ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Ασκήση Εργαστηρίου Ανάξτησης ΠΛηροφορίας

Ιωάννης Παπαγιάννης
20390174
Ομάδα:Δευτέρα 19:00-20:00

ΑΝΑΦΟΡΑ ΚΑΙ ΤΕΚΜΗΡΙΩΣΗ ΓΙΑ ΤΗΝ ΜΗΧΑΝΗ ΑΝΑΖΗΤΗΣΗΣ ΑΚΑΔΗΜΑΙΚΩΝ ΕΡΓΑΣΙΩΝ

ΕΙΣΑΓΩΓΗ

Η παρακάτω αναφορά και τεκμηρίωση παρουσιάζει το σχεδιασμό,την υλοποίηση, και την αξιολόγηση μιας μηχανής αναζήτησης ακαδημαϊκών εργασιών. Η μηχανή αναζήτησης επιτρέπει στους χρήστες να αναζητούν ακαδημαϊκά άρθρα με βάση λέξεις κλειδιά,συγγραφείς, και ημερομηνίες δημοσίευσης και η υλοποίηση της μηχανής βασίζεται σε επεξεργασία κειμένου,αλγόριθμους αναζήτησης και κατάταξης.

ΣΧΕΔΙΑΣΜΟΣ ΚΑΙ ΥΛΟΠΟΙΗΣΗ ΣΤΑΧΥΟΛΟΓΗΤΗΣ(WEB CRAWLER)

Αρχικά αφού επέλεξα το αποθετήριο ακαδημαϊκών εργασιών PubMed, υλοποίησα έναν web crawler σε Python με την βοήθεια της βιβλιοθήκης BeutifulSoup για την συλλογή μεταδεδομένων ακαδημαϊκών εργασιών όπως τίτλος,συγγραφείς,περίληψη,ημερομηνία δημοσίευσης κ.λπ. από την επιλεγμένη πηγή και αποθήκευσα τα δεδομένα που συλλέγονται σε δομημένη μορφή json.

ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΚΕΙΜΕΝΟΥ(Text Processing)

Στην συνέχεια έκανα την προεπεξεργασία του κειμενικού περιεχομένου των ακαδημαϊκών εργασιών για την προετοιμασία τους για ευρετηρίαση και αναζήτηση. Για να συμβεί αυτό έκανα εργασίες όπως tokenization,stemming/lemmatization και stop-word removal και αφαίρεση ειδικών χαρακτήρων και χρησιμοποίησα τις βιβλιοθήκες nltk.tokenize, nltk.corpus, nltk.stem, re.

ΕΥΡΕΤΗΡΙΟ(Indexing)

Επίσης δημιούργησα μια ανεστραμμένη δομή δεδομένων ευρετηρίου (inverted index) για την αποτελεσματική αντιστοίχιση όρων στα έγγραφα στα οποία εμφανίζονται και εφάρμοσα μια δομή δεδομένων για την αποθήκευση του ευρετηρίου και χρησιμοποίησα την βιβλιοθήκη json και requests.

ΜΗΧΑΝΗ ΑΝΑΖΗΤΗΣΗΣ(Search Engine)

Έπειτα δημιούργησα μια διεπαφή χρήστη την οποία υλοποίηση με χρήση του web framework Flask. To frontend αποτελείται από δυο html σελίδες:μια σελίδα αναζήτησης και μια σελίδα αποτελεσμάτων. Ο χρήστης μπορεί να εισάγει τον όρο αναζήτησης,τον συγγραφέα και την ημερομηνία δημοσίευσης αλλά και να επιλέξει τους αλγορίθμους αναζήτησης και κατάταξης που θα χρησιμοποιηθούν. Οι αλγόριθμοι αναζήτησης που θα μπορεί να χρησιμοποιήσει είναι οι Boolean retrieval,Vector Space Model και Probalistic retrieval models BM25.Οι βιβλιοθήκες που χρησιμοποίησα για την δημιουργία των παραπάνω αλγορίθμων ανάκτησης είναι οι sklearn.feature_extraction.text, sklearn.metrics.pairwise, gensim.summarization.bm25.

ΕΠΕΞΕΡΓΑΣΙΑ ΕΡΩΤΗΜΑΤΟΣ(Query Processing)

Εδώ δημιούργησα ένα Module επεξεργασίας ερωτημάτων το οποίο λαμβάνει τα ερωτήματα των χρηστών,τα αναλύει και ανακτά σχετικά έγγραφα χρησιμοποιώντας ανεστραμμένο ευρετήριο. Το module υλοποιεί λειτουργίες Boolean(AND,OR,NOT) για την ανάκτηση των εγγράφων.

ΚΑΤΑΤΑΞΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ(Ranking)

Τέλος εφάρμοσα έναν απλό αλγόριθμο TF-IDF(Term Frequency-Inverse Document Frequency) για την κατάταξη των αποτελεσμάτων και έχω ενσωματώσει και πιο προηγμένες τεχνικές κατάταξης, όπως οι αλγόριθμοι Vector Space Model(VSM) και BM25.

