Martin Mao

Xi Chen

Report

NDCG@5 values before improvements:

mondego:0.26

machine learning:0.17

software engineering:0

security:0

student affair:0

graduate courses:0

crista lopes:0.35

REST:0

computer games:0.35

information retrieval:0

Improvements

* Added Weights based on Title and Anchor Text. These weights will add more weight to documents that contain the query terms in their Title and Anchor texts.
  + We noted that the first few results returned would be results containing the title or anchor text of queries. The remaining results might not contain the query terms in either the title or anchor text.
* Add 2-Gram, 3-Gram, and 4-Gram weights per documents. Documents that contain a permutation of the query terms in a 2-gram will have their weights adjusted. The weights of the N-grams are based on the TFIDF of the N-grams among documents.
* Every file we searched (had a different file for each n-gram type, a file for all titles, a file for all anchor texts, a file for TFIDF numbers, and a file for all positions) was run on a separate thread. Since each file was independent, we could speed up search by doing file searches on all different threads.
* The positions were used to find the relevant text snippets related to the query terms.
* Modified Tokenizer to treat Capitalized letters as a split between words, in case the user or a document forgot to put a space to separate out the words in the document.
* Used the hyperlinks from each file as a guide for a hub for each file. Each link that is targeted will have an authority associated with how many links point from each hub file. We will run hubs and authority computation up to 3 times to get the authority score for each link. The authority will add a weight to the links.
* The total weight of each document is computed by:
  + First find the TFIDF of the N-Gram for the query
  + Add the TFIDF scores of each 2-Gram permutation found in the query
  + Add the TFIDF values computed from normal Term Frequency
  + Then we compute the cosine similarity of each document’s current weight
  + Add in a weight determined by authority of each docID
    - Weight is computed by link’s authority divided by the average authority of all searched documents; then we take the logarithm of that value.
  + We then add more weight based on the title of the page
    - We take the average weight of all found documents and add double that weight to all documents with the query terms in the document title
  + We then add more weight based on the anchor text of the page
    - We take the average weight of all found documents and add that weight to all documents with the query terms in the document’s anchor text
  + We then return the documents with the highest remaining weight values

Information

* The search engine and NDCG computation are on separate projects and are started separately from each other.
* Time to pull all ten documents with text snippets
  + 1874 ms
* Time to pull all ten documents without text snippets
  + 776 ms
* Time to pull a one word query with text snippets
  + 240 ms
* Time to pull a one word query without text snippets
  + 210 ms
* Time to pull a four word query with text snippets
  + 384 ms
* Time to pull a four word query without text snippets
  + 302 ms