Information Retrieval

Project

Dataset : github issues

Introduction

Collected dataset consisting of 100,000 documents of github issues with the url and title. On executing the program, for a query it retrieves the top 10 docs with the github urls that matches the given query.

Observations from dataset

Contains most repetitive data so the no of terms retained after preprocessing are less

Execution:

- Preprocesses the data
- Calculates the inverted index
- Updates the posting list
- Implements cosine similarity in giving the results

Algorithms/ Approaches

1. Preprocessing the data
    a. Removes punctuations from the text data : using regex
    b. Removes the numbers from the data : uses regex
    c. Converts the data to lowercase : uses .lower() function
    d. Tokenizes the words from the text : uses nltk word_tokenize library
2. Inverted Index
    a. Calculates the inverted index by traversing to each token in the word and placing in the dict

---

b. After every 50 docs updates the dict
c. Updates the posting list

3. Cosine Similarity

a. Uses term at a scoring for calculating the cosine similarity
b. Taking the common tokens which are present in query and doc, it first calculates the term frequency in the doc, document frequency, then computes the tf-idf score and adds to the score of the doc
c. Finally divides it by the length of the doc (normalization)
d. Gives the top 10 retrieved results with their score