Evaluating and Comparing Performance of

Information Retrieval Systems

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

CS6200 Fall 2016 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 BeutifulSoup, jSoup, 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 for cleaning the given files, making the Lucene IR system, evaluation of search engines and the documentation. Meghna was involved in making the BM25 algorithm, query expansion, removing stop words from the corpus, and snippet generation. Athul coded the indexer, tf-idf and smoothed query likelihood IR systems.

1. Literature and Resources
2. Implementation and Discussion
3. Results

1. Conclusion and Outlook
2. Bibliography