

News Search Engine

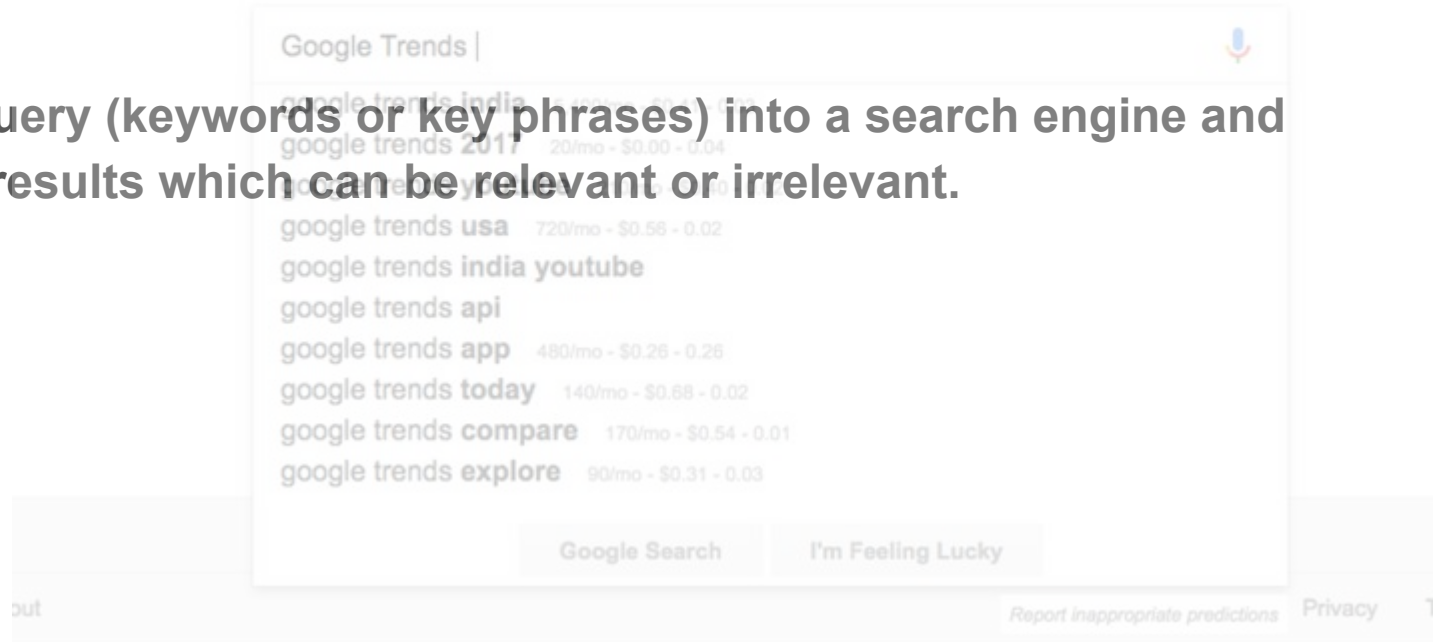


Deepak Kumar (S20160010022)
Kaustubh Pandey(S20160010041)
Prashant Mahanta (S20160010066)
Wasim Ishaq Khan (S20160010107)

What is a Search Engine?

A search engine is a service that allow users to search for content over a collection.

A user enters a query (keywords or key phrases) into a search engine and receives a list of results which can be relevant or irrelevant.



Ranking of Retrieved Documents

For a query, there can be large number of results. So we need to sort the results in decreasing order of their relevance.

