CS242 Information Retrieval and Web Search

Project Report Part-B
Hadoop Indexing & Search Engine Interface

Aditi Lokhande
Department of Computer Science and Engineering
University of California, Riverside
Riverside, California, The US
alokh001@ucr.edu

Siddhant Purohit
Department of Computer Science and Engineering
University of California, Riverside
Riverside, California, The US
spuro001@ucr.edu

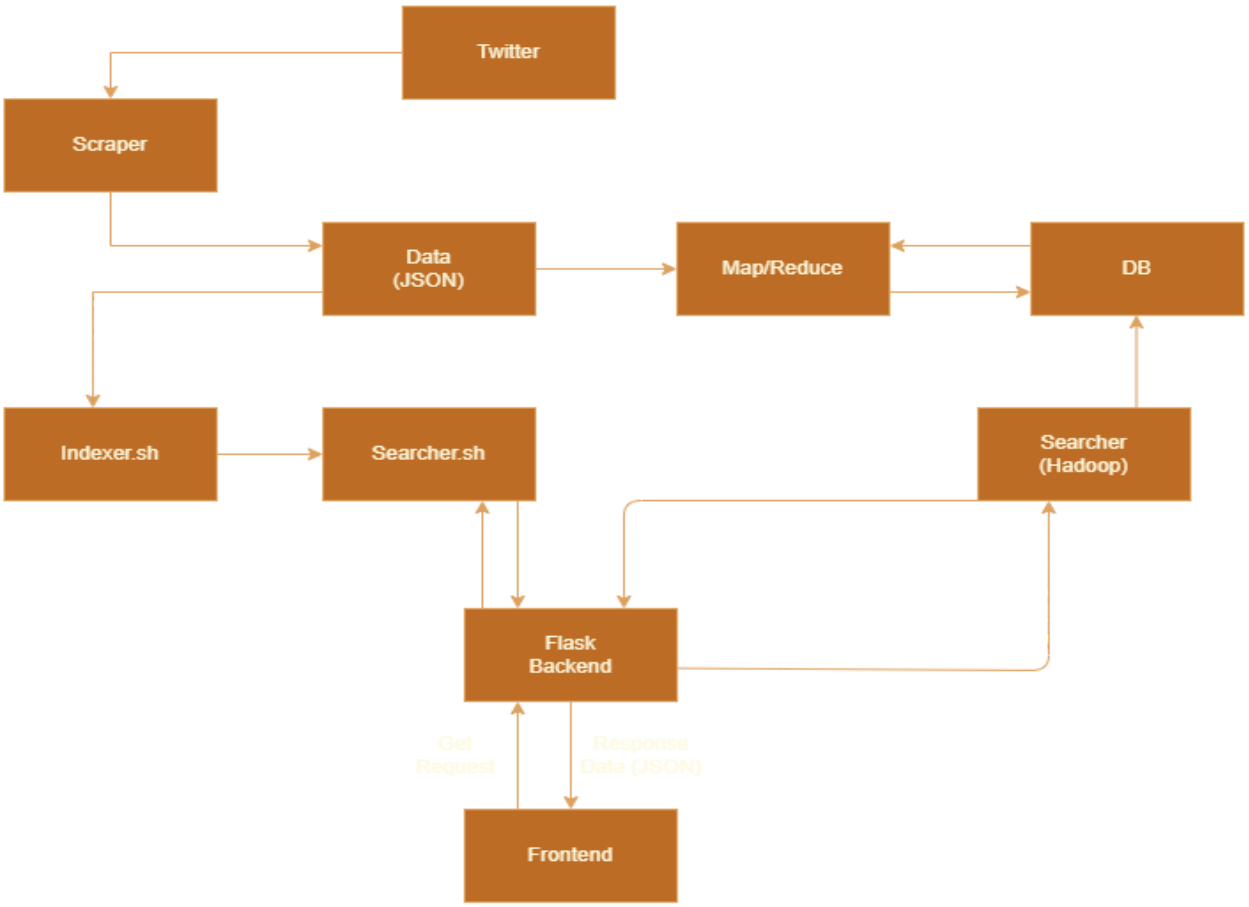
Bipin Dhoddamane Ravi
Department of Computer Science and Engineering
University of California, Riverside
Riverside, California, The US
bdhod001@ucr.edu