ΑΞΙΟΛΟΓΗΣΗ

Για την αξιολόγηση της μηχανής αναζήτησης,θα χρησιμοποιούσα διάφορα σύνολα δεδομένων που περιέχουν ακαδημαϊκά άρθρα. Όπως φαίνεται και παρακάτω αρχικά παρατηρώ ότι το σύστημα για διάφορα ερωτήματα χρηστών καθυστερεί στο χρόνο απόκρισης του, άλλα μας επιστρέφει τουλάχιστον 10 διαφορετικά έγγραφα για κάθε ερώτημα που βάζουμε. Οι μετρικές που μπορούν να χρησιμοποιηθούν περιλαμβάνουν την ακρίβεια (precision) όπου παρατηρούμε ότι η μηχανή αναζήτησης πράγματι μας επιστρέφει έγγραφα σχετικά με τα ερωτήματα που του δώσαμε και μας δίνει τα σωστά αποτελέσματα αλλά όταν χρησιμοποιούμε φίλτρα για συγγράφεις και ημερομηνία δημοσιεύσεις δεν μας εμφανίζει τα κατάλληλα αποτελέσματα. Όσον αφορά την αξιολόγηση συνολικής απόδοσης η μηχανή έχει κάποιες καθυστερήσεις αλλά είναι σταθερό στα αποτελέσματα που δίνει. Στην συνέχεια όσον αφορά τους αλγορίθμους παρατήρησα ότι και οι 3 αλγόριθμοι μου εμφανίζουν τα ίδια αποτελέσματα για διάφορα είδη ερωτημάτων όπως φαίνεται και στις παρακάτω εικόνες.

Μηχανή Αναζήτησης Ακαδημαϊκών Εργασιών

Εισάγετε τον όρο αναζήτησης: cancer Συγγραφέας: Ημερομηνία Δημοσίευσης: Αναζήτηση

Χρήση Boolean Retrieval Χρήση Vector Space Model Χρήση BM25 Χρήση TF-IDF

Αποτελέσματα Αναζήτησης για "cancer"

Boolean Retrieval

•

cancer cure critic analysi

Συγγραφείς: p, , r, o, y, ,, , ,,

Περίληψη: abstract cancer one dread diseas th centuri spread continu increas incid st centuri situat alarm everi fourth person lifetim risk cancer india regist lakh new case cancer everi year wherea figur million worldwid cancer curabl short answer question ye fact cancer curabl caught earli enough cancer cell continu grow unless one four thing occur cancer mass remov surgic use chemotherapi anoth type cancerspecif medic hormon therapi use radiat therapi cancer cell shrink disappear

Ημερομηνία Δημοσίευσης: 2016 Jul-Sep;53(3):441-442.

Σύνδεσμος

•

global cancer incid mortal rate trendsan updat

Συγγραφείς: l, i, n, d, s, e, y, , t, o, r, r, ,, , ,,

Περίληψη: abstract limit publish data recent cancer incid mortal trend worldwid use intern agenc research cancer cancermondi clearinghous present agestandard cancer incid death rate also present trend incid mortal select countri five contin highincom countri hic continu highest incid rate site well lung colorect breast prostat cancer although low middleincom countri lmic count among highest rate mortal rate cancer declin mani hic increas lmic lmic highest rate stomach liver esophag cervic cancer although rate remain high hic plateau decreas common cancer due decreas known risk factor screen earli detect improv treatment mortal contrast rate sever lmic increas cancer due increas smoke excess bodi weight physic inact lmic also disproportion burden infectionrel cancer appli cancer control measur need reduc rate hic arrest grow burden lmic

Ημερομηνία Δημοσίευσης: 2016 Jan;25(1):16-27.

Σύνδεσμος

•

tumor microenviron recent advanc variou cancer treatment

Μηχανή Αναζήτησης Ακαδημαϊκών Εργασιών

Εισάγετε τον όρο αναζήτησης: cancer Συγγραφέας: anna dart Ημερομηνία Δημοσίευσης: Αναζήτηση
☐ Χρήση Boolean Retrieval ☑ Χρήση Vector Space Model ☐ Χρήση BM25 ☐ Χρήση TF-IDF

Vector Space Model

•

cancer cure critic analysi

Συγγραφείς: p, , r, o, y, ,, , ,,

Περίληψη: abstract cancer one dread diseas th centuri spread continu increas incid st centuri situat alarm everi fourth person lifetim risk cancer india regist lakh new case cancer everi year wherea figur million worldwid cancer curabl short answer question ye fact cancer curabl caught earli enough cancer cell continu grow unless one four thing occur cancer mass remov surgic use chemotherapi anoth type cancerspecif medic hormon therapi use radiat therapi cancer cell shrink disappear

Ημερομηνία Δημοσίευσης: 2016 Jul-Sep;53(3):441-442.

