

Τμήμα Αρχειονομίας, Βιβλιοθηκονομίας
& Μουσειολογίας

Πτυχιακή εργασία:

Ανάπτυξη Web App εφαρμογής με σκοπό την θεματική ευρετηρίαση με βάση τον αλγόριθμο
TF-IDF (Term Frequency – Inverse Document Frequency)

Γεράσιμος Παπαϊωάννου

AM: B2019118

Περιεχόμενα

1. Περίληψη
2. Εισαγωγή
 - 2.1. Η δεικτοδότηση
 - 2.2. Ευρετήριο
 - 2.3. Υλοποιώντας
3. Θεματική Ευρετηρίαση
 - 3.1. Εισαγωγή
 - 3.2. Επεξήγηση Διαδικασίας Θεματικής Ευρετηρίασης
 - 3.3. Πλεονεκτήματα της Θεματικής Ευρετηρίασης
 - 3.4. Παραδείγματα Εφαρμογών της Θεματικής Ευρετηρίασης
4. TF IDF αξία της διαδικασίας αλγοριθμικός
5. Βιβλιοθήκη
6. Ανάπτυξη Συστήματος / Συστατικά μέρη και τεχνολογίες
 - 6.1. Η βάση δεδομένων
 - 6.2. API και Web εφαρμογή
7. Δημιουργία βάσης
8. Δημιουργία ευρετηρίου
9. Αναζήτηση
10. Αξιολόγηση των αποτελεσμάτων και sorting
11. Filtering Results
12. Αποτελέσματα
13. Βιβλιογραφία

1. Περίληψη

Στην συγκεκριμένη πτυχιακή εργασία αναπτύσσεται web εφαρμογή που χρησιμοποιεί τον αλγόριθμο TF-IDF για τη θεματική ευρετηρίαση μικρού μεγέθους βιβλιοθήκης. Η εφαρμογή αυτή επιτρέπει την αποθήκευση των περιγραφικών στοιχείων των τεκμηρίων στο cloud, παρέχοντας έτσι ευελιξία και προσβασιμότητα είτε σε cloud ή σε κλειστό δίκτυο. Ο χρήστης έχει τη δυνατότητα να εκτελεί αναζητήσεις και κατηγοριοποιήσεις μέσω της web πλατφόρμας, η οποία εκμεταλλεύεται τον TF-IDF αλγόριθμο για τη βελτιστοποίηση της ευρετηρίασης. Μέσω αυτής της συνδυασμένης προσέγγισης cloud και TF-IDF, η εργασία προσφέρει μια σύγχρονη και το σημαντικότερο ευέλικτη λύση για τη διαχείριση και αναζήτηση περιεχομένου σε βιβλιοθήκες μικρού μεγέθους.

2. Εισαγωγή

2.1 Δεικτοδότηση

Η δεικτοδότηση (indexing) είναι μια διαδικασία κατά την οποία οι πληροφορίες οργανώνονται με τρόπο που να διευκολύνει την αναζήτηση και την ανάκτηση. Οι μέθοδοι δεικτοδότησης ποικίλλουν ανάλογα με τον τύπο των δεδομένων και την εφαρμογή. Παρακάτω παρουσιάζονται οι πιο συνηθισμένες μέθοδοι δεικτοδότησης:

1. Αλφαβητική Δεικτοδότηση: Οι εγγραφές οργανώνονται αλφαβητικά, συνήθως βάσει του ονόματος ή του τίτλου. Είναι απλή και εύκολη στη χρήση, ιδιαίτερα για βιβλιοθήκες ή αρχεία.

2. Θεματική Δεικτοδότηση: Οι πληροφορίες οργανώνονται βάσει θεματικών κατηγοριών ή λέξεων-κλειδιών. Αυτή η μέθοδος χρησιμοποιείται ευρέως σε βάσεις δεδομένων και βιβλιοθήκες για την ταξινόμηση βιβλίων και άρθρων βάσει θεμάτων.

3. Δεικτοδότηση Κλειδιών (Keyword Indexing): Χρησιμοποιούνται συγκεκριμένες λέξεις-κλειδιά για να περιγράψουν τα περιεχόμενα ενός εγγράφου ή δεδομένου. Οι λέξεις-κλειδιά μπορεί να είναι προκαθορισμένες (π.χ. από ένα θησαυρό) ή να επιλέγονται ελεύθερα.

4. Δεικτοδότηση Βάσει Περιεχομένου (Full-Text Indexing): Όλα τα λόγια σε ένα κείμενο καταχωρούνται στο ευρετήριο. Αυτός ο τύπος δεικτοδότησης είναι χρήσιμος για τη γρήγορη αναζήτηση μέσα σε μεγάλα κείμενα και χρησιμοποιείται ευρέως στις μηχανές αναζήτησης.

5. Ευρετήριο Αντιστρόφου (Inverted Indexing): Δημιουργείται ένα ευρετήριο που αντιστρέφει τον κανονικό τρόπο δεικτοδότησης, επιτρέποντας την αναζήτηση όχι μόνο βάσει τίτλων ή λέξεων-κλειδιών αλλά βάσει της παρουσίας οποιασδήποτε λέξης μέσα στο κείμενο. Αυτή η μέθοδος είναι κοινή στις μηχανές αναζήτησης στο διαδίκτυο.

6. Δεικτοδότηση Εικόνων (Image Indexing): Οι εικόνες καταχωρούνται βάσει χαρακτηριστικών όπως το χρώμα, το σχήμα, ή οι μεταδεδομένες (metadata) πληροφορίες. Είναι χρήσιμη για βάσεις δεδομένων εικόνων και εφαρμογές αναγνώρισης εικόνων.

7. Δεικτοδότηση Χωρικών Δεδομένων (Spatial Indexing): Χρησιμοποιείται για την καταχώρηση γεωγραφικών ή χωρικών δεδομένων, επιτρέποντας την αναζήτηση βάσει τοποθεσίας. Συχνά χρησιμοποιείται σε GIS (Geographic Information Systems).

8. Δεικτοδότηση Βάσεων Δεδομένων (Database Indexing): Βελτιστοποιεί την αναζήτηση και την ανάκτηση δεδομένων από βάσεις δεδομένων. Περιλαμβάνει την χρήση ευρετηρίων όπως B-trees και hash indexes. Τα B- trees στις βάσεις δεδομένων έχουν την ιδιότητα ότι διατηρούν την ισορροπία τους, δηλαδή η μέγιστη διαφορά στο μήκος των μονοπατιών από τη ρίζα προς τα φύλλα είναι περιορισμένη. Οι hash indexes (δείκτες κατακερματισμού) είναι μια μορφή δομής δεδομένων που χρησιμοποιείται για γρήγορη αναζήτηση δεδομένων. Η ιδέα πίσω από τους δείκτες κατακερματισμού είναι να χρησιμοποιήσουν μια συνάρτηση κατακερματισμού για να αντιστοιχίσουν τα κλειδιά δεδομένων (όπως πχ. αριθμούς ή συμβολοσειρές) σε διευθύνσεις αποθηκευμένων δεδομένων. Και οι 2 τεχνικές δίνουν λύσεις στις προκλήσεις των big data.

2.2 Εισαγωγή στο Ευρετήριο Βιβλιοθήκης

Το ευρετήριο σε μια βιβλιοθήκη αποτελεί ένα από τα πιο κρίσιμα εργαλεία για την οργάνωση, τη διαχείριση και την ανάκτηση πληροφοριών. Σε μια εποχή όπου η πληροφορία είναι πιο πλούσια και πιο προσιτή από ποτέ, η ύπαρξη ενός καλά οργανωμένου ευρετηρίου εξασφαλίζει την αποδοτική πρόσβαση σε γνώσεις και πόρους. Η βασική λειτουργία του ευρετηρίου είναι να παρέχει ένα σύστημα κατάταξης και ταξινόμησης των βιβλίων και άλλων υλικών της βιβλιοθήκης, διευκολύνοντας έτσι την εύρεση και τη χρήση τους από τους επισκέπτες της βιβλιοθήκης.