Utkarsh Neema
Department of Computer Science and Engineering
University of California, Riverside
Riverside, California, The US
uneem001@ucr.edu

1. Problem Statement

The objective of this project is to design a Search Engine using Web Crawlers, Lucene Indexing, Hadoop, and GUI design. In Part-A we go over the architecture, design & implementation of Web Crawler, Lucene Indexing of the collected data, and design a Query-Parser for the indexed data. In order to implement the above we choose to gather Geo-Tagged Twitter data along with other relevant fields. We aim to retrieve information from these tweets and plot it on a map to gain better insights from the data.

2. System Architecture



3. Hadoop Indexing

The Apache Hadoop software library is a framework that uses basic programming principles to enable the distributed processing of massive data volumes across clusters of machines. It's built to expand from a single server to thousands of devices, each with its own computation and storage capabilities. Rather than relying on hardware to provide high availability, the library is designed to identify and handle problems at the application layer, allowing a highly available service to be delivered on top of a cluster of computers that may all fail.

2.1. Indexing Strategy

Hadoop Indexing:

We have used Hadoop MapReduce Job on python in order to index collected data from our scraper. We have created an Inverse Index for tweets data against its id. We cleaned and pre-process our tweets data by removing stopwords, punctuations and converting them to lowercase terms. Next, Inverse Index only on the pre-processed data words which are sorted based on their occurrences in each tweet. This Index is stored in a database and can be used for effective and efficient retrieval by the Searcher. We have used tf.idf ranking algorithm and cosine similarity checker for each term query to search results.

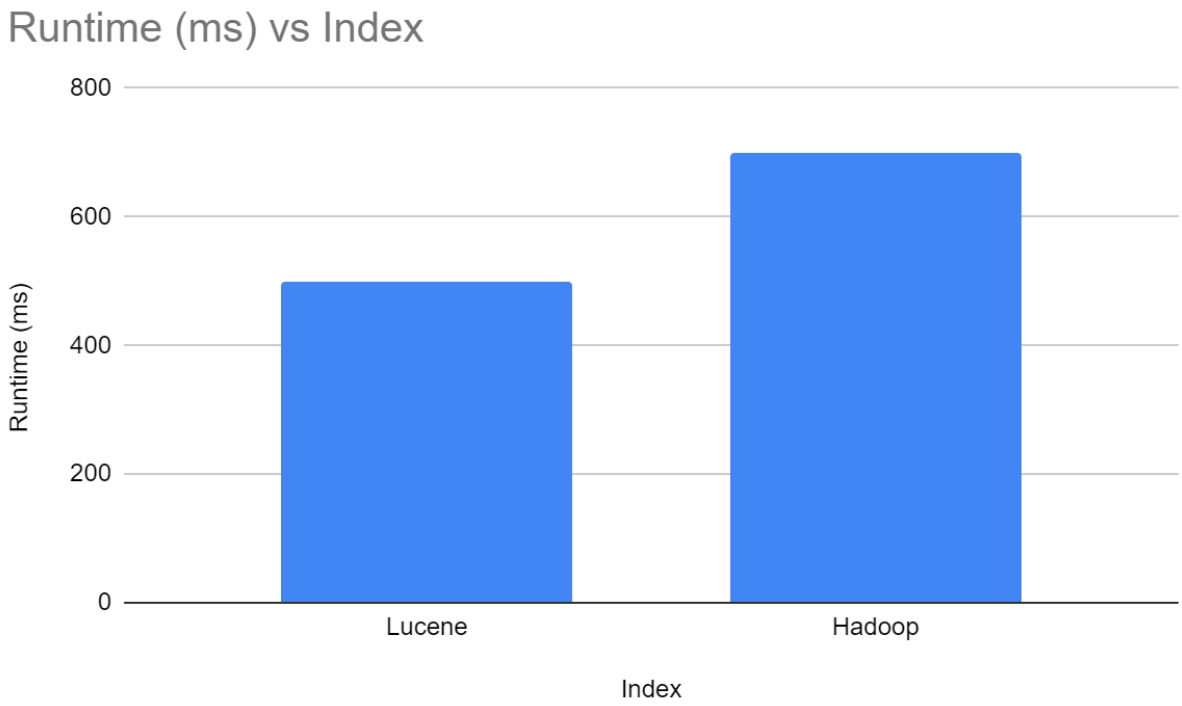
Lucene Indexing:

