Κωνσταντής Παπαντώνιου-Χατζηγιώσης 4769, Βασίλειος Σωμάκος 4806, 23 Μαίου 2024

Αναφορά Εργασίας: Σχεδιασμός και Λειτουργικότητα του Συστήματος Αναζήτησης με Lucene

Εισαγωγή

Ο στόχος του συστήματος εύρεσης και αναζήτησης που αναπτύχθηκε είναι η δημιουργία ενός εργαλείου που επιτρέπει την αποδοτική αποθήκευση και αναζήτηση πληροφοριών από ένα σύνολο δεδομένων σε μορφή CSV. Το σύστημα χρησιμοποιεί τη βιβλιοθήκη Apache Lucene για τη δημιουργία index και την εκτέλεση αναζητήσεων. Παράλληλα, παρέχεται ένα γραφικό περιβάλλον χρήστη (GUI) για την αλληλεπίδραση με το σύστημα, διευκολύνοντας τους χρήστες να πραγματοποιούν αναζητήσεις και να προβάλλουν τα αποτελέσματα των αναζητήσεων αυτών.

Λειτουργικότητα του Συστήματος

Το σύστημα αποτελείται από τρία βασικά μέρη:

1. Indexer: Εργαλείο index που διαβάζει τα δεδομένα από ένα αρχείο CSV και δημιουργεί ευρετήρια χρησιμοποιώντας την Apache Lucene.
2. Searcher: Εργαλείο αναζήτησης που επιτρέπει την αναζήτηση με διάφορα κριτήρια.
3. LuceneGui: Γραφικό περιβάλλον χρήστη που επιτρέπει την αλληλεπίδραση με το σύστημα ς εύρεσης και αναζήτησης.

Συλλογή (corpus)

Περιγραφή της Συλλογής

Η συλλογή (corpus) του συστήματος αποτελείται από ένα αρχείο CSV το οποίο περιέχει επιστημονικά άρθρα. Το αρχείο περιλαμβάνει διάφορα πεδία που αντιπροσωπεύουν τις πληροφορίες των άρθρων. Η συλλογή αυτή είναι σημαντική για την αξιολόγηση και τον πειραματισμό με το σύστημα ανάκτησης πληροφορίας που έχουμε υλοποιήσει.

Αριθμός Αρχείων και Μέγεθος

Η συλλογή αποτελείται από ένα μόνο αρχείο CSV το οποίο περιέχει 400 εγγραφές. Κάθε εγγραφή αντιπροσωπεύει ένα άρθρο με τα αντίστοιχα πεδία του. Αυτό το μέγεθος είναι επαρκές για να πραγματοποιηθούν δοκιμές και να αξιολογηθεί η απόδοση του συστήματος.

Πηγή Δεδομένων

Τα δεδομένα προέρχονται από ένα σύνολο επιστημονικών άρθρων που έχουν συλλεχθεί και αποθηκευτεί σε μορφή CSV. Η πηγή αυτών των δεδομένων είναι η παρακάτω συλλογή από το Kaggle:

https://www.kaggle.com/datasets/rowhitswami/nips-papers-1987-2019-updated?select=papers.csv

Πεδία της Συλλογής

Τα πεδία που περιλαμβάνονται στην συλλογή είναι τα εξής:

1. source_id: Ένας μοναδικός αναγνωριστικός αριθμός για κάθε άρθρο.
2. year: Το έτος δημοσίευσης του άρθρου.
3. title: Ο τίτλος του άρθρου.
4. abstract: Η περίληψη του άρθρου.
5. full_text: Το πλήρες κείμενο του άρθρου.

<u>Ανάλυση Κειμένου και Κατασκευή Ευρετηρίου</u>

Προεπεξεργασία Άρθρων

Η διαδικασία της προεπεξεργασίας είναι κρίσιμη για τη δημιουργία ενός αποδοτικού ευρετηρίου. Ακολουθεί μια λεπτομερής περιγραφή της προεπεξεργασίας των άρθρων:

1. Ανάγνωση και Καθαρισμός Δεδομένων:

Τα δεδομένα διαβάζονται από ένα αρχείο CSV χρησιμοποιώντας τη βιβλιοθήκη opencsv.

Ελέγχονται για τη μορφοποίησή τους και αφαιρούνται οποιεσδήποτε μη αναγκαίες πληροφορίες.

2. Εξαγωγή και Ανάλυση Περιεχομένου:

Το περιεχόμενο κάθε άρθρου αναλύεται στα εξής πεδία: source_id, year, title, abstract, full_text.

Τα πεδία title, abstract και full_text αναλύονται για την δημιουργία των αντιστοίχων πεδίων στο ευρετήριο της Lucene.

3. Καθαρισμός Κειμένου:

Αφαίρεση ειδικών χαρακτήρων και συμβόλων που δεν είναι χρήσιμα για την αναζήτηση.

Εφαρμογή κανονικοποίησης για να διασφαλιστεί η ομοιομορφία των κειμένων.

4. Φιλτράρισμα και Ανάλυση Λέξεων:

Χρησιμοποιείται ένας StandardAnalyzer για την ανάλυση των κειμένων.