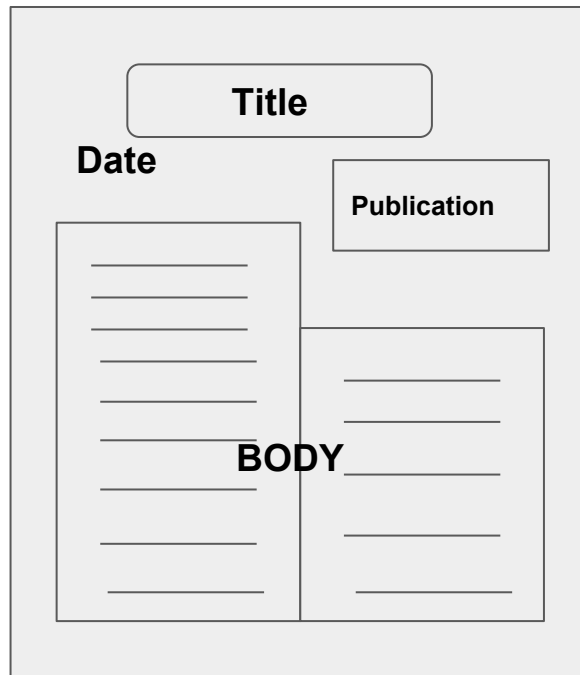
There are many algorithms which can be used to calculate relevance of a document for a given query like cosine similarity, tf-idf score based, BM25, BM25F, etc.

In this project we focus on BM25 (lucene's default ranking algorithm) and BM25F.

We then compare the performance of these ranking algorithms.

Dataset : All News ([Kaggle](https://www.kaggle.com/snapcrack/all-the-news))

File structure:



Each document in our collection is structured having 3 fields: **Date of Publication, Title and Body.**

BM25F Relevance Algorithm

BM25F which is an extension of the BM25 ranking function adapted to score structured documents.

BM25F performs per-field BM25 calculation, but uses the shared document frequency across multiple fields. So we assign different weights to different sections of the document. b controls to what degree document length normalizes

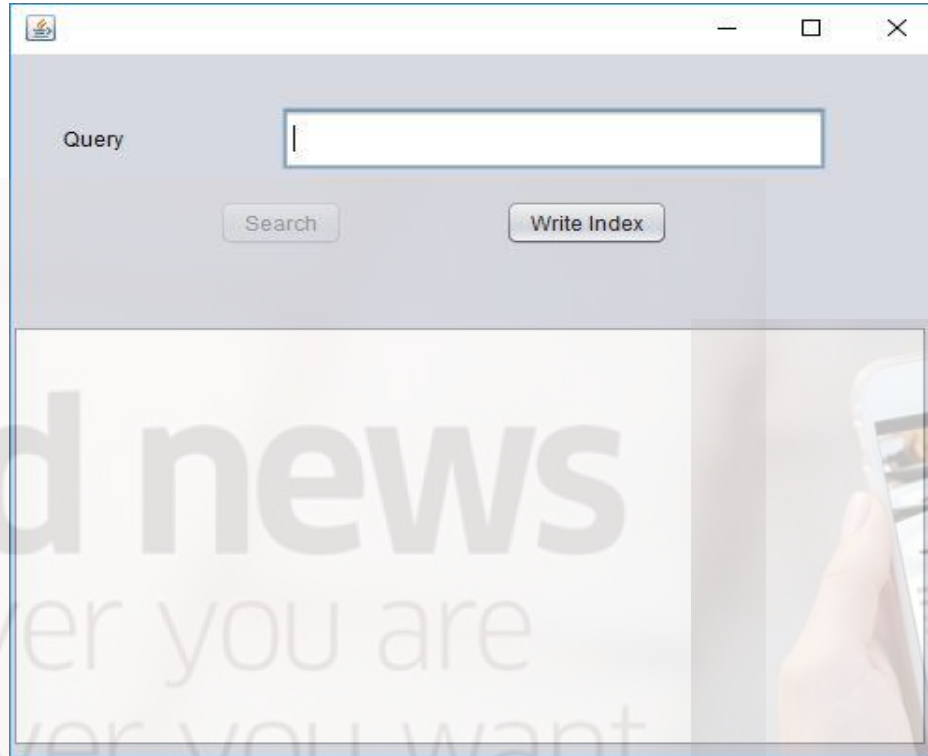
tf values. k_1 is a constant that allows us to control the non-linear growing term frequency function.