For lucene indexing we used the following fields:

1. Tweets: In order to search relevant tweets based on keywords.
2. User: To search a specific twitter user name and their tweets.
3. Hashtags: In order to search relevant tweets based on hashtags.
4. Geo Location: To search tweets around a specific geographical location.

We removed punctuations and stop words and processed the data of any stop words befor passing it to the query parser.

4. **Run Time**



Observations:

We observe that obtaining search query results using Lucene index takes lesser time than that of Hadoop index. Lucene searches through the text faster. Since hadoop is running on a single machine it cannot parallelize hence it is slower.

5. **Limitations**

- Scope can be increased: As of now we are only considering major cities in the US.
- Safe Search Filter can be added on further analysis of the tweets.

6. **Obstacles**

The major obstacles we faced during this phase of the project are:

- Running the Search Engine on local machine
- Lucene setup is difficult on local machine
- Setting up server is not possible on the assigned machine
- Data cannot be directly transferred between local and assigned machine
- The package mrjob is not installed and it is difficult to add it with the limited permissions on the assigned machine/bolt server
- Setting up Java Environment and path files is difficult to achieve with the limited permissions on the assigned machine/bolt server
- Debugging on the assigned machine is difficult due to the limited editing options on the terminal

7. Instructions to Deploy the System

A. Instructions to build Lucene Index

➢ Since simpleJson package is missing in the existing JRE, use the following code before running the executable files:

```
export  JSON_JAVA=/opt/home/cs242-w22/    #(path to simple json)
export CLASSPATH=$CLASSPATH:$JSON_JAVA/json-simple-1.1.1.jar:.
```

➢ We have created two separate .sh files:

1. Indexer.sh - Executable file for indexer, takes index directory and data file as command line arguments.

2. Searcher.sh - Executable file for searcher, takes query term as a command line argument.

B. **Instructions to build Hadoop Index**

1) Run MRJob.py on the scraper data to get Indexed File

2) Run Searcher_MR.py with the Indexed file and data File as parameters3) Enter Search query on command Line

C. Instructions to Run flask server

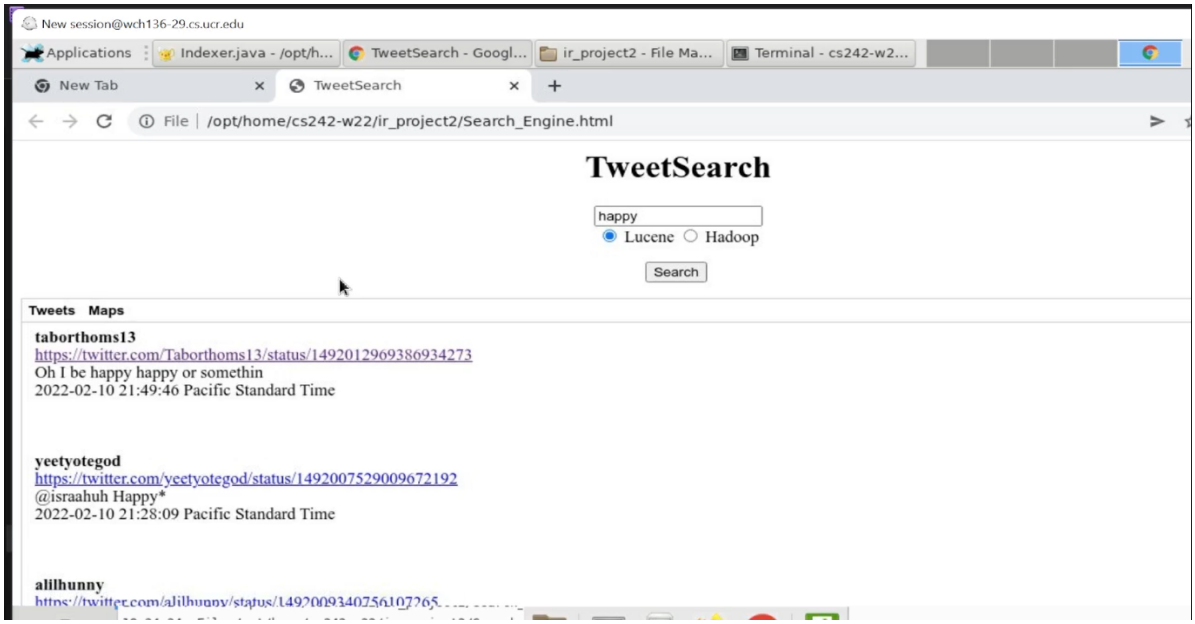
To run the backend flask server use the following command:

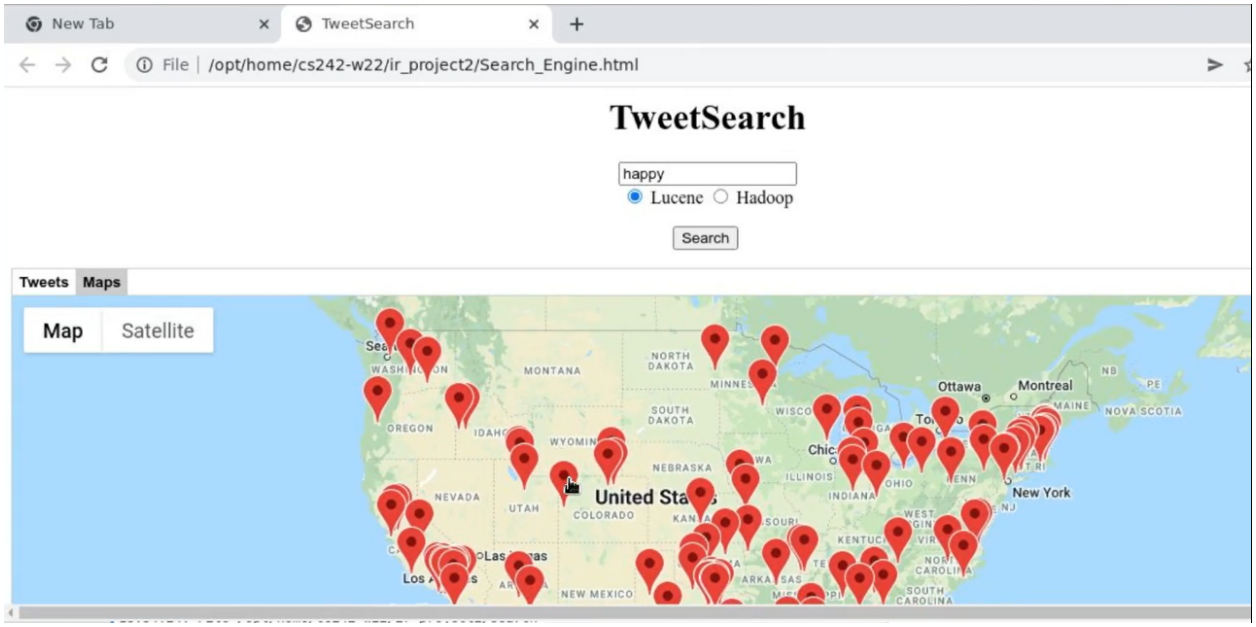
flask run -h localhost -p 5002

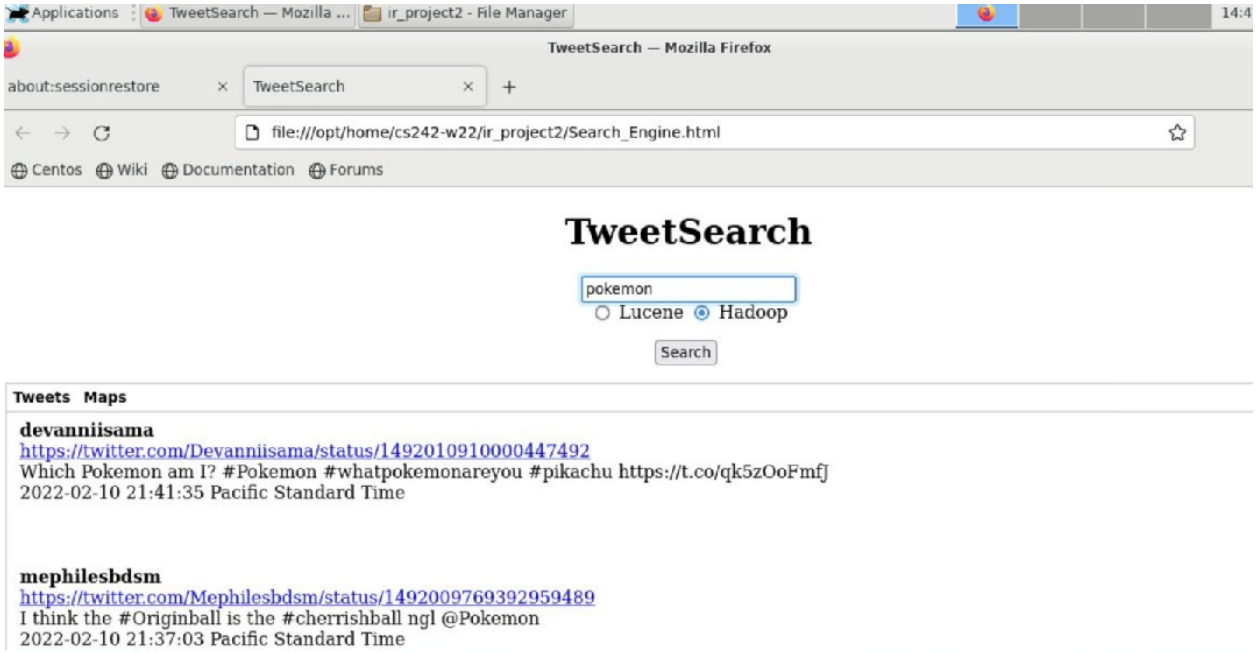
To run the frontend run the html file 'Search_Engine.html' in any browser.

