**Information Retrieval Systems**

Evaluation and Retrieval Effectiveness Comparison

(Submitted in partial fulfillment of the requirements of Northeastern University’s

CS6200 Fall 2017 course, taught by Prof. Nada Naji)

Athul **Muralidharan**

Meghna **Venkatesha**

Sachin **Haldavanekar**

1. Introduction

Throughout the Fall of 2017, the authors were exposed to various core Information Retrieval concepts. In this project, an attempt has been made to put the concepts into practice by building, evaluating and comparing different search engines.

The goal of this project is to build IR systems, make variations to improve the baseline runs, and compare their performance levels in terms of retrieval effectiveness.

The project is mainly coded in Python and Java. Excel has been used for analysis and data visualization. Libraries that are used include BeautifulSoup and Lucene. PyCharms and IntelliJ by JetBrains are the IDEs that were used to aid in coding. Github has been used for version control.

This work on this project has been planned and executed by Athul Karthik, Meghna Venkatesha and Sachin Haldavanekar. All of them actively studied various resources to implement IR systems.

Sachin was responsible making the Lucene IR system, evaluation of search engines and the documentation. Meghna was involved in for cleaning the given files, making the tf-idf Retrieval System, query expansion, removing stop words from the corpus and snippet generation. Athul coded the indexer, BM25 algorithm and smoothed query likelihood IR systems.

1. Literature and Resources

The techniques used in this project are:

1. Indexing using unigrams
2. Retrieval Modelling with BM25, tf-idf, Smoothed Query Likelihood and Lucene
3. Stopping using a stop list
4. Stemming using a stem list
5. Query Expansion using pseudo-relevance feedback
6. Snippet Generation using static summary

The tools that were used in this project are:

1. PyCharms by JetBrains
2. IntelliJ by JetBrains
3. Eclipse by Eclipse Foundation, Inc.
4. Github Desktop by GitHub, Inc.
5. Lucene Core by The Apache Software Foundation
6. Excel by Microsoft

The scholarly work and research articles (refer Bibliography for URLs) that were referenced include:

1. Selecting Effective Expansion Terms for Better IR
2. Query Expansion by Pseudo Relevance Feedback
3. Query Expansion Using Term Distribution and Term Association
4. Query Expansion using Local and Global Document Analysis
5. A Probabilistic Analysis of the Rocchio Algorithm with TF-IDF for text Categorization
6. Generation of Document Snippets Based on Queries and Search Results
7. Implementation and Discussion
8. Results

1. Conclusion and Outlook
2. Bibliography

Iman, Hazra, and Aditi Sharan. "Selecting Effective Expansion Terms for Better Information Retrieval." International Journal of Computer Science and Applications 7.2 (2010): 52-64. Web. 4 Dec. 2016. <http://www.tmrfindia.org/ijcsa/v7i24.pdf>.

Feng, Zheyun. "Query Expansion by Pseudo Relevance Feedback." N.p., 27 Aug. 2012. Web. <http://fengzheyun.github.io/downloads/projects/before2015/documen t/DocRetri.pdf>.

Pal, Dipasree, Mandar Mitra, and Kalyankumar Datta. "Query Expansion Using Term Distribution and Term Association." N.p., 4 Mar. 2013. Web. 5 Dec. 2016. <https://arxiv.org/abs/1303.0667>.

Xu, Jinxi, and W. Bruce Croft. "4. Query Expansion Using Local and Global Document Analysis." N.p., 02 Oct. 2010. Web. 05 Dec. 2016. <http://www.eng.utah.edu/~cs7961/papers/XuCroft-SIGIR96.pdf>.

Joachims, Thorsten. "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization." N.p., 1997. Web. 4 Dec. 2016. <https://www.cs.cornell.edu/people/tj/publications/joachims\_97a.pdf>.

Verstak, Alexandre A., and Anurag Acharya. Generation of Document Snippets Based on Queries and Search Results. Patent US 8145617 B1. 27 Mar. 2012. Print.

Carr, Ian, and Lucy Vasserman. "SnipIt - A Real Time Dynamic Programing Approach to Snippet Generation for HTML Search Engines." (n.d.): n. pag. Web.<http://www.cs.pomona.edu/~dkauchak/ir\_project/whitepapers/S nippet-IL.pdf>.