Οι βιβλιοθήκες χρησιμοποιούν διάφορους τύπους ευρετηρίων για την οργάνωση του υλικού τους, όπως:

1. Κατάλογος Καρτών (Card Catalog): Παραδοσιακός τρόπος οργάνωσης, όπου οι πληροφορίες για κάθε βιβλίο καταχωρούνται σε κάρτες και ταξινομούνται αλφαβητικά ή θεματικά.

2. Ηλεκτρονικό Ευρετήριο (OPAC - Online Public Access Catalog): Ψηφιακή βάση δεδομένων που επιτρέπει στους χρήστες να αναζητούν υλικό μέσω υπολογιστών ή άλλων συσκευών. Προσφέρει μεγαλύτερη ευελιξία και ταχύτητα στην αναζήτηση.

Το ευρετήριο δεν περιορίζεται μόνο στην απλή καταγραφή των τίτλων και των συγγραφέων. Περιλαμβάνει επίσης θεματικές κατηγορίες, περιλήψεις και άλλες σχετικές πληροφορίες που βοηθούν τους χρήστες να εντοπίσουν με ακρίβεια το υλικό που χρειάζονται. Επιπλέον, τα σύγχρονα ευρετήρια ενσωματώνουν λειτουργίες αναζήτησης που επιτρέπουν πολύπλοκες και εξειδικευμένες αναζητήσεις, διευκολύνοντας έτσι τους ερευνητές και τους σπουδαστές.

Η ύπαρξη ενός καλοσχεδιασμένου ευρετηρίου συμβάλλει επίσης στην:

1. Εξοικονόμηση Χρόνου: Οι χρήστες μπορούν γρήγορα και εύκολα να εντοπίσουν το υλικό που αναζητούν χωρίς να χρειάζεται να εξετάσουν φυσικά ράφια και ράφια βιβλίων.

2. Βελτίωση της Πρόσβασης: Το υλικό γίνεται προσβάσιμο σε μεγαλύτερο αριθμό χρηστών, συμπεριλαμβανομένων και αυτών που κάνουν αναζήτηση από απόσταση μέσω διαδικτύου.

3. Διαχείριση Συλλογής: Οι βιβλιοθηκονόμοι μπορούν να παρακολουθούν την κατάσταση και τη χρήση του υλικού, διευκολύνοντας τη διαχείριση και την ανανέωση της συλλογής.

Ένα ευρετήριο δεν είναι απλώς μια λίστα βιβλίων ή άρθρων· είναι ένας απαραίτητος μηχανισμός που υποστηρίζει τη δομή και τη λειτουργία της βιβλιοθήκης. Μέσα από τη συνεχή εξέλιξή του, ειδικά με την εισαγωγή των ψηφιακών τεχνολογιών, το ευρετήριο παραμένει ένα από τα πιο πολύτιμα εργαλεία για την προώθηση της γνώσης και της εκπαίδευσης.

Η χρησιμότητα ενός ευρετηρίου σε μια βιβλιοθήκη είναι πολυδιάστατη και εξαιρετικά σημαντική για την αποτελεσματική διαχείριση και την πρόσβαση στο υλικό της βιβλιοθήκης. Παρακάτω παρατίθενται οι κύριοι λόγοι για τους οποίους το ευρετήριο είναι χρήσιμο σε μια βιβλιοθήκη:

1. Οργάνωση Υλικού: Το ευρετήριο βοηθά στην οργάνωση του υλικού της βιβλιοθήκης με συστηματικό τρόπο, επιτρέποντας την εύκολη κατηγοριοποίηση των βιβλίων, περιοδικών και άλλων υλικών.

2. Εύκολη Αναζήτηση: Επιτρέπει στους χρήστες να βρίσκουν εύκολα το υλικό που αναζητούν. Μέσω του ευρετηρίου, οι χρήστες μπορούν να αναζητήσουν βιβλία και άλλα υλικά βάσει τίτλου, συγγραφέα, θεματικής κατηγορίας ή λέξεων-κλειδιών.

3. Ταχεία Πρόσβαση: Μειώνει τον χρόνο που απαιτείται για την αναζήτηση ενός συγκεκριμένου βιβλίου ή υλικού, διευκολύνοντας την άμεση πρόσβαση στις πληροφορίες.

4. Διαχείριση Συλλογής: Βοηθά τους βιβλιοθηκονόμους να διαχειρίζονται καλύτερα τη συλλογή της βιβλιοθήκης, να εντοπίζουν εύκολα ποια βιβλία είναι διαθέσιμα, ποια είναι δανεισμένα και ποια χρειάζονται συντήρηση ή αντικατάσταση.

5. Διευκόλυνση Δανεισμού: Το ευρετήριο επιτρέπει την παρακολούθηση και τη διαχείριση του δανεισμού βιβλίων και άλλων υλικών, εξασφαλίζοντας ότι τα αντικείμενα επιστρέφονται έγκαιρα και είναι διαθέσιμα για άλλους χρήστες.

6. Υποστήριξη Έρευνας: Είναι πολύτιμο εργαλείο για ερευνητές και φοιτητές, παρέχοντας τους τη δυνατότητα να εντοπίζουν υλικό που σχετίζεται με τα ερευνητικά τους θέματα ή τα μαθήματά τους.

7. Ενημέρωση και Επέκταση Συλλογής: Επιτρέπει τη συνεχή ενημέρωση και τον εμπλουτισμό της συλλογής με νέα υλικά, διατηρώντας την βιβλιοθήκη επίκαιρη και χρήσιμη για τους χρήστες της.

8. Διατήρηση Ιστορικού: Κρατά ιστορικό των υλικών που περιλαμβάνονται στη συλλογή της βιβλιοθήκης, παρέχοντας σημαντικές πληροφορίες για την εξέλιξη και την ανάπτυξη της συλλογής με την πάροδο του χρόνου.

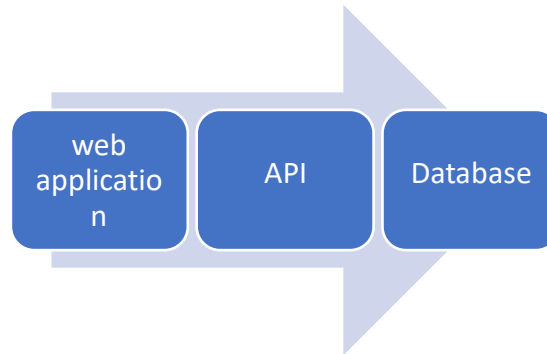
2.3 Υλοποιώντας

Η αναπτυξιακή πορεία της τεχνολογίας και η εξέλιξη της ψηφιακής πληροφορίας έχουν οδηγήσει σε μια συνεχή αύξηση της ποσότητας των δεδομένων που διαχειριζόμαστε. Σε αυτό το πλαίσιο, η αποτελεσματική οργάνωση και αναζήτηση πληροφοριών είναι κρίσιμης σημασίας. Η παρούσα πτυχιακή εργασία επικεντρώνεται στην ανάπτυξη μιας καινοτόμου λύσης για τη θεματική ευρετηρίαση μικρού μεγέθους βιβλιοθήκης, βασισμένη σε μια εξελιγμένη τεχνική ανάλυσης κειμένου.

Συγκεκριμένα, εξετάζουμε την εφαρμογή του αλγορίθμου TF-IDF (Term Frequency-Inverse Document Frequency) σε συνδυασμό με μια web πλατφόρμα που επιτρέπει τη διαχείριση των εγγράφων στο cloud. Η χρήση του TF-IDF σκοπεύει στην αυτοματοποίηση της ευρετηρίασης, βελτιώνοντας την απόδοση της αναζήτησης και την οργάνωση του περιεχομένου.

Μέσω αυτής της προσέγγισης, παρέχουμε μια πρακτική λύση για τη διαχείριση μικρών βιβλιοθηκών, επιτρέποντας στους χρήστες την εύκολη πρόσβαση και αναζήτηση πληροφοριών από οπουδήποτε μέσω της web πλατφόρμας που στηρίζεται σε cloud υποδομή.

Τα στοιχεία που συνθέτουν το σύστημα συνολικά είναι η βάση δεδομένων, το API (Application Programming Interface) και η Web εφαρμογή.



