Walking through the projects

Document Retrieval with tf-idf and BM25 ranking algorithm

Smita Bhattacharya
M.Sc. in Data Science & Artificial Intelligence
Saarland University

Document Retrieval with TF-IDF and BM25 ranking algorithm

- **Goal:** Retrieve relevant document based on given query

- **Dataset:**
  - **Corpus**: TREC dataset
  - **Query:** A text file that contains the query
  - **Helper file:** A text file contains regular expression for each query for evaluation
- **Method:** TF- IDF
- **Process:**
  - The corpus is **parsed** using **beautifulsoup library** to create a data frame(Pandas) contains document id and corresponding text
  - Preprocessing of the text is done by
    - Tokenization(using NLTK library)
    - Lower-casing
    - Removing punctuation tokens
  - Inverse document frequency is computed for each term(Using Math library to calculate the log value)

Document Retrieval with TF-IDF and BM25 ranking algorithm(Cont.)

- **Term frequency** is calculated for each term of a document and stored along with doc id
- For query terms, the tf-idf weights for these terms as the product of the term's idf and the tf-value of the term in the respective document is computed and stored
- Each corpus document and each query document is represented as vector of tf-idf score
- **Cosine similarity** is calculated to get the relevance
- The relevance score is **sorted and top 50** documents are output as result

Evaluation:

- Gold standard relevant documents are fetched by using the regex in the pattern file
- **Evaluation metric:**
  - Precision@50: 65%

Disadvantage:

- Does not capture position in text, semantics, co-occurrences in different documents

Document Retrieval with TF-IDF and BM25 ranking algorithm(Cont.)

- **Method:** Okapi BM25
- **Process:**
  - Calculation of tf and idf weight is similar as before
  - 2 new hyperparameters:
    - K = controls the impact of term frequency(1.2)
    - B = controls the impact of document frequency(0.75)
    - The formula is different.
    -

$$\text{BM25}(D,Q) = \sum_{i=1}^{n} IDF(q_i, D) \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i) + k_1 \cdot (1 - b + b \cdot |D|/d_{avg}))}$$

- Precision @50 improves to 79 %