

# Information Retrieval and Extraction

## Mini Project - Phase 2

### (Wikipedia Search Engine)

Name: Akshett Rai Jindal

Roll Number: 2019114001

## Directory Structure

```
.
├── index.sh          -- script to run indexer
├── LICENSE
├── README.pdf
├── README.wiki
├── requirements.txt  -- python requirements to run the program
├── search.sh         -- script to search from index
├── src
│   ├── indexer.py    -- code for indexing
│   └── searcher.py   -- code for searching
```

## Optimizations

- Removed stop words from the data
- Removed all the symbols from the data
- Stemmed all the tokens to get low number of unique tokens
- Used Base-64 encoding for numbers
- Removed unnecessary meta articles from index
- Removed all the tokens which were too small or too lengthy
- Created different index files for each field type to optimize field searches
- Rounded the IDF values to reduce the number of characters
- Used f-strings and join methods to reduce string concatenation time
- Minimum use of regex
- Split data into different numbered files with preindexes and offsets to make it faster to search and reduce memory usages
- Use heaps while merging (while indexing) and ranking (while searching) for faster sorting

## Benchmarking

**Note: The benchmarking has been done on dataset provided for phase 1**

- For indexing, the code takes 371 seconds on Quadcore Intel i5-7th gen PC
- The total size of the inverted index is : 2,38,200 bytes which is within the limit
- It contains 112 files with 15,85,174 total unique tokens
- For searching, with the provided queries, it takes the following times for searching:

S.No.	Queries	Time (seconds)
1	Billie Jean michael jackson	0.24301073600054224
2	b:Marc Spector i:Marvel Comics c:1980 comics debuts	0.17570688999967388
3	Allan Rune Pettersson	0.15535681000073964
4	b:angus r:tories i:James Lee	0.14201764199970057
5	uss conolly	0.12140283200005797
6	b:Speed of Light c:Concepts in Physics	0.15087084400056483
7	The Capture 1950	0.13463407199924404
8	t:birdseye b:tarantula	0.11396531499940465
9	4 July 1776	0.23054767000030552

## Index Format

S.No.	File Name	Description	Format
1	article_titles_*.txt	articles ids and their titles	{doc_id} {doc_title}
2	idf_*.txt	IDF values for all the tokens	{token} {idf}
3	index_b_*.txt	the documents ids and count for all the tokens in bodies of documents	{token} {doc1_id:count1} {doc2_id:count2} ...
4	index_c_*.txt	the documents ids and count for all the tokens in categories section of documents	{token} {doc1_id:count1} {doc2_id:count2} ...
5	index_i_*.txt	the documents ids and count for all the tokens in infobox of documents	{token} {doc1_id:count1} {doc2_id:count2} ...
6	index_l_*.txt	the documents ids and count for all the tokens in external links of documents	{token} {doc1_id:count1} {doc2_id:count2} ...
7	index_r_*.txt	the documents ids and count for all the tokens in references section of documents	{token} {doc1_id:count1} {doc2_id:count2} ...
8	index_t_*.txt	the documents ids and count for all the tokens in title of documents	{token} {doc1_id:count1} {doc2_id:count2} ...
9	offsets_{field_type}_*.txt	the offsets of each line in bytes for each field type and index file	{offset}
10	pre_index_{field_type}.txt	the first lines of each index file for each type	{line}
11	pre_index_idf.txt	the first lines of each idf file	{line}
12	pre_index_titles.txt	the first lines of each title	{line}

## Running the Code

1. For creating and merging the index:

```
$ bash index.sh <path_to_xml_dump_file> <path_of_directory_to_create_index_in> <path_of_stats_file>
```

2. For searching in the index:

```
$ bash search.sh <path_to_queries_file> <path_of_index_directory> <path_of_output_file>
```