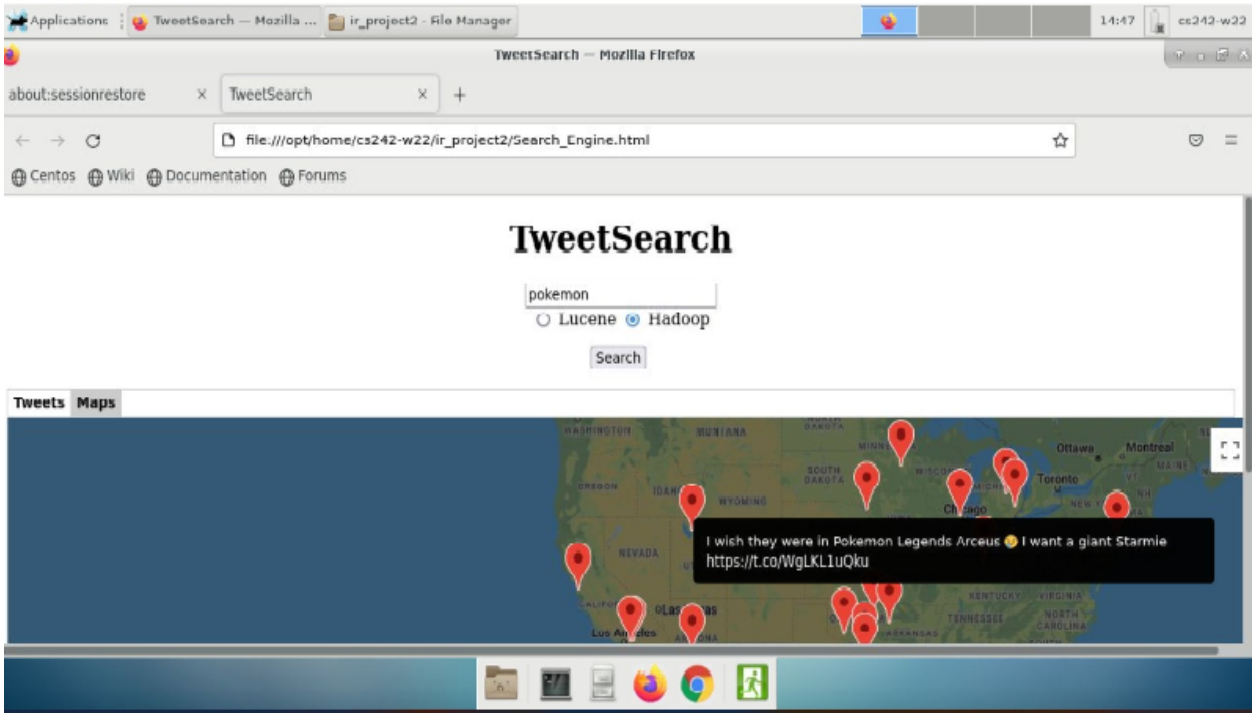
8. Output Screenshots

Below is the Screenshot of the output that is obtained on searching the keyword "happy" after indexing. Also we have plotted the tweet location on a map.


New session@wch136-29.cs.ucr.edu
Applications
Indexer.java - /opt/h...
TweetSearch - Googl...
ir_project2 - File Ma...
Terminal - cs242-w2...
New Tab
TweetSearch
File | /opt/home/cs242-w22/ir_project2/Search_Engine.html
TweetSearch
happy
Lucene
Hadoop
Search
Tweets
Maps
taborthoms13
https://twitter.com/Taborthoms13/status/1492012969386934273
Oh I be happy happy or somethin
2022-02-10 21:49:46 Pacific Standard Time
yeetyotegod
https://twitter.com/yeetyotegod/status/1492007529009672192
@israahuh Happy*
2022-02-10 21:28:09 Pacific Standard Time
alilhunny


New Tab
TweetSearch
File | /opt/home/cs242-w22/ir_project2/Search_Engine.html
TweetSearch
happy
Lucene
Hadoop
Search
Tweets
Maps
Map
Satellite
MONTANA
NORTH DAKOTA
SOUTH DAKOTA
OREGON
WYOMING
NEBRASKA
NEVADA
UTAH
COLORADO
NEW MEXICO
Ottawa
Montreal
New York


Applications
TweetSearch — Mozilla ...
ir_project2 - File Manager
14:4
TweetSearch — Mozilla Firefox
about:sessionrestore
TweetSearch