3. Θεματική ευρετηρίαση

3.1. Εισαγωγή

Η θεματική ευρετηρίαση αναφέρεται στη διαδικασία οργάνωσης και κατηγοριοποίησης πληροφοριών, κυρίως κειμένου, βάσει των θεμάτων ή των εννοιών που περιλαμβάνει. Στο πλαίσιο ενός συστήματος θεματικής ευρετηρίασης, αναζητούνται και αναγνωρίζονται τα κύρια θέματα ή εννοιολογικά πεδία που σχετίζονται με ένα κείμενο.

Οι τεχνικές θεματικής ευρετηρίασης χρησιμοποιούνται για να επιτύχουν αυτό τον στόχο. Συγκεκριμένα, αλγόριθμοι επεξεργασίας φυσικής γλώσσας (Natural Language Processing - NLP) και ανάλυσης κειμένου μπορούν να εξάγουν σημαντικές λέξεις-κλειδιά, θέματα ή εννοιολογικές συνδέσεις από ένα κείμενο. Στη συνέχεια, αυτές οι πληροφορίες χρησιμοποιούνται για τη δημιουργία ενός ευρετηρίου που οργανώνει το κείμενο με βάση τα θέματα.

Η θεματική ευρετηρίαση είναι χρήσιμη σε πολλούς τομείς, όπως η ανάκτηση πληροφοριών, η οργάνωση περιεχομένου σε βάσεις δεδομένων, η κατηγοριοποίηση ειδήσεων και η αυτόματη σύνοψη κειμένου. Σε βιβλιοθήκες, επικεντρώνεται στον τρόπο οργάνωσης και προσπέλασης των πληροφοριών που περιέχονται σε ένα σύνολο εγγράφων. Αντί να αναφέρεται απλώς σε τίτλους ή συγγραφείς, η θεματική ευρετηρίαση επιδιώκει να αναδείξει τα κύρια θέματα και τις εννοιολογικές συνδέσεις που απαρτίζουν τα έγγραφα.

Με αυτόν τον τρόπο, οι χρήστες μπορούν να αναζητούν πληροφορίες με βάση τα θέματα που τους ενδιαφέρουν, επιτυγχάνοντας πιο ακριβή και συναφή αποτελέσματα. Επιπλέον, η θεματική ευρετηρίαση επιτρέπει τη διαχείριση μεγάλου όγκου πληροφοριών, βοηθώντας στον αποτελεσματικότερο εντοπισμό και χρήση των εγγράφων σε μια βιβλιοθήκη.

3.2. Επεξήγηση Διαδικασίας Θεματικής Ευρετηρίασης

Ορισμός Θησαυρού (Thesaurus): Οι θησαυροί είναι οργανωμένες συλλογές όρων και σχέσεων μεταξύ αυτών των όρων που χρησιμοποιούνται για την κατηγοριοποίηση των θεμάτων. Η δημιουργία ενός θησαυρού περιλαμβάνει την επιλογή και την οργάνωση των όρων που σχετίζονται με συγκεκριμένα θέματα.

Χρήση Θησαυρών: Οι θησαυροί χρησιμοποιούνται για την τυποποίηση των θεματικών όρων και για την ενίσχυση της συνέπειας στη θεματική ευρετηρίαση.

Συλλογή Δεδομένων: Η συλλογή εκπαιδευτικών συνόλων δεδομένων (Training Datasets) που περιέχουν προσημασμένα παραδείγματα είναι κρίσιμη για την εκπαίδευση αλγορίθμων NLP.

Εκπαίδευση Αλγορίθμων: Οι αλγόριθμοι NLP εκπαιδεύονται χρησιμοποιώντας αυτά τα δεδομένα για να αναγνωρίζουν και να εξάγουν θεματικούς όρους και εννοιολογικές σχέσεις από κείμενα.

Τεχνικές NLP: Χρησιμοποιούνται διάφορες τεχνικές NLP, όπως η ανάλυση συναισθήματος, η ανάλυση πλαισίου (context analysis) και η αναγνώριση οντοτήτων (entity recognition) για την εξαγωγή θεμάτων από το κείμενο.

Αλγόριθμοι Μηχανικής Μάθησης: Οι αλγόριθμοι μηχανικής μάθησης μπορούν να εκπαιδευτούν να αναγνωρίζουν και να ταξινομούν τα θέματα μέσα από τη συνεχή μάθηση και βελτίωση.

3.3. Πλεονεκτήματα της Θεματικής Ευρετηρίασης

1. **Αναλυτική Αναζήτηση:** Οι χρήστες μπορούν να πραγματοποιούν αναζητήσεις βασισμένες σε συγκεκριμένα θέματα ή εννοιολογικές σχέσεις, γεγονός που βελτιώνει την ακρίβεια και τη συνάφεια των αποτελεσμάτων.

2. **Οργάνωση Περιεχομένου:** Η θεματική ευρετηρίαση διευκολύνει την οργάνωση και διαχείριση μεγάλου όγκου δεδομένων, επιτρέποντας στους χρήστες να εντοπίζουν σχετικό περιεχόμενο πιο αποδοτικά.

3. **Στοχευμένη Πρόσβαση:** Οι ερευνητές και οι φοιτητές μπορούν να βρουν γρήγορα και αποτελεσματικά τα έγγραφα που σχετίζονται με τα συγκεκριμένα θέματα που τους ενδιαφέρουν.

4. **Επιστημονική Ανάλυση:** Η θεματική ευρετηρίαση υποστηρίζει την επιστημονική ανάλυση και την εύρεση συσχετίσεων μεταξύ διαφορετικών θεμάτων και εννοιών.

3.4. Παραδείγματα Εφαρμογών της Θεματικής Ευρετηρίασης

1. Οι ακαδημαϊκές βιβλιοθήκες χρησιμοποιούν θεματική ευρετηρίαση για να βοηθήσουν τους φοιτητές και τους ερευνητές να βρίσκουν επιστημονικά άρθρα, βιβλία και εργασίες που σχετίζονται με τα πεδία σπουδών τους.

2. Στις ψηφιακές βιβλιοθήκες, η θεματική ευρετηρίαση βοηθά στην οργάνωση και ανάκτηση ψηφιακού περιεχομένου, όπως ηλεκτρονικά βιβλία, άρθρα και αρχεία πολυμέσων.

3. Οι επιχειρήσεις χρησιμοποιούν τη θεματική ευρετηρίαση για την οργάνωση και ανάκτηση εγγράφων, αναφορών και άλλων πληροφοριών που σχετίζονται με τις επιχειρηματικές τους δραστηριότητες.

Η θεματική ευρετηρίαση αποτελεί ουσιαστικό στοιχείο για τη σύγχρονη βιβλιοθηκονομία και την ανάκτηση πληροφοριών. Βοηθά στην καλύτερη οργάνωση και πρόσβαση στο περιεχόμενο, εξασφαλίζοντας ότι οι χρήστες μπορούν να βρουν εύκολα και γρήγορα τις πληροφορίες που χρειάζονται. Η συνεχής εξέλιξη των τεχνολογιών NLP και μηχανικής μάθησης αναμένεται να βελτιώσει περαιτέρω τις δυνατότητες της θεματικής ευρετηρίασης, καθιστώντας την ακόμα πιο αποδοτική και χρήσιμη για την κοινότητα των χρηστών των βιβλιοθηκών.

4. TF-IDF αλγόριθμος

Ο αλγόριθμος TF-IDF (Term Frequency-Inverse Document Frequency) είναι ένα εργαλείο της επεξεργασίας φυσικής γλώσσας (NLP) που χρησιμοποιείται για την αξιολόγηση της σημαντικότητας μιας λέξης σε ένα κείμενο σε σχέση με μια συλλογή εγγράφων. Ο αλγόριθμος αυτός είναι ευρέως χρησιμοποιούμενος σε εφαρμογές όπως η αναζήτηση πληροφοριών, η ομαδοποίηση κειμένων και η εξόρυξη γνώσης.

