

Final project report

Andrew ID: haichuax

0.How to run these applications

Step 1: Change external IP: 34.106.64.114 (it will change every time I restart the clusters) in:

sample\src\pages\landing.js

sample\src\component\TopN.js

sample\src\component\MiniSearchClient.js

Example:

← hw3-m EDIT RESET + CREATE MACHINE IMAGE + CREATE SIMILAR ⋮ OPERATIONS ▾

DETAILS OBSERVABILITY OS INFO SCREENSHOT

Network tags

dataproc-notebook-vm

Network interfaces

| Name ↑ | Network | Subnetwork | Primary internal IP address | Alias IP ranges | Stack Type | External IP address | Netw |
|--------|---------|------------|-----------------------------|-----------------|------------|------------------------------|-------|
| nic0 | default | default | 10.180.0.3 | | IPv4 | 34.106.64.114 (Ephemeral) | Prent |

```
axios.post("http://34.106.64.114:5000/upload", formData)
```

Step2: Start the first application on your PC:

```
cd sample
```

```
docker run -it --rm -v %cd%:/app -v /app/node_modules -p 3001:3000 -e
```

```
CHOKIDAR_USEPOLLING=true tonyrays/dockerhub:projtimagepush2
```

Step3: Start the second application on GCP:

Connect to compute engine via SSH, move folder flask_second_app to the compute engine.

```
pip install Flask
```

```
pip install flask_cors
```

```
python app.py
```

Note1 If TCP connection is blocked:

Create a firewall rule to allow traffic to port 5000:

Firewall + CREATE FIREWALL POLICY + CREATE FIREWALL RULE

Notifications

✓ Create firewall rule "allow-flask-port-5000" Just now
new pro

✓ Start VM instance "hw3-w-1" 1 hour ago

Note2 Data folder:

Create a 'Data' folder and insert the necessary data into it. Then, compress the folder into a zip file for uploading. Alternatively, you can use the zip file provided in the extra credit quiz.("./test_data/Data.zip")

1. Brief introduction to the first application

The first application contains four frontend pages:

- (1) Landing.js: where you can upload a zip file to construct Inverted Indices
- (2) MiniSearchIndex.js: index page, proceed to (3) or (4), or go back to (1)
- (3) MiniSearchClient.js: Search For Term
- (4) Top_N.js: TOP-N Frequent Terms
- (5) dropdown.js: implement a dropdown for better user experience
- (6) Docker image on dockerhub

```
sample > Dockerfile > ...
You, 2 months ago | 1 author (You)
1 # pull official base image
2 FROM node:13.12.0-alpine
3
4 # set working directory
5 WORKDIR /app
6
7 # add `/app/node_modules/.bin` to $PATH
8 ENV PATH /app/node_modules/.bin:$PATH
9
10 # install app dependencies
11 COPY package.json ./
12 COPY package-lock.json ./
13 RUN npm install --silent
14 RUN npm install react-scripts@3.4.1 -g --silent
15 RUN npm install axios
16 RUN npm install @mui/material @emotion/react @emotion/styled
17 RUN npm install react-router-dom
18 RUN npm install react-bootstrap
19 # add app
20 COPY . ./
21
22 # start app
23 CMD ["npm", "start"]
```

Watch the video for more details.

2. How I create the second application step by step

Since I have completed extra credit quiz, so I just modify the mapper.py and reducer.py I used for extra credits.

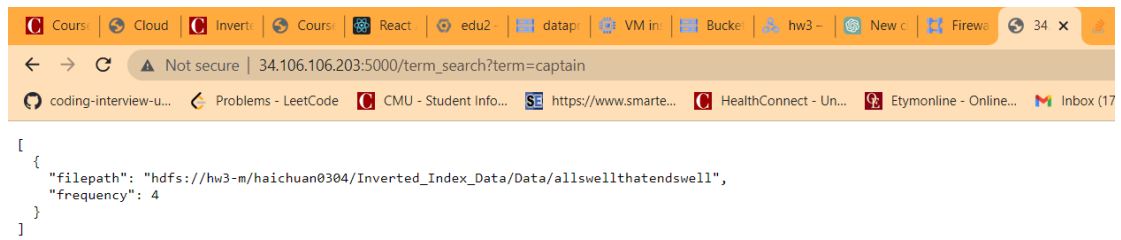
The second application contains:

- (1) App.py
- (2) Mapper_q.py
- (3) Reducer_q.py
- (4) Term_search.py
- (5) Top_n_search.py

2.1 Test manually generate json file:

```
hadoop jar /usr/lib/hadoop/hadoop-streaming.jar -file mapper_q.py -mapper 'python mapper_q.py' -
file reducer_q.py -reducer 'python reducer_q.py' -input /haichuan0304/Inverted_Index_Data/Data/ -
output /haichuan0304/output_inverted_final6
```

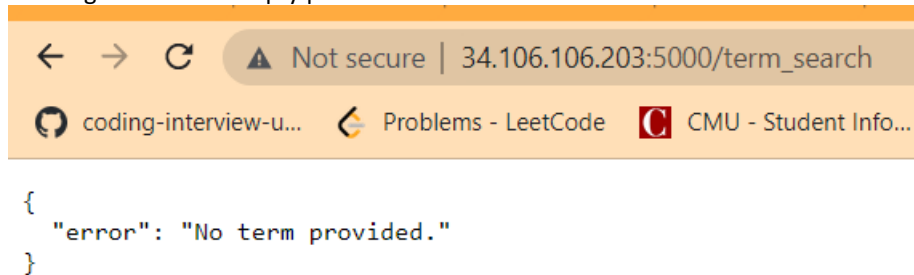
```
hadoop fs -getmerge /haichuan0304/output_inverted_final6 ./flask/inverted_index.json
```



A screenshot of a web browser window. The address bar shows the URL `34.106.106.203:5000/term_search?term=captain`. The browser tabs include 'Cours', 'Cloud', 'Invert', 'Cours', 'React', 'edu2', 'datap', 'VM in', 'Bucke', 'hw3', 'New c', and 'Firew'. The page content displays a JSON object:

```
[ { "filepath": "hdfs://hw3-m/haichuan0304/Inverted_Index_Data/Data/allswellthatendswell", "frequency": 4 } ]
```

I also handle edge cases like 'empty params':



A screenshot of a web browser window. The address bar shows the URL `34.106.106.203:5000/term_search`. The browser tabs include 'coding-interview-u...', 'Problems - LeetCode', and 'CMU - Student Info...'. The page content displays a JSON object:


```
{ "error": "No term provided." }
```

2.2 Test react app:

`docker run -it --rm -v %cd%:/app -v /app/node_modules -p 3001:3000 -e CHOKIDAR_USEPOLLING=true tonyrays/dockerhub:projimagepush2`

localhost:3001/miniSearchIndex/miniSearchClient

coding-interview-u... Problems - LeetCode CMU - Student Info... https://www.smarte... HealthConnect



Mini Search Engine: Search For Term

[Back to miniSearchIndex](#) show per page


| Id | filepath | frequency |
|----|---|-----------|
| 1 | hdfs://hw3-m/haichuan0304/Inverted_Index_Data/Data/cymbeline | 15 |
| 2 | hdfs://hw3-m/haichuan0304/Inverted_Index_Data/Data/3kinghenryvi | 3 |
| 3 | hdfs://hw3-m/haichuan0304/Inverted_Index_Data/Data/1kinghenryiv | 11 |
| 4 | hdfs://hw3-m/haichuan0304/Inverted_Index_Data/Data/2kinghenryiv | 18 |
| 5 | hdfs://hw3-m/haichuan0304/Inverted_Index_Data/Data/asyoulikeit | 16 |
| 6 | hdfs://hw3-m/haichuan0304/Inverted_Index_Data/Data/allswellthatendswell | 13 |
| 7 | hdfs://hw3-m/haichuan0304/Inverted_Index_Data/Data/comedyoferrors | 11 |

Also test out top_n_search:

CoursCloudInvertCoursRe: xedu2-datapVM in:Buckethw3

localhost:3001/miniSearchIndex/topN

coding-interview-u...Problems - LeetCodeCMU - Student Info...https://www.smarte...HealthConnect -



