BM25 Algorithm

Outline

1. Recap NLP Basics
2. Project Discussion
3. BM25 Algorithm

Where to find dataset to work with?

1. https://paperswithcode.com/datasets
2. https://www.kaggle.com/datasets
3. https://datasetsearch.research.google.com/
4. https://archive.ics.uci.edu/ml/datasets.php
5. https://www.tensorflow.org/datasets
6. https://huggingface.co/docs/datasets/
7. https://registry.opendata.aws/
8. Everywhere …

Okapi BM25

Introducing BM25

- BM25 is a ranking function used by search engines to rank documents by relevance to a query.
- It's a bag-of-words retrieval function, meaning it considers individual terms in a document without regard to grammar or word order.
- BM25 improves on TF-IDF by incorporating document length and term frequency saturation.
- Key components of BM25 include term frequency (TF), inverse document frequency (IDF), and document length normalization.
- It's known for its effectiveness and efficiency in information retrieval tasks.

BM25: The Math

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

$$\text{IDF}(q_i) = \ln\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right)$$

where parameters k_1 and b are tunable.

$f(q_i, D)$ # of times q_i appears in the doc D

$n(q_i)$ # of docs containing q_i

https://en.wikipedia.org/wiki/Okapi_BM25

Information Theoretic BM25

- BM25 can be interpreted as a probabilistic relevance model.
- It estimates the probability that a document is relevant to a query.
- This is based on term frequencies and document lengths.
- BM25 incorporates term frequency saturation, unlike TF-IDF.
- It assumes documents with higher term frequencies are more likely to be relevant.
- The model also considers document length to avoid bias towards longer documents.

BM25 Extensions: BM25+ and BM25L

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \left[\frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)} + \delta \right]$$

$$\text{score}(d,t) = \text{IDF}(t) \cdot \frac{f'(t,d) \cdot (k_1 + 1)}{f'(t,d) + k_1}$$

$$f'(t,d) = \frac{\text{tf}'(t,d)}{1 - b + b \cdot \text{dl}'(d)}$$

$$\text{dl}'(d) = \frac{\text{dl}(d)}{\text{avdl}}$$

$$\text{tf}'(t,d) = \text{tf}(t,d) \cdot \log_2 \left(1 + c \cdot \frac{\text{avdl}}{\text{dl}(d)}\right)$$

- BM25+: primary focus is on improving the scoring of documents with low term frequencies for a given query term.
- BM25+: Introduces an additional parameter that provides a baseline score, ensuring that even documents with infrequent query term occurrences contribute to the overall relevance score
- BM25L: Addresses issues with long and short documents in BM25.
- BM25L: Normalizes term frequency based on document length, promoting fairness.
- Both extensions aim to enhance the effectiveness of the original BM25 algorithm.