$$tf(t, d) = \sum_{c \in d} w_c * tf_c(t, d)$$

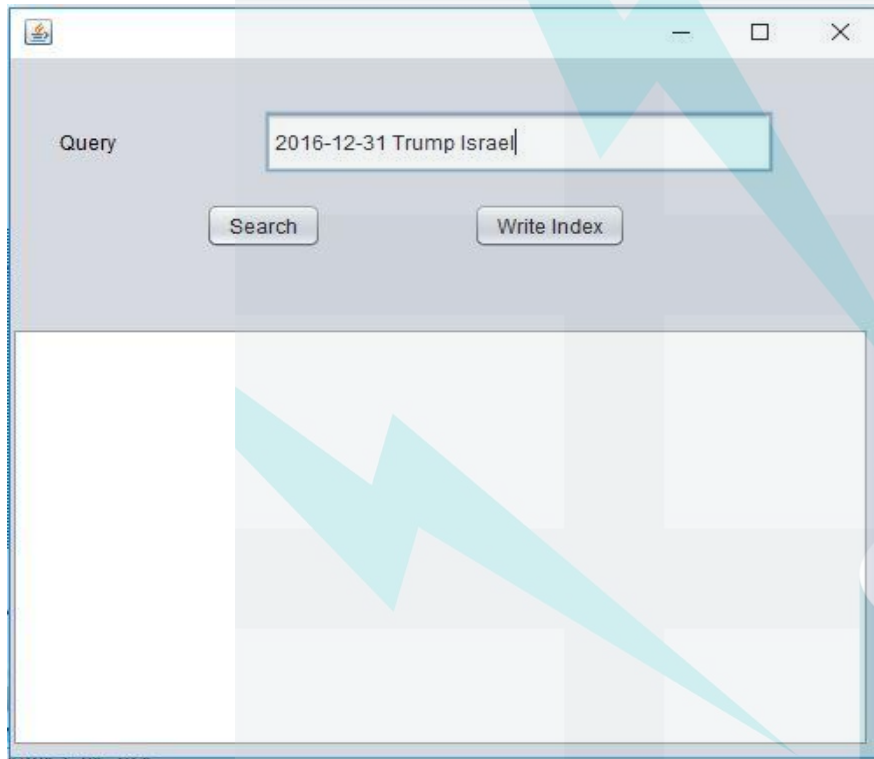
$$tf_c(t, d) = \frac{occurs_c(t, d)}{1 + b_c((\frac{l_{d,c}}{l_c} - 1))}$$