Η λειτουργία του αλγορίθμου TF-IDF βασίζεται σε δύο βασικές έννοιες: το Term Frequency (TF) και το Inverse Document Frequency (IDF).

Term Frequency (TF):

Αναφέρεται στη συχνότητα εμφάνισης μιας συγκεκριμένης λέξης σε ένα έγγραφο. Υπολογίζεται με τον τύπο:

$TF(\text{word}, \text{document}) = (\text{Συνολικός αριθμός λέξεων στο έγγραφο}) / (\text{Συχνότητα εμφάνισης λέξης στο έγγραφο})$

Inverse Document Frequency (IDF):

Μετρά τη σπανιότητα μιας λέξης σε όλη τη συλλογή των εγγράφων. Υπολογίζεται με τον τύπο:

$IDF(\text{word}) = \log((\text{Συνολικός αριθμός εγγράφων}) / (\text{Αριθμός εγγράφων που περιέχουν τη λέξη}))$

Τέλος, ο συνολικός TF-IDF για μια λέξη σε ένα έγγραφο υπολογίζεται ως το γινόμενο του TF και του IDF:

$TF-IDF(\text{word}, \text{document}) = TF(\text{word}, \text{document}) \times IDF(\text{word})$

Ο μεγάλος TF-IDF υποδεικνύει ότι η συγκεκριμένη λέξη είναι σημαντική για το συγκεκριμένο έγγραφο. Η διαδικασία επαναλαμβάνεται για όλες τις λέξεις και όλα τα έγγραφα, προσδίδοντας βαθμολογίες σημαντικότητας σε κάθε λέξη.

5. Βιβλιοθήκη

Σκοπός είναι η ανάπτυξη ενός εργαλείου που θα είναι αξιόπιστο βοήθημα σε ανθρώπους ή οργανισμούς που δεν έχουν τη δυναμική για την ενασχόληση με τη βιβλιοθήκη ως πρώτο

μέλημα και δεν θα μπορούν να δαπανούν συστηματικά χρόνο για τη ανάπτυξη της βιβλιοθήκης.

Στην δικιά μας περίπτωση οι εγγραφές της βιβλιοθήκης που θα μπορούσαμε να επεξεργαστούμε ήταν περίπου 800 από ένα σύνολο περίπου 2000 βιβλίων. Τα περισσότερα ήταν στην ελληνική γλώσσα και είχαν ξεκινήσει με ένα πολύ βασικό τρόπο περιγραφής που ανανεώνεται με τον καιρό σε μια πολύ πιο αναλυτική περιγραφή για το καθένα ξεχωριστά. Με βάση τα νέα πεδία η περιγραφή του κάθε βιβλίου αποτελείται από τα εξής πεδία:

- Κύριος Τίτλος
- Υπότιτλος(ένας η περισσότεροι)
- Λέξη κλειδί (συνώνυμο του tag – μηχανισμός που δεν χρησιμοποιείτε πια από την βιβλιοθήκη)
- Συγγραφέας
- Εκδόσεις
- Χρονολογία έκδοσης
- Τόμος αριθμός (στην περίπτωση που ανήκει σε κάποια σειρά)
- Κωδικός (με σκοπό να μπορούμε να το εντοπίσουμε στη φυσική του θέση)
- Αριθμός Σελίδων
- Γλώσσα
- Link αποθετηρίου (αν υπάρχει σε ψηφιακή μορφή)
- Αριθμός αντιτύπων
- Μεταφραστής (αν υπάρχει)
- Εικονογράφος φωτογράφος (αν υπάρχει)
- Επιμέλεια (αν υπάρχει)
- Περιεχόμενα
- Γενικά
- Οπισθόφυλλο
- Όλο το βιβλίο
- Εικόνα (+alternative text ώστε να εμφανίζεται στα αποτελέσματα στη google)

Τα υποχρεωτικά πεδία για την δημιουργία εγγραφής είναι:

- Κύριος Τίτλος
- Λέξη κλειδί

- Αριθμός έκδοσης
- Το alternative text εφόσον υπάρχει εικόνα
- Αριθμός αντιτύπων
- Γλώσσα

Τα σημαντικά όμως πεδία που θα στηριχθεί ο αλγοριθμός μας είναι τα

- Περιεχόμενα
- Γενικά
- Οπισθόφυλλο και
- Όλο το βιβλίο

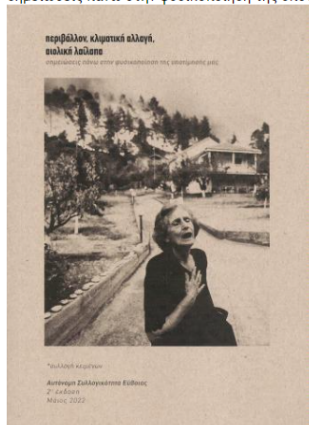
Τα «Περιεχόμενα» και το «Οπισθόφυλλο» ουσιαστικά είναι μεταφορά αυτούσια από τα αντίστοιχα τμήματα του βιβλίου εφόσον αυτά υπάρχουν. Τα «Γενικά» είναι κείμενο που αφορά το βιβλίο και μπορεί να είναι από ένα απόσπασμα από την εισαγωγή μέχρι κάποια αξιολογη κριτική και «Όλο το βιβλίο» αποτελεί όλο το περιεχόμενο του βιβλίου ώστε να μπορούμε να κάνουμε πιο εκτενή ανάλυση κτλ . Σχεδόν ο κανόνας είναι ότι αποτελεί άδειο πεδίο.

Ένα παράδειγμα από τη βιβλιοθήκης μας αποτελεί η παρακάτω εγγραφή που απεικονίζεται και στις ακόλουθες φωτογραφίες:

<https://arxeiokin.gr/periballon-klimatiki-allagi-aioliki-lailapa>

Περιβάλλον, κλιματική αλλαγή, αιολική λαίλαπα

Υπότιτλοι
σημειώσεις πάνω στην φυσικοποίηση της υποτίμησής μας



Περιεχόμενα

Εισαγωγή

Απολογισμός και επόμενη μέρα μετά την καταστροφή στη Βόρεια Εύβοια, Αυτόνομη Συλλογικότητα Εύβοιας

Δήμοι, δήμαρχοι και δήμιοι, Αυτόνομη Συλλογικότητα Εύβοιας

Τα καταστραμμένα δίκτυα ύδρευσης, η έλλειψη θέρμανσης, η ρύπανση του αέρα και η κρατική έγνοια για το περιβάλλον, Αυτόνομη Συλλογικότητα Εύβοιας

Να σταματήσουμε την καταλήστευσή μας μέσω των λογαριασμών ρεύματος, Αυτόνομη Συλλογικότητα Εύβοιας

Καταργώντας τους υδρογονάνθρακες με κολοσσιαίες επενδύσεις στους... υδρογονάνθρακες, Αυτόνομη Συλλογικότητα Εύβοιας

Πιο πιθανό είναι να μπαίνουν σε παγετωνική περίοδο, Συνέντευξη του καθηγητή Πέτρου Ιωάννου στον Κώστα Δεληγιάννη, περιοδικό Οίκo της

Καθημερινής

Αντιμετώπιση της κλιματικής αλλαγής, η συνεργασία Κίνας- ΗΠΑ είναι σημαντική αλλά και εύθραυστη, Global Times

Κλιματική αλλαγή- κατασκευάζοντας κρίσεις, scorthedearthstragedy

Μουντή αρμονία, θεϊκοί κατακλυσμοί και σωτήρες της συμφοράς- η πολιτική διαχείριση της κλιματικής αλλαγής, tetartoskosmos

Ο κοινωνικός ανταγωνισμός ενόχλησε την επιστήμη και η επιστημονική αλήθεια, Πρωτοβουλία Αναρχικών Κυψέλης

Λέξη/κλειδί

ΟΙΚΟΛΟΓΙΑ

ΣΥΓΓΡΑΦΕΑΣ

ΣΥΛΛΟΓΙΚΟ

ΧΡΟΝΟΛΟΓΙΑ

Sun, 05/01/2022 - 12:00

ΚΩΔΙΚΟΣ

SYL00390

ΣΕΛΙΔΕΣ

90

Γλώσσα

Ελληνικά

Αριθμός αντιτύπων

2

Ψηφιακό αρχείο

Off

Αριθμός έκδοσης

2

Οπισθόφυλλο

Τέλος, έχουμε χρέος να θωρακίσουμε την πατρίδα μας απέναντι στην πραγματικότητα της Κλιματικής Αλλαγής.

