

CS 221 – Project 3 Milestone 3

TEAM: Priyanka Ravi (33246700) and Rahul Sridhar (41608676)

APPROACH:

As far as Milestone 2 was concerned, the ranking was based on only the document URL, title, each query term's tf-idf and finally PageRank. The incremental changes from the previous milestone to this milestone with respect to ranking, retrieval and user interface design are:

- Included scores for all fancy tags – meta, h1, h2, h3, h4, th, b, a
- Scores for bigrams
- Overall parameter tuning based on the standard query set of 10 terms and a few other query terms
- Show the relevant text snippet for each document in the results
- Minor aesthetic changes to the interface

Our rationale behind choosing the above set of heuristics for ranking documents:

1. URL, title and other fancy tags:
 - a. For any search term, it is quite natural to expect results that describe the search term entirely. This is the intuition behind using the condition of presence of the term within the URL and/or tags – title, h1, h2, etc. If the search term is present within the URL itself or one of these tags, there is a high probability for the documents to be relevant to the search term.
2. Meta tag:
 - a. Some websites explicitly populate the “*keywords*” attribute of the “*meta*” tag for search engines to make use of. This method provides valuable information for ranking albeit being susceptible to keyword stuffing and hence increasing the number of false positives in the displayed results
 - b. Given that we are indexing and ranking the ICS domain, we expect it to be mostly free of the keyword stuffing problem
3. Bigrams:
 - a. Intuitively, terms/tokens of the query that appear together in documents should be weighted more than terms appearing independently
4. tf-idf:
 - a. One of the most important term and document-based heuristics that is used to give more weight to rare terms that appear frequently within a document
5. PageRank:
 - a. Although its use has been limited by Google to avoid link spamming, we believe this method has a lot of merit within the context of this search engine
 - b. As expected, www.ics.uci.edu has the highest PageRank amongst all the websites that we have used for building this search engine

Additionally, although we had stored the font size and capitalization information of terms while building the indices, when we included these as factors in the scoring process, we noticed that the information they were capturing was highly correlated with the information that fancy tags such as title, h1, h2, etc. were capturing. To reduce the risk of overfitting, we did not include them in the final scoring process.

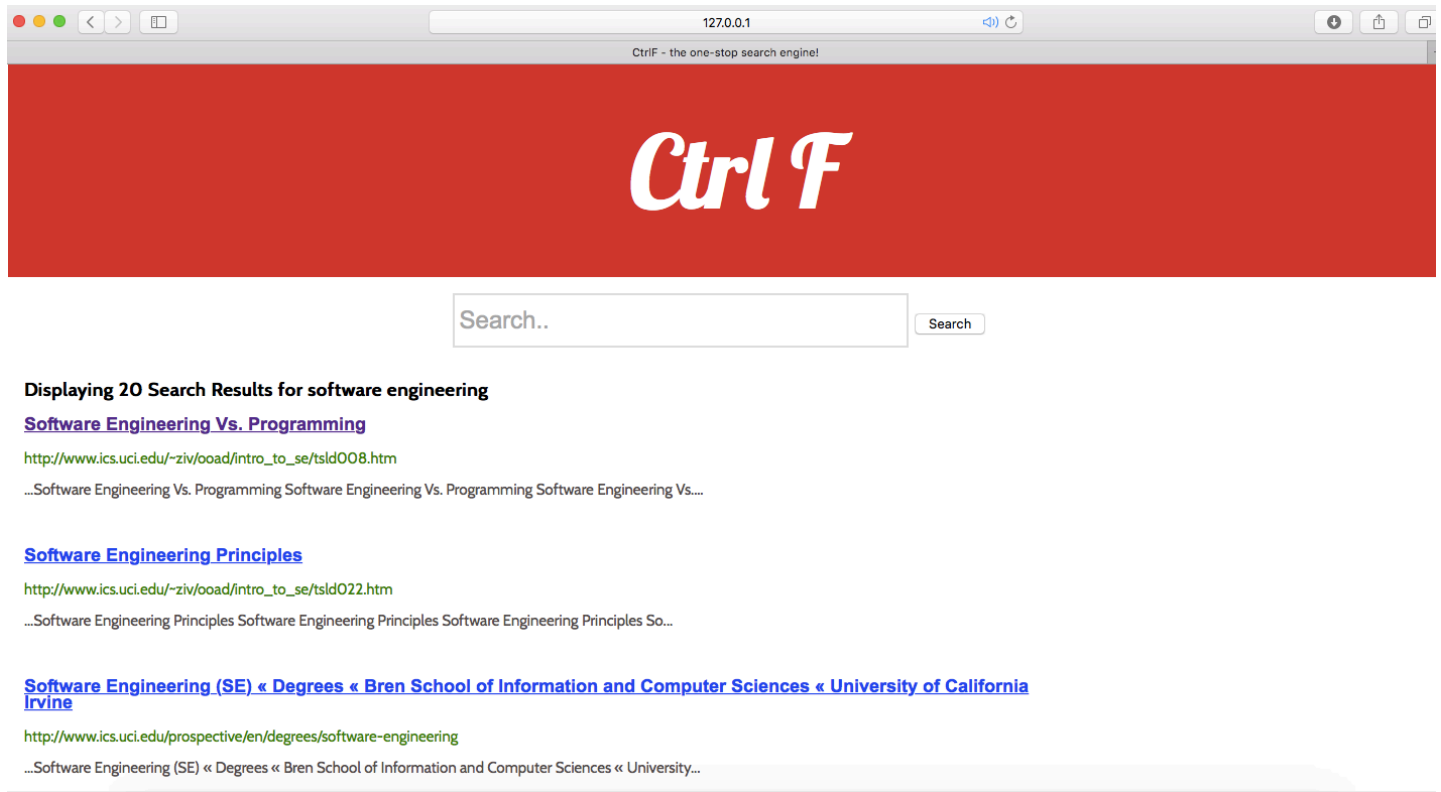
Other experiments that were performed (but weren't used in the final results):