$$BM25F_D = \sum_{t \in q \cap d} \frac{tf(t, d)}{tf(t, d) + k_1} * idf(t)$$

Writing Index from Collection



Searching Query



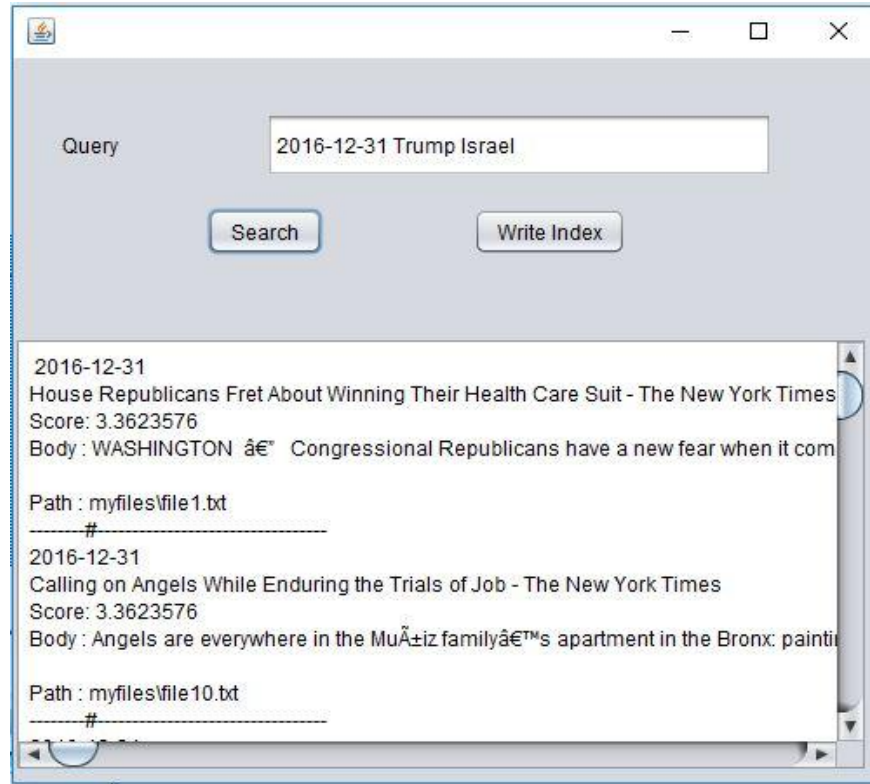
Query

2016-12-31 Trump Israel

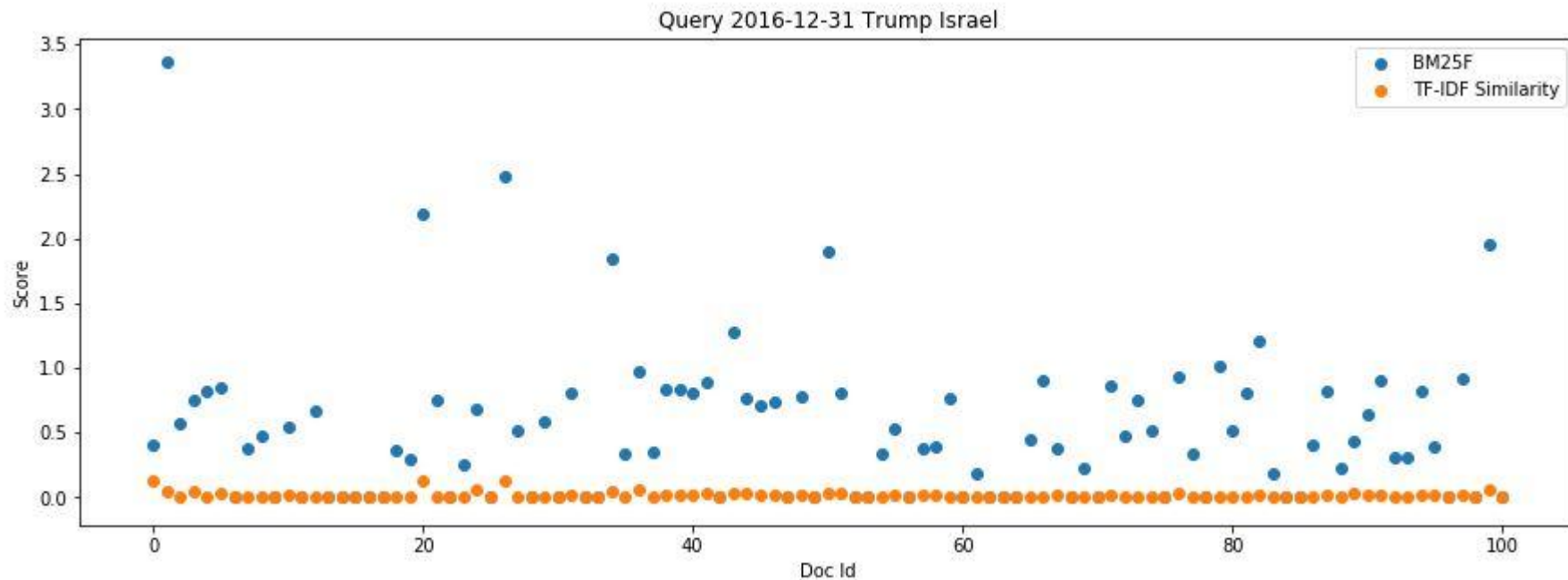
Search

Write Index

Retrieved Results



Comparison between BM25F and TF-IDF Similarity



Comparison between BM25 and BF25F Similarity