Μητσοτάκης, 05.08.2021, Διάγγελμα από το Μαξίμου την ώρα που καιγόταν η μισή Βόρεια Εύβοια.

Λίγες μόλις δεκαετίες πριν, διάφοροι δημοσιολογούντες έλεγαν πως οι γυναίκες είναι εκ φύσεως κατώτερες και για αυτό θα έπρεπε να μην ψηφίζουν, αλλά και ούτε να λαμβάνουν μέρος στη δημόσια ζωή... Λίγες μόλις δεκαετίες πριν, διάφοροι δημοσιολογούντες ή και επιστήμονες έλεγαν πως οι κάτοικοι της Αφρικής που είχαν αιμαλωτιστεί από τους Ευρωπαίους και πωλούνταν ως δούλοι, ήταν εκ φύσεως κατώτερη φυλή... Λίγες μόλις δεκαετίες πριν, διάφοροι επιστήμονες ή και οικονομολόγοι έλεγαν ότι οι φτωχοί ήταν εκ φύσεως κατώτεροι άνθρωποι και δεν πείραζε να πεθαίνουν κατά χιλιάδες στους πολέμους, γιατί έτσι η φύση βρίσκει μια ισορροπία πάνω στο ζήτημα του υπερπληθυσμού. Και όλοι αυτοί οι φρικαλέοι λόγοι, όχι μόνο ξεστομίζονταν και γράφονταν, αλλά ήταν και κυρίαρχοι και περιφέρονταν ως η αλήθεια.

Σήμερα, τηρουμένων των αναλογιών και της δυναμικής του κάθε ζητήματος, διάφοροι δημοσιολογούντες ή και πρωθυπουργοί, την ώρα που τα μηδενικά λειτουργικά μέσα πυρόσβεσης τους, μεθόδευαν το κάψιμο του δάσους της Βόρειας Εύβοιας, την γενοκτονία εκατομμυρίων ζώων και την καταστροφή των ζώων χιλιάδων ανθρώπων, μας είπαν πως για όλα αυτά φταίει ένα φυσικό φαινόμενο: η κλιματική αλλαγή. Άλλωστε για τον ίδιο λόγο λένε πως πρέπει να υπομένουμε πάνω από τα κεφάλια μας να λειτουργούν αιολικά εργοστάσια ύψους 185 μέτρων ή να υπομένουμε την καταστροφή παρθένων δασών και ανέγγιχτων από τον άνθρωπο οικοσυστημάτων...

Powered by Drupal

Δεύτερο παράδειγμα από τη βιβλιοθήκη αποτελεί η παρακάτω εγγραφή που απεικονίζεται και στις ακόλουθες φωτογραφίες:

<https://arxeiokin.gr/synenteyxeis-me-rizospastries-palaistinies-gynaikes>

Συνεντεύξεις με ριζοσπάστριες παλαιστίνιες γυναίκες



Περιεχόμενα
Πρόλογος από τις εκδόσεις Καμιονέτα
Εισαγωγή από την συλλογή Shoal
ΣΥΝΕΝΤΕΥΞΕΙΣ
Lina Nabulsi
Siren Khudairy
Shahd Abusalama
Lama Suleiman
Faiza Abu Shamsiyah
Amal Awad

Rana Abu Rahmah
Ghada Hamdan
Mona Al-Farra
Izdihar Afyouni
ΓΛΩΣΣΑΡΙ
Γεγονότα
Τοποθεσίες
Ορολογία
Παλαιστινιακές Οργανώσεις

Λέξη/κλειδί
ΠΑΛΑΙΣΤΙΝΗ
ΕΚΔΟΣΕΙΣ
Καμιονέτα
ΣΥΓΓΡΑΦΕΑΣ
ΣΥΛΛΟΓΙΚΟ
ΧΡΟΝΟΛΟΓΙΑ
Fri, 09/01/2023 · 12:00
ΚΩΔΙΚΟΣ
SYL00451
ΣΕΛΙΔΕΣ
184
Γλώσσα
Ελληνικά
Αριθμός αντιτύπων
2
Ψηφιακό αρχείο
Off
Τόπος έκδοσης
Ιωάννινα
Αριθμός έκδοσης
1

Οπισθόφυλλο

Η ιδέα για αυτό το βιβλίο γεννήθηκε από μια αίσθηση ανησυχίας ότι τα κινήματα αλληλεγγύης συνηθίζουν να εμπλέκονται με τις Παλαιστίνιες μόνο όσον αφορά τον αγώνα τους ενάντια στην ισραηλινή κατοχή. Μερικές φορές, αυτό μπορεί να αποτελέσει τροχοπέδη στο να τις βλέπουμε ως συντρόφισσες-επαναστάτριες και γενικά συντρόφισσες στους κοινούς μας αγώνες για έναν καλύτερο κόσμο, κόντρα στον απολυταρχισμό, τη λευκή υπεροχή, την πατριαρχία, τον καπιταλισμό και τον κρατικό έλεγχο.

Ελπίζουμε ότι το εύρος των θεμάτων που καλύφθηκαν στις συζητήσεις θα βοηθήσει τους ανθρώπους εκτός Παλαιστίνης να αντιληφθούν τον παλαιστινιακό αγώνα σε μια τρισδιάστατη μορφή και να δουν τις Παλαιστίνιες ριζοσπάστριες ως συντρόφισσες στον παγκόσμιο αγώνα μας για ελευθερία.

Κολλεκτίβα Shoal

6. Ανάπτυξη Συστήματος / Συστατικά μέρη και τεχνολογίες

Το σύστημα μας ουσιαστικά αποτελείται από 3 βασικά μέρη.

- Την βάση δεδομένων
- το Application Programming Interface (API) και
- την Web εφαρμογή.

6.1. Η βάση δεδομένων

Η βάση δεδομένων MongoDB είναι μια NoSQL (Not Only SQL) βάση δεδομένων που βασίζεται σε έγγραφα, όπου τα δεδομένα αποθηκεύονται σε μορφή BSON (Binary JSON). Κάθε έγγραφο αποτελείται από ένα σύνολο πεδίων-τιμών και μπορεί να περιέχει άλλα έγγραφα, πίνακες και λίστες.

Τα πλεονεκτήματα της MongoDB περιλαμβάνουν την ευελιξία, την απλότητα και την αποδοτικότητα. Η ευελιξία προκύπτει από το γεγονός ότι δεν απαιτεί προκαθορισμένο σχήμα, επιτρέποντας την εύκολη προσαρμογή στις ανάγκες των εφαρμογών. Η απλότητα προέρχεται από τη χρήση JSON-παρόμοιας σύνταξης για τα δεδομένα, που καθιστά ευανάγνωστη και κατανοητή τη βάση δεδομένων.

Επιπλέον, η MongoDB υποστηρίζει ανταγωνιστική απόδοση σε εφαρμογές με μεγάλο όγκο δεδομένων και προσφέρει χαρακτηριστικά όπως αυτόματη ισορροπία φορτίου, αντιγραφή δεδομένων για αντοχή σε αναλύσεις, και δυνατότητες διαχείρισης σε ενσωματωμένα συστήματα.

Συνολικά, η MongoDB αποτελεί ένα ισχυρό εργαλείο για την αποθήκευση και διαχείριση δεδομένων σε σύγχρονες εφαρμογές, παρέχοντας ευελιξία, απλότητα και αποδοτικότητα.

7.2. API και Web εφαρμογή

Το API είναι ένα σύνολο κανόνων και προτύπων που επιτρέπουν σε διάφορα λογισμικά να επικοινωνούν μεταξύ τους. Συγκεκριμένα, αποτελεί μια διεπαφή που επιτρέπει σε εφαρμογές να αλληλεπιδρούν με μια υπηρεσία ή ένα σύστημα.

