσιάζουν για συγκεκριμένους λόγους κάποιο ιδιαίτερο ενδιαφέρον και παρουσιάζεται το EspaceNet και ο ρόλος του στο οικοσύστημα πατεντών.

Στο κεφάλαιο 7, είναι ο επίλογος της εργασίας με προτάσεις και σκέψεις για το μέλλον τόσο όσον αφορά την παρούσα εργασία όσο και γενικά την συνδρομή της Τεχνητής Νοημοσύνης γενικά όσον αφορά τις πατέντες.

Κεφάλαιο 8 είναι η βιβλιογραφία.

2

Πατέντες και ο ρόλος της Τεχνητής Νοημοσύνης

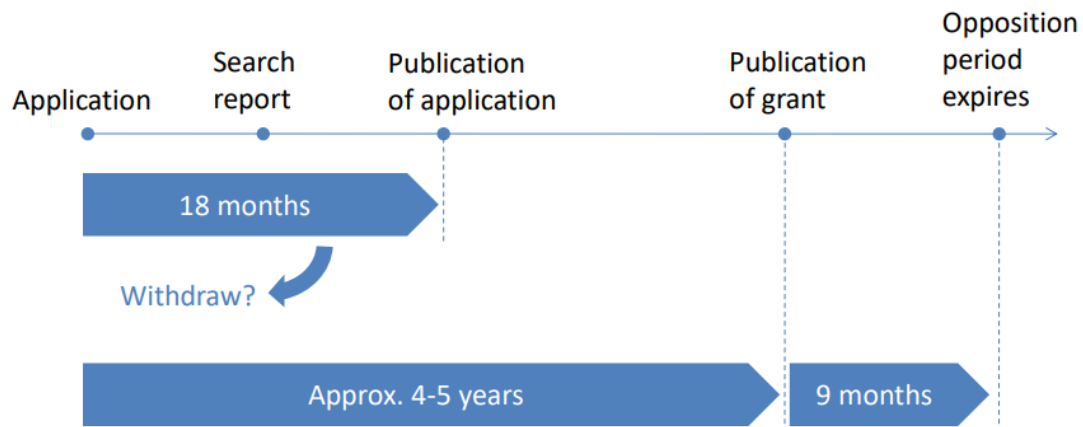
2.1 Η παρούσα κατάσταση σχετικά με τις πατέντες

Μια πατέντα αντιπροσωπεύει το δικαίωμα προστασίας ενός νέου προϊόντος, μιας νέας μεθόδου, ή μιας νέας διαδικασίας που έχει επινοηθεί από ένα άτομο ή μια εταιρεία. Αυτή η προστασία παρέχει στον κάτοχο της το δικαίωμα να αποκτήσει αποκλειστική χρήση και εκμετάλλευση της εφευρέσής του για έναν ορισμένο χρονικό διάστημα, κάνοντας την εφεύρεση οικονομικά ανταγωνιστική και προστατεύοντας την από αντιγραφή ή κλοπή.

Παρότι οι πατέντες παρέχουν προστασία στους εφευρέτες, η διαδικασία αναζήτησης και αντιστοίχισης πατεντών μπορεί να αποτελέσει μια πρόκληση. Ο όγκος των υαρχόντων πατεντών είναι τεράστιος και συνεχώς αυξάνεται, καθιστώντας δύσκολη την εύρεση της ακριβούς πληροφορίας που χρειάζεται ένα άτομο ή μια εταιρεία για να προχωρήσει με την εφεύρεσή τους. Επιπλέον, η ανάγκη για μη επικάλυψη με άλλες πατέντες απαιτεί μια εκτεταμένη αναζήτηση πριν από την υποβολή μιας νέας πατέντας.

Η διαδικασία υποβολής πατεντών αποτελεί ένα σύνθετο και χρονοβόρο έργο που απαιτεί προσεκτική προετοιμασία και τεκμηρίωση. Οι εφευρέτες ή οι εταιρείες που επιθυμούν να υποβάλουν μια πατέντα πρέπει να ακολουθήσουν ένα συγκεκριμένο σύνολο βήματα για να εξασφαλίσουν ότι η εφεύρεσή τους πληροί τις απαιτήσεις του νόμου περί πατεντών και είναι έτοιμη για υποβολή.

The patent procedure at the EPO



Εικ2. Τυπικό χρονοδιάγραμμα υποβολής πατέντας [1]

Η διαδικασία ξεκινά με την κατάθεση μιας αίτησης πατέντας σε κάποιο αρμόδιο γραφείο πατεντών ή οργανισμό προστασίας πνευματικής ιδιοκτησίας. Σε αυτό το στάδιο, η εφεύρεση περιγράφεται με λεπτομέρεια και συνοδεύεται από σχέδια ή διαγράμματα που επιδεικνύουν τον τρόπο λειτουργίας της. Η αίτηση πρέπει επίσης να περιλαμβάνει δηλώσεις που να περιγράφουν το πεδίο εφαρμογής της εφεύρεσης και την καινοτομία που αυτή προσφέρει στον τομέα της.

Μετά την υποβολή, η αίτηση περνά από μια σειρά από ελέγχους και διαδικασίες εξέτασης από το γραφείο πατεντών, όπου εξετάζεται η νομιμότητα και η νεότητα της εφεύρεσης. Αν όλες οι απαιτήσεις πληρούνται, η πατέντα εκδίδεται και ο υποβάλλων αποκτά το δικαίωμα προστασίας για μια περίοδο χρόνου, συνήθως για περίπου 20 χρόνια από την ημερομηνία υποβολής.

Σε παγκόσμιο επίπεδο, υπάρχουν πολλά γραφεία πατεντών που αναλαμβάνουν την υποβολή και εξέταση αιτήσεων πατεντών. Αυτά τα γραφεία προσφέρουν εξειδικευμένες υπηρεσίες για την υποστήριξη εφευρετών και εταιρειών σε ολόκληρη τη διαδικασία

2.2 Αυτοματοποιημένες διαδικασίες στην υποβολή μια νέας πατέντας

Γίνεται εύκολα κατανοητό ότι το παρόν τοπίο αποτελεί μια «αναπόφευκτη» τροχοπέδη στην ανθρώπινη εξέλιξη καθώς μια λύση από την στιγμή που προτείνεται μέχρι την εφαρμογή της μπορεί να απαιτηθούν χρόνια μέχρι και ο χρόνος είναι πάντα εχθρός στην πρόοδο.

Ήταν φυσικό συνεπακόλουθο καθώς οι παραδοσιακές μέθοδοι υποβολής πατεντών χωλαίνουν να υπάρχει έντονο ερευνητικό ενδιαφέρον στον χώρο της Τεχνητής Νοημοσύνης για τη βελτίωση και την υποστήριξη αυτής της διαδικασίας. Η χρήση τεχνικών Τεχνητής Νοημοσύνης, όπως οι αλγόριθμοι μηχανικής μάθησης και η επεξεργασία φυσικής γλώσσας, έχει ανοίξει νέους ορίζοντες για τη βελτίωση της διαδικασίας υποβολής πατεντών.

Με τη χρήση τεχνικών μηχανικής μάθησης, μπορεί να δημιουργηθούν αλγόριθμοι που αναλύουν τον όγκο των υπαρχόντων πατεντών και παρέχουν αυτόματα προτάσεις για την κατηγοριοποίηση και την αξιολόγηση της νεότερης εφεύρεσης.

Το ερευνητικό έργο στον τομέα της Τεχνητής Νοημοσύνης για την υποστήριξη στη διαδικασία υποβολής πατεντών αναμένεται να συνεχιστεί και να εξελιχθεί περαιτέρω, προσφέροντας νέες και καινοτόμες λύσεις που θα βελτιώσουν την αποτελεσματικότητα και την ακρίβεια της διαδικασίας υποβολής πατεντών και θα συμβάλλουν στην επιτάχυνση της.

3

Γνωσιακό Υπόβαθρο

3.1 Εισαγωγή

Η χρήση βαθιάς μάθησης για αναζήτηση πατεντών βάσει λέξεων-κλειδιών αποτελεί μια σύγχρονη και καινοτόμα προσέγγιση που μπορεί να βελτιώσει σημαντικά την ακρίβεια και την αποτελεσματικότητα των αναζητήσεων σε βάσεις δεδομένων πατεντών. Ακολουθούν ορισμένα βασικά σημεία για την υλοποίηση και τα οφέλη αυτής της τεχνολογίας:

Υλοποίηση

1. Συλλογή Δεδομένων:

Συλλογή μεγάλου όγκου δεδομένων πατεντών από δημόσιες και ιδιωτικές βάσεις δεδομένων.

Προεπεξεργασία των κειμένων πατεντών για καθαρισμό και κανονικοποίηση δεδομένων.

2. Εκπαίδευση Μοντέλου:

Χρήση προκατασκευασμένων γλωσσικών μοντέλων, όπως το BERT ή το GPT, που έχουν ήδη εκπαιδευτεί σε μεγάλο όγκο κειμένου.

Πρόσθετη εκπαίδευση (fine-tuning) αυτών των μοντέλων χρησιμοποιώντας δεδομένα πατεντών για να μάθουν τις ιδιαιτερότητες της γλώσσας που χρησιμοποιείται στις πατέντες.

3. Δημιουργία Λεξικών και Ανάκτηση Πληροφορίας:

Χρήση λέξεων-κλειδιών και φράσεων για δημιουργία δυναμικών αναζητήσεων μέσω ενσωματωμένων διανυσμάτων (embeddings).

Ανάκτηση σχετικών πατεντών με βάση την σημασιολογική ομοιότητα των περιεχομένων τους με τις λέξεις-κλειδιά.

4. Βελτιστοποίηση και Αξιολόγηση:

Χρήση μεθόδων βελτιστοποίησης για τη βελτίωση της ακρίβειας του μοντέλου.

Αξιολόγηση της απόδοσης του συστήματος με μετρικές όπως η ακρίβεια (precision), η ανάκληση (recall), και το F1-score.

Οφέλη

1. Αυξημένη Ακρίβεια:

Τα μοντέλα βαθιάς μάθησης μπορούν να κατανοήσουν την σημασιολογία των κειμένων και να αναγνωρίσουν σχετικές πατέντες ακόμα και αν χρησιμοποιούνται διαφορετικές λέξεις για την περιγραφή παρόμοιων εννοιών.

2. Αυτοματοποιημένη Επεξεργασία:

Αυτοματοποίηση της διαδικασίας αναζήτησης, μείωση του χρόνου και της ανθρώπινης προσπάθειας που απαιτείται για την εύρεση σχετικών πατεντών.

3. Καλύτερη Αντιστοίχιση:

Χρήση εξελιγμένων τεχνικών επεξεργασίας φυσικής γλώσσας για την αντιστοίχιση λέξεων-κλειδίων με τα πιο σχετικά αποτελέσματα.

4. Συνεχής Μάθηση και Βελτίωση:

Τα μοντέλα μπορούν να βελτιώνονται συνεχώς με νέες δεδομένες πατέντες και ανατροφοδότηση από χρήστες, βελτιώνοντας την απόδοσή τους με την πάροδο του χρόνου.

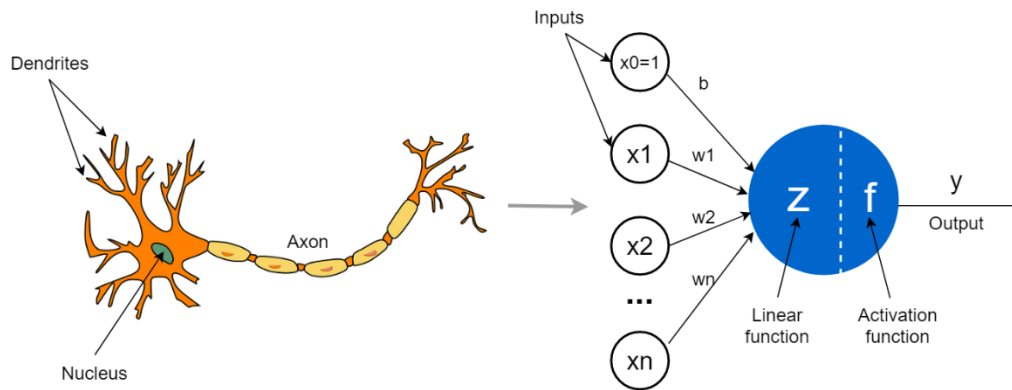
Η χρήση βαθιάς μάθησης για την αναζήτηση πατεντών βάσει λέξεων-κλειδίων είναι ένα πολλά υποσχόμενο πεδίο που μπορεί να φέρει σημαντικές βελτιώσεις στην ταχύτητα και την ακρίβεια της έρευνας πατεντών, επιτρέποντας στους ερευνητές και τις εταιρείες να εντοπίζουν πιο αποτελεσματικά καινοτόμες ιδέες και εφευρέσεις.

3.1.1 Νευρωνικά Δίκτυα

Τα νευρωνικά δίκτυα αναπτύχθηκαν εμπνευσμένα από τον τρόπο λειτουργίας του ανθρώπινου εγκεφάλου και αποτελούν ένα ισχυρό εργαλείο στον χώρο της τεχνητής

νοημοσύνης και της μηχανικής μάθησης. Είναι μαθηματικά μοντέλα που αποτελούνται από συνδεδεμένους κόμβους, γνωστούς ως νευρώνες, τα οποία λειτουργούν παράλληλα και αλληλεπιδρούν για την επεξεργασία πληροφοριών.

Η βασική μονάδα του νευρωνικού δικτύου είναι ο νευρώνας.



Εικ3 Φυσικός Νευρώνας και ο αντίστοιχος σε Αλγόριθμο [4]

Κάθε νευρώνας λαμβάνει εισόδους, εκτελεί υπολογισμούς με βάση αυτές τις εισόδους και το συνδυάζει με ένα βάρος, παράγοντας ένα ενδιάμεσο αποτέλεσμα. Αυτό το ενδιάμεσο αποτέλεσμα στη συνέχεια περνάει από μια συνάρτηση ενεργοποίησης, η οποία καθορίζει εάν και πόσο θα ενεργοποιηθεί ο νευρώνας. Το τελικό αποτέλεσμα εξάγεται και μεταδίδεται στους νευρώνες της επόμενης στρώσης.

Τα νευρωνικά δίκτυα χωρίζονται συνήθως σε δύο κατηγορίες βάσει της δομής τους: τα τροφοδοτούμενα με δεδομένα εμπρός (feedforward) και τα αναδρομικά (recurrent). Στα τροφοδοτούμενα με δεδομένα εμπρός, η πληροφορία κινείται μόνο προς τη μία κατεύθυνση, από την είσοδο προς την έξοδο, ενώ στα αναδρομικά δίκτυα, η πληροφορία μπορεί να κινείται και προς τις δύο κατευθύνσεις, επιτρέποντας την αναδρομική συνδεσιμότητα μεταξύ νευρώνων και στρωμάτων.

Οι νευρωνικές αρχιτεκτονικές έχουν εξελιχθεί σε μια πληθώρα μοντέλων που καλύπτουν διάφορες εφαρμογές και προβλήματα μηχανικής μάθησης. Ακολουθούν ορισμένες από τις πιο γνωστές αρχιτεκτονικές νευρωνικών δικτύων:

Τροφοδοτούμενα Προς Εμπρός Δίκτυα (Feedforward Neural Networks): Αυτή η αρχιτεκτονική αποτελείται από επίπεδα νευρώνων που συνδέονται με γραμμικές ή μη γραμμικές συναρτήσεις ενεργοποίησης. Τα δεδομένα ρέουν μόνο προς μία κατεύθυνση από το επίπεδο εισόδου προς το επίπεδο εξόδου.

Συνελκτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks - CNNs): Τα CNNs έχουν σχεδιαστεί για την αναγνώριση προτύπων σε εικόνες και βίντεο. Χρησιμοποιούν συνελκτικά επίπεδα για να ανιχνεύσουν χαρακτηριστικά σε διαφορετικές περιοχές της εικόνας.

Αναδρομικά Νευρωνικά Δίκτυα (Recurrent Neural Networks - RNNs): Τα RNNs είναι σχεδιασμένα για την επεξεργασία ακολουθιών δεδομένων, όπως χρονοσειρές και φυσική γλώσσα. Κάθε νευρώνας σε ένα RNN διατηρεί μια εσωτερική κατάσταση που λειτουργεί ως μνήμη κατά την επεξεργασία νέων εισόδων.

Δίκτυα Μνήμης Μακράς Βραχείας Προθεσμίας (Long Short-Term Memory - LSTM): Τα LSTM είναι ένα είδος αναδρομικών νευρωνικών δικτύων που διατηρούν μια μακροχρόνια μνήμη και είναι ιδανικά για εφαρμογές που απαιτούν την κατανόηση μακροπρόθεσμων εξαρτήσεων στα δεδομένα.

Δίκτυα Επαναλαμβανόμενων Πυλών (Gated Recurrent Units - GRUs) είναι μια παραλλαγή των αναδρομικών νευρωνικών δικτύων (RNNs) και ανήκουν στην κατηγορία των μοντέλων μηχανικής μάθησης που χρησιμοποιούνται για την ανάλυση ακολουθιακών δεδομένων. Τα GRUs παρόμοια με τα LSTM, αποτελούν μια πιο προηγμένη μορφή RNNs, η οποία έχει σχεδιαστεί για να αντιμετωπίζει ορισμένα προβλήματα απόδοσης που αντιμετωπίζουν τα LSTM, και ταυτόχρονα έχουν απλούστερη δομή. Τα GRUs αποτελούνται από μεμονωμένες "πύλες" που ελέγχουν τη ροή των πληροφοριών εντός του δικτύου. Οι πύλες αυτές περιλαμβάνουν μια πύλη ενημέρωσης και μια πύλη επαναφοράς. Η πύλη ενημέρωσης (update gate) αποφασίζει πόσο από την προηγούμενη κατάσταση θα διατηρηθεί. Η πύλη επαναφοράς (reset gate) αποφασίζει πόσο από την προηγούμενη κατάσταση θα ξεχαστεί. Αυτές οι πύλες επιτρέπουν στα GRUs να μάθουν να εκπροσωπούν αλληλεπιδράσεις μεταξύ των στοιχείων της ακολουθίας, παρέχοντας ταυτόχρονα μια λύση στο πρόβλημα της εξαφάνισης ή της εκρηκτικής αύξησης της κλίμακας των αναδρομικών συνελκτικών δικτύων. Τα GRUs χρησιμοποιούνται ευρέως σε εφαρμογές όπως η αναγνώριση φωνής, η ανάλυση ακολουθιών κειμένου, η μετάφραση γλωσσών, η ανάλυση συναισθήματος κ.λπ. Και η απλότητά τους και η λιγότερη πολυπλοκότητα σε σύγκριση με τα LSTM τα καθιστούν πολύ αποδοτικά σε πολλές εφαρμογές.