Σύνδεσμος

•

global cancer incid mortal rate trendsan updat

Συγγραφείς: l, i, n, d, s, e, y, , t, o, r, r, ,, , ,,

Περίληψη: abstract limit publish data recent cancer incid mortal trend worldwid use intern agenc research cancer cancermondi clearinghous present agestandard cancer incid death rate also present trend incid mortal select countri five contin highincom countri hic continu highest incid rate site well lung colorect breast prostat cancer although low middleincom countri lmic count among highest rate mortal rate cancer declin mani hic increas lmic lmic highest rate stomach liver esophag cervic cancer although rate remain high hic plateau decreas common cancer due decreas known risk factor screen earli detect improv treatment mortal contrast rate sever lmic increas cancer due increas smoke excess bodi weight physic inact lmic also disproportion burden infectionrel cancer appli cancer control measur need reduc rate hic arrest grow burden lmic

Ημερομηνία Δημοσίευσης: 2016 Jan;25(1):16-27.

Σύνδεσμος

•

tumor microenviron recent advanc variou cancer treatment

Συγγραφείς: i, i, , w, a, n, g

Μηχανή Αναζήτησης Ακαδημαϊκών Εργασιών

Εισάγετε τον όρο αναζήτησης: headache Συγγραφέας: levi filler Ημερομηνία Δημοσίευσης: Αναζήτηση
☐ Χρήση Boolean Retrieval ☐ Χρήση Vector Space Model ☐ Χρήση BM25 ☐ Χρήση TF-IDF

Αποτελέσματα Αναζήτησης για "headache"

Boolean Retrieval

•

evalu manag emerg depart headach

Συγγραφείς: l, e, v, i, , f, i, l, l, e, r, ,, , ,,

Περίληψη: abstract acut headach emerg depart ed pose diagnost dilemma may overwhelm provid attempt weigh cost advanc workup risk miss seriou patholog major headach concern benign primari headach disord identifi lifethreaten secondari caus headachewhich may broadli categor structur infecti vascular causesi primari focu evalu ed secondari headach associ high morbid mortal requir strict scrutini histori physic examin adequ riskstratifi patient innov emerg technolog may assist provid diagnosi headach challeng previou goldstandard diagnost evalu herein present gener overview workup manag headach ed special section diagnost consider evalu acut mening subarachnoid hemorrhag acut angleclosur glaucoma

Ημερομηνία Δημοσίευσης: 2019 Feb;39(1):20-26.

Σύνδεσμος

•

headach elderli

Συγγραφείς: r, o, b, e, r, t, , g, , k, a, n, i, e, e, k, i, ,, , ,,

Περίληψη: abstract headach common neurolog symptom affect nearli half world popul given time although preval declin age headach remain common neurolog complaint among elderli popul headach divid primari secondari caus primari headach compris twothird headach among elderli defin clinic criteria diagnos base symptom pattern exclus secondari caus primari headach includ migrain tensiontyp trigemin autonom cephalalgia hypnic headach secondari headach defin suspect etiolog higher index suspicion secondari headach disord warrant older patient newonset headach roughli time like seriou underli caus frequent differ symptomat present compar younger adult variou imag laboratori evalu indic presenc red flag sign symptom head ct procedur choic acut headach present brain mri chronic headach complaint manag headach elderli popul challeng due presenc multipl medic comorbid polypharmaci differ drug metabol clearanc keyword elderli facial pain headach migrain

Ημερομηνία Δημοσίευσης: 2019:167:511-528.

Σύνδεσμος

•

ΔΥΣΚΟΛΙΕΣ ΚΑΙ ΒΕΛΤΙΩΣΕΙΣ

Κατά τη διάρκεια της υλοποίησης,οι δυσκολίες που αντιμετώπισα είναι σχετικές με τον αποδοτικό σχεδιασμό των αλγορίθμων κατάταξης και την βελτιστοποίηση της απόδοσης της αναζήτησης καθώς δεν μπόρεσα να υλοποιήσω με τον σωστό τρόπο τον αλγόριθμο BM25 αλλά ούτε και τον αλγόριθμο TF-IDF και οι αλγόριθμοι που χρησιμοποιώ συνήθως μου βγάζουν τα ίδια αποτελέσματα μεταξύ τους. Επομένως οι βελτιώσεις που θα πρότεινα είναι η σωστή υλοποίηση των αλγορίθμων BM25 και TF-IDF και η βελτίωση των αλγορίθμων VSM και Boolean Retrieval για την εμφάνιση των κατάλληλων αποτελεσμάτων καθώς επίσης και βελτίωση στην αναζήτηση και στην εφαρμογή των φίλτρων για συγγράφεις και ημερομηνία δημοσίευσης.

ΣΥΜΠΕΡΑΣΜΑΤΑ

Η μηχανή αναζήτησης ακαδημαϊκών εργασιών πρόκειται για μια λειτουργική λύση για την ανάκτηση σχετικών άρθρων. Με την υλοποίηση τον παραπάνω βελτιώσεων και αξιολογήσεων,είναι δυνατόν να επιτευχθεί πολύ μεγαλύτερη ακρίβεια ως προς τα αποτελέσματα της αναζήτησης και τις λειτουργίες που πρέπει να πραγματοποιεί η μηχανή αναζήτησης.