INFORMATION RETRIEVAL

https://www.geeksforgeeks.org/inverted-index/

Introduction to Information Retrieval

1.1 The impact of the web on IR, unstructured and semi-structured text

https://www.geeksforgeeks.org/difference-between-structured-semi-structured-and-unstructured-data/

1.2 Basic IR Models Inverted index and Boolean queries, Boolean and vector-space retrieval models

https://www.tutorialspoint.com/natural_language_processing/natural_language_processing_information_retrieval.htm#:~:text=Boolean%2C%20Vector%20and%20Probabilistic%20are%20the%20three%20classical%20IR%20models.

- **R** − A document is predicted as relevant to the query expression if and only if it satisfies the query expression as −

$((text \vee information) \wedge rerieval \wedge \tilde{} \ theory)$

Cosine Similarity Measure Formula

Cosine is a normalized dot product, which can be calculated with the help of the following formula −

$$Score(\vec{d}\vec{q}) = \frac{\sum_{k=1}^{m} d_k \cdot q_k}{\sqrt{\sum_{k=1}^{m}(d_k)^2} \cdot \sqrt{\sum_{k=1}^{m} m(q_k)^2}}$$

$$Score(\vec{d}\vec{q}) = 1 \ when \ d = q$$

$$Score(\vec{d}\vec{q}) = 0 \ when \ d \ and \ q \ share \ no \ items$$

1.3 Ranked retrieval; text-similarity metrics; TF-IDF (term frequency/inverse document frequency) weighting; cosine similarity

https://www.cse.iitk.ac.in/users/nsrivast/HCC/ranked%20retrieval.pdf

The log frequency weight of term t in d is

$$w_{t,d} = \begin{cases} 1 + \log_{10} \mathrm{tf}_{t,d}, & \text{if } \mathrm{tf}_{t,d} > 0 \\ 0, & \text{otherwise} \end{cases}$$

1.4 Basic Tokenizing, Indexing, and Implementation of Vector-Space Retrieval: Simple tokenizing

1.5 Stop-word removal, and stemming; inverted indices; efficient processing with sparse vectors

Retrieval Models

2.1 Boolean, vector space

2.2 TFIDF, Okapi, probabilistic

Okapi BM25: A Nonbinary Model

- The simplest score for document d is just idf weighting of the query terms present in the document:

$$RSV_d = \sum_{t \in q} \log \frac{N}{\mathrm{df}_t}$$

- Improve this formula by factoring in the term frequency and document length:

$$RSV_d = \sum_{t \in q} \log \left[\frac{N}{\mathrm{df}_t} \right] \cdot \frac{(k_1 + 1)\mathrm{tf}_{td}}{k_1((1 - b) + b \times (L_d / L_{\mathrm{ave}})) + \mathrm{tf}_{td}}$$

2.3 language modeling, latent semantic indexing

2.4 Vector space scoring. The cosine measures. Efficiency considerations

https://www.techopedia.com/definition/30336/link-analysis

https://nlp.stanford.edu/IR-book/pdf/20crawl.pdf

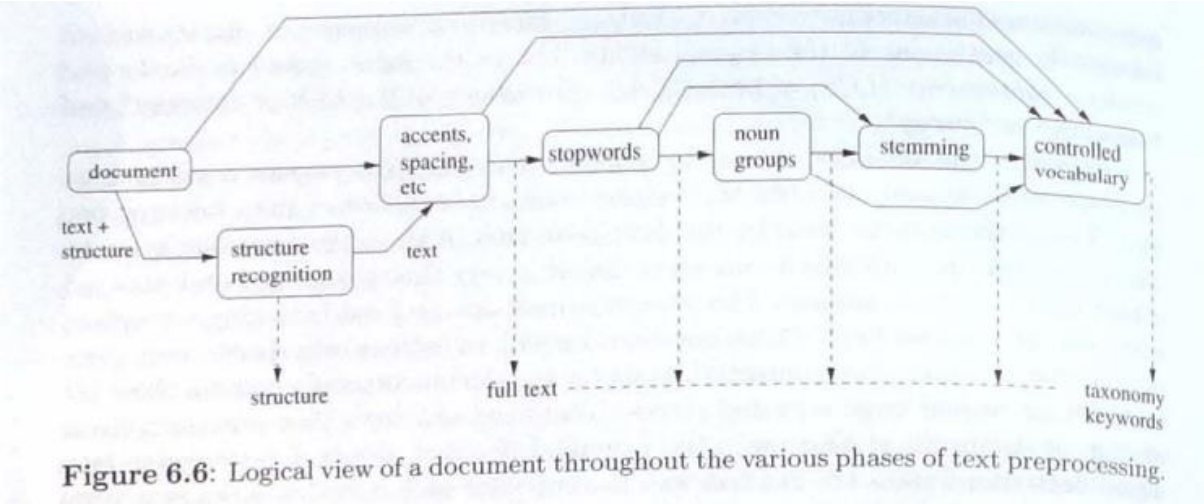

Figure 6.6: Logical view of a document throughout the various phases of text preprocessing.

pg 255

Pg 258 Stemming, thesaurus,