3.1.2 Σύντομη παρουσίαση των Νευρωνικών δικτύων που χρησιμοποιούνται στην εργασία που εξυπηρετεί η διπλωματική

Θα γίνει σύντομη παρουσίαση των νευρωνικών δικτύων[5] που χρησιμοποιήθηκαν στην εφαρμογή που υλοποιεί την λειτουργικότητα ανεύρεσης των πατεντών

3.1.3 Δίκτυα Μνήμης Μακράς Βραχείας Προθεσμίας (Long Short-Term Memory - LSTM)

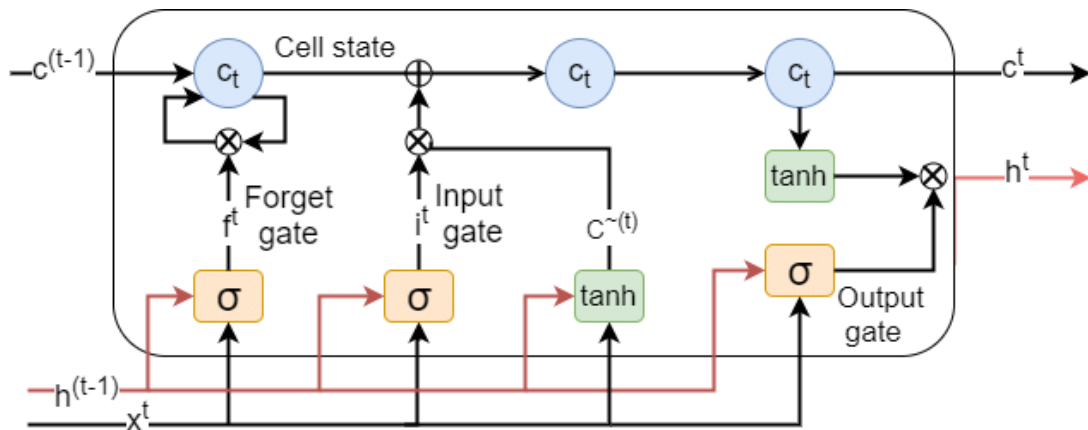
Τα Δίκτυα Μνήμης Μακράς Βραχείας Προθεσμίας (Long Short-Term Memory - LSTM) είναι ένα είδος αναδρομικών νευρωνικών δικτύων (RNNs) και αποτελούν μια ισχυρή αρχιτεκτονική για την επεξεργασία ακολουθιών δεδομένων. Αναπτύχθηκαν για να αντιμετωπίσουν το πρόβλημα της εξαφάνισης της κλίμακας στα απλά RNNs, το οποίο εμποδίζει την αποτελεσματική μάθηση μακροπρόθεσμων εξαρτήσεων σε ακολουθίες δεδομένων.

Οι LSTM διαθέτουν ένα μηχανισμό πύλης που αποτελείται από τρεις κύριες πύλες: την πύλη εισόδου (input gate), την πύλη εξόδου (output gate), και την πύλη λήθης (forget gate). Αυτές οι πύλες ελέγχουν τη ροή των πληροφοριών εντός του δικτύου και επιτρέπουν στα LSTM να διατηρούν μια μακροχρόνια μνήμη και να μάθουν μακροπρόθεσμες εξαρτήσεις στα δεδομένα.

Η πύλη εισόδου αποφασίζει ποιες νέες πληροφορίες θα εισαχθούν στην εσωτερική κατάσταση του κελιού μνήμης. Η πύλη εξόδου αποφασίζει ποια μέρος της εσωτερικής κατάστασης θα εξαχθεί ως έξοδος του κελιού μνήμης. Η πύλη λήθης ελέγχει ποιες πληροφορίες θα διατηρηθούν ή θα ξεχαστούν από την εσωτερική κατάσταση του κελιού μνήμης.

Ένας σημαντικός λόγος για τη δημοτικότητα των LSTM είναι η ικανότητά τους να αντιμετωπίζουν προβλήματα που αφορούν την αναγνώριση, την πρόβλεψη και τη δημιουργία ακολουθιών δεδομένων, όπως η φυσική γλώσσα, η μουσική σύνθεση, η αναγνώριση προτύπων σε εικόνες και βίντεο, κ.λπ.

Συνολικά, τα LSTM έχουν ευρεία εφαρμογή σε πολλούς τομείς, όπου απαιτείται η αναγνώριση και η πρόβλεψη πολύπλοκων ακολουθιών δεδομένων.



Εικ4. Τοπική δομή του Νευρωνικού Δικτύου LSTM

3.1.4 Δίκτυα Επαναλαμβανόμενων Πυλών (Gated Recurrent Units - GRUs)

Τα Δίκτυα Επαναλαμβανόμενων Πυλών (Gated Recurrent Units - GRUs) είναι μια παραλλαγή των αναδρομικών νευρωνικών δικτύων (RNNs) και ανήκουν στην κατηγορία των μοντέλων μηχανικής μάθησης που χρησιμοποιούνται για την ανάλυση ακολουθιακών δεδομένων.

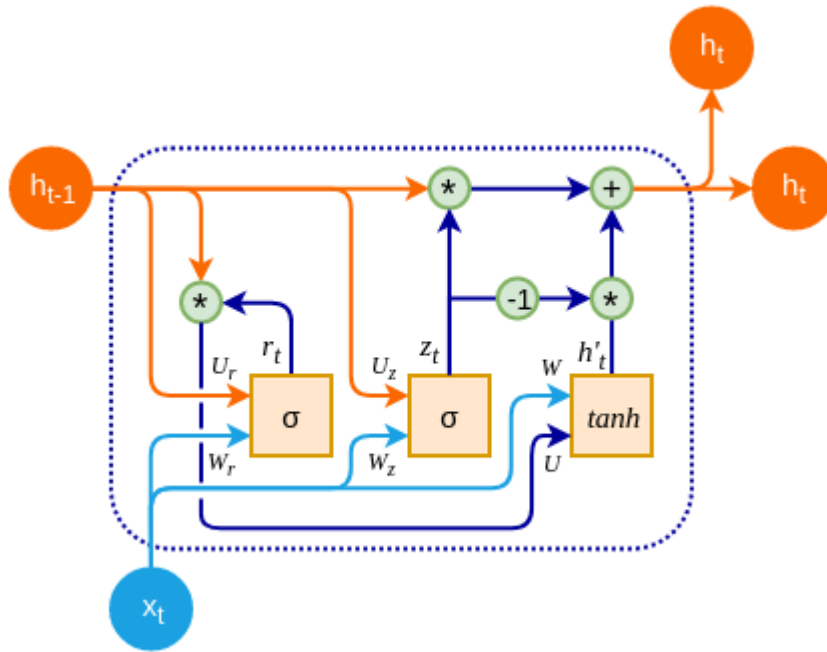
Τα GRUs παρόμοια με τα LSTM, αποτελούν μια πιο προηγμένη μορφή RNNs, η οποία έχει σχεδιαστεί για να αντιμετωπίζει ορισμένα προβλήματα απόδοσης που αντιμετωπίζουν τα LSTM, και ταυτόχρονα έχουν απλούστερη δομή. Τα GRUs αποτελούνται από μεμονωμένες "πύλες" που ελέγχουν τη ροή των πληροφοριών εντός του δικτύου. Οι πύλες αυτές περιλαμβάνουν μια πύλη ενημέρωσης και μια πύλη επαναφοράς.

Η πύλη ενημέρωσης (update gate) αποφασίζει πόσο από την προηγούμενη κατάσταση θα διατηρηθεί.

Η πύλη επαναφοράς (reset gate) αποφασίζει πόσο από την προηγούμενη κατάσταση θα ξεχαστεί.

Αυτές οι πύλες επιτρέπουν στα GRUs να μάθουν να εκπροσωπούν αλληλεπιδράσεις μεταξύ των στοιχείων της ακολουθίας, παρέχοντας ταυτόχρονα μια λύση στο πρόβλημα της εξαφάνισης ή της εκρηκτικής αύξησης της κλίμακας των αναδρομικών συνελκτικών δικτύων.

Τα GRUs χρησιμοποιούνται ευρέως σε εφαρμογές όπως η αναγνώριση φωνής, η ανάλυση ακολουθιών κειμένου, η μετάφραση γλωσσών, η ανάλυση συναισθήματος κ.λπ. Και η απλότητά τους και η λιγότερη πολυπλοκότητα σε σύγκριση με τα LSTM τα καθιστούν πολύ αποδοτικά σε πολλές εφαρμογές.

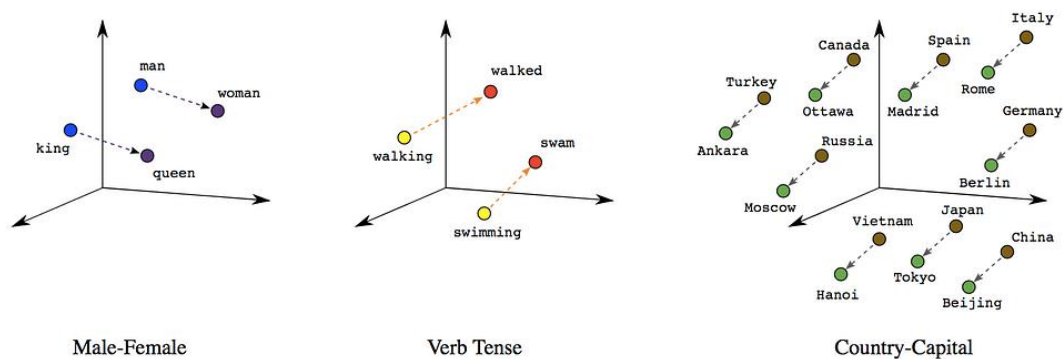


Εικ5. Τυπική δομή του Νευρωνικού δικτύου GRU

3.2 Γλωσσικά Μοντέλα

3.2.1 Word Embeddings

Τα Word embeddings [6] είναι μαθηματικές αναπαραστάσεις των λέξεων σε ένα συνεχές, πολυδιάστατο διανυσματικό χώρο. Αυτές οι αναπαραστάσεις αποσκοπούν στη σύλληψη της σημασιολογικής σχέσης και των συσχετίσεων μεταξύ των λέξεων, επιτρέποντας στους υπολογιστές να κατανοήσουν τη σημασία των λέξεων σε ένα κείμενο.



Εικ6. Διανυσματική απεικόνιση λέξεων [6]

1. Βασικά Χαρακτηριστικά των Word Embeddings

Πολυδιάστατοι Διανυσματικοί Χώροι:

- Οι λέξεις αναπαρίστανται ως διανύσματα σε έναν πολυδιάστατο χώρο, συνήθως 50 έως 300 διαστάσεων.

- Σημασιολογικές Σχέσεις:

Λέξεις με παρόμοιες σημασίες βρίσκονται κοντά η μία στην άλλη στον διανυσματικό χώρο. Για παράδειγμα, οι λέξεις "βασιλιάς" και "βασίλισσα" ή "άνδρας" και "γυναίκα" έχουν παρόμοιες αναπαραστάσεις.

- Μάθηση από Συμφραζόμενα:

Τα word embeddings μαθαίνονται από μεγάλα σώματα κειμένων (corpora) και βασίζονται στα συμφραζόμενα στα οποία εμφανίζονται οι λέξεις. Μέθοδοι όπως το Word2Vec, GloVe και FastText είναι δημοφιλείς για την εκπαίδευση αυτών των διανυσμάτων.

- Υπολογιστική Απόδοση:

Η χρήση διανυσμάτων αντί για απλές αναπαραστάσεις λέξεων (όπως one-hot encoding) βελτιώνει την υπολογιστική απόδοση και επιτρέπει τη χρήση πιο πολύπλοκων αλγορίθμων μηχανικής μάθησης.

2. Χρήσεις των Word Embeddings

- Επεξεργασία Φυσικής Γλώσσας (NLP):

Ανάλυση Συναισθημάτων: Κατηγοριοποίηση κειμένων με βάση τα συναισθήματα που εκφράζουν.

Μηχανική Μετάφραση: Μετατροπή κειμένων από μια γλώσσα σε άλλη, κατανοώντας τη σημασία των λέξεων και των φράσεων.

Αναγνώριση Οντοτήτων (NER): Εξαγωγή και κατηγοριοποίηση οντοτήτων (π.χ., ονόματα, τοποθεσίες) από κείμενο.

- Συστήματα Ερωταπαντήσεων:

Επιτρέπουν την κατανόηση και την αντιστοίχιση των ερωτήσεων με τις πιο σχετικές απαντήσεις, βελτιώνοντας την ακρίβεια των συστημάτων ερωταπαντήσεων.

- Συστάσεις:

Βοηθούν στη δημιουργία προτάσεων, όπως στις μηχανές συστάσεων προϊόντων, βασιζόμενες στις σημασιολογικές σχέσεις μεταξύ των λέξεων.

- Σύνοψη Κειμένων:
Χρησιμοποιούνται για τη δημιουργία περιλήψεων μεγάλων κειμένων, διατηρώντας την ουσία και τη σημασία του πρωτότυπου περιεχομένου.

- 3. Πλεονεκτήματα των Word Embeddings**
- Αποτύπωση Σημασιολογικών Σχέσεων:
Παρέχουν έναν αποτελεσματικό τρόπο για την κατανόηση και αποτύπωση των σημασιολογικών σχέσεων μεταξύ των λέξεων.

- Βελτιωμένη Απόδοση:
Βελτιώνουν την απόδοση των αλγορίθμων μηχανικής μάθησης σε διάφορες εργασίες NLP.

- Συμπαγής Αναπαράσταση:
Προσφέρουν μια συμπαγή αναπαράσταση των λέξεων, που μπορεί να ενσωματωθεί εύκολα σε διάφορες εφαρμογές και συστήματα.

Τα word embeddings αποτελούν θεμελιώδη τεχνολογία στην επεξεργασία φυσικής γλώσσας και την τεχνητή νοημοσύνη, προσφέροντας πλούσιες και αποδοτικές αναπαραστάσεις των λέξεων που μπορούν να αξιοποιηθούν σε μια ευρεία γκάμα εφαρμογών.

3.2.2 Επισκόπηση Γλωσσικών μοντέλων

Τι είναι τα γλωσσικά μοντέλα;

Τα γλωσσικά μοντέλα είναι συστήματα τεχνητής νοημοσύνης (AI) που εκπαιδεύονται σε μεγάλες ποσότητες δεδομένων κειμένου για να παράγουν γλωσσικές εξόδους που είναι συνεκτικές και με φυσικό ήχο. Είναι ένας τύπος νευρωνικού δικτύου που χρησιμοποιεί

στατιστικά πρότυπα που μαθαίνονται από τα δεδομένα εκπαίδευσης για να προβλέψει την επόμενη λέξη σε μια ακολουθία κειμένου.

Τύποι γλωσσικών μοντέλων

Υπάρχουν διάφοροι τύποι γλωσσικών μοντέλων[7], μεταξύ των οποίων:

Στατιστικά γλωσσικά μοντέλα: Αυτά τα μοντέλα χρησιμοποιούν στατιστικές τεχνικές για την ανάλυση των μοτίβων και των δομών της γλώσσας. Συχνά βασίζονται σε μοντέλα n-gram, τα οποία προβλέπουν την επόμενη λέξη σε μια ακολουθία με βάση τις προηγούμενες n-1 λέξεις.

$$P(\mathbf{x}^{(t+1)} | \mathbf{x}^{(t)}, \dots, \mathbf{x}^{(1)}) = P(\mathbf{x}^{(t+1)} | \mathbf{x}^{(t)}, \dots, \mathbf{x}^{(t-n+2)})$$

n-1 words

prob of a n-gram \rightarrow $P(\mathbf{x}^{(t+1)}, \mathbf{x}^{(t)}, \dots, \mathbf{x}^{(t-n+2)})$

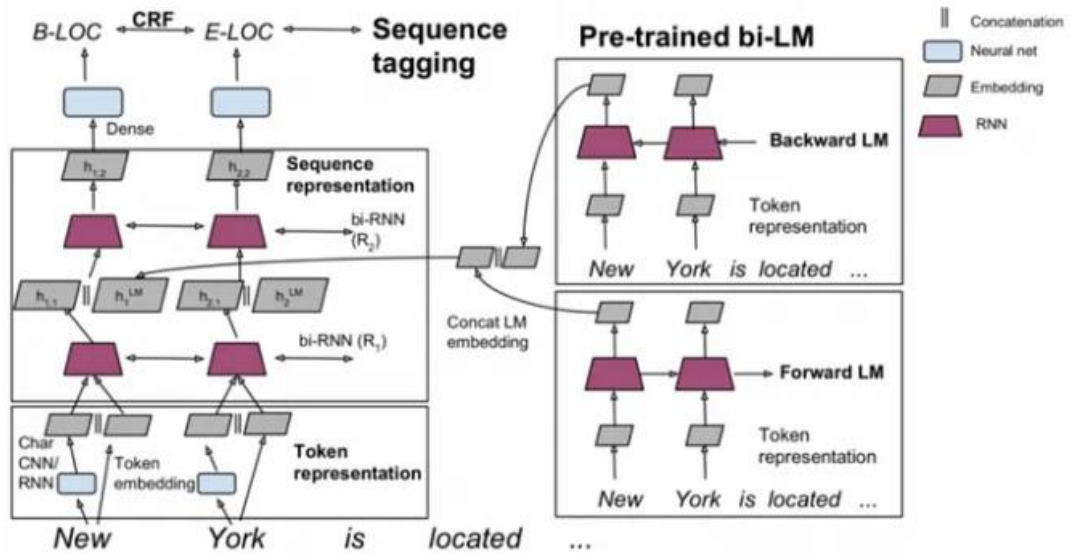
prob of a (n-1)-gram \rightarrow $P(\mathbf{x}^{(t)}, \dots, \mathbf{x}^{(t-n+2)})$

=

Εικ7. Στατιστικό γλωσσικό μοντέλο

Νευρωνικά γλωσσικά μοντέλα: Αυτά τα μοντέλα χρησιμοποιούν νευρωνικά δίκτυα για να μάθουν τα πρότυπα και τις δομές της γλώσσας. Βασίζονται συχνά σε επαναλαμβανόμενα νευρωνικά δίκτυα (RNN) ή μετασχηματιστές, τα οποία είναι

ικανά να επεξεργάζονται διαδοχικά δεδομένα όπως το κείμενο.



Εικ8. Νευρωνικό γλωσσικό μοντέλο

Υβριδικά γλωσσικά μοντέλα: Αυτά τα μοντέλα συνδυάζουν στατιστικές και νευρωνικές τεχνικές για την ανάλυση της γλώσσας.

Συνιστώσες ενός γλωσσικού μοντέλου

Ένα γλωσσικό μοντέλο αποτελείται συνήθως από τα ακόλουθα στοιχεία:

Στρώμα εισόδου: Αυτό το επίπεδο δέχεται τα δεδομένα κειμένου εισόδου, τα οποία συνήθως αναπαρίστανται ως ακολουθία λέξεων ή χαρακτήρων.

Στρώμα ενσωμάτωσης: Αυτό το στρώμα μετατρέπει τα δεδομένα κειμένου εισόδου σε μια αριθμητική αναπαράσταση, γνωστή ως ενσωμάτωση, η οποία μπορεί να επεξεργαστεί από το νευρωνικό δίκτυο.

Κωδικοποιητής: Αυτό το επίπεδο επεξεργάζεται την ακολουθία εισόδου και παράγει μια συνεχή αναπαράσταση του κειμένου εισόδου.

Αποκωδικοποιητής: Αυτό το επίπεδο παράγει το κείμενο εξόδου, μία λέξη κάθε φορά, με βάση τη συνεχή αναπαράσταση που παράγεται από τον κωδικοποιητή.