Mini Search Engine: TOP-N Frequent Terms

Back to miniSearchIndex

Choose your N value: 10

Search

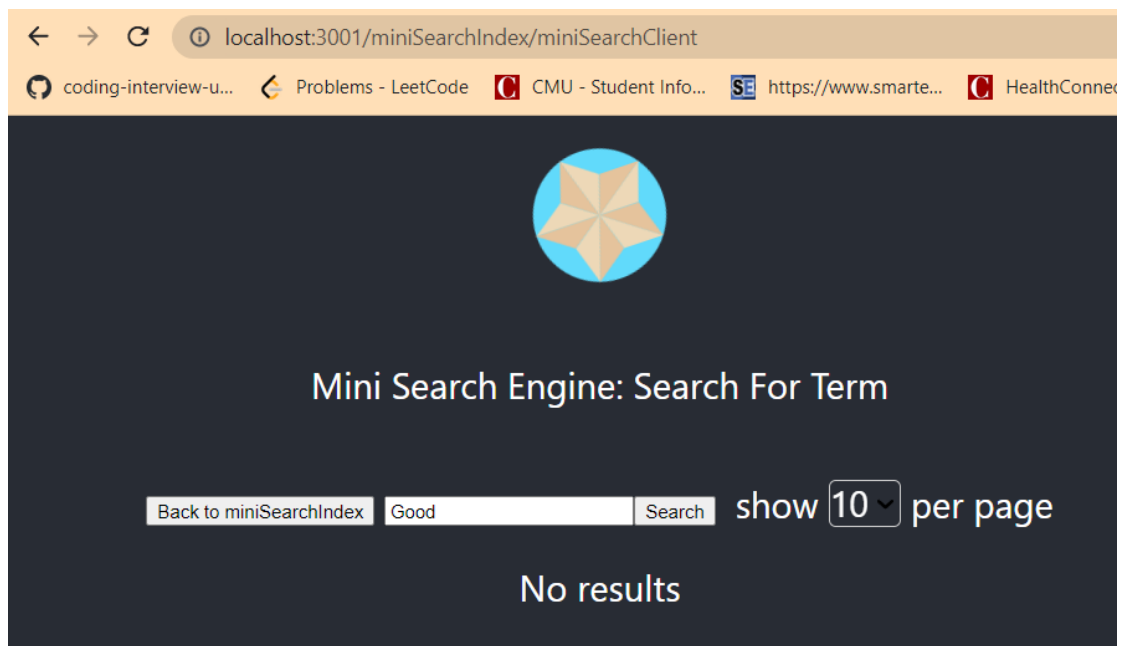
| Id | Term | Frequency |
|----|------|-----------|
| 1 | the | 4818 |
| 2 | I | 4131 |
| 3 | and | 3717 |
| 4 | of | 3114 |
| 5 | to | 3039 |
| 6 | a | 2642 |
| 7 | my | 2145 |
| 8 | in | 1893 |

2.3 Add stop word list

```
JS TopN.js M JS MiniSearchClient.js M mapper_q.py X reducer_q.py
C: > Users > haich > Downloads > quiz_inverted > mapper_q.py > ...
1  #!/usr/bin/env python
2
3  import sys
4  import os
5  import re
6
7  # stop words list
8  stop_words = set(["a", "an", "and", "are", "as", "at", "be", "by", "
9
10 # input comes from standard input (stdin)
11 for line in sys.stdin:
12     # remove leading and trailing whitespace
13     line = line.strip()
```

2.4 Regenerate json file with stop words:

```
hadoop jar /usr/lib/hadoop/hadoop-streaming.jar -file mapper_q.py -mapper 'python mapper_q.py' -
file reducer_q.py -reducer 'python reducer_q.py' -input /haichuan0304/Inverted_Index_Data/Data/ -
output /haichuan0304/output_inverted_final7
hadoop fs -getmerge /haichuan0304/output_inverted_final7 ./inverted_index.json
```



localhost:3001/miniSearchIndex/miniSearchClient

coding-interview-u...


Problems - LeetCode

CMU - Student Info...

SE

https://www.smarte...

HealthCon



Mini Search Engine: Search For Term

Back to miniSearchIndex

monkey

Search

show 10 per page

| Id | filepath | frequency |
|----|---|-----------|
| 1 | hdfs://hw3-m/haichuan0304/Inverted_Index_Data/Data/2kinghenryiv | 1 |
| 2 | hdfs://hw3-m/haichuan0304/Inverted_Index_Data/Data/asyoulikeit | 1 |

← → ↻ ⓘ localhost:3001/miniSearchIndex/topN

coding-interview-u... Problems - LeetCode CMU - Student Info... SE https://www.smarte... Health

Mini Search Engine: TOP-N Frequent Terms

[Back to miniSearchIndex](#) Choose your N value:

| Id | Term | Frequency |
|----|------|-----------|
| 1 | i | 4738 |
| 2 | you | 2789 |
| 3 | my | 2442 |
| 4 | not | 1649 |
| 5 | me | 1601 |
| 6 | s | 1474 |
| 7 | his | 1384 |
| 8 | your | 1278 |
| 9 | but | 1277 |
| 10 | this | 1222 |

The top-N results are not satisfactory; therefore, it is necessary to include additional stop-words.