- Acronym expansion (discarded due to no/insignificant improvement to results)
- Scraping Google's results:

- We wrote a few scripts using Google's API to retrieve Google's results for the 10 queries in the standard set
- The problem we identified was that some of the results retrieved by the API were either not present in the web page when we manually entered the search term in Google or were in a different order than that shown by Google
- On further research, we gathered that this is how the Google API was intended to be used and there was no workaround to retrieve the actual results that are seen by a real-world user
- We hence extracted the required links manually from Google

RESULTS:

Final Interface:



NDCG@5 Calculation and Comparison:

Note: NDCG was calculated on a 1-5 relevance scale (5 being the highest). The formulation used for DCG:

$$DCG_5 = rel_1 + \sum_{i=2}^5 \frac{rel_i}{\log_2 i}$$

DCG for Google's results:

Index	Relevance	DG	DCG
1	5	5	5
2	4	4	9
3	3	1.8298	10.8928
4	2	1	11.8928
5	1	0.4307	12.3235

NDCG@5 = DCG@5 (our search engine)/ DCG@5 (Google)

- Along with this report, we have also attached two files showing our:
 - Results for all the queries, and
 - NDCG calculation methodology

Query	NDCG@5 - Old	NDCG@5 - Final	Comments
mondego	0.4057	0.4057	Same as previous
machine learning***	0.4057	0.4057	Same as previous
software engineering	0.0699	0.4183	Improvement
security**	0.3246	0.3246	Same as previous
student affairs	0.3246	0.3758	Improvement
graduate courses	0.4057	0.4057	Same as previous
Crista Lopes	0.6492	0.6492	Same as previous
REST	0.5680	0.7303	Improvement
computer games*	0.4057	0.5455	Improvement
information retrieval**	0.5274	0.5806	Improvement
Average	0.4087	0.4841	Improvement

*We were able to match only two of Google's results with the valid links in bookkeeping. The NDCG@5 value shown here is a worst-case extrapolation of what would have been the NDCG assuming none of the other three (hypothetically valid) links shown in Google were retrieved by us

** Similar to above, but we were able to match four Google results

*** The top link shown by Google and our search engine are different as far as the identifier (URL) is concerned, but redirect to the same page. We have considered this to be a match in our calculation.

KEY OBSERVATIONS AND CONCLUSION:

At an overall level, there is an improvement in the average NDCG@5 value for the standard query set. If we break down the values at a query level, from the previous milestone to this milestone, there is an improvement in five queries (out of which 4 are bigrams) and the value stays the same for the other five queries (out of which even though 3 are bigrams, one of them is a name of a person).

The improvement can be attributed to the incremental additions made from last milestone to this one, especially the bigram analysis. We hypothesize that the improvement is not too high due to a few reasons:

- Our scoring technique in the previous milestone was non-trivial and performed quite well on its own
- Bigram analysis is one of the biggest attributes for the improvement, but as noted above, for the five queries in which no improvement was observed, "machine learning", "Crista Lopes" and "graduate courses" are the only 3 bigrams, and one can reasonably adjudge that the nature of the query terms is such that bigram analysis might not bring about a great improvement to their results (especially for the first two terms)

Finally, we believe that this comparison of our search engine's performance with Google's must be taken with a pinch of salt. When one looks at Google's results objectively, it is easy to discern that Google does not use only query-based features as inputs to its ranking model. Some of the features it seems to be using are the age of the webpage, how frequently it is updated, user visit frequency, advanced parameter tuning and much more, which are all features that we are not privy to. Taking all the above into consideration, we treat our search engine's performance as a success, although we acknowledge the fact that there is still room for improvement. Latent Semantic Indexing, machine learning-based parameter tuning, etc. are potential enhancements that could be performed to realize further improvement.