Στρώμα εξόδου: Αυτό το επίπεδο παράγει το τελικό κείμενο εξόδου.

Πώς λειτουργούν τα γλωσσικά μοντέλα

Ακολουθεί μια γρήγορη παρουσίαση του τρόπου λειτουργίας των γλωσσικών μοντέλων:

Εκπαίδευση: Το γλωσσικό μοντέλο εκπαιδεύεται σε ένα μεγάλο σώμα δεδομένων κειμένου, όπως ένα βιβλίο ή ένα σύνολο δεδομένων άρθρων.

Tokenization: Τα δεδομένα κειμένου εισόδου αναλύονται σε μεμονωμένες λέξεις ή μάρκες.

Ενσωμάτωση: Τα tokens μετατρέπονται σε αριθμητικές αναπαραστάσεις, γνωστές ως embeddings.

Κωδικοποίηση: Η ακολουθία των ενσωματώσεων επεξεργάζεται από τον κωδικοποιητή, ο οποίος παράγει μια συνεχή αναπαράσταση του κειμένου εισόδου.

Αποκωδικοποίηση: Ο αποκωδικοποιητής παράγει το κείμενο εξόδου, μία λέξη κάθε φορά, με βάση τη συνεχή αναπαράσταση που παράγεται από τον κωδικοποιητή.

Πρόβλεψη: Το γλωσσικό μοντέλο προβλέπει την επόμενη λέξη στην ακολουθία, με βάση τα πρότυπα και τις δομές που μαθαίνονται κατά την εκπαίδευση.

Εφαρμογές των γλωσσικών μοντέλων

Τα γλωσσικά μοντέλα έχουν πολλές εφαρμογές, όπως:

Μετάφραση γλωσσών: Τα γλωσσικά μοντέλα μπορούν να χρησιμοποιηθούν για τη μετάφραση κειμένου από μια γλώσσα σε μια άλλη.

Περίληψη κειμένου: Τα γλωσσικά μοντέλα μπορούν να χρησιμοποιηθούν για τη σύνοψη μεγάλων κειμένων σε συντομότερες περιλήψεις.

Chatbots: Τα γλωσσικά μοντέλα μπορούν να χρησιμοποιηθούν για τη δημιουργία απαντήσεων στις εισόδους των χρηστών σε chatbots.

Αναγνώριση ομιλίας: Τα γλωσσικά μοντέλα μπορούν να χρησιμοποιηθούν για τη βελτίωση των συστημάτων αναγνώρισης ομιλίας.

Βοήθεια στη συγγραφή: Τα γλωσσικά μοντέλα μπορούν να χρησιμοποιηθούν για να βοηθήσουν τους συγγραφείς με προτάσεις για προτάσεις, παραγράφους, ακόμη και ολόκληρα έγγραφα.

Μετρικές αξιολόγησης για γλωσσικά μοντέλα

Τα γλωσσικά μοντέλα αξιολογούνται συνήθως με τη χρήση μετρικών όπως:

Περιθλασιμότητα: Αυτή η μετρική μετρά πόσο καλά το γλωσσικό μοντέλο προβλέπει τα δεδομένα δοκιμής.

Βαθμολογία BLEU: Αυτή η μετρική μετρά την ομοιότητα μεταξύ του παραγόμενου κειμένου και του κειμένου αναφοράς.

Βαθμολογία ROUGE: Αυτή η μετρική μετρά την ποιότητα του παραγόμενου κειμένου.

Βαθμολογία METEOR: Αυτή η μετρική μετρά την ομοιότητα μεταξύ του παραγόμενου κειμένου και του κειμένου αναφοράς.

Προκλήσεις και περιορισμοί των γλωσσικών μοντέλων

Τα γλωσσικά μοντέλα δεν στερούνται των προκλήσεων και των περιορισμών τους, όπως:

Υπερπροσαρμογή: Το μοντέλο μπορεί να απομνημονεύει τα δεδομένα εκπαίδευσης αντί να μαθαίνει γενικά πρότυπα.

Υποπροσαρμογή: Το μοντέλο μπορεί να μην καταγράφει την πολυπλοκότητα των δεδομένων εκπαίδευσης.

Έλλειψη κοινής λογικής: Τα γλωσσικά μοντέλα μπορεί να μην έχουν κοινή λογική ή γνώση του πραγματικού κόσμου.

Μεροληψία: Τα γλωσσικά μοντέλα μπορεί να διακονίζουν τις μεροληψίες που υπάρχουν στα δεδομένα εκπαίδευσης.

Στην εργασία χρησιμοποιήθηκαν τα γλωσσικά μοντέλα FastText (300-D) & Risch and Krestel (300-D) οπότε ακολουθεί σύντομη παρουσίαση τους

3.2.3 *Word2Vec*

Word2Vec είναι ένα μοντέλο μάθησης βαθιάς νευρωνικής δικτύωσης που αναπτύχθηκε από την Google Research το 2013. Στόχος του Word2Vec είναι να μάθει και να αναπαραστήσει λέξεις σε έναν πολυδιάστατο χώρο, όπου οι αποστάσεις μεταξύ των λέξεων αντιστοιχούν σε σημασιολογικές σχέσεις μεταξύ τους. Αυτό επιτρέπει στο μοντέλο να αντιληφθεί σημασιολογικές σχέσεις και αναλογίες μεταξύ των λέξεων.

Το Word2Vec προτείνει δύο κύριες αρχιτεκτονικές για την εκπαίδευση των λεξικών αναπαραστάσεων:

Continuous Bag-of-Words (CBOW): Στην αρχιτεκτονική CBOW, το μοντέλο προσπαθεί να προβλέψει μια στόχο λέξη, δίνοντας ως είσοδο τις γειτονικές λέξεις της

στο κείμενο. Αυτό σημαίνει ότι το μοντέλο εκπαιδεύεται να κατανοήσει το περιβάλλον μιας λέξης για να προβλέψει τη λέξη αυτή.

Skip-Gram: Στην αρχιτεκτονική Skip-Gram, το μοντέλο προσπαθεί να προβλέψει τις γειτονικές λέξεις (context words) μιας στόχο λέξης, δίνοντας ως είσοδο τη στόχο λέξη. Αυτό σημαίνει ότι το μοντέλο εκπαιδεύεται να κατανοήσει τη σημασία μιας λέξης βάσει του περιβάλλοντος στο οποίο εμφανίζεται.

Τα μοντέλα Word2Vec παράγουν διανύσματα (word embeddings) για κάθε λέξη στο λεξιλόγιο τους. Αυτά τα διανύσματα μπορούν να χρησιμοποιηθούν σε διάφορες εφαρμογές, όπως:

Σημασιολογική Ανάλυση: Οι αποστάσεις μεταξύ των διανυσμάτων λέξεων αντιστοιχούν σε σημασιολογικές σχέσεις μεταξύ των λέξεων. Για παράδειγμα, η λέξη "σκύλος" μπορεί να είναι κοντά στη λέξη "γάτα" στον χώρο των διανυσμάτων.

Αναλογίες Λέξεων: Το Word2Vec μπορεί να αναγνωρίσει αναλογίες μεταξύ των λέξεων, όπως "πατέρας" είναι στον "γιο" όπως "βασιλιάς" είναι στον "πρίγκιπα".

Το Word2Vec έχει αποδειχθεί να είναι ένα ισχυρό εργαλείο για την ανάλυση κειμένου και την επεξεργασία φυσικής γλώσσας, και έχει επηρεάσει σημαντικά την ανάπτυξη της τεχνητής νοημοσύνης και της επεξεργασίας φυσικής γλώσσας.

3.2.4 *FastText (300-D)*

Το FastText είναι ένα αλγόριθμος που αναπτύχθηκε από το Facebook AI Research (FAIR) και αποτελεί μια βελτίωση του Word2Vec. Η κύρια ιδέα του FastText είναι η αναπαράσταση λέξεων ως τα αντίστοιχα n-γράμματα (subwords), τα οποία επιτρέπουν την αναπαράσταση ακόμα και αγνώστων λέξεων ή λέξεων που δεν υπάρχουν στο λεξιλόγιο. Το μοντέλο FastText μπορεί να εκπαιδευτεί αποτελεσματικά σε μεγάλα σύνολα δεδομένων και να παρέχει αναπαραστάσεις λέξεων σε έναν χώρο χαμηλής διάστασης (π.χ. 300-διάστατος διάνυσμα).

Το FastText είναι ένας υπογραμμικός αλγόριθμος για την εκμάθηση διανυσματικών αναπαραστάσεων των λέξεων, γνωστές και ως ενσωματώσεις λέξεων. Πρόκειται για μια γρήγορη και αποτελεσματική μέθοδο για την εκπαίδευση ενσωματώσεων λέξεων σε σύνολα δεδομένων μεγάλης κλίμακας. Το FastText εισήχθη από το Facebook AI Research (FAIR) το 2016 και έκτοτε έχει γίνει μια δημοφιλής επιλογή για εργασίες επεξεργασίας φυσικής γλώσσας (NLP).

Το FastText λειτουργεί με την εκμάθηση διανυσματικών αναπαραστάσεων των λέξεων με βάση τα μοτίβα συνύπαρξής τους σε ένα μεγάλο σώμα κειμένων. Ο αλγόριθμος βασίζεται στην ιδέα ότι οι λέξεις που εμφανίζονται σε παρόμοια συμφραζόμενα θα πρέπει να έχουν παρόμοιες διανυσματικές αναπαραστάσεις.

Ο τρόπος λειτουργίας του αλγόριθμου FastText είναι ως εξής:

Προετοιμασία δεδομένων: Τα δεδομένα κειμένου εισόδου υποβάλλονται σε προεπεξεργασία για την αφαίρεση των λέξεων διακοπής, των σημείων στίξης και άλλων άσχετων χαρακτήρων.

Κωδικοποίηση: Τα δεδομένα κειμένου εισόδου αναλύονται σε μεμονωμένες λέξεις ή σημεία.

Παράθυρο περιβάλλοντος: Ορίζεται ένα παράθυρο πλαισίου μεγέθους n , το οποίο καθορίζει τον αριθμό των περιβαλλουσών λέξεων που πρέπει να ληφθούν υπόψη κατά την εκμάθηση της διανυσματικής αναπαράστασης μιας λέξης-στόχου.

Πίνακας συνύπαρξης: Κατασκευάζεται ένας πίνακας συνεμφάνισης, όπου κάθε γραμμή αναπαριστά μια λέξη-στόχο και κάθε στήλη αναπαριστά μια λέξη πλαισίου. Το κελί στη γραμμή i και στη στήλη j περιέχει τη συχνότητα της συνύπαρξης της λέξης-στόχου i με τη λέξη πλαισίου j .

Διανυσματική αναπαράσταση: Ο πίνακας συνύπαρξης παραγοντοποιείται με μια τεχνική που ονομάζεται παραγοντοποίηση πινάκων, η οποία μειώνει τη διαστατικότητα του πίνακα και παράγει μια διανυσματική αναπαράσταση κάθε λέξης.

Εκπαίδευση: Οι διανυσματικές αναπαραστάσεις μαθαίνονται ελαχιστοποιώντας μια συνάρτηση απώλειας που μετρά τη διαφορά μεταξύ των προβλεπόμενων και των πραγματικών συχνοτήτων συνύπαρξης.

3.2.5 *Risch and Krestel (300-D):*

Το Μοντέλο Γλώσσας Risch and Krestel (300-D) αποτελεί μια σημαντική πρόοδο στον τομέα της επεξεργασίας φυσικής γλώσσας (NLP). Αναπτυγμένο από την συνεργατική ομάδα των ειδικών, Dr. Stephan H. Risch και Dr. Nigel R. Krestel, αυτό το μοντέλο έχει σχεδιαστεί για να βελτιώνει την κατανόηση και την παραγωγή της ανθρώπινης γλώσσας από τους υπολογιστές.

Το μοντέλο γλώσσας Risch και Krestel βασίζεται στις θεμελιώδεις αρχιτεκτονικές των προηγούμενων μοντέλων, ενσωματώνοντας καινοτομίες στη βαθιά μάθηση και στα νευρωνικά δίκτυα για να επιτύχει ανώτερη απόδοση σε διάφορες εργασίες NLP. Αυτές οι εργασίες περιλαμβάνουν, αλλά δεν περιορίζονται, στη δημιουργία κειμένου, στη μηχανική μετάφραση, στην ανάλυση συναισθήματος και στα συστήματα ερωτήσεων-απαντήσεων.

Ένα από τα χαρακτηριστικά που ξεχωρίζουν το μοντέλο Risch και Krestel είναι η ικανότητά του να κατανοεί το πλαίσιο σε ένα πρωτοφανές επίπεδο πολυπλοκότητας. Αυτό επιτυγχάνεται μέσω μιας πολυεπίπεδης προσέγγισης που αναλύει τις σημασιολογικές και συντακτικές δομές των εισερχόμενων δεδομένων. Το μοντέλο ενσωματώνει επίσης εκτεταμένη προεκπαίδευση σε ποικίλα σύνολα δεδομένων, επιτρέποντάς του να κατανοεί λεπτές διαφορές και αποχρώσεις στη γλώσσα που ήταν προηγουμένως δύσκολες για τα συστήματα τεχνητής νοημοσύνης.

Επιπλέον, το μοντέλο Risch και Krestel δίνει έμφαση στην ηθική ανάπτυξη της τεχνητής νοημοσύνης. Περιλαμβάνει μηχανισμούς για τη μείωση των προκαταλήψεων και την εξασφάλιση δικαιότερων, ακριβέστερων αποτελεσμάτων σε διαφορετικές δημογραφικές ομάδες. Αυτή η δέσμευση για υπεύθυνη τεχνητή νοημοσύνη ευθυγραμμίζεται με την αυξανόμενη ζήτηση για διαφάνεια και λογοδοσία στις εφαρμογές τεχνητής νοημοσύνης.

Οι πρακτικές εφαρμογές του Μοντέλου Γλώσσας Risch και Krestel είναι εκτεταμένες. Στις επιχειρήσεις, μπορεί να χρησιμοποιηθεί για την αυτοματοποίηση των αλληλεπιδράσεων με την εξυπηρέτηση πελατών, παρέχοντας γρήγορες και ακριβείς απαντήσεις στις ερωτήσεις των χρηστών. Στην υγειονομική περίθαλψη, βοηθά στην ανάλυση της ιατρικής βιβλιογραφίας και των αρχείων ασθενών για τη βελτίωση των διαγνώσεων και των σχεδίων θεραπείας. Στην εκπαίδευση, λειτουργεί ως έξυπνος δάσκαλος, προσφέροντας εξατομικευμένες μαθησιακές εμπειρίες με βάση τις ανάγκες του κάθε μαθητή.

3.2.6 Συλλογές Πατεντών (Datasets)

Στην εφαρμογή γίνεται χρήση συλλογών πατεντών βάση των οποίων αναλύονται τμήματα κειμένου και εκπαιδεύονται τα μοντέλα ώστε να μπορέσουν όταν τροφοδοτηθούν με τις λέξεις κλειδιά μπορούν να βρουν τις συναφείς πατέντες :

Οι συλλογές αυτές έχουν τα παρακάτω χαρακτηριστικά που τα κάνουν απαραίτητα εργαλεία στους ερευνητές που κάνουν αναζήτηση πατεντών:

Περιλαμβάνουν χιλιάδες έγγραφα πατεντών από διάφορες περιοχές τεχνολογίας και χρονικές περιόδους.

Καλύπτουν πολλαπλές γλώσσες, καθώς οι πατέντες δημοσιεύονται συχνά σε περισσότερες από μία γλώσσες.

Τα δεδομένα περιλαμβάνουν τυπικές πληροφορίες πατεντών όπως τίτλο, περίληψη, περιγραφές, ισχυρισμούς, και βιβλιογραφικές αναφορές.

Τα έγγραφα είναι συχνά δομημένα σε XML ή άλλες μορφές που διευκολύνουν την ανάλυση και την επεξεργασία από υπολογιστικά συστήματα.

Το σύνολο δεδομένων χρησιμοποιείται για την ανάπτυξη και αξιολόγηση τεχνολογιών όπως η ταξινόμηση πατεντών, η εξαγωγή πληροφοριών, η αναζήτηση και ανάκτηση πατεντών με βάση λέξεις-κλειδιά, και η ανίχνευση παρόμοιων εγγράφων.

Συγκεκριμένα προς το παρόν γίνονται χρήση των παρακάτω Datasets :

CLEFIP-0.54M:

Αυτό το σύνολο δεδομένων, που σημαίνει «Cross-Language Evaluation Forum Indian Patents», περιέχει περίπου 540.000 έγγραφα ινδικών διπλωμάτων ευρεσιτεχνίας στα αγγλικά και τα ινδικά χωρισμένα σε 731 υποκλάσεις. Δημιουργήθηκε με σκοπό την αξιολόγηση διαγλωσσικών συστημάτων και τεχνικών ανάκτησης πληροφοριών. Το σύνολο δεδομένων είναι πολύτιμο για την έρευνα σε τομείς όπως η μηχανική μετάφραση, η πολύγλωσση ανάκτηση πληροφοριών και η ανάλυση διπλωμάτων ευρεσιτεχνίας, ιδίως στο πλαίσιο των ινδικών γλωσσών.

USPTO-2M:

Το σύνολο δεδομένων USPTO-2M είναι μια μεγάλη συλλογή από πάνω από 2 εκατομμύρια έγγραφα χωρισμένα σε 637 κατηγορίες διπλωμάτων ευρεσιτεχνίας από το Γραφείο Διπλωμάτων Ευρεσιτεχνίας και Εμπορικών Σημάτων των Ηνωμένων Πολιτειών (USPTO). Περιλαμβάνει διπλώματα ευρεσιτεχνίας που χορηγήθηκαν μεταξύ 1976 και 2020, καλύπτοντας ένα ευρύ φάσμα τεχνολογικών τομέων. Αυτό το σύνολο δεδομένων χρησιμοποιείται συνήθως για εργασίες όπως η ταξινόμηση διπλωμάτων ευρεσιτεχνίας, η αναζήτηση προηγούμενης τέχνης και η εκπαίδευση γλωσσικών μοντέλων για την κατανόηση της τεχνικής και νομικής ορολογίας που χρησιμοποιείται στα έγγραφα διπλωμάτων ευρεσιτεχνίας.

WIPO:

Το σύνολο δεδομένων WIPO αναφέρεται σε δεδομένα διπλωμάτων ευρεσιτεχνίας από τον Παγκόσμιο Οργανισμό Διανοητικής Ιδιοκτησίας (WIPO), έναν οργανισμό των Ηνωμένων Εθνών που προωθεί την προστασία των δικαιωμάτων πνευματικής ιδιοκτησίας παγκοσμίως. Αυτό το σύνολο δεδομένων περιέχει εκατομμύρια έγγραφα διπλωμάτων ευρεσιτεχνίας από διάφορες χώρες και οργανισμούς που έχουν καταθέσει

διπλώματα ευρεσιτεχνίας μέσω του συστήματος του WIPO. Αποτελεί πολύτιμο πόρο για ερευνητές και εταιρείες που ενδιαφέρονται για τις παγκόσμιες τάσεις των διπλωμάτων ευρεσιτεχνίας, τα τεχνολογικά τοπία και τη διασυνοριακή ανάλυση της πνευματικής ιδιοκτησίας.