file:///opt/home/cs242-w22/ir_project2/Search_Engine.html
Centos
Wiki
Documentation
Forums
TweetSearch
pokemon
Lucene
Hadoop
Search
Tweets
Maps
devanniisama
https://twitter.com/Devanniisama/status/1492010910000447492
Which Pokemon am I? #Pokemon #whatpokemonareyou #pikachu https://t.co/qk5zOoFmfJ
2022-02-10 21:41:35 Pacific Standard Time
mephilesbdsm
https://twitter.com/Mephilesbdsm/status/1492009769392959489
I think the #Originball is the #cherrishball ngl @Pokemon
2022-02-10 21:37:03 Pacific Standard Time


Applications
TweetSearch — Mozilla ...
ir_project2 - File Manager
14:47
cs242-w22
TweetSearch — Mozilla Firefox
about:sessionrestore
TweetSearch
file:///opt/home/cs242-w22/ir_project2/Search_Engine.html
Centos
Wiki
Documentation
Forums
TweetSearch
pokemon
Lucene
Hadoop
Search
Tweets
Maps
Ottawa
Montreal
Toronto
I wish they were in Pokemon Legends Arceus I want a giant Starmie
https://t.co/WgLKL1uQku

9. Collaboration Detail

Team Member	Contribution
Aditi Lokhande	Frontend, Documentation, Presentation
Bipin Dhoddamane Ravi	Backend, Indexing, Documentation
Siddhant Dushyant Purohit	Hadoop Indexing, Presentation
Utkarsh Neema	Backend, Presentation, Documentation