2.5 Regenerate json file with even more stop words:

Here I used stop words from library nltk.

```
import nltk
from nltk.corpus import stopwords
print(stopwords.words('english'))
```

```
{'ourselves', 'hers', 'between', 'yourself', 'but', 'again', 'there', 'about', 'once', 'during', 'out', 'very',
'having', 'with', 'they', 'own', 'an', 'be', 'some', 'for', 'do', 'its', 'yours', 'such', 'into', 'of', 'most', 'itself',
'other', 'off', 'is', 's', 'am', 'or', 'who', 'as', 'from', 'him', 'each', 'the', 'themselves', 'until', 'below', 'are',
'we', 'these', 'your', 'his', 'through', 'don', 'nor', 'me', 'were', 'her', 'more', 'himself', 'this', 'down',
'should', 'our', 'their', 'while', 'above', 'both', 'up', 'to', 'ours', 'had', 'she', 'all', 'no', 'when', 'at', 'any',
'before', 'them', 'same', 'and', 'been', 'have', 'in', 'will', 'on', 'does', 'yourselves', 'then', 'that',
'because', 'what', 'over', 'why', 'so', 'can', 'did', 'not', 'now', 'under', 'he', 'you', 'herself', 'has', 'just',
'where', 'too', 'only', 'myself', 'which', 'those', 'i', 'after', 'few', 'whom', 't', 'being', 'if', 'theirs', 'my',
'against', 'a', 'by', 'doing', 'it', 'how', 'further', 'was', 'here', 'than'}
```


localhost:3001/miniSearchIndex/topN

Mini Search Engine: TOP-N Frequent Terms

[Back to miniSearchIndex](#) Choose your N value:

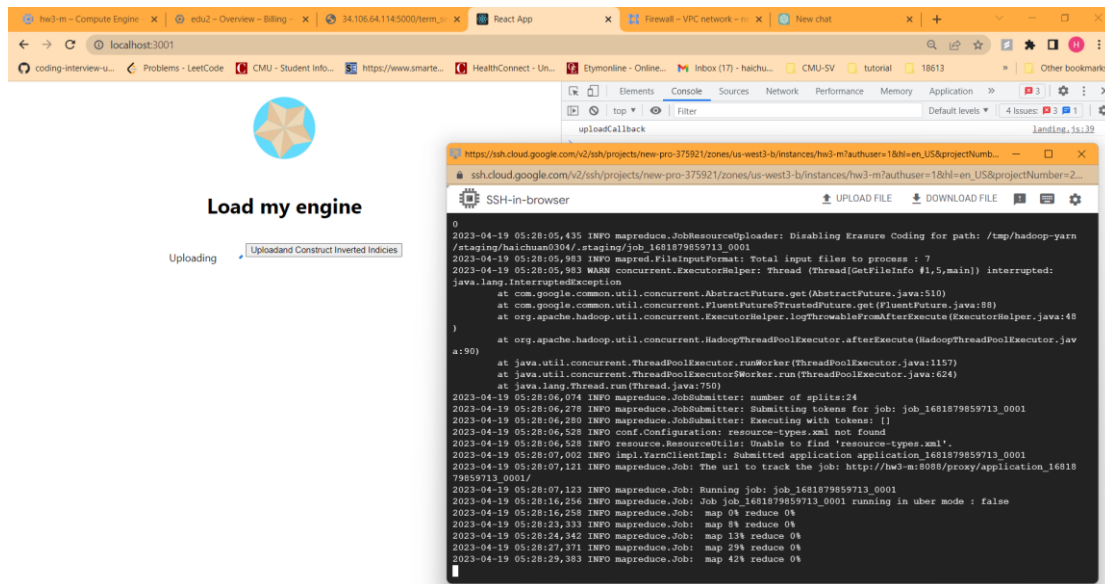
| Id | Term | Frequency |
|----|-------|-----------|
| 1 | thou | 1190 |
| 2 | d | 976 |
| 3 | king | 858 |
| 4 | lord | 788 |
| 5 | thy | 740 |
| 6