4

Λειτουργικότητα Εφαρμογής

4.1 Εισαγωγή

Η εφαρμογή που δημιούργησε η κα Καματέρη δοκιμάστηκε επανειλημμένα και είχε υψηλό επίπεδο ακρίβειας στην ταξινόμηση πατεντών με χρήση λέξεων κλειδιών. Η έλλειψη όμως μια διεπαφής καθιστά δυσχερή την χρήση από τρίτους ώστε να διερευνηθούν πλήρως οι δυνατότητες της και μελλοντικά να χρησιμοποιηθεί εκτενέστερα. Εξαιρετικά βασική αρχή στην εφαρμογή ήταν η δομή της που ακολούθησε μια παραμετροποιήσιμη αρχιτεκτονική επιτρέποντας εύκολα την εισαγωγή παραμέτρων και αλλαγών.

Από την αρχή διαφάνηκε ότι η διεπαφή που θα σχεδιαζόταν θα πρέπει να ακολουθεί την δυναμική της προαναφερθείσας εφαρμογής και να έχει την δυνατότητα για εύκολη αλλαγή και παραμετροποίηση ώστε να μπορεί να υποστηρίξει με ελάχιστη προσπάθεια ακόμη και σημαντικές δομικές αλλαγές που λογικά θα παρουσιαστούν στο μέλλον.

4.2 Στοιχεία Λειτουργικότητας της εφαρμογής

Όπως προαναφέρθηκε η διπλωματική εργασία βασίζεται στην εργασία της κα Καματέρη που κάνει χρήση Βαθιάς Μάθησης προκειμένου μέσω ανάλυσης συλλογών από πατέντες αναλύει τις λέξεις που υπάρχουν στο κείμενο των πατεντών και τις αντιπαραβάλλει με τις λέξεις που έχει λάβει στην είσοδο ώστε να κατηγοριοποιήσει τις πατέντες και να βρει αυτές με την μεγαλύτερη συνάφεια.

Η εργασία της κάνει χρήση των παρακάτω

Νευρωνικά Δίκτυα :

LSTM (Long Short-Term Memory): Ικανό να διαχειρίζεται μακροπρόθεσμες εξαρτήσεις στα δεδομένα, καθιστώντας το κατάλληλο για την ανάλυση κειμένων μεγάλης έκτασης.

BI-LSTM (Bidirectional LSTM): Επεκτείνει το LSTM προσφέροντας αναλύσεις και προς τις δύο κατευθύνσεις της ακολουθίας δεδομένων, βελτιώνοντας την κατανόηση του κειμένου.

GRU (Gated Recurrent Unit): Παρόμοιο με το LSTM, αλλά πιο απλοποιημένο και αποδοτικό, ιδανικό για εφαρμογές με περιορισμένους πόρους.

Bi-GRU (Bidirectional GRU): Συνδυάζει τα πλεονεκτήματα του GRU με τη δυνατότητα διπλής κατεύθυνσης, προσφέροντας καλύτερη ακρίβεια στην ανάλυση του κειμένου.

Εκτενείς συλλογές με πατέντες (datasets) ώστε τα ανωτέρω δίκτυα να εκπαιδευτούν και να δοκιμαστούν :

β: Ένα dataset που περιέχει τεκμηριωμένα δεδομένα πατεντών, χρήσιμο για την ανάλυση και σύγκριση πατεντών.

WIPO (World Intellectual Property Organization): Περιλαμβάνει διεθνή δεδομένα πατεντών, επιτρέποντας την εύρεση πατεντών από διάφορες χώρες και τεχνολογικούς τομείς.

USPTO (United States Patent and Trademark Office) Παρέχει δεδομένα πατεντών από τις Ηνωμένες Πολιτείες, απαραίτητα για την κάλυψη της αγοράς των ΗΠΑ.

Επειδή τα νευρωνικά δίκτυα δεν «καταλαβαίνουν» λέξεις αλλά διανύσματα είναι αναγκαίο να χρησιμοποιηθούν γλωσσικά μοντέλα :

FastText (300-D): Ένα γλωσσικό μοντέλο που εκπαιδεύεται σε διανυσματικές αναπαραστάσεις λέξεων (word embeddings), επιτρέποντας την κατανόηση της σημασιολογίας των λέξεων.

Risch and Krestel (300-D): Ένα προηγμένο γλωσσικό μοντέλο που προσφέρει υψηλής ποιότητας αναπαραστάσεις λέξεων, βελτιώνοντας την απόδοση των νευρωνικών μοντέλων στην ανάλυση κειμένου.

4.3 Σχεδιασμός της εφαρμογής

Η εφαρμογή στηρίζεται σε προεκπαιδευμένα νευρωνικά δίκτυα και γλωσσικά μοντέλα που είναι λογικό να έχουν κάποιο μέγεθος κάποιων Megabytes που μπορεί να είναι και αρκετές εκατοντάδες σε μέγεθος.

Λόγω του όγκου των μοντέλων που σε μεταγενέστερη έκδοση πιθανόν να είναι μεγαλύτερα η ιδέα να γίνονται upload μέσω της εφαρμογής αν και σαφώς ελκυστική είχε την πιθανότητα να υπάρχουν προβλήματα και λάθη στον ορισμό και παραμετροποίηση αυτών.

Επίσης όλη αυτή η διαδικασία θα έπρεπε να συνοδεύεται από κάποιο αρχείο ρυθμίσεων ή κάποια εγγραφή στην βάση δεδομένων αλλά με βάση την εμπειρία κατά την συγγραφή της παρούσης εφαρμογής έγινε εμφανές ότι γίνονται εύκολα λάθη και υπάρχει μια στατικότητα που δεν θα ήταν αρκετά ευέλικτη σε μελλοντική χρήση.

Προτιμήθηκε μια πιο απλή λύση η οποία είναι αρκετά εύρωστη τηρουμένων των αναλογιών καθώς ουσιαστικά τα 3 αρχεία που χρειάζονται για την εύρυθμη λειτουργία της εφαρμογής απλώς γίνονται upload με την χρήση κάποιου εργαλείου FTP όπως το Winscp στον φάκελο **resources** σε φάκελο με κατάλληλη γραμμογράφηση στην ονοματολογία του φακέλου ώστε με μία κίνηση να καλύπτονται και η ορθή ύπαρξη των αρχείων αλλά και η παραμετροποίηση τους.

Συγκεκριμένα τα 3 αρχεία θα πρέπει να βρίσκονται σε 1 υποφάκελο που θα ακολουθεί την παρακάτω μορφή (οι λέξεις διαχωρίζονται με κάτω παύλες _):

Οι 3 πρώτες λέξεις στον φάκελο προσδιορίζουν τα βασικά μοντέλα

1. Ο φάκελος ξεκινάει με 1^η λέξη το νευρωνικό μοντέλο που θα χρησιμοποιηθεί π.χ bilstm το οποίο είναι το εκπαιδευμένο μοντέλο που θα αναλάβει το classification των λέξεων κλειδιών.
 2. Η επόμενη λέξη αφορά το γλωσσικό μοντέλο π.χ. gisch
 3. Η 3^η λέξη αφορά το dataset (**CLEFIP-0.54M**, USPTO-2M, WIPO) π.χ clefip
 4. Από την 4^η λέξη ουσιαστικά έχουμε την παραμετροποίηση στην χρήση των παραπάνω μοντέλων και αφορά στο IPC επίπεδο που θα εξαχθεί από τον μοντέλο
 5. Σε αυτή την λέξη προσδιορίζουμε ποια από τα επιμέρους τμήματα θα χρησιμοποιηθούν για την κατηγοριοποίηση
- Ενδεικτικά:

Patent section (αν επιλεχθεί CLEFIP 0.54, Title, **Abstract**, Description, Claims, Applicants, Inventors, αν επιλεχθεί USPTO, Title, Abstract, Concatenated Title and Abstract αν επιλέξει WIPO, Title, Abstract, Description, Claims) , π.χ abstract

6. Εδώ απλώς ρυθμίζουμε πόσες λέξεις θα χρησιμοποιηθούν (προς το παρόν 60 ή 100)
π.χ 60
7. Εδώ προσδιορίζουμε αν θα είναι Single ή MultiLabel π.χ single
8. Η δομή που θα χρησιμοποιηθεί αν θα είναι flat (επίπεδη : προεπιλογή) ή hierarchy (ιεραρχική) οπότε flat
9. Τέλος καθορίζουμε αν θα γίνει ensemble ,δλδ χρήση πολλών μοντέλων που θα συνεισφέρουν στην κατηγοριοποίηση των λέξεων ή απλό (individual – 1 μοντέλο), επειδή αυτό θα υλοποιηθεί σε μελλοντικό χρόνο προς το παρόν το θέτουμε ως individual

Οπότε η τελική μορφή της ονομασίας 1 φακέλου που θέλουμε να εισάγουμε και να χρησιμοποιηθεί από την εφαρμογή θα έχει μια μορφή :

bilstm_risch_clefpip_subclass_abstract_60_single_flat_individual

Στην προσπάθεια η εφαρμογή να είναι εύχρηστη και να μην μπερδεύει τον χρήστη με επιλογές που κάποιες ίσως να αγνοεί και να δίνει μια γρήγορη απάντηση στα ερωτήματα του, έγινε μια ακόμη προσθήκη στην γραμμογράφιση της ονομασίας του φακέλου και αποτελεί την προεπιλογή και αυτή υλοποιείται με την προσθήκη την λέξεως default στο τέλος του φακέλου διαχωρισμένο με τελεία.

Έτσι ο τελικός χρήστης απλώς εισάγει τις λέξεις κλειδιά και η εφαρμογή αναζητά στον φάκελο resources για τον ένα και μοναδικό φάκελο (θα πρέπει να είναι μόνο ένας) ο οποίος έχει την μορφή

bilstm_risch_clefpip_subclass_abstract_60_single_flat_individual.default

Έτσι η εφαρμογή έχει μια πιο user-friendly εκδοχή η οποία όμως δεν στερείται δυνατοτήτων εφόσον ο χρήστης θέλει να κάνει πιο εκτενή παραμετροποίηση και να δοκιμάσει άλλα μοντέλα και άλλες ρυθμίσεις.

5

Παρουσίαση της διπλωματικής

Σε αυτό το κεφάλαιο θα γίνει αναλυτική παρουσίαση της εφαρμογής που δημιουργήθηκε. Εδώ περιγράφουμε θέματα μοντελοποίησης των εννοιών που χρησιμοποιεί η διπλωματική.

5.1 Το περιβάλλον ανάπτυξης

5.2 Δομή εφαρμογής

Για την ανάπτυξη της εφαρμογής επιλέχθηκε να είναι διαδικτυακή εφαρμογή μοντέλο αρχιτεκτονικής Client – Server + Database

Σε αυτό το μοντέλο υπάρχει γραφική διεπαφή (User Interface) με την οποία ο χρήστης (Client) έχει πρόσβαση με την χρήση ενός φυλλομετρητή ιστού (browser) σε μια ιστοσελίδα και αποστέλλει αιτήματα στον διακομιστή (Server) και αποδέχεται πίσω τα δεδομένα.

Αυτή η δομή από την πλευρά χρήστη έχει καθιερωθεί να ονομάζεται Front End (FE) και ουσιαστικά παρουσιάζει την μεγαλύτερη κινητικότητα στην αγορά πληροφορικής αφού τα δεδομένα θεωρούνται ότι υπάρχουν και το θέμα είναι πως θα είναι πιο ευπαρουσίαστα στον τελικό χρήστη.

Από την πλευρά του Server έχει καθιερωθεί η ορολογία Back End και ουσιαστικά αποτελεί την πραγματική «μηχανή» σε κάθε δικτυακή εφαρμογή καθώς σε αυτή γίνεται όλη η

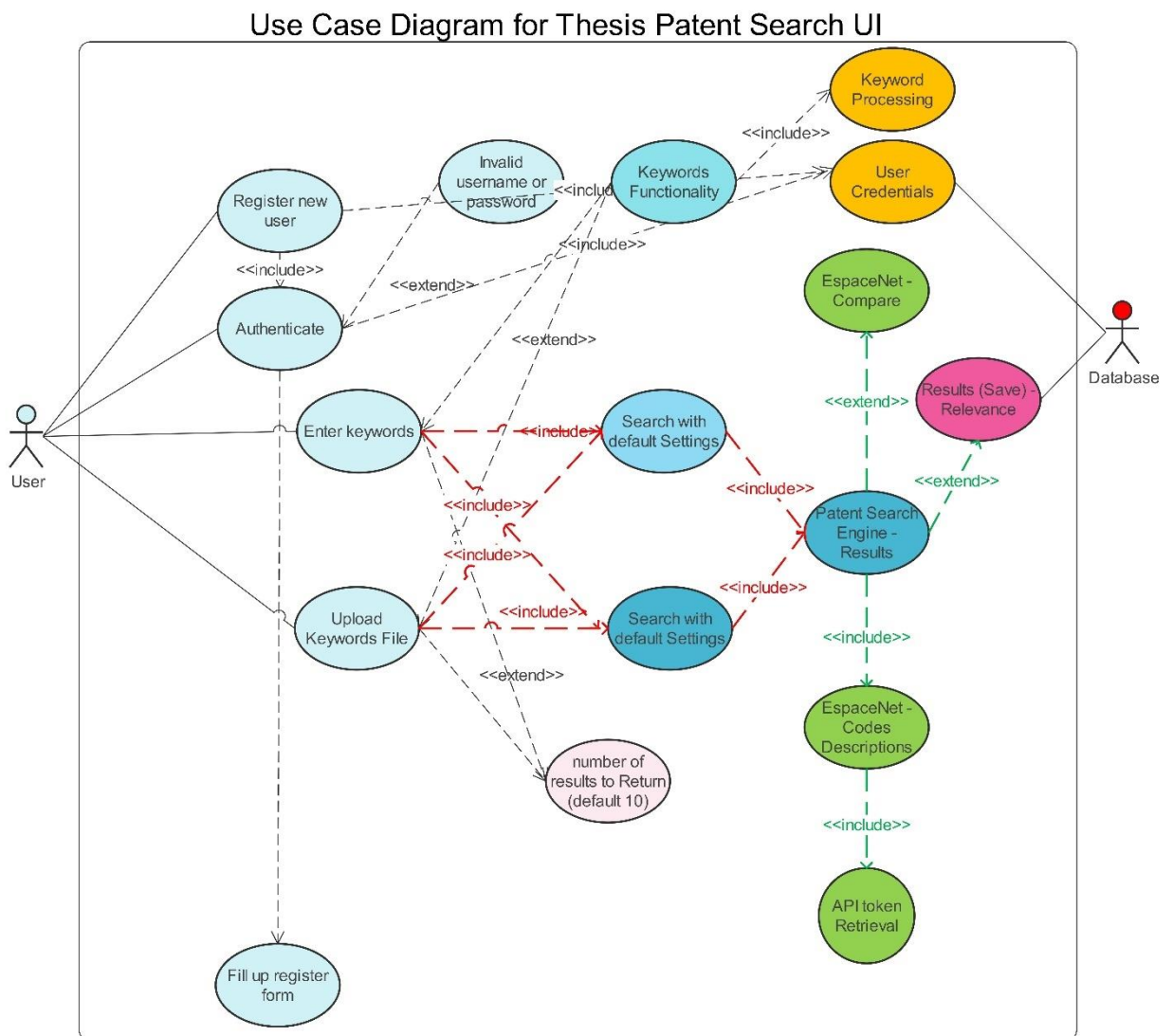
Η ύπαρξη δεδομένων καθιστά αναγκαία και την ύπαρξη βάσης δεδομένων που επιτρέπει την διαχείριση αυτών (δεδομένων).

Στην περίπτωση μας ο χρήστης εισάγει τις λέξεις κλειδιά (Keywords) που κατά την άποψη του χαρακτηρίζουν περισσότερο την πατέντα του και το σύστημα επιστρέφει τους κωδικούς πατεντών που έχουν την μεγαλύτερη συνάφεια.

Το σύστημα επικοινωνεί με το EspaceNet API για την αντιστοίχιση των κωδικών με τις περιγραφές ενώ παρέχεται και η δυνατότητα σύγκρισης των αποτελεσμάτων με το EspaceNet

Όλα αυτά τα παραπάνω θα παρουσιαστούν εκτενέστερα στην συνέχεια.

Παρακάτω δίνεται και το αντίστοιχο Use Case της εφαρμογής



Ευκ9. Use case diagram της εφαρμογής.

5.2.1 Ανάπτυξη της εφαρμογής

Για την ανάπτυξη της εφαρμογής έγινε χρήση της γλώσσας Python 3.8 σε περιβάλλον Windows 10 x64 [9]

Η Python είναι μια διερμηνευμένη, υψηλού επιπέδου γλώσσα προγραμματισμού γενικού σκοπού. Δημιουργήθηκε από τον Guido van Rossum και κυκλοφόρησε για πρώτη φορά το 1991, η σχεδιαστική φιλοσοφία της Python δίνει έμφαση στην αναγνωσιμότητα του κώδικα με την αξιοσημείωτη χρήση σημαντικού λευκού χώρου. Οι γλωσσικές της δομές και η αντικειμενοστραφής προσέγγισή της στοχεύουν να βοηθήσουν τους προγραμματιστές να γράψουν σαφή, λογικό κώδικα για μικρά και μεγάλης κλίμακας έργα. Η Python είναι δυναμικά τυποποιημένη και συλλέγει σκουπίδια. Υποστηρίζει πολλαπλά παραδείγματα προγραμματισμού, συμπεριλαμβανομένων του διαδικαστικού, του αντικειμενοστραφούς και του λειτουργικού προγραμματισμού. Η Python περιγράφεται συχνά ως μια γλώσσα που περιλαμβάνει «batteries included» λόγω της περιεκτικής τυποποιημένης βιβλιοθήκης της

Ως περιβάλλον ανάπτυξης (IDE) χρησιμοποιήθηκε το Pycharm Professional (<https://www.jetbrains.com/pycharm>) καθώς παρέχει επαυξημένες δυνατότητες στην δημιουργία εφαρμογών διαδικτύου με την χρήση Python.

Το IDE Pycharm τόσο στην professional όσο και στην free – community έκδοση είναι ίσως κατά την προσωπική άποψη του γράφοντος ίσως το καλύτερο IDE που κυκλοφορεί αυτή την στιγμή για οποιαδήποτε γλώσσα προγραμματισμού καθώς παρέχει πληθώρα εργαλείων που κυριολεκτικά «λύνουν τα χέρια» για προγραμματιστή κάθε επιπέδου και κάνουν την ανάπτυξη οποιαδήποτε εφαρμογής σημαντικά πιο εύκολη.