Οι λέξεις φιλτράρονται και αναλύονται για την δημιουργία όρων που θα χρησιμοποιηθούν στο ευρετήριο.

Δημιουργία Ευρετηρίου

Η κατασκευή του ευρετηρίου πραγματοποιείται με την χρήση της Apache Lucene. Ακολουθούν οι λεπτομέρειες της διαδικασίας:

1. Δημιουργία IndexWriter:

Δημιουργείται ένας IndexWriter με χρήση ενός StandardAnalyzer για την ανάλυση των κειμένων.

Ο IndexWriter είναι υπεύθυνος για την προσθήκη εγγράφων στο ευρετήριο.

2. Προσθήκη Εγγράφων:

Κάθε άρθρο προστίθεται ως ένα έγγραφο (Document) στο ευρετήριο.

Τα πεδία του άρθρου προστίθενται στο έγγραφο χρησιμοποιώντας τις αντίστοιχες κλάσεις πεδίων (StringField για τα πεδία source_id και year, TextField για τα πεδία title, abstract, και full_text).

3. Αποθήκευση και Κλείσιμο:

Μόλις όλα τα έγγραφα έχουν προστεθεί, ο IndexWriter κλείνει και το ευρετήριο αποθηκεύεται στον καθορισμένο κατάλογο.

Υποστήριξη Διαφόρων Τρόπων Αναζήτησης

Η δομή του ευρετηρίου και τα πεδία του υποστηρίζουν διάφορους τρόπους αναζήτησης:

1. Αναζήτηση Κλειδιών (Keyword Search):

Μπορεί να γίνει αναζήτηση σε πολλαπλά πεδία (title, abstract, full_text) χρησιμοποιώντας έναν MultiFieldQueryParser.

2. Αναζήτηση Πεδίου (Field Search):

Η αναζήτηση μπορεί να περιοριστεί σε συγκεκριμένο πεδίο (π.χ. μόνο στον τίτλο ή το full_text) χρησιμοποιώντας έναν QueryParser για το αντίστοιχο πεδίο.

3. Ταξινόμηση Αποτελεσμάτων:

Τα αποτελέσματα μπορούν να ταξινομηθούν με βάση το έτος δημοσίευσης ή αλφαβητικά με βάση τον τίτλο χρησιμοποιώντας την κλάση ScoreDoc για την ανάκτηση και ταξινόμηση των εγγράφων από το ευρετήριο.

Με αυτήν την ανάλυση και κατασκευή ευρετηρίου, το σύστημα εξασφαλίζει την αποδοτική και ακριβή αναζήτηση επιστημονικών άρθρων, προσφέροντας ταυτόχρονα ευελιξία στους χρήστες για διαφορετικούς τρόπους αναζήτησης και παρουσίασης των αποτελεσμάτων.

Είδη Ερωτήσεων Πέρα από την Αναζήτηση με Λέξεις Κλειδιά

Το σύστημα υποστηρίζει διάφορα είδη ερωτήσεων πέρα από την απλή αναζήτηση με λέξεις κλειδιά, τα οποία είναι τα εξής:

1. Αναζήτηση με Φράσεις (Phrase Search):

Οι χρήστες μπορούν να αναζητούν συγκεκριμένες φράσεις μέσα σε εισαγωγικά, π.χ., "machine learning applications".

2. Αναζήτηση με Λογικούς Τελεστές (Boolean Search):

Χρήση λογικών τελεστών όπως AND, OR, NOT για την δημιουργία πιο σύνθετων ερωτημάτων.

Π.χ., machine AND learning, machine OR learning, machine NOT learning.

Παρουσίαση Αποτελεσμάτων Αναζήτησης

Δομή Παρουσίασης Αποτελεσμάτων

Τα αποτελέσματα της αναζήτησης παρουσιάζονται με τον ακόλουθο τρόπο:

1. Αριθμός Αποτελεσμάτων Ανά Σελίδα:

Τα αποτελέσματα παρουσιάζονται ανά 10 ανά σελίδα. Οι χρήστες μπορούν να πλοηγηθούν στις επόμενες σελίδες χρησιμοποιώντας κουμπιά πλοήγησης.

2. Τονισμός Όρων Αναζήτησης:

Οι όροι αναζήτησης τονίζονται μέσα στα αποτελέσματα για να είναι εύκολα αναγνωρίσιμοι. Αυτό βοηθά τους χρήστες να εντοπίσουν γρήγορα τα τμήματα των άρθρων που είναι σχετικά με την αναζήτησή τους.

3. Πληροφορίες για το Άρθρο:

Περιλαμβάνονται βασικές πληροφορίες όπως ο τίτλος, το όνομα του συγγραφέα, η ημερομηνία δημοσίευσης και η πηγή του άρθρου.

Δυνατότητες Αναδιάταξης Αποτελεσμάτων

Οι χρήστες έχουν τη δυνατότητα να αναδιατάξουν τα αποτελέσματα της αναζήτησης βάσει δύο κριτηρίων:

1. Χρονολογική Σειρά:

Ταξινόμηση βάσει της ημερομηνίας δημοσίευσης (μίκροτερη προς μεγαλύτερη)

2. Αλφαβητική Σειρά