Το σημαντικό που μας παρέχον τα API είναι ότι επιτρέπουν σε διαφορετικά συστήματα και υπηρεσίες να αλληλεπιδρούν και να συνεργάζονται, επιτρέποντας τη δημιουργία πιο σύνθετων εφαρμογών και υπηρεσιών.

Το API είναι γραμμένο σε Angular και η web εφαρμογή γραμμένη σε Node.js και ουσιαστικά αποτελούν τον πυρήνα του ολοκληρωμένου συστήματος. Η web εφαρμογή δεν αποτελεί μια ολοκληρωμένη λύση που θα έχει τις λειτουργίες ξεχωριστά για τον admin ή τον χρήστη. Δεν υπάρχει login και διαφορετικοί ρόλοι. Για αυτό το να περάσουμε όλες τις λέξεις που είναι για το ευρετήριο και να θέσουμε τη μορφή που θέλουμε ή να τις απορρίψουμε από το ευρετήριο ήταν μια διαδικασία που γινόταν μέσα από ρυθμίσεις στον κώδικα. Στην τελική όμως φάση η εφαρμογή προορίζεται για τον τελικό χρήστη με σκοπό την αναζήτηση και μόνο.

7. Δημιουργία βάσης

Με βάση την Web εφαρμογή δημιουργήσαμε την βάση και επιλέξαμε τους κατάλληλους όρους για τον αλγόριθμο TF-IDF. Αυτό έγινε αφού εισάγαμε το σύνολο των βιβλίων από την ήδη δημιουργημένη βάση που υπήρχε ήδη στη βιβλιοθήκη που χρησιμοποιήσαμε.

Η βάση ήταν εύκολο να την εξάγουμε γιατί το site για τη χρήση της βιβλιοθήκης ήταν δημιουργημένο σε Drupal και αυτό μας έδωσε εύκολα τα εργαλεία που χρειαζόμασταν ώστε να μπορέσουμε να εξάγουμε τα δεδομένα μέσω των έτοιμων εργαλείων που προσφέρει. Αυτό γίνεται μέσω συγκεκριμένων endpoints που μας έφεραν αποτελέσματα σε «πακέτα» των 50 βιβλίων.

Έτσι, το πρώτο που οφείλαμε να κάνουμε ήταν να περάσουμε όλα τα βιβλία στην δική μας βάση. Το επόμενο βήμα ήταν να πάρουμε όλες τις λέξεις από τα πεδία που θα εφαρμόσουμε τον αλγόριθμο TF-IDF («περιεχόμενα», «οπισθόφυλο», «γενικά» και «όλο το βιβλίο») από όλα τα βιβλία και να πάρουμε μια φορά την κάθε λέξη ώστε με την χρήση της Web Εφαρμογής να επιλέξουμε αν έχει κάποια αξία στο να συγκαταλέγεται στο ευρετήριο και αν ναι ποια είναι μορφή που αντιστοιχίζονται με σκοπό να βρίσκουμε τις ίδιες λέξεις. Για παράδειγμα τα ουσιαστικά τα αντιστοιχίζαμε με την ονομαστική πτώση (κλήση) και τα επίθετα με το αρσενικό φύλο.

Για παράδειγμα στην περίπτωση των τριών μορφών της λέξης «είναι» βρήκαμε τουλάχιστον 4 διαφορετικές μορφές : «Είναι», «Είναι», «είναι», «είναι». Με κεφαλαίο το πρώτο γράμμα καθώς και με ή χωρίς τόνο. Όμως, σε αυτήν την περίπτωση δεν υπάρχει κάποια ιδιαίτερη αξία για την ένταξή της στο ευρετήριο.

Σε ένα άλλο παράδειγμα η περίπτωση της λέξης πολίτης βρέθηκε με διάφορες μορφές. Ενδεικτικά μερικές είναι: «πολίτης», «πολιτών», «πολίτην». Ήταν λέξη που έπρεπε να υπάρχει στο ευρετήριο και η αντίστοιχη μορφή της για όλες τις εκδοχές που την εντοπίσαμε ήταν η λέξη «πολίτης»,

Σε αυτό το στάδιο αντιμετωπίσαμε ιδιαίτερα πρόβλήματα κυρίως εξαιτίας του πολυτονικού συστήματος και των γραμμένων λάθος λέξεων (ορθογραφικά λάθη/ αναγραμματισμοί κ.α.). Τα λάθη μπορεί να προέρχονται από τη διαδικασία του data entry που αρχικά είχε στηριχθεί στο να περνούν τα δεδομένα αυτά από μέλη της βιβλιοθήκης ενώ τα σήμερα τα λάθη είναι πολύ λιγότερα και είναι παράγωγα από λογισμικό OCR (Optical Character Recognition).

8. Δημιουργία ευρετηρίου

Έτσι μόλις υπολογίσαμε ποιες λέξεις έχουν σημασία για το ευρετήριο και ποια η κυρίαρχη μορφή της λέξης υπολογίσαμε το ποσοστό επί του κειμένου που είχαν. Δηλαδή αν για παράδειγμα στο συγκεντρωτικό κείμενο του ενός βιβλίου υπάρχουν 96 διαφορετικές λέξεις αλλά η μία 5 φορές και όλες οι υπόλοιπες από μια φορά τότε το ποσοστό θα ήταν περίπου 1% για όλες τις λέξεις εκτός από μία με ποσοστό 5%. Έτσι, κάθε βιβλίο συνοδεύεται από ευρετήριο όρων και ποσοστών του κάθε όρου επί του κειμένου ώστε να μπορούμε να χρησιμοποιήσουμε αυτά τα μεταδεδομένα (που δεν αντιλαμβάνεται ο χρήστης που κάνει την αναζήτηση) στον αλγόριθμο TF-IDF.

9. Αναζήτηση

Η αναζήτηση είναι πιο πολύπλοκη διαδικασία μιας και στηρίζεται σε μια λέξη ή μια φράση στην οποία ο χρήστης θα γράψει για να επικεντρωθεί στον περιεχόμενο που τον ενδιαφέρει. Το ζήτημα όμως είναι ότι με κάποιο τρόπο πρέπει να γίνει ένας έλεγχος για αυτές τις λέξεις, και ποιες να κρατήσουμε για να υπάρχει μια αξία στην αναζήτηση.

Με βάση τα περίπου 780 βιβλία που απαρτίζουν την βάση δεδομένων μας συγκεντρώσαμε περίπου 50.000 λέξεις. Αυτές οι λέξεις δεν είναι μοναδικές αλλά λέξεις που έχουν διάφορες μορφές ως προς την κλίση, αν είναι σε ενικό ή πληθυντικό κτλ. Όμως στην βάση έχουμε κρατήσει και λέξεις που δεν έχουν κάποια αξία μέσα από αυτές τις 50.000 καθώς και κάποιες που μπορεί να εκπίπτουν σε σφάλματα (όπως η λέξη «είναι» που έχει γραφτεί λαθεμένα ως «είνια»).

Με αυτόν τον τρόπο στην αναζήτηση με βάση το API και τις λέξεις που έχουν αξία η αναζήτηση «δικτατορία στα Ιωάννινα και νεολαία» θα αντιστοιχούσε σε αναζήτηση 3 διαφορετικών λέξεων : «δικτατορία», «Ιωάννινα» και «νεολαία». Πάλι με τη χρήση του API βλέπουμε ότι και οι 3 λέξεις έχουν σημασία. Το θέμα όμως τώρα είναι να παρουσιαστούν τα αποτελέσματα με βάση κάποια αξιολόγηση.

10. Αξιολόγηση των αποτελεσμάτων και sorting

Στο παράδειγμα «δικτατορία στα Ιωάννινα και νεολαία» βλέπουμε μέσα από τον πίνακα της κάθε εγγραφής αν οι λέξεις «δικτατορία», «Ιωάννινα» και «νεολαία» υπάρχουν και προσθέτουμε στο συνολικό σκορ της κάθε εγγραφής το ποσοστό από την κάθε λέξη. Έτσι σε αυτήν την περίπτωση βλέπουμε ότι οι πρώτες εγγραφές περιέχουν και τους 3 όρους και ακολουθούν οι εγγραφές που περιέχουν τους δύο και μετά τον έναν.