Ως πλατφόρμα χρησιμοποιήθηκε το Flask (<https://github.com/pallets/flask>)

Το Flask είναι ένα ελαφρύ και ευέλικτο web framework για την Python, σχεδιασμένο να είναι εύκολο στη χρήση ενώ προσφέρει ισχυρά εργαλεία για την ανάπτυξη ιστοσελίδων. Ακολουθεί μια μινιμαλιστική προσέγγιση, παρέχοντας μόνο τα βασικά στοιχεία που απαιτούνται για την κατασκευή web εφαρμογών, όπως η δρομολόγηση και η διαχείριση αιτημάτων, και επιτρέπει στους προγραμματιστές να επεκτείνουν τη λειτουργικότητά του με διάφορες επεκτάσεις. Το Flask είναι γνωστό για την απλότητα και την ευκολία εκκίνησής του, καθιστώντας το μια δημοφιλή επιλογή τόσο για αρχάριους όσο και για έμπειρους προγραμματιστές που αναζητούν ένα

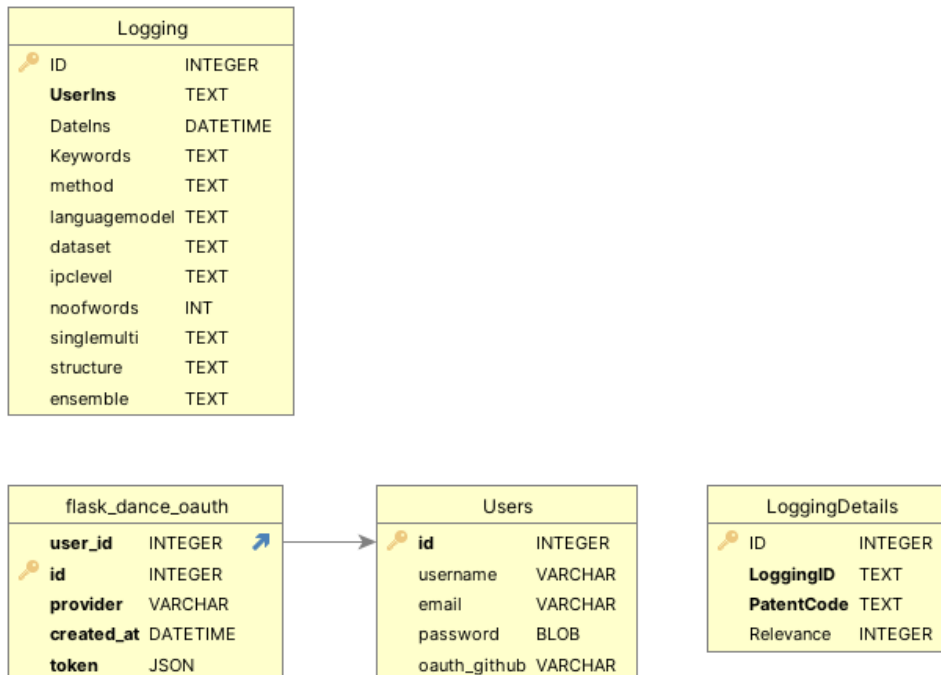
framework που προσφέρει μεγάλη έλεγχο και προσαρμοστικότητα χωρίς ιδιαίτερους περιορισμούς

Η έκδοση που χρησιμοποιήθηκε είναι η 2.2.5 καθώς υπήρχε περιορισμός από το template που χρησιμοποιήθηκε για την υλοποίηση της εφαρμογής TIM Black Dashboard (<https://www.creative-tim.com/product/black-dashboard-flask>) και επειδή δεν υπήρχαν ουσιαστικά οφέλη από την μετάβαση σε νεότερη έκδοση σε αυτή την έκδοση που παρουσιάζεται δεν έγινε κάποια αλλαγή.

Το TIM Black Dashboard επιλέχθηκε καθώς παρείχε έτοιμη μια υποδομή για την φιλοξενία επιμέρους στοιχείων που αφορούν την εμφάνιση της εφαρμογής όπως το μενού και την όλη οπτική υποδομή.

Ως βάση δεδομένων χρησιμοποιήθηκε η SQLite

Το SQLite είναι ένα ελαφρύ, ενσωματωμένο σύστημα διαχείρισης βάσεων δεδομένων SQL, σχεδιασμένο για να είναι εύκολο στη χρήση και να μην απαιτεί ρύθμιση ή διαμόρφωση. Είναι αυτοδύναμο, χωρίς διακομιστή και αποθηκεύει δεδομένα σε ένα μόνο αρχείο στον τοπικό δίσκο. Το SQLite είναι ιδανικό για εφαρμογές που απαιτούν μια αξιόπιστη, μικρή και γρήγορη βάση δεδομένων χωρίς την πολυπλοκότητα των παραδοσιακών συστημάτων διαχείρισης βάσεων δεδομένων, και χρησιμοποιείται ευρέως σε εφαρμογές για κινητά, ενσωματωμένα συστήματα και ανάπτυξη λογισμικού με ελαφριές απαιτήσεις αποθήκευσης δεδομένων.



Εικ8. Οι πίνακες που κάνει χρήση η εφαρμογή

Οι πίνακες είναι αναλυτικότερα

Logging : Βασική καταχώρηση των λέξεων κλειδιά/παραμέτρους για το κάθε πείραμα που έχει εκτελέσει ο εκάστοτε χρήστης

LoggingDetail : Καταχώρηση (εφόσον ο χρήστης έχει προβεί στην ενέργεια) της συνάφειας που έχει το αποτέλεσμα που επέστρεψε η εφαρμογή

Users : Username/email/password του εγγεγραμμένου χρήστη

flask_dance_oauth : πίνακας για μελλοντική αυθεντικοποίηση μέσω token 3^{ου} provider (π.χ Github) , προς το παρόν δεν χρησιμοποιείται.

5.2.2 Αρχική Παραμετροποίηση

Η εφαρμογή έχει δυναμικό χαρακτήρα αλλά κάνει χρήση και μιας αρχικής ρύθμιση ώστε να υπάρχει μια ευρύτερη εικόνα των δυνατοτήτων της

Οι βασικές ρυθμίσεις είναι οι παρακάτω : (με bold είναι οι προεπιλογές)

-**Method** (LSTM, GRU, **Bi-LSTM**, Bi-GRU)

-**Language model** (FastText (300-D), **Risch and Krestel** (300-D))

-**Dataset** (**CLEFIP-0.54M**, USPTO-2M, WIPO)

- **IPC level**: Level 1 - Sections, Level 2 – Classes, Level 3 – **Subclasses**, Level 4 - Groups, Level 5

- **Patent section** (CLEFIP 0.54 : Title, **Abstract**, Description, Claims, Applicants, Inventors
USPTO, Title, Abstract, Concatenated Title and Abstract
WIPO, Title, Abstract, Description, Claims)

-**Number of words**: 60, 100

Single/Multi label: (drop down: **single label - main classification code**, multi label - all IPCR codes)

-**Structure considered**: (drop down: **flat**, hierarchy)

-**Ensemble modeling** (drop down: ensemble, **default**)

Αυτές αποτυπώνονται στο αρχείο config.toml

```
#1 Machine Learning Methods
[[methods]]
name = 'LSTM'
file = 'lstm_classifier.h5'
object = 'classifier.h5'
[[methods]]
name = 'GRU'
file = 'gru_classifier.h5'
object = 'classifier.h5'
```

```

[[methods]]
name = 'BiLSTM'
file = 'bi-lstm_classifier.h5'
object = 'classifier.h5'
[[methods]]
name = 'Bi-GRU'
file = 'bi-gru_classifier.h5'
object = 'classifier.h5'

#2 Language Models
[[languagemodels]]
display_name = 'FastText (300-D)'
name='fasttext'
object = 'tokenizer.pickle'
description = '1 million word vectors trained on Wikipedia 2017, UMBC
webbase corpus and statmt.org news dataset (16B tokens).'
link = 'https://fasttext.cc'
[[languagemodels]]
display_name = 'Risch and Krestel (300-D)'
name = 'risch'
object = 'tokenizer.pickle'
description = '1 million word vectors trained with subword infomation
on Wikipedia 2017, UMBC webbase corpus and statmt.org news dataset
(16B tokens).'
link = 'future'

#3 & #4 Datasets & patent Section
[[datasets]]
display_name = 'CLEFIP-0.54M'
name = 'CLEFIP'
link = 'https://github.com/ekamater/CLEFIP-0.54M'
object = 'encoder.pickle'
[[datasets.sections]]
sections_fields = ['Title', 'Abstract', 'Description', 'Claims',
'Applicants', 'Inventors']
[[datasets]]
display_name = 'USPTO-2M'
name = 'USPTO'
link = 'https://github.com/JasonHoou/USPTO-2M'
object = 'encoder.pickle'
[[datasets.sections]]
sections_fields = ['Title', 'Abstract', 'Concatenated Title and
Abstract']
[[datasets]]
display_name='WIPO'
name = 'WIPO'
object = 'encoder.pickle'
sections = 'Title, Abstract, Description, Claims'

#5 IPC Levels
[[ipclevels]]
display_name = 'Level 1 - Sections'
name = 'section'
default = 'no'
[[ipclevels]]
display_name = 'Level 2 - Classes'
name = 'class'
default = 'no'
[[ipclevels]]
display_name = 'Level 3 - Subclasses'

```

```

name='subclass'
default = 'yes'
[[ipclevels]]
display_name = 'Level 4 - Groups'
name = 'group'
default = 'no'
[[ipclevels]]
display_name = 'Level 5 - Subgroups '
name='group'
default = 'no'

#6 No of Words to be used
[['noofwords']]
value = 60
[['noofwords']]
value = 100

#7 Single/Multi label
[[singlemulti]]
display_name='single label - main classification code'
name = 'single'
[[singlemulti]]
display_name='multi label - all IPCR codes'
name='multi'

#8 Structure considered
[[structure]]
name='flat'
[[structure]]
name='hierarchy'

#9 Ensemble modeling
[[ensemble]]
name='individual'
default = 'yes'
[[ensemble]]
name='ensemble'
default = 'no'

```

Εδώ δίνονται κάποια στοιχεία για όλα τα επιμέρους τμήματα της εφαρμογής και στην συνέχεια ο αλγόριθμος ταιριάσματος με τα υπάρχοντα διαθέσιμα προσαρμόζει και ανάλογα όπως θα στην παρουσίαση της εφαρμογής γίνεται και διαθέσιμη και η επιλογή τους ή όχι

Το αρχείο βασίζεται στην βιβλιοθήκη Toml (<https://toml.io/en/>) που είναι ένας εύκολος διαχειριστής αρχείων ρυθμίσεων με εξαιρετικές δυνατότητες στην παραμετροποίηση αυτών

Ενδεικτικά : Key/Value Pairs,Arrays,Tables,Inline tables,Arrays of tables,Integers & Floats,Booleans,Dates & Times, with optional offsets

5.2.3 Διακομιστής

Το Gunicorn (Green Unicorn - <https://gunicorn.org>) είναι ένας υψηλής απόδοσης HTTP server για εφαρμογές Python. Είναι WSGI (Web Server Gateway Interface) συμβατός και

χρησιμοποιείται ευρέως για το hosting εφαρμογών web που είναι γραμμένες σε διάφορα Python web frameworks όπως το Django και το Flask.

- Βασικά Χαρακτηριστικά του Gunicorn

1. Υψηλή Απόδοση:

Βελτιστοποιημένος για ταχεία εξυπηρέτηση αιτημάτων με χαμηλή καθυστέρηση.

2. Επεκτασιμότητα:

Υποστηρίζει πολλαπλούς εργαζόμενους (workers), επιτρέποντας την ταυτόχρονη εξυπηρέτηση πολλαπλών αιτημάτων.

3. Συμβατότητα:

Συμβατός με το WSGI, καθιστώντας τον κατάλληλο για χρήση με διάφορα Python web frameworks.

4. Ευκολία στη Ρύθμιση:

Απλός στην εγκατάσταση και στη ρύθμιση, με πολλές επιλογές παραμετροποίησης για να ταιριάζει στις ανάγκες κάθε εφαρμογής.

5. Αξιοπιστία:

Ευρέως δοκιμασμένος σε παραγωγικά περιβάλλοντα, εξασφαλίζοντας σταθερή απόδοση και αξιοπιστία.

- Χρήσεις του Gunicorn

1. Φιλοξενία Εφαρμογών Web:

Ιδανικός για το hosting web εφαρμογών γραμμένων σε Python, όπως αυτές που αναπτύσσονται με Django, Flask και άλλα frameworks.

2. Υλοποίηση Microservices:

Χρησιμοποιείται για την ανάπτυξη και εξυπηρέτηση microservices, παρέχοντας γρήγορη και αξιόπιστη επικοινωνία μεταξύ υπηρεσιών.

3. Εφαρμογές Cloud:

Συχνά χρησιμοποιείται σε cloud-based αρχιτεκτονικές λόγω της επεκτασιμότητας και της συμβατότητάς του με υπηρεσίες όπως το Kubernetes.

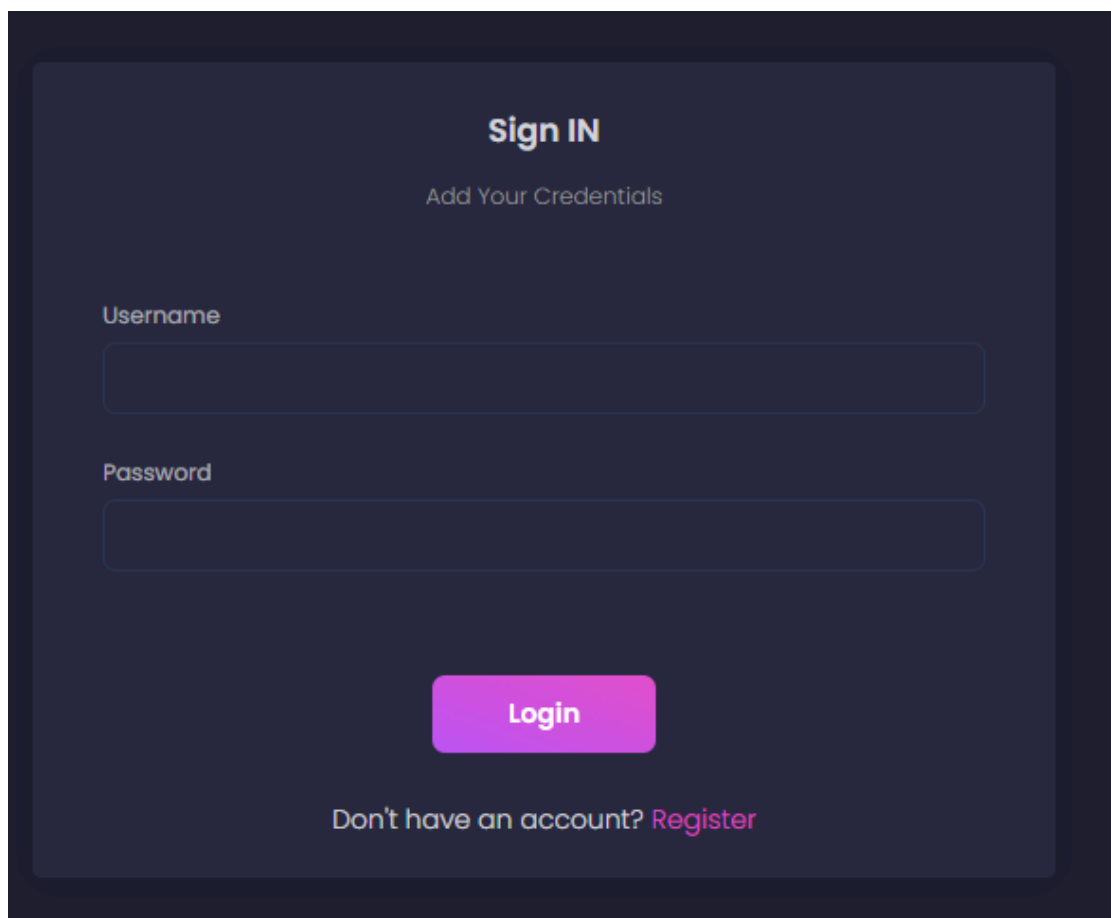
Το Gunicorn είναι ένα αξιόπιστο και αποδοτικό εργαλείο για τη φιλοξενία και την εξυπηρέτηση Python web εφαρμογών, προσφέροντας υψηλή απόδοση, επεκτασιμότητα και ευκολία στη χρήση. Παρατίθεται η γραμμή που ενεργοποιεί την εφαρμογή :

```
gunicorn -w 2 -b 0.0.0.0:5000 run:app
```

Βασικά στοιχεία είναι η παράμετρος **-w** που ορίζει πόσους workers θα έχει (καθορίζεται με βάση την ισχύ του μηχανήματος που τρέχει τον Gunicorn Server) και παράμετρος **-b** που ορίζει σε ποιες IP «ακούει» ο server και φυσικά σε ποια πόρτα (στο συγκεκριμένο παράδειγμα η εφαρμογή μας είναι διαθέσιμη από όλο τον κόσμο και είναι προσβάσιμο στην πόρτα **5000**)

5.3 Σύστημα αυθεντικοποίησης

Το σύστημα αυθεντικοποίησης που χρησιμοποιείται είναι σχετικά απλό . Ο χρήστης εφόσον έχει γραφτεί στην εφαρμογή απλώς εισάγει αναγνωριστικό χρήστη (username) & κωδικό πρόσβασης (password)



The image shows a dark-themed user interface for a sign-in page. At the top, the text "Sign IN" is displayed in a light color, followed by the subtitle "Add Your Credentials". Below this, there are two input fields: "Username" and "Password", each with a light blue border. A prominent blue button labeled "Login" is centered below the fields. At the bottom, there is a link that says "Don't have an account? Register", where "Register" is highlighted in blue.

Φυσικά εάν δεν είναι ήδη εγγεγραμμένος μπορεί να κάνει εγγραφή χρησιμοποιώντας το Register και θα μεταφερθεί στην σελίδα εγγραφής.

