Web Search Engine

Project Description

The objective of this project is to build a web search engine for the UIC domain. This includes building a web crawler, preprocessing & indexing the web pages, and implement the vector space model to retrieve top 10 relevant webpages based on input query.

Weighting Scheme

Out of the various weighting schemes available, the technique used for this project is ***TF-IDF*** (term frequency – inverse document frequency) where the value increases proportionally to the number of times a word appears in a document and how many documents contain that word.

Similarity Measure

The similarity measure used for this project is ***Cosine Similarity.*** Cosine similarity uses document vector and a query vector; the vectors represent each unique term with an index, and the value at that index is some measure of how important that term is to the document. It is calculated by defining the cosine of the angle between them i.e. dot product of vectors divided by product of their lengths.

Possible Alternatives

The other similarity measures are Dice's Coefficient, Jaccard's Coefficient, and Overlap Coefficient. Cosine similarity is usually used in the context of text mining for comparing documents whereas Jaccard's Coefficient is usually used for binary cases and Dice Coefficient just checks for the existence of a word in the document.

The other weighting schemes available are Word2Vec and Bag-Of-Words. Word2Vec is a technique where words from the vocabulary are mapped to vectors of real numbers. These vectors are calculated from the probability distribution for each word appearing before or after another. Bag-Of-Words builds a vocabulary from a corpus of documents and counts how many times the words appear in each document.

Another technique used for ranking is Page Rank where PageRank assigns a score to a document based upon the documents it links to, and the documents which link to it. The score does not vary depending on the query used.