CSCI 6370: Information Retrieval Tuesday July 7, 2015

Search Engine Phase Three

**Team and Search Engine name**: Boom

|  |  |  |
| --- | --- | --- |
| **Team member** | **Student ID** | **Role** |
| Alex Campos | 10234425 | Python Developer |
| Alvaro Leal | 10123676 | Code Reviewer, Testing and Design Improvements |
| Divya Vuppala | 20331466 | Technology research – Python Libraries and Framework |
| Akhil Choppadandi | 20331723 | Functional Testing of search engine |

**Objective**: To customize our current inverted index so that we calculate the total number of words contained in it, as well as their corresponding frequency. The idea is that once we have this information, we go ahead and compute the most significant index terms and modify our inverted index so that only the most significant words are part of it.

The algorithm to compute the most significant words is as follows:

1. Calculate the inverted index as normal
2. Once we have the inverted index, calculate its length to get total number of words
3. For each key (term) in our inverted index count calculate the term frequency and store it on a temporary array.
4. Sort the temporary array by decreasing word frequency order (from most frequent to least frequent word).
5. Remove the top and bottom 10% words. Everything in between are considered the most significant terms.

Figure : We remove the most and least frequent terms and end up with the most significant terms (middle)

With the most significant inverted index already calculated, we proceed to build a set of 10 queries Q as follow: computer AND science, computer BUT science, web OR developer, computer OR technology, graduate OR resource, undergraduate AND fall, utrgv, student AND government, advisement, compass AND accuplacer

Given these queries, we calculated the following recall and precision for both 1) a search engine using the traditional inverted index and 2) the modified most significant inverted index getting the following results:

In conclusion, we can say that a simple Inverted Index does provide the maximum amount of recall but the precision is not very accurate which is the exact opposite of what happened when we have a most significant term inverted index, which with our precision but in our small corpus, it severely impacted our recall (very few documents were retrieved).