Στην περίπτωση του συγκεκριμένου παραδείγματος τα βιβλία που έχουμε ως αποτελέσματα είναι στην παρακάτω εικόνα:

Ιόνιο Πανεπιστήμιο Τμήμα Αρχειονομίας, Βιβλιοθηκονομίας & Μουσειολογίας		
δικτατορία στα Ιωάννινα και νεολαία		
No	Τίτλος	Κωδικός / Link
1	ΓΙΑΝΝΙΝΑ 1967-1973,	THE01001
2	Η ΔΙΚΤΑΤΟΡΙΑ ΜΕΤΑΞΑ	SYL00045
3	Ο Ελληνικός Φασισμός στον Μεσοπόλεμο	SYL00454
4	Η ΜΑΥΡΗ ΒΙΒΛΟΣ ΤΗΣ ΧΡΥΣΗΣ ΑΥΓΗΣ	PSA01001
5	Αναμνήσεις	STI01002
6	Αναμνήσεις	STI01001
7	50 Κείμενα για την κρίση	SYL00010
8	PENTZAIΟΙ	PAN02001

Προηγούμενη

Αποτελέσματα 8 Σελίδα 1 / 1

Επόμενη

Εκκαθάριση

Υπολογισμός

Στην περίπτωση της αναζήτησης «Ιταλία και αυτονομία» οι δύο πρώτες σελίδες αποτελεσμάτων είναι οι ακόλουθες ενώ τα αποτελέσματα είναι σχεδόν 80 μιας και η αναζήτηση αποτελείται μόνο από δύο λέξεις.

Ιόνιο Πανεπιστήμιο
Τμήμα Αρχειονομίας, Βιβλιοθηκονομίας
& Μουσειολογίας

ιταλία και αυτονομία

No	Τίτλος	Κωδικός / Link
1	Επισφάδα ?	SYL00136
2	Η Κόκκινη Διετία 1968-1969	SCAN01001
3	Σημείωση για την εργατική αυτονομία	SYL00264
4	Μιλώντας για τον ένοπλο αγώνα	SYL00007
5	ΑΥΤΟΝΟΜΙΑ	SYL00082
6	ΕΠΑΝΑΣΤΑΣΗ ΚΑΙ ΑΝΤΕΠΑΝΑΣΤΑΣΗ ΣΤΗΝ ΙΤΑΛΙΑ	SYL00020
7	Η μεγάλη ιστορία του 20 αιώνα	SYL00249
8	ΑΝΤΙΠΛΗΡΟΦΟΡΗΣΗ 1977 - 1978	SYL00004
9	ΑΝΤΙΠΛΗΡΟΦΟΡΗΣΗ 1978 - 1979	SYL00005
10	Μαθαίνουμε Ακούγοντας	SYL00539

Προηγούμενη
Αποτελέσματα 79 Σελίδα 1 / 8
Επόμενη
Εκκαθάριση
Υπολογισμός

Ιόνιο Πανεπιστήμιο
Τμήμα Αρχειονομίας, Βιβλιοθηκονομίας
& Μουσειολογίας

ιταλία και αυτονομία

No	Τίτλος	Κωδικός / Link
11	ANTI media	SYL00097
12	το κουρδικό πείραμα για τη δημοκρατική αυτονομία	SYL00103
13	ΧΡΟΝΟΣ ΚΑΙ ΧΕΙΡΑΦΕΤΗΣΗ	TIS01001
14	AZIONE RIVOLUZIONARIA	SYL00384
15	Πρόταση για μια ελεύθερη Καταλονία: Κοινοτιστική, Δημοκρατική και Συνομοσπονδιακή	SYL00487
16	Για τον Καστοριάδη και το περιοδικό "Σοσιαλισμός ή Βαρβαρότητα" της περιόδου 1952 - 1956	SYL00368
17	Β παγκόσμιος πόλεμος Υπουργείο Αμύνης Ε.Σ.Σ.Δ. 1939-1945	SYL00124
18	Η μεγάλη ιστορία του 20 αιώνα	SYL00252
19	1914 Η ΑΥΤΟΝΟΜΙΑ ΤΗΣ ΒΟΡΕΙΟΥ ΗΠΕΙΡΟΥ	SYL00100
20	Πολιτική εξουσία και κοινωνικές τάξεις	POU02002

Προηγούμενη
Αποτελέσματα 79 Σελίδα 2 / 8
Επόμενη
Εκκαθάριση
Υπολογισμός

Ακολουθεί ένα ακόμα παράδειγμα με τη φράση «Βοσνία και διάλυση της Γιουγκοσλαβίας» και τα αποτελέσματα είναι στην ακόλουθη φωτογραφία.

Βοσνία και διάλυση της Γιουγκοσλαβίας

No	Τίτλος	Κωδικός / Link
1	The fixer - A story from Sarajevo	SAC01001
2	ΒΟΣΝΙΑ, ΕΡΖΕΓΟΒΙΝΗ	SYL00069
3	Η κατάρρευση της Γιουγκοσλαβίας και οι ελληνικές φαντασιώσεις	TZI01002
4	Οι απτίες για τις ήττες τη διάσπαση του ΚΚΕ και της ελληνικής αριστεράς Τόμος Α	VON01001

[Προηγούμενη](#)

Αποτελέσματα 4 Σελίδα 1 / 1

[Επόμενη](#)

[Εκκαθάριση](#)

[Υπολογισμός](#)

Εξαιτίας της περιορισμένης βάσης σε αριθμό λέξεων αντιμετωπίσαμε το πρόβλημα ότι ο χρήστης στην αναζήτηση ωφείλει να θέσει όρους ακριβώς με τη μορφή που είναι περασμένοι στην βάση. Σε αντίθετη περίπτωση δεν επιστρέφει αποτέλεσμα. Τα αποτελέσματα λοιπόν περιορίζονται σε πολύ μικρό αριθμό και αυτό είναι ζήτημα που χρήζει επίλυσης. Μια διαφορά στο τονισμό, μια μορφή της λέξης που δεν υπάρχει είναι συνθήκες που μπορούν να θέσουν τον όρο εκτός αναζήτησης.

Για αυτή την περίπτωση ειλέξαμε την λύση με τον αλγόριθμο Levenshtein Distance. Είναι ένας αλγόριθμος που χρησιμοποιείται για τη μέτρηση της απόστασης μεταξύ δύο ακολουθιών χαρακτήρων. Η απόσταση Levenshtein μετράει πόσες μεταβολές (εισαγωγές, διαγραφές ή αντικαταστάσεις ενός χαρακτήρα) απαιτούνται για να μετατραπεί μια ακολουθία σε μια άλλη.

Πιο συγκεκριμένα, αν έχουμε δύο ακολουθίες χαρακτήρων, τον έναν "source" και τον άλλον "target", ο αλγόριθμος Levenshtein Distance υπολογίζει το ελάχιστο αριθμό επεξεργασιών που απαιτούνται για να μετατραπεί η ακολουθία "source" στην ακολουθία "target".

Ο αλγόριθμος είναι χρήσιμος σε πολλές εφαρμογές, όπως η αυτόματη διόρθωση ορθογραφικών λαθών, η αναζήτηση και η ταξινόμηση κειμένων, καθώς και σε προβλήματα που αφορούν την αναγνώριση προτύπων.

