CS172 Project: Index and Web Search Project Report

Team Members - Jackie Chan (861158919)
Abe Hu (861148832)
Daniel Nguyen (861156246)

Collaboration Details:

Jackie Chan:

- Implemented indexing and search functions

Abe Hu:

- Created the backend Server endpoints
- Created frontend user interface

Daniel Nguyen:

- Implemented indexing and search functions
- Complete 4-5 page report

Overall Collaboration Splits:

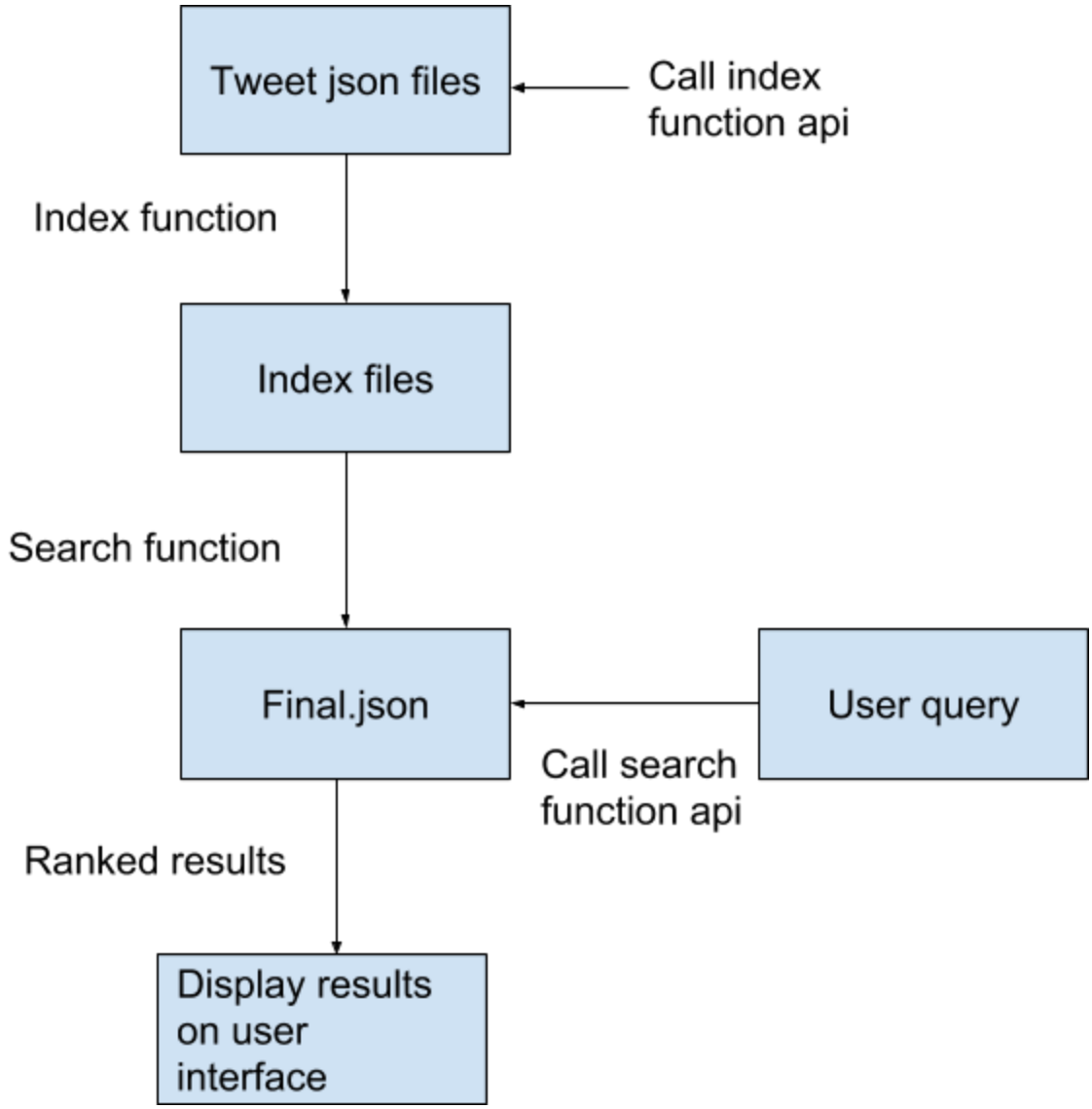
Jackie Chan - 33.33%

Abe Hu - 33.33%

Daniel Nguyen - 33.33%

Overview of systems:

a. Architecture:



b. Data indexing and searching:

We used pylucene to iterate through every json file in the data directory and index them. The fields that we included in the index were username, text, time, and location. For the search, we used the standard analyzer, to do a very simple search. After inputting a query, the function will return the top 10 results.

c. Data Structures and Architecture Explained:

For the frontend user interface we chose to use react and redux. Bootstrap was used for styling the DOM elements.