Sign UP
Add Your Credentials

Username

Email

Password

Register

Have an account? [Login](#)

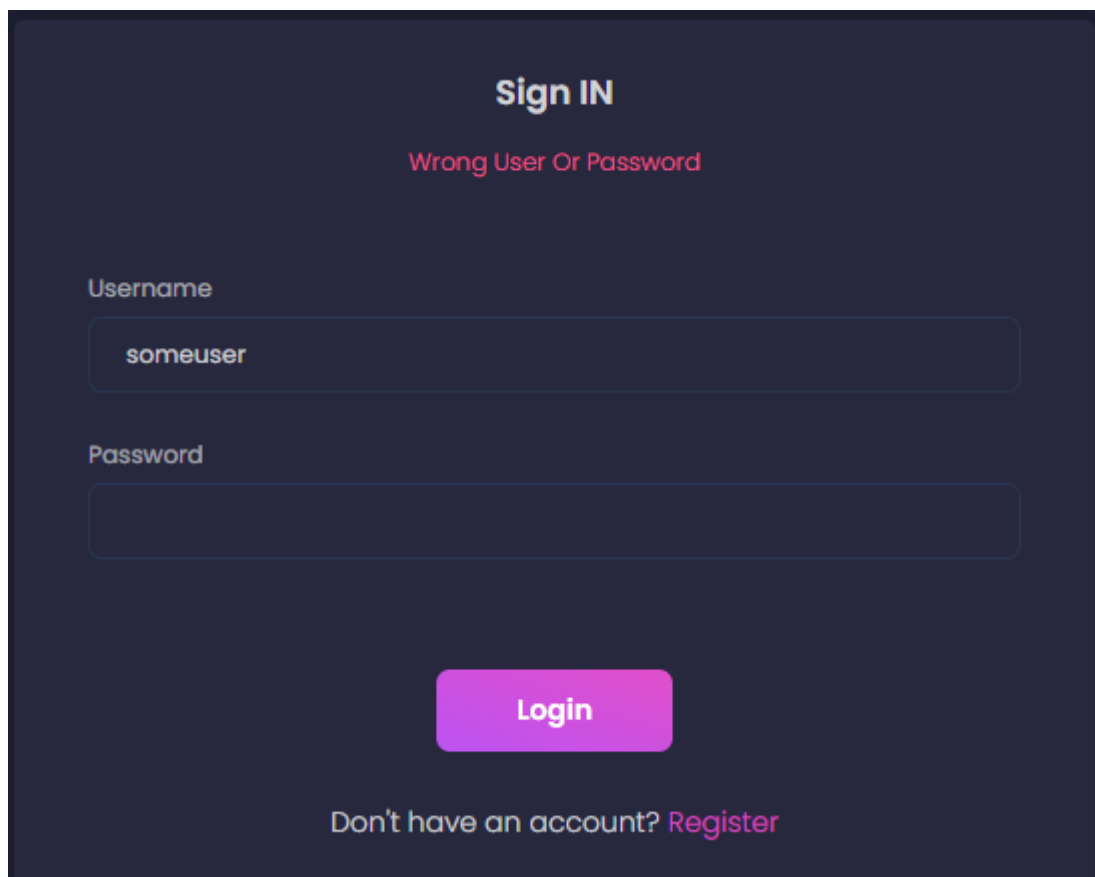
Εκεί θα δηλώσει το επιθυμητό όνομα χρήστη το email του και τον κωδικό που θα θέλει να χρησιμοποιήσει

Ο κώδικας που χρησιμοποιείται είναι σχετικά τετριμμένος καθώς ο κωδικός που εισάγει ο χρήστης κρυπτογραφείται με την δημιουργία salt με χρήση αλγόριθμου κρυπτογράφησης SHA256 και hashing που δημιουργείται με τον αλγόριθμο κρυπτογράφησης SHA512

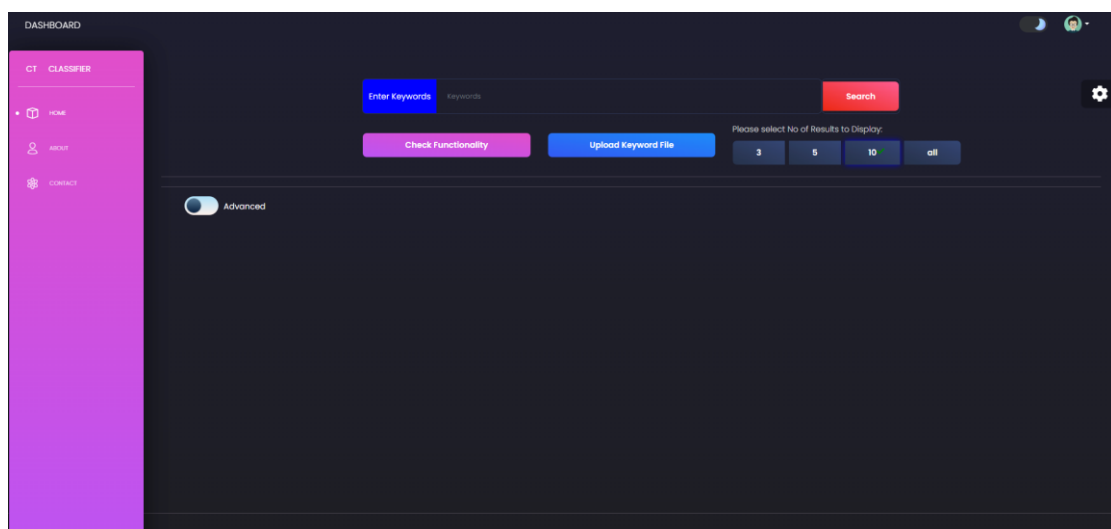
Αυτά ενώνονται σε 1 Byte array το οποίο αποθηκεύεται στην βάση δεδομένων

Στην αντίστροφη διαδικασία ο κωδικός ανακτάται , τοποθεείται σε 1 Byte array και αποκρυπτογραφείται.

Εφόσον υπάρχει κάποιο σφάλμα ο χρήστης ενημερώνεται με την εμφάνιση μηνύματος



5.4 Η κύρια οθόνη



Σε αυτή την κύρια οθόνη διακρίνουμε 2(+1) τμήματα

1. Το μενού στο αριστερό που έχει 3 επιλογές :

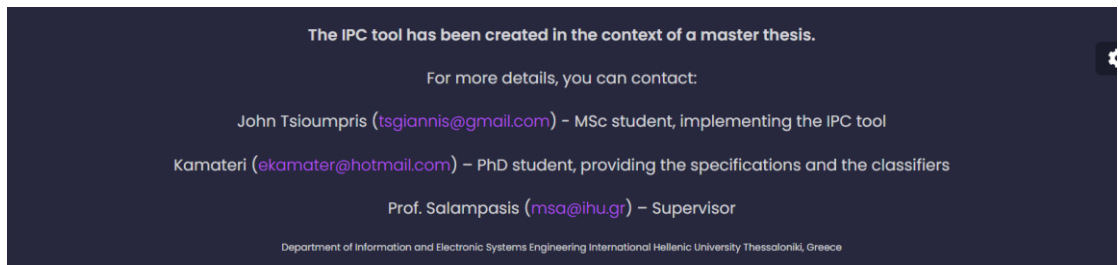
HOME : Με αυτή την επιλογή επιστρέφουμε στην αρχική οθόνη εφόσον έχουμε προχωρήσει στην εύρεση αποτελεσμάτων εφόσον θέλουμε να υποβάλλουμε καινούργιες λέξεις κλειδιά ή να αλλάξουμε κάποιες παραμέτρους

ABOUT : Γίνεται μια γρήγορη εμφάνιση των αρχικών Deep Learning μεθόδων, Language Models και Datasets που χρησιμοποιήθηκαν στο project

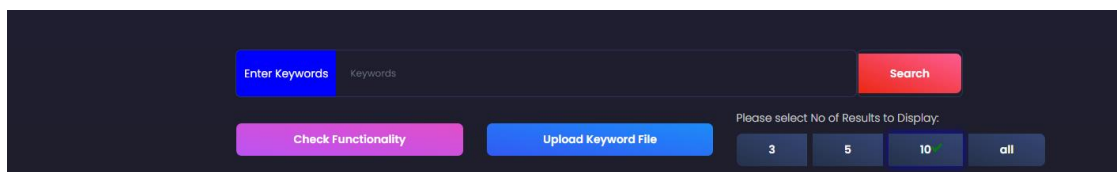
The screenshot shows three sections of the application interface:

- Datasets:** A table with columns: DISPLAY_NAME, NAME, LINK, OBJECT, SECTIONS. It lists datasets like CLEFIP-0.54M, USPTO-2M, and WIPO.
- Language Models:** A table with columns: DISPLAY_NAME, NAME, OBJECT, DESCRIPTION, LINK. It lists models like FastText (300-D) and Rich and Krestal (300-D).
- Deep Learning Classifiers:** A table with columns: NAME, FILE, OBJECT. It lists classifiers like LSTM and GRU.

CONTACT : Γίνεται μια παρουσίαση των συντελεστών στην δημιουργία αυτής της εφαρμογής

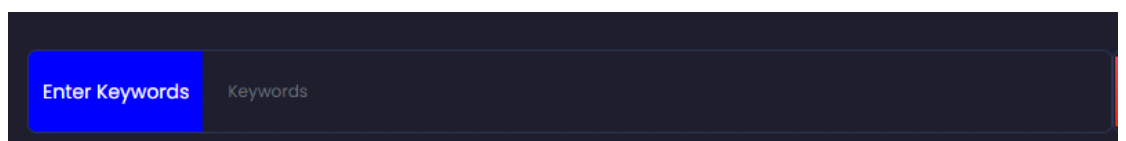


2. Στην κύρια οθόνη έχουμε την βασική λειτουργικότητα αλληλεπίδρασης με τον χρήστη

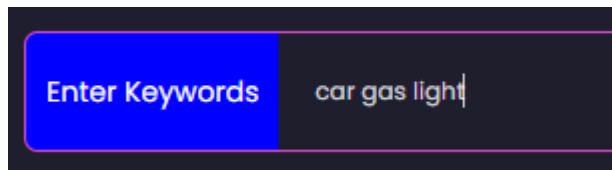


Εδώ ο χρήστης έχει τις εξής επιλογές

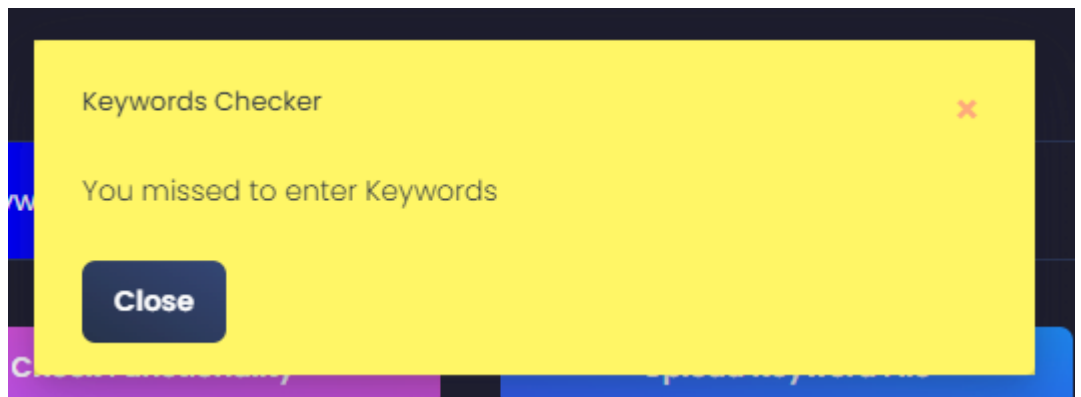
- Να εισάγει χειροκίνητα τις λέξεις κλειδιά αφήνοντας 1 κενό κάθε φορά



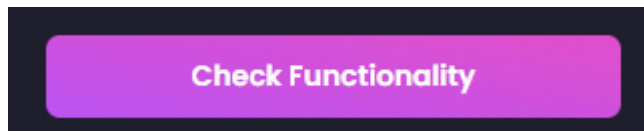
- Να κάνει κατευθείαν αναζήτηση (εφόσον έχει κάνει εισαγωγή των λέξεων)



Ειδάλλως θα εμφανιστεί το ενημερωτικό pop up ότι έγινε προσπάθεια αναζήτησης χωρίς πρώτα να γίνει εισαγωγή των λέξεων κλειδιά

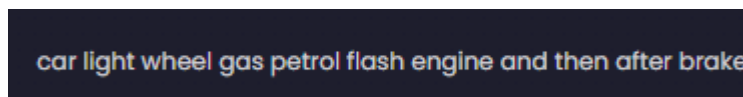


- Να ανεβάσει 1 αρχείο .txt στο οποίο έχουν προεισαχθεί λέξεις κλειδιά. Το αρχείο θα πρέπει να έχει την εξής απλή μορφή, 1 λέξη ανά γραμμή και διαβάζει μόνο τις πρώτες 60 λέξεις
- Την δυνατότητα με την χρήση του κουμπιού Check Functionality



να ελέγξει ποιες από τις λέξεις που εισήγαγε είναι λέξεις που μπορούν να χρησιμοποιηθούν στην εφαρμογή και να φιλτράρει τις λέξεις που είναι κοινότητες, η εφαρμογή έτσι και αλλιώς κάνει αυτή την διαδικασία κατά την τροφοδοσία του μοντέλου Βαθιάς Μάθησης αλλά μπορεί να το ελέγξει και ο χρήστης.

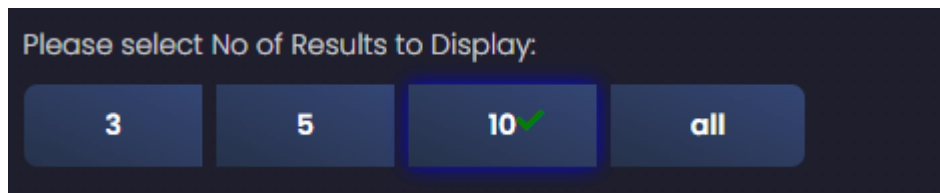
Π.χ για τις παρακάτω λέξεις :



Παρουσιάζει pop up πίνακα στις οποίους παρουσιάζει τις λέξεις και αν θα συμπεριληφθούν στην αναζήτηση ή όχι



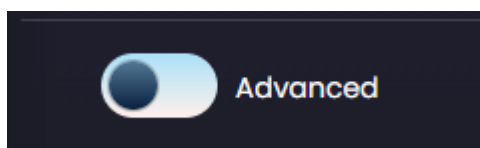
- Επίσης έχει την δυνατότητα να επιλέξει πόσα αποτελέσματα να γυρίσει : 3,5,10 ή όλα τα αποτελέσματα που θα επιστρέψει το νευρωνικό μοντέλο



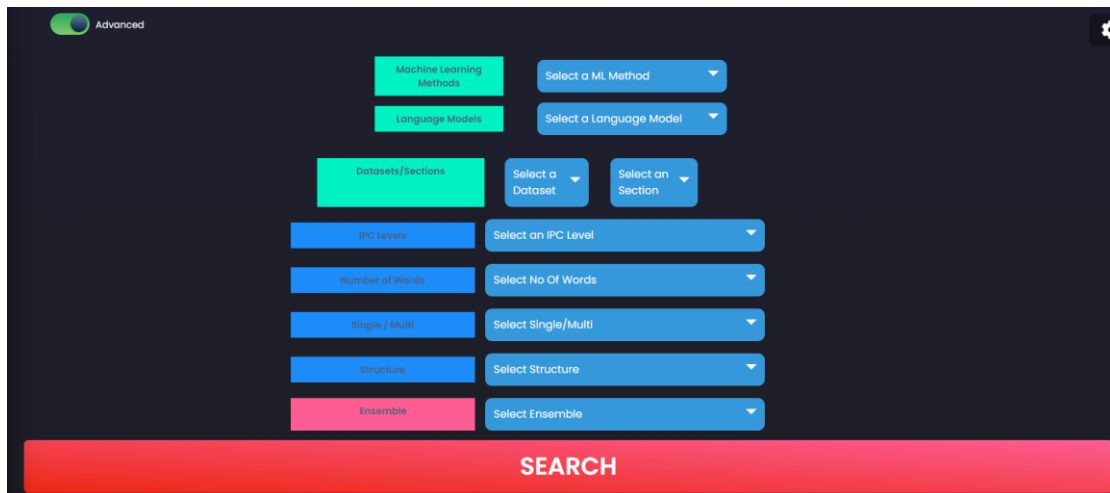
Η προεπιλογή είναι στα 10

Σε αυτό το σημείο η εφαρμογή έχει ολοκληρώσει την λειτουργικότητα της σε βασικό επίπεδο όπου ο χρήστης θέλει να κάνει αναζήτηση με τις προεπιλεγμένες ρυθμίσεις

Εάν όμως θέλει να δοκιμάσει και άλλες ρυθμίσεις αυτό γίνεται δυνατό με την χρήση του toggle Advanced

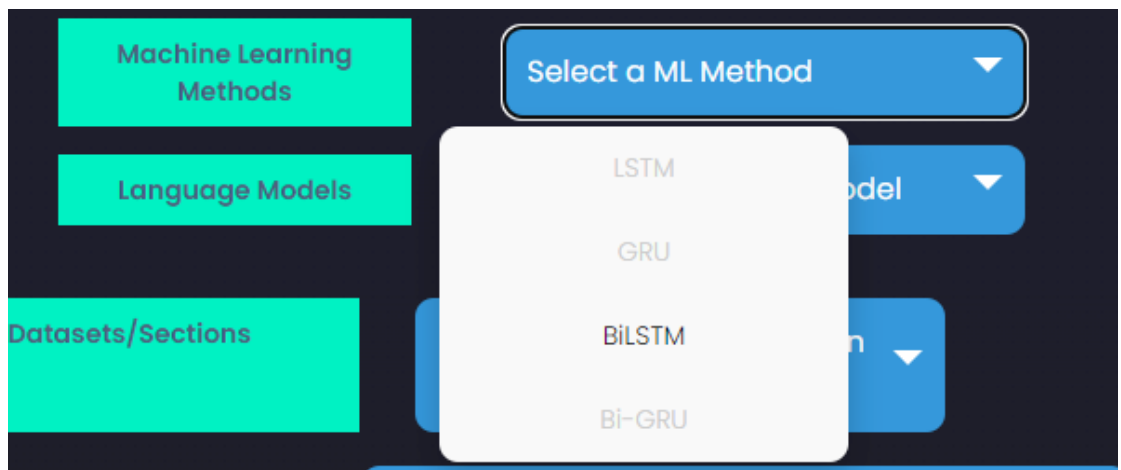


Επιλέγοντας το toggle γίνονται διαθέσιμες όλες οι δυνατές ρυθμίσεις



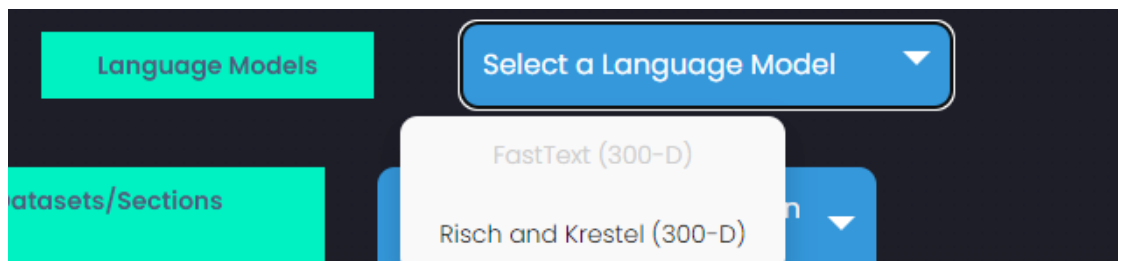
Οι επιλογές είναι οι παρακάτω και είναι πλήρως δυναμικά αφού διαβάζουν με το σύστημα αναζήτησης των διαθέσιμων φακέλων :

- Machine Learning Methods



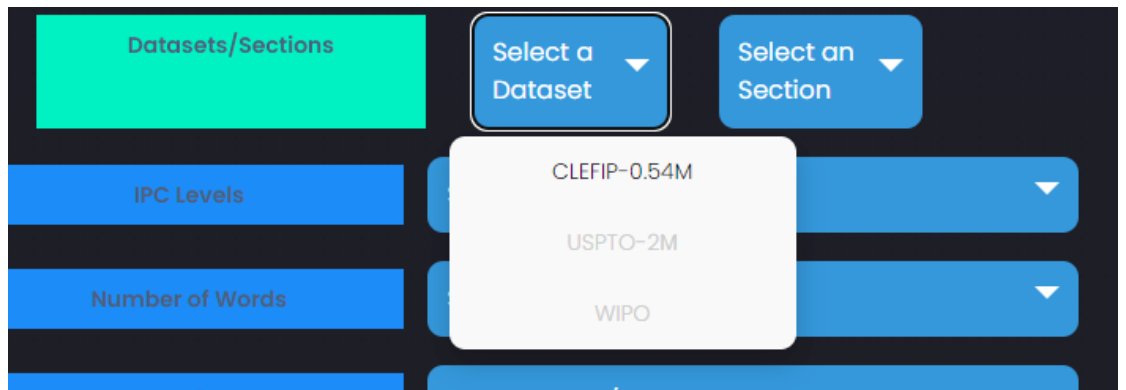
Υπάρχουν ήδη κάποιες προεπιλογές αλλά δεν ενεργοποιούνται αν δεν υπάρχουν διαθέσιμες, όπως φαίνεται και στην ανωτέρω εικόνα η αναζήτηση βρήκε μόνο διαθέσιμη μοντέλο BiLSTM