Στην δική μας περίπτωση λοιπόν με βάση τον συγκεκριμένο αλγόριθμο εντοπίσαμε την λέξη που είναι πιο κοντά στον όρο της αναζήτησης. Για να γίνει αυτό κάνουμε υπολογισμούς στο API όπου γίνονται όλοι οι υπολογισμοί μεταξύ αρχικών των μορφών λέξεων και της μορφής που επιλέγουμε για αυτούς ώστε να είναι η κεντρική με τον όρο αναζήτησης και επιστρέφουμε την λέξη με την μικρότερη απόσταση. Ο αλγόριθμος με μια απλουστευμένη μορφή σε javascript που χρησιμοποιούμε είναι στις παρακάτω λίγες γραμμές:

```

levenshteinDistance(str1:String, str2:String) {
  const len1 = str1.length + 1;
  const len2 = str2.length + 1;

  // Create a matrix to store distances
  const matrix = Array.from({ length: len1 }, (_, i) => Array.from({ length: len2 }, (_, j) => (i === 0 ? j : j === 0 ? i : 0)));

  // Fill in the matrix
  for (let i = 1; i < len1; i++) {
    for (let j = 1; j < len2; j++) {
      const cost = str1.charAt(i - 1) === str2.charAt(j - 1) ? 0 : 1;
      matrix[i][j] = Math.min(
        matrix[i - 1][j] + 1, // Deletion
        matrix[i][j - 1] + 1, // Insertion
        matrix[i - 1][j - 1] + cost // Substitution
      );
    }
  }

  // The bottom-right cell contains the Levenshtein distance
  return matrix[len1 - 1][len2 - 1];
}

```

11. Filtering Results

Ουσιαστικά η παραπάνω διαδικασία ταξινομεί όλα τα περιεχόμενα μέσα από μία διαδικασία από το κοντινότερο στην αναζήτηση μέχρι αποτελέσματα που δεν προκύπτει να έχουν καμμία σχέση με τα αποτελέσματα.

Αυτό θα μπορούσε να εξυπηρετήσει από την άποψη ότι πάντα έχουμε κάποια αποτελέσματα σε οποιαδήποτε αναζήτηση αλλά αυτό θα θέσει σε κίνδυνο την αξιοπιστία της αναζήτησης καθώς και ο χρήστης θα περιηγείται σε αποτελέσματα χωρίς κανέναν λόγο για κάτι τέτοιο.

Για αυτό θέσαμε έναν κανόνα όπου τα αποτελέσματα που θα έχει πρόσβαση ο χρήστης θα είναι κατά το ήμισυ τουλάχιστον των επιτυχών όρων που υπάρχουν στα βιβλία. Δηλαδή στην περίπτωση της αναζήτησης «Γιάννενα και κατοχή» τα αποτελέσματα που θα εμφανιστούν θα πρέπει οπωσδήποτε να εμπεριέχουν τον ένα από τους δύο όρους (προφανώς το και δεν συμπεριλαμβάνεται στο λεξιλόγιό μας ως χρήσιμη λέξη).

Στην περίπτωση όπου οι όροι εμπεριέχει τρεις βασικούς όρους «Αντιφασισμός και νεολαία στη μεταπολίτευση» τότε το αποτέλεσμα θα βασίζεται σε τουλάχιστον έναν όρο με καλό ποσοστό του όρου εντός του κειμένου ή σε τουλάχιστον δύο όρους.

Με αυτό τον τρόπο μειώνεται δραστικά ο τεράστιος αριθμός αποτελεσμάτων που δεν έχουν αξία προς τον χρήστη. Βέβαια ένα από τα σημαντικά αποτελέσματα είναι και ο όγκος αποτελεσμάτων που θα πρέπει να παραδώσει το API στο web application και να παρουσιαστεί με κάποιον κατανοητό εύκολα τρόπο στην χρήση. Παρακάτω παρατίθεται η συνάρτηση που χρησιμοποιήσαμε.

```

filterResultsByScores(){
  var minScore = 0.5 * allWords.length * 100;
  let fR = [];
  for(var i=0;i<results.length;i++){
    if(results.score>minScore){
      fR.push(results[i]);
    }
  }
  results = fR;
}

```

13. Αποτελέσματα

Τα αποτελέσματα από την διαδικασία ήταν αρκετά ικανοποιητικά. Σχεδόν ποτέ δεν χανόντουσαν στις αναζητήσεις οι κοντινές περιγραφές. Συνήθως σε ένα σύνολο 780 βιβλίων τα βιβλία που θα έπρεπε να συμπεριληφθούν, υπήρχαν πάντα ως αποτέλεσμα του αλγορίθμου. Μερικές φορές σε λάθος σειρά με βάση την αξιολόγηση από τον αλγόριθμο.

Ο αλγόριθμος όμως μερικές φορές επιστρέφει βιβλία με μια συνάφεια που δεν έχει άμεση σχέση με την αναζήτηση. Δηλαδή στην αναζήτηση η περιοχή πολλές φορές παρακάμπτεται με βάση τους υπόλοιπους όρους στην αναζήτηση. Δεν έχει υιοθετηθεί κάποια βαρύτητα στους όρους και αυτό κάνει πολλές φορές την αναζήτηση να κυμαίνεται σε διαφορετικά επίπεδα.

Ο αλγόριθμος θα ήταν πολύ πιο αποδοτικός και πιο εύκολα αξιολογίσιμος αν τα βιβλία ήταν της τάξης των 20.000 αντί για 780. Μέχρι τώρα οι λέξεις που έπρεπε να διαλέξουμε από το σύνολο του περιεχομένου των βιβλίων ήταν της τάξης των 50.000 και οι όροι που επιλέξαμε ότι έχουν μια σημασία περίπου 10.000. Για να γίνει αυτό δαπανήθηκε εργασία περίπου μιας εβδομάδας. Όμως, μετά από αυτό το σημείο η αναζήτηση/εύρεση νέων όρων προς επιλογή ή μη για την χρησιμότητα τους στο ευρετήριο γίνεται πιο δύσκολη και συνεπώς ευκολότερη στον όγκο εργασίας για τον χρήστη που εισάγει νέα δεδομένα στο σύστημα.

Θα είναι μια διαδικασία που το ανάλογο ερώτημα σχετικά με τους όρους θα υιοθετηθεί ως διαδικασία κατά την εισαγωγή του κάθε βιβλίου στο site με σκοπό την διευκόλυνση του χρήστη. Τέλος, η διαδικασία της χρησιμοποίησης του αλγορίθμου TF-IDF έχει αξία σε μια βιβλιοθήκη μεγαλύτερη αλλά η ανάπτυξη του συστήματος μέχρι τώρα μπορεί να αποτελέσει μια εύκολη βάση για την σύγκριση διαφορετικών διαδικασιών πάνω στην ευρετηρίαση καθώς και με απλές διαδικασίες να επιλεγεί ένας πιο σύνθετος τρόπος αναζήτησης που να στηρίζεται σε διαφορετικούς τρόπους και να δίνει αποτελέσματα με καλύτερη ακρίβεια στην ιεράρχηση αλλά και στον αριθμό με βάση την συνάφεια.

14. Βιβλιογραφία

- 1."Thesaurus Construction and Use: A Practical Manual", Jean Aitchison, Alan Gilchrist
- 2."Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works", Patricia Harpring
- 3."Introduction to Information Retrieval", Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze
- 4."Indexing: A Nuts-and-Bolts Guide for Technical Writers", Kurt Ament
- 5."Τα Βασικά της Βιβλιοθηκονομίας και της Πληροφόρησης", Κωνσταντίνος Κυριάκος
- 6."Introduction to Algorithms", Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein
- 7."Database Management Systems", Raghu Ramakrishnan, Johannes Gehrke
- 8."The Art of Computer Programming, Volume 3: Sorting and Searching", Donald E. Knuth
- 9."Database Systems: The Complete Book", Hector Garcia-Molina, Jeffrey D. Ullman, Jennifer Widom
- 10."Natural Language Processing with Python", Steven Bird, Ewan Klein, Edward Loper
- 11."Algorithms", Robert Sedgewick, Kevin Wayne
- 12."MongoDB - The Complete Developer's Guide", Maximilian Schwarzmüller
- 13."An Effective TF/IDF-Based Text-to-Text Semantic Similarity Measure for Text Classification", Shereen Albitar, Sébastien Fournier, and Bernard Espinasse
- 14."Research and Application to Automatic Indexing", Lei Wang , Shui-cai Shi, Xue-qiang Lv, Yu-qin Li
15. "Text Mining Approach Using TF-IDF and Naive Bayes for Classification of Exam Questions Based on Cognitive Level of Bloom's Taxonomy", Annisa Aninditya, Muhammad Azani Hasibuan, Edi Sutoyo