The server is designed to run continuously.We created two endpoints: one for indexing, and one for searching. The index route is called automatically every time the server runs. It will only need to be run once since the index files will be saved to storage after the first call. Entering a query and pressing the submit button will take the search endpoint and run the function.

The index function grabs each line and writes it to a document with specific fields. The new created files will be stored in a directory called lucene_index.

The search function scores each element based on the text's relevance to the query. Once it has completed, it will return the top ten results as a json object to the front end. The client side user interface will then display the user, text, time, location, and other information.

Limitations of Systems:

- We did not implement stemming
- Limited in the user relevance of results, due to the fact that we only used the standard analyzer
- Did not implement the extra credit gps coordinate portion

Instructions for system deployment:

1. Install python
2. Install pip (python package manager)
3. Install all python packages
    a. pylucene
    b. Flask
    c. Requests
    d. React
    e. Redux
4. To start the backend
    a. Npm start
    b. Run React app
    c. Go to src/public directory
    d. To run flask input - `FLASK_APP=server.py flask run`
    e. Input query

System in Action:

Index function

```
Current file: tweets32.json
Current file: tweets24.json
Current file: tweets12.json
Current file: tweets28.json
Current file: tweets288.json
Current file: tweets68.json
Current file: tweets44.json
Current file: tweets52.json
Current file: tweets72.json
Current file: tweets292.json
Current file: tweets284.json
Current file: tweets64.json
Current file: tweets48.json
Current file: tweets124.json
Current file: tweets132.json
Current file: tweets80.json
Current file: tweets260.json
Current file: tweets276.json
Current file: tweets108.json
Current file: tweets96.json
Current file: tweets256.json
Current file: tweets128.json
Current file: tweets240.json
Current file: tweets112.json
Current file: tweets0.json
Current file: tweets104.json
Current file: tweets152.json
Current file: tweets144.json
Current file: tweets328.json
Current file: tweets168.json
Current file: tweets304.json
Current file: tweets216.json
Current file: tweets200.json
Current file: tweets312.json
Current file: tweets220.json
Current file: tweets332.json
Current file: tweets324.json
Current file: tweets148.json
Current file: tweets236.json
Current file: tweets308.json
Current file: tweets164.json
Current file: tweets172.json

Number of indexed documents: 232592
Finished

------------------------------------------------------
```

```
Enter a blank line to quit.
Searching for: Information retrieval
10 total matching documents in 0:00:00.002219:

{'username': u'saraasalimi', 'text': u'Information not given accurately and completely.'
'score': 11.135669708251953, 'time': u'Thu Jun 07 02:53:27 +0000 2018'}


{'username': u'disasteraware', 'text': u'New Alert: Tsunami Information (Hawaiian Island
 - Off The Hamakua Coast Of The Big Island - 4.1, Severity: INFORMATION', 'score': 9.961
0932922363, 'time': u'Wed Jun 06 20:40:22 +0000 2018'}


{'username': u'andrewshaylie', 'text': u'If anyone would like to give blood, HMU for the
nformation !!!\u2764\ufe0f\U0001f489\u2764\ufe0f', 'score': 9.577493667602539, 'time': u
hu Jun 07 03:06:08 +0000 2018'}


{'username': u'LazyEyezz', 'text': u'@Solo_Kalin Aye, u not tellin no false information
l', 'score': 9.577493667602539, 'time': u'Thu Jun 07 03:49:02 +0000 2018'}


{'username': u'wirtzbill', 'text': u"@ijclark All the information is in the article you
dn't read.", 'score': 9.577493667602539, 'time': u'Thu Jun 07 03:57:53 +0000 2018'}


{'username': u'dipeshkhadye', 'text': u'#vtasindia2018 new data classification - Veritas
nformation Classifier', 'score': 9.577493667602539, 'time': u'Thu Jun 07 05:57:11 +0000
18'}


{'username': u'JohnDiiorio4', 'text': u'@renato_mariotti @realDonaldTrump I suppose you
ve access to government information', 'score': 8.39596176147461, 'time': u'Wed Jun 06 21
9:52 +0000 2018'}


{'username': u'richrad', 'text': u'there is some information perhaps i\u2019d rather not
ave https://t.co/vvJGlvFQVZ', 'score': 8.39596176147461, 'time': u'Thu Jun 07 04:06:45 +
00 2018'}


{'username': u'_StormMae', 'text': u'@CoachT_SFU For your information, I got salted cara
l. https://t.co/LWAARFcjw9', 'score': 8.39596176147461, 'time': u'Thu Jun 07 05:02:36 +0
0 2018'}


{'username': u'SEASHORE46', 'text': u'@markgstrong @RealDonaldTrFan Where did you get yo
 information from playboy', 'score': 8.39596176147461, 'time': u'Wed Jun 06 22:13:53 +00
 2018'}
```