- Language Model

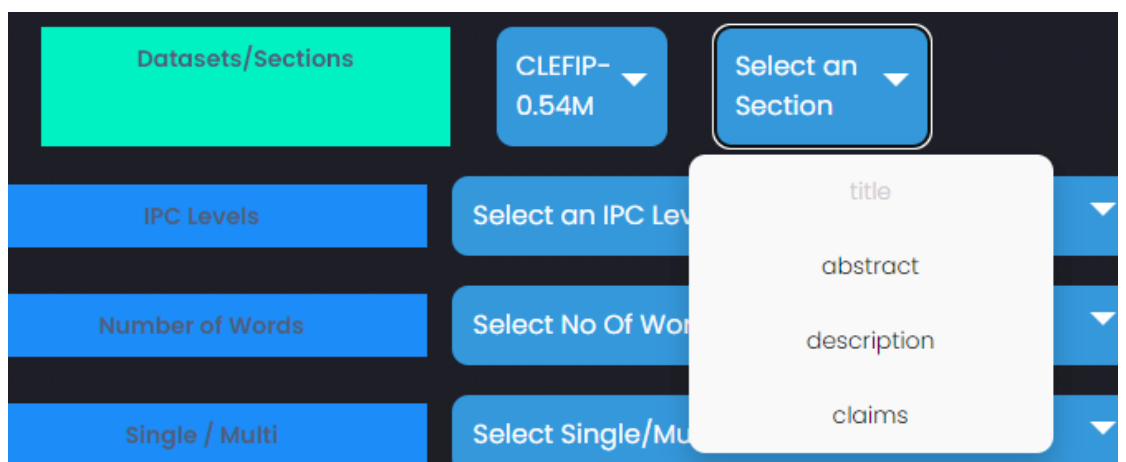


Εδώ βλέπουμε ότι είναι διαθέσιμο μόνο το Γλωσσικό Μοντέλο Risch and Krestel (300 -D)

- Datasets



Όμοια βλέπουμε ότι είναι διαθέσιμο μόνο το CLEFIP Dataset, με την επιλογή του Dataset γίνεται διαθέσιμο για επιλογή και ποια από τα Sections θα χρησιμοποιηθούν



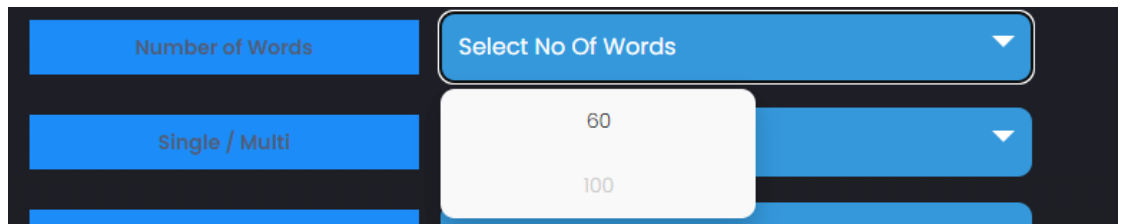
Εδώ βλέπουμε ότι το σύστημα βρήκε διαθέσιμα 3 διαφορετικά Sections, οπότε ο χρήστης μπορεί να επιλέξει αυτό που πιθανόν να του δώσει καλύτερα αποτελέσματα

- IPC Levels



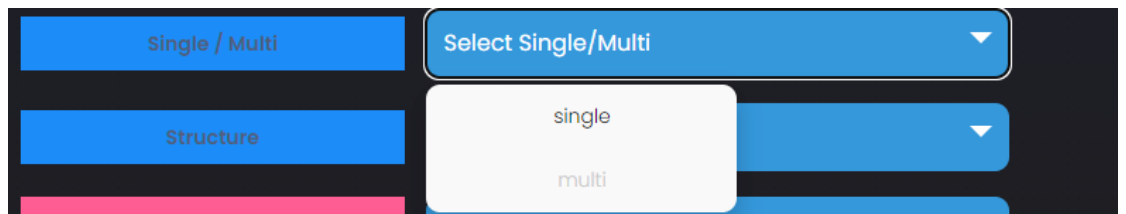
Προς το παρόν έχουμε διαθέσιμο μόνο το Level 3 τα Subclasses

- Number of words

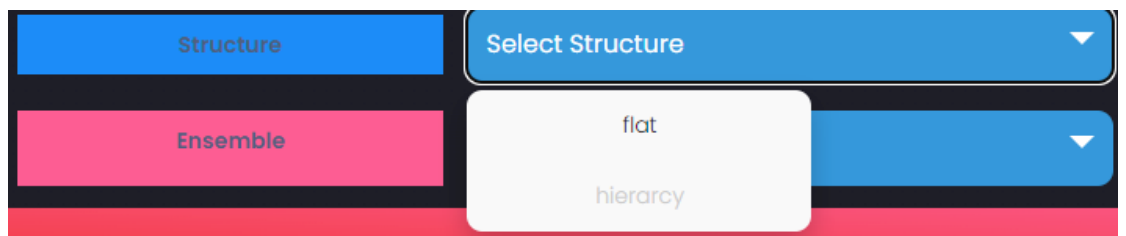


Με βάση τις διαθέσιμες επιλογές είναι ενεργό μόνο το 60

- Single / Multi Label

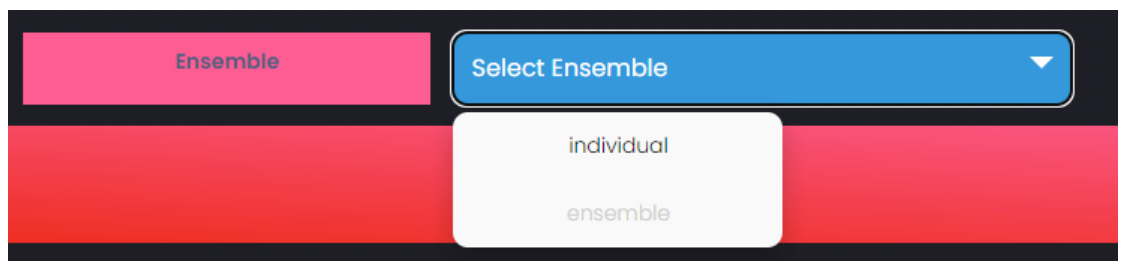


- Structure



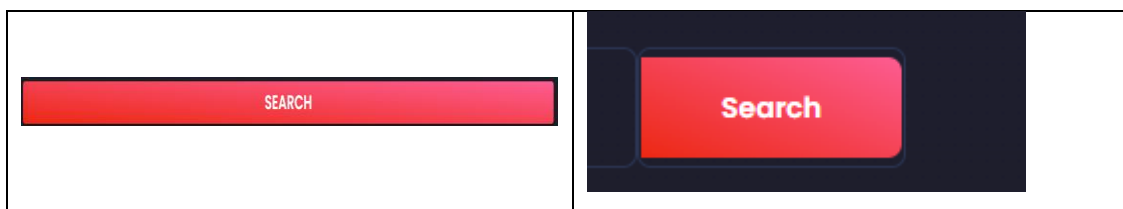
Διαθέσιμο προς το παρόν μόνο το flat

- Ensemble



Παρόμοια η τελευταία επιλογή είναι το individual, ensemble με διαθέσιμο μόνο το individual

Μετά ο χρήστης μπορεί να πατήσει το Search είτε αυτό που είναι στο τμήμα των προχωρημένων ρυθμίσεων είτε αυτό που είχε στο Compact (User friendly)



5.4.1 Αποτελέσματα

Εφόσον ο χρήστης έχει κάνει την εισαγωγή των λέξεων και έχει πατήσει την Αναζήτηση (Search) θα μεταφερθεί στην επόμενη οθόνη:

CODES	DESCRIPTIONS	RELEVANT/IRRELEVANT
B60T	Vehicle Brake Control Systems Or Parts Thereof Brake Control Systems Or Parts Thereof, In General Arrangement Of Braking Elements On Vehicles In General Portable Devices For Preventing Unwanted Movement Of Vehicles Vehicle Modifications To Facilitate Cooling Of Brakes	Relevant
F02D	Controlling Combustion Engines	Irrelevant
B60Q	Arrangement Of Signalling Or Lighting Devices, The Mounting Or Supporting Thereof Or Circuits Therefor, For Vehicles In General	Irrelevant
B60K	Arrangement Or Mounting Of Propulsion Units Or Of Transmissions In Vehicles Arrangement Or Mounting Of Plural Diverse Prime-movers In Vehicles Auxiliary Drives For Vehicles Instrumentation Or Dashboards For Vehicles Arrangements In Connection With Cooling, Air Intake, Gas Exhaust Or Fuel Supply Of Propulsion Units In Vehicles	Relevant
C10M	Lubricating Compositions Use Of Chemical Substances Either Alone Or As Lubricating Ingredients In A Lubricating Composition	Irrelevant
B60R	Vehicles, Vehicle Fittings, Or Vehicle Parts, Not Otherwise Provided For	Relevant
F16C	Shafts Flexible Shafts Elements Or Crankshaft Mechanisms Rotary Bodies Other Than Gearing Elements Bearings	Relevant
F16H	Gearing	Relevant

Εδώ παρατηρούμε τα εξής υποτιμήματα ξεκινώντας από πάνω αριστερά:

- Ο αύξων αριθμός του πειράματος

Experiment NO 295

- Οι λέξεις κλειδιά που χρησιμοποιήθηκαν (είναι φιλτραρισμένες και δεν περιλαμβάνουν κοινότυπες λέξεις)
- Το κουμπί «Save to database» με το οποίο θα ασχοληθούμε αργότερα
- Το κουμπί «Export to CSV» στο οποίο θα υπάρχει αναφορά αργότερα
- Ακριβώς απέναντι είναι το κουμπί «Compare to EspaceNet» με το οποίο θα ασχοληθούμε αργότερα
- Ο κύριος πίνακας αποτελεσμάτων στον οποίο γίνεται πλήρης παρουσίαση των κωδικών που βρήκε αντιστοίχιση το νευρωνικό μοντέλο σύμφωνα με τις ρυθμίσεις που υπέβαλε ο χρήστης (προεπιλεγμένες ή επιλεγμένες από τον χρήστη)

CODES	DESCRIPTIONS	RELEVANT/IRRELEVANT
B60T	Vehicle Brake Control Systems Or Parts Thereof Brake Control Systems Or Parts Thereof, In General Arrangement Of Braking Elements On Vehicles In General Portable Devices For Preventing Unwanted Movement Of Vehicles Vehicle Modifications To Facilitate Cooling Of Brakes	<input checked="" type="checkbox"/> RELEVANT
F02D	Controlling Combustion Engines	<input checked="" type="checkbox"/> RELEVANT
B60Q	Arrangement Of Signalling Or Lighting Devices, The Mounting Or Supporting Thereof Or Circuits Therefor, For Vehicles In General	<input checked="" type="checkbox"/> RELEVANT
B60K	Arrangement Or Mounting Of Propulsion Units Or Of Transmissions In Vehicles Arrangement Or Mounting Of Plural Diverse Prime-movers In Vehicles Auxiliary Drives For Vehicles Instrumentation Or Dashboards For Vehicles Arrangements In Connection With Cooling, Air Intake, Gas Exhaust Or Fuel Supply Of Propulsion Units In Vehicles	<input checked="" type="checkbox"/> RELEVANT
C10M	Lubricating Compositions Use Of Chemical Substances Either Alone Or As Lubricating Ingredients In A Lubricating Composition	<input checked="" type="checkbox"/> RELEVANT
B60R	Vehicles, Vehicle Fittings, Or Vehicle Parts, Not Otherwise Provided For	<input checked="" type="checkbox"/> RELEVANT
F16C	Shafts Flexible Shafts Elements Or Crankshaft Mechanisms Rotary Bodies Other Than Gearing Elements Bearings	<input checked="" type="checkbox"/> RELEVANT
F16H	Gearing	<input checked="" type="checkbox"/> RELEVANT

Ο πίνακας έχει 3 στήλες

- ❖ Στην πρώτη στήλη έχουμε τον κωδικό της πατέντας σύμφωνα με το IPC Level και παράλληλα πατώντας πάνω στον κωδικό έχει την δυνατότητα να μεταφερθεί στην ιστοσελίδα του EspaceNet και να πάρει περισσότερες και αναλυτικότερες πληροφορίες π.χ. για τον κωδικό πατέντας C10M

Cooperative Patent Classification

Search for Search View section | Index | A | B | **C** | D | E | F | G | H | Y |

« C10L2300/00 C10M1/00 »

Symbol	Classification and description
<input type="checkbox"/> C	CHEMISTRY; METALLURGY s i

CHEMISTRY

<input type="checkbox"/> C10	PETROLEUM, GAS OR COKE INDUSTRIES; TECHNICAL GASES CONTAINING CARBON MONOXIDE; FUELS; LUBRICANTS; PEAT
<input type="checkbox"/> C10M	LUBRICATING COMPOSITIONS (well drilling compositions C09K 8/02); USE OF CHEMICAL SUBSTANCES EITHER ALONE OR AS LUBRICATING INGREDIENTS IN A LUBRICATING COMPOSITION (lubricants for medical use A61); mould release, i.e. separating, agents for metals B22C 3/00 , for plastics or substances in a plastic state, in general B29C 33/56 , for glass C03B 40/02 ; use of particular substances in particular apparatus or conditions, see F16N or the relevant groups for the application, e.g. A21D 8/08 , B21C 9/00 , H01B 3/18 ; immersion oils for microscopy G02B 21/33) s d i !

IPC3 groups

<input type="checkbox"/> C10M 1/00	{Liquid compositions essentially based on mineral lubricating oils or fatty oils; Their use as lubricants}
<input type="checkbox"/> C10M 3/00	{Liquid compositions essentially based on lubricating components other than mineral lubricating oils or fatty oils and their use as lubricants; Use as lubricants of single liquid substances (compositions in general essentially based on macromolecular compounds C08L)}
<input type="checkbox"/> C10M 5/00	{Solid or semi-solid compositions containing as the essential lubricating ingredient mineral lubricating oils or fatty oils and their use}
<input type="checkbox"/> C10M 7/00	{Solid or semi-solid compositions essentially based on lubricating components other than mineral lubricating oils or fatty oils and their use as lubricants; Use as lubricants of single solid or semi-solid substances (compositions in general essentially based on macromolecular compounds C08L)}

Base-Materials

<input type="checkbox"/> C10M 101/00	Lubricating compositions characterised by the base-material being a mineral or fatty oil (containing more than 10% water C10M 173/00) d
<input type="checkbox"/> C10M 103/00	Lubricating compositions characterised by the base-material being an inorganic material (containing more than 10% water C10M 173/00) d
<input type="checkbox"/> C10M 105/00	Lubricating compositions characterised by the base-material being a non-macromolecular organic compound d
<input type="checkbox"/> C10M 107/00	Lubricating compositions characterised by the base-material being a macromolecular compound d
<input type="checkbox"/> C10M 109/00	Lubricating compositions characterised by the base-material being a compound of unknown or incompletely defined constitution (C10M 101/00 takes precedence) d
<input type="checkbox"/> C10M 111/00	Lubrication compositions characterised by the base-material being a mixture of two or more compounds covered by more than one of the main groups C10M 101/00 - C10M 109/00 , each of these compounds being essential d

- ❖ Στην δεύτερη στήλη έχουμε την περιγραφή της πατέντας σύμφωνα με το Espacenet
- ❖ Στην 3^η στήλη υπάρχει η δυνατότητα ο χρήστης να κάνει μια εκτίμηση των αποτελεσμάτων και στην συνέχεια να επιλέξει να τα σώσει στην βάση δεδομένων ώστε να υπάρχει έλεγχος στην απόδοσης της εφαρμογής και συνεπακόλουθα σε μελλοντικό χρόνο να υπάρχει δυνατότητα για βελτίωση αυτής μέσω της επιλογής RELEVANT (σχετική) ή IRRELEVANT (μη σχετική)



CODES	DESCRIPTIONS	RELEVANT/IRRELEVANT
B60T	Vehicle Brake Control Systems Or Parts Thereof Brake Control Systems Or Parts Thereof, In General Arrangement Of Braking Elements On Vehicles In General Portable Devices For Preventing Unwanted Movement Of Vehicles Vehicle Modifications To Facilitate Cooling Of Brakes	<input checked="" type="checkbox"/>
F02D	Controlling Combustion Engines	<input type="checkbox"/>
B60Q	Arrangement Of Signalling Or Lighting Devices, The Mounting Or Supporting Thereof Or Circuits Therefor, For Vehicles In General	<input checked="" type="checkbox"/>
B60K	Arrangement Or Mounting Of Propulsion Units Or Of Transmissions In Vehicles Arrangement Or Mounting Of Plural Diverse Prime-movers In Vehicles Auxiliary Drives For Vehicles Instrumentation Or Dashboards For Vehicles Arrangements In Connection With Cooling, Air Intake, Gas Exhaust Or Fuel Supply Of Propulsion Units In Vehicles	<input type="checkbox"/>
C10M	Lubricating Compositions Use Of Chemical Substances Either Alone Or As Lubricating Ingredients In A Lubricating Composition	<input checked="" type="checkbox"/>
B60R	Vehicles, Vehicle Fittings, Or Vehicle Parts, Not Otherwise Provided For	<input checked="" type="checkbox"/>
F16C	Shafts Flexible Shafts Elements Or Crankshaft Mechanisms Rotary Bodies Other Than Gearing Elements Bearings	<input type="checkbox"/>
F16H	Gearing	<input checked="" type="checkbox"/>

(Η επιλογή έγινε τυχαία καθαρά για την παρουσίαση της λειτουργικότητας)
 Παράλληλα με την βάση μπορεί να γίνει και εξαγωγή σε CSV στο οποίο
 μπορεί ο χρήστης να κάνει εξαγωγή σε CSV(Μαρκκαρίστηκαν με bold οι
 κωδικοί για ευκολότερη απεικόνιση στο Word)

CODES,DESCRIPTIONS,RELEVANT/IRRELEVANT
B60T ,Vehicle Brake Control Systems Or Parts Thereof Brake Control Systems Or Parts Thereof, In General Arrangement Of Braking Elements On Vehicles In General Portable Devices For Preventing Unwanted Movement Of Vehicles Vehicle Modifications To Facilitate Cooling Of Brakes,1
F02D ,Controlling Combustion Engines,2
B60Q ,Arrangement Of Signalling Or Lighting Devices, The Mounting Or Supporting Thereof Or Circuits Therefor, For Vehicles In General,2
B60K ,Arrangement Or Mounting Of Propulsion Units Or Of Transmissions In Vehicles Arrangement Or Mounting Of Plural Diverse Prime-movers In Vehicles Auxiliary Drives For Vehicles Instrumentation Or Dashboards For Vehicles Arrangements In Connection With Cooling, Air Intake, Gas Exhaust Or Fuel Supply

Of Propulsion Units In Vehicles,1
C10M,Lubricating Compositions Use Of Chemical Substances Either Alone Or As Lubricating Ingredients In A Lubricating Composition,2
B60R,Vehicles, Vehicle Fittings, Or Vehicle Parts, Not Otherwise Provided For,1
F16C,Shafts Flexible Shafts Elements Or Crankshaft Mechanisms Rotary Bodies Other Than Gearing Elements Bearings,1
F16H,Gearing,1
F02N,Starting Of Combustion Engines Starting Aids For Such Engines, Not Otherwise Provided For,1
B60B,Vehicle Wheels Castors Axles For Wheels Or Castors Increasing Wheel Adhesion,1

Τέλος ο χρήστης μπορεί να επιλέξει να συγκρίνει τα αποτελέσματα που έχουν επιστραφεί με το σύστημα Αναζήτησης του EspaceNet το οποίο όμως έχει περιορισμό ότι μπορεί να δεχτεί έως 4 λέξεις οπότε στην παρούσα φάση γίνεται χρήση μόνο των πρώτων 4 λέξεων από τις λέξεις που έδωσε ο χρήστης.

CODES	DESCRIPTIONS	RELEVANT/IRRELEVANT	Comparison results for the first 4 keywords		
CODES	PERCENTAGE	DESCRIPTIONS			
B60R	Vehicle Brake Control Systems Or Parts Thereof Brake Control Systems Or Parts Thereof, In General Arrangement Of Braking Elements On Vehicles In General Portable Devices For Preventing Unwanted Movement Of Vehicles Vehicle Modifications To Facilitate Cooling Of Brakes	RELEVANT	Y02T10/00	0.000000	Road transport of goods or passengers
F02N	Controlling Combustion Engines	IRRELEVANT	B62K11/00	0.000000	Motorcycles, engine-assisted cycles or motor scooters with one or two wheels
B60R	Arrangement Of Signalling Or Lighting Devices, The Mounting Or Supporting Thereof Or Circuits Thereof, For Vehicles In General	RELEVANT	B60K31/00	0.000000	Vehicle fittings, acting on a single sub-unit only, for automatically controlling vehicle speed, i.e. preventing speed from exceeding an arbitrarily established velocity or maintaining speed at a particular velocity, as selected by the vehicle operator
B60R	Arrangement Or Mounting Of Propulsion Units Or Of Transmissions In Vehicles Arrangement Or Mounting Of Plural Diverse Prime-movers In Vehicles Auxiliary Drives For Vehicles Instrumentation Or Dashboards For Vehicles Arrangements In Connection With Cooling, Air Intake, Gas Exhaust Or Fuel Supply Of Propulsion Units In Vehicles	IRRELEVANT	F02C3/00	0.000000	Gas-turbine plants characterised by the use of combustion products as the working fluid
C10M	Lubricating Compositions Use Of Chemical Substances Either Alone Or As Lubricating Ingredients In A Lubricating Composition	RELEVANT	C23CB/00	0.000000	Solid state diffusion of only non-metal elements into metallic material surfaces
B60R	Vehicles, Vehicle Fittings, Or Vehicle Parts, Not Otherwise Provided For	RELEVANT	A47J37/00	0.000000	Baking
F16C	Shafts Flexible Shafts Elements Or Crankshaft Mechanisms Rotary Bodies Other Than Gearing Elements Bearings	IRRELEVANT	B62K2202/00	0.000000	Motorised scooters
			B62J15/00	0.000000	Mud-guards for wheels
			Y04S30/00	0.000000	Systems supporting specific end-user applications in the sector of transportation

Σε αυτό έχει προστεθεί η λειτουργικότητα να εξετάζεται ο βαθμός συνοχής και ανάλογα να χρωματίζονται οι πατέντες που επέστρεψε η δικιά μας εφαρμογή σε σχέση με το EspaceNet , όσο μεγαλύτερη συνάφεια υπάρχει τόσο πιο έντονο είναι το χρώμα (στην παρούσα φάση καθώς το EspaceNet χρησιμοποιεί πιο χαμηλό IPC level δεν είναι δυνατόν να παρουσιαστούν όλα τα διαθέσιμα χρώματα αλλά υπάρχει η πρόβλεψη αν προστεθεί μοντέλο που να κάνει πιο στοχευμένη αναζήτηση στο μέλλον να υπάρχει καλύτερη αντιστοίχιση)

6

Λοιπές Αναφορές

6.1 Εισαγωγή

Στο συγκεκριμένο κεφάλαιο θα γίνει αναφορά σε σημεία κώδικα που παρουσιάζουν ιδιαίτερο ενδιαφέρον και επίσης μια αναφορά στο EspaceNet καθώς παίζει ιδιαίτερο ενδιαφέρον καθώς παρέχει την περιγραφή των κωδικών των πατεντών.

6.2 Στοιχεία Κώδικα

Οι ιδιαίτερες αναφορές στον κώδικα είναι :

- **JINJA**

Η επικοινωνία του BE με το FE πραγματοποιείται μέσω του Jinja που είναι μια ισχυρή και ευέλικτη μηχανή προτύπων (template engine) που χρησιμοποιείται σε συνδυασμό με το Flask και επιτρέπει στους προγραμματιστές να δημιουργούν δυναμικά HTML έγγραφα, εισάγοντας δεδομένα από το backend της εφαρμογής με ευκολία και ευελιξία και μπορούν να αποδώσουν (render) πρότυπα χρησιμοποιώντας τη μέθοδο `render_template` του Flask. Ένα παράδειγμα βασικής χρήσης έχει ως εξής:

```
return render_template('home/index.html', selected_option=None,
segment='index', **locals())
```

Βασικά Χαρακτηριστικά του Jinja

1. Ενσωμάτωση Δεδομένων: Το Jinja επιτρέπει την ενσωμάτωση δεδομένων σε HTML πρότυπα χρησιμοποιώντας διακριτική σύνταξη. Για παράδειγμα, οι μεταβλητές μπορούν να ενσωματωθούν χρησιμοποιώντας διπλές αγκύλες `{{ variable }}`, που καθιστούν τα δεδομένα από το Python backend εύκολα προσβάσιμα στο HTML.
2. Λογικοί Έλεγχοι και Βρόχοι: Το Jinja υποστηρίζει λογικούς ελέγχους (if statements) και βρόχους (for loops), επιτρέποντας την εκτέλεση δυναμικών λειτουργιών μέσα στα πρότυπα. Για παράδειγμα:

```
{% if selected_option %}
<div class="alert alert-success" role="alert">
    {{ selected_option }}
</div>
{% endif %}
```

3. Κληρονομικότητα Προτύπων**: Το Jinja υποστηρίζει την κληρονομικότητα προτύπων, επιτρέποντας στους προγραμματιστές να δημιουργούν βασικά πρότυπα (base templates) από τα οποία μπορούν να κληρονομήσουν άλλα πρότυπα. Αυτό βοηθά στη διατήρηση συνεπούς εμφάνισης και αίσθησης σε ολόκληρη την εφαρμογή.

```
{% extends "layouts/base.html" %}
```

Το Jinja, ως μηχανή προτύπων για το Flask, προσφέρει ευέλικτες και δυναμικές δυνατότητες για την ανάπτυξη web εφαρμογών. Επιτρέποντας την ενσωμάτωση δεδομένων, τη χρήση λογικών ελέγχων και βρόχων, την εφαρμογή φίλτρων και μακροεντολών, και την κληρονομικότητα προτύπων, το Jinja βοηθά τους προγραμματιστές να δημιουργούν καθαρό, επαναχρησιμοποιήσιμο και συντηρήσιμο κώδικα.

- **Flask Session object**

Η επικοινωνία μεταξύ FE & BE γίνεται με την χρήση του αντικειμένου Flask Session που επιτρέπει την αποθήκευση δεδομένων σε μορφή ζευγαριού : κλειδί + τιμή και επιτρέπει τη διατήρηση δεδομένων μεταξύ των αιτημάτων ενός χρήστη.

Στην εφαρμογή χρησιμοποιείται για να αποθηκεύονται οι επιλογές του χρήστη και παράλληλα να τροφοδοτείται η εφαρμογή με προεπιλογές

```
session['methods'], session['languagemodels'],
session['datasets'], session['ipclevels'], \
session['noofwords'], session['singlemulti'],
session['structures'], session['ensemble'] = load_config()
```

▪ Επικοινωνία με EspaceNet

○ EspaceNet Token

Το EspaceNet θα παρουσιαστεί εκτενώς στην επόμενη ενότητα αυτό του κεφαλαίου αλλά είναι ενδιαφέρον να γίνει μια γρήγορη αναφορά καθώς η επικοινωνία είχε αρκετές μικροπροκλήσεις καθώς η βιβλιογραφία δεν βοηθάει ιδιαίτερα

Το πιο απλό αλλά και παράλληλο «μυστηριώδες» ήταν η δημιουργία του EPO_token του μοναδικού «κλειδιού» που επιτρέπει την επικοινωνία με το API του EspaceNet

Η εγγραφή στο site της EspaceNet παρέχει στον τελικό ενδιαφερόμενο χρήστη 2 κλειδιά το consumer_key και το consumer_secret.

Δυστυχώς σε αυτό το σημείο δεν υπάρχει ευκρινής πληροφόρηση σχετικά με την χρήση αυτών για την δημιουργία του EPO_token, αλλά με αναζήτηση η λύση ήταν ιδιαίτερα απλή και παρουσιάζεται στο παρακάτω snippet κώδικα :
(τα κλειδιά έχουν αντικατασταθεί)

```
def get_EPO_key():
    consumer_key = "*****"
    consumer_secret = "+++++++"

    # Concatenate consumer key and consumer secret with a colon
    combined = f"{consumer_key}:{consumer_secret}"

    # Encode the combined string to Base64
    encoded = base64.b64encode(combined.encode()).decode()
    return encoded
```

○ XML response

Το API του EspaceNet επιστρέφει στα requests μας XML περιεχόμενο αλλά η μορφή του δεν είναι «formatted» με ευκολία στην εξαγωγή των πληροφοριών που χρειάζονται με αποτέλεσμα αναγκαστικά να στηρίζεται σε συγκεκριμένες λέξεις κλειδιά αλλά η εξαγωγή δεν είναι πάντα επιτυχημένη κυρίως όσον αφορά την πλήρη έκταση της πληροφορίας

Παρατίθεται παράδειγμα του XML

```
<cpc:title-part>\n\n      <cpc:text>VEHICLE BRAKE CONTROL SYSTEMS OR\n      PARTS THEREOF</cpc:text>\n\n\n\n      </cpc:title-part>\n      <cpc:title-part>\n\n      <cpc:text>BRAKE CONTROL SYSTEMS OR PARTS\n      THEREOF, IN GENERAL </cpc:text>\n\n\n      <cpc:explanation>\n\n      <cpc:text>electrodynamic brake systems for vehicle, in\n      general\n\n      <cpc:class-ref scheme="cpc">B60L</cpc:class-ref>; brakes\n\n      <cpc:u>per se</cpc:u>, i.e. devices where braking effect\n      occurs, including ultimate brake actuators,\n\n      <cpc:class-ref scheme="cpc">F16D</cpc:class-ref>\n\n      </cpc:text>\n\n      </cpc:explanation>\n\n\n      </cpc:title-part>
```

6.3 *EspaceNet*

Το Espacenet είναι μια δωρεάν διαδικτυακή υπηρεσία που παρέχεται από το Ευρωπαϊκό Γραφείο Διπλωμάτων Ευρεσιτεχνίας (European Patent Office - EPO), η οποία επιτρέπει την αναζήτηση και την πρόσβαση σε παγκόσμια πατενταρισμένα έγγραφα. Χρησιμοποιείται ευρέως από ερευνητές, εφευρέτες, και επαγγελματίες της βιομηχανίας για να εξερευνήσουν

υπάρχουσες ευρεσιτεχνίες, να αξιολογήσουν την καινοτομία και να διασφαλίσουν τη μοναδικότητα των εφευρέσεων τους.

Τα Βασικά Χαρακτηριστικά του Espacenet :

- Δωρεάν Πρόσβαση:
Παρέχει ελεύθερη πρόσβαση σε ένα τεράστιο όγκο πατενταρισμένων πληροφοριών, χωρίς την ανάγκη συνδρομής ή πληρωμής.
- Διεθνής Κάλυψη:
Καλύπτει πατέντες από πολλές χώρες και περιφέρειες, παρέχοντας μια παγκόσμια προοπτική για την κατάσταση της τεχνολογίας και της καινοτομίας.
- Ευκολία στη Χρήση:
Εύκολη πλοήγηση και χρήση, καθιστώντας το προσβάσιμο σε χρήστες με διάφορα επίπεδα εμπειρίας στην αναζήτηση πατεντών.
- Εκτενής Βάση Δεδομένων:
Προσφέρει πρόσβαση σε περισσότερα από 120 εκατομμύρια πατενταρισμένα έγγραφα από όλο τον κόσμο, καλύπτοντας διάφορες τεχνολογίες και βιομηχανίες.
- Πολλαπλές Επιλογές Αναζήτησης:
 - Απλή Αναζήτηση: Κατάλληλη για αρχάριους, επιτρέπει την αναζήτηση με λέξεις-κλειδιά, ονόματα εφευρετών ή αιτούντων, και αριθμούς πατεντών.
 - Σύνθετη Αναζήτηση: Προσφέρει πιο εξελιγμένες επιλογές, όπως λογικούς τελεστές, αναζήτηση ανά ημερομηνία, και εξειδικευμένα φίλτρα για ακριβέστερα αποτελέσματα.
- Πρόσβαση σε Πλήρη Κείμενα:
Δυνατότητα προβολής των πλήρων κειμένων των πατεντών, συμπεριλαμβανομένων των περιγραφών, των απαιτήσεων (claims), και των σχεδίων.
- Διεπαφή Χρήστη:
Φιλική προς τον χρήστη διεπαφή με υποστήριξη πολλών γλωσσών, που διευκολύνει την πλοήγηση και την αναζήτηση.
- Εργαλείο Έρευνας και Ανάπτυξης:
Χρήσιμο εργαλείο για την έρευνα και ανάπτυξη, βοηθώντας τους ερευνητές να εντοπίσουν υπάρχουσες τεχνολογίες, να αποφύγουν παραβιάσεις πατεντών, και να εμπνευστούν από υπάρχουσες καινοτομίες.
- Παροχή Αναφορών:
Δυνατότητα δημιουργίας και αποθήκευσης αναφορών πατεντών, βοηθώντας τους χρήστες να οργανώσουν και να διαχειριστούν τις πληροφορίες που βρίσκουν.

Πρακτικές Χρήσεις του Espacenet

- Αναζήτηση Υφιστάμενων Πατεντών: Βοηθάει στον εντοπισμό ήδη υπάρχοντων πατεντών για να αποφευχθεί η επικάλυψη με νέες αιτήσεις.
- Αξιολόγηση Τεχνολογίας: Διευκολύνει την ανάλυση των τεχνολογικών τάσεων και της κατάστασης της τέχνης σε συγκεκριμένους τομείς.
- Νομική Προστασία: Βοηθά στην προστασία της πνευματικής ιδιοκτησίας μέσω της κατανόησης των ήδη υπάρχουσων πατεντών και της νομολογίας που σχετίζεται με αυτές.

Το Espacenet αποτελεί έναν αναντικατάστατο πόρο για όσους ασχολούνται με την καινοτομία, την έρευνα, και την προστασία της πνευματικής ιδιοκτησίας, προσφέροντας εύκολη πρόσβαση σε μια εκτενή συλλογή πατενταρισμένων δεδομένων από όλο τον κόσμο.

7

Επίλογος

Η δημιουργία μιας διεπαφής χρήστη για μια τόσο ενδιαφέρουσα προϋπάρχουσα εφαρμογή είναι σίγουρα μια πρόκληση καθώς πρέπει να επιτύχει να δημιουργήσει μια γέφυρα ανάμεσα στον μέσο χρήστη που λογικά έχει πλήρη άγνοια επί της λειτουργικότητας της βασικής εφαρμογής και απλώς θέλει να «δει» κάποια αποτελέσματα και φυσικά την εφαρμογή καθαυτό που απλώς πρέπει να τροφοδοτηθεί σωστά για να λειτουργήσει ορθά.

7.1 Σύνοψη και συμπεράσματα

Με την καθοδήγηση που έλαβα καθ' όλη τη διάρκεια συγγραφής αυτής της εργασίας, κατάφερα να αναπτύξω και να παρουσιάσω μια εφαρμογή που, παρά την απλότητά της στη χρήση, διαθέτει την ικανότητα να ανταποκρίνεται αποτελεσματικά τόσο στις τρέχουσες όσο και στις μελλοντικές ανάγκες των χρηστών. Η διαδικασία αυτή με βοήθησε να δημιουργήσω ένα εργαλείο που συνδυάζει την ευχρηστία με τη λειτουργικότητα, εξασφαλίζοντας την ικανοποίηση και την υποστήριξη των χρηστών σε βάθος χρόνου.

7.2 Μελλοντικές επεκτάσεις

Καμία εφαρμογή όσο πλήρης και αν φαντάζει κατά την ολοκλήρωση της ποτέ δεν είναι «πλήρης» και φυσικά η παρούσα δεν θα μπορούσε να είναι διαφορετική. Θεωρώ ότι σε 1^η φάση δεν στερείται σημαντικής λειτουργικότητας με επιφύλαξη ίσως στην ενεργοποίηση π.χ στην εγγραφή χρήστη με χρήση του Google ή Github account.

Επίσης πιθανόν κάποια περαιτέρω παραμετροποίηση στην οπτική παρουσίαση των αποτελεσμάτων ίσως με κάποιες περισσότερες δυνατότητες στην ιεραρχία των κλάσεων των πατεντών.

Πιστεύω ότι μετά από εκτενέστερη την χρήση της εφαρμογής θα φανούν οι όποιες πιθανόν ελλείψεις και θα αποτελέσουν αντικείμενο συγγραφής κάποιας άλλης διπλωματικής.

8

Βιβλιογραφία

- [1] https://eclass.upatras.gr/modules/document/file.php/MECH1278/TINEYMA_TIKH%20IDIOKTHSIA.pdf (Πρόσβαση 06-06-2024)
<https://towardsdatascience.com/the-concept-of-artificial-neurons-perceptrons-in-neural-networks-fab22249cbfc>

- [2] <https://ieeexplore.ieee.org/document/8695493>
- [3] <https://arxiv.org/html/2404.08668v1>
- [4] <https://towardsdatascience.com/the-concept-of-artificial-neurons-perceptrons-in-neural-networks-fab22249cbfc> (Πρόσβαση 06-06-2024)
- [5] <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10202128>
- [6] <https://towardsdatascience.com/a-guide-to-word-embeddings-8a23817ab60f>
- [7] <https://www.comet.com/site/blog/understanding-language-models-in-nlp/>
- [8] <https://ceur-ws.org/Vol-3604/paper7.pdf> (Πρόσβαση 06-06-2024)
- [9] <https://peps.python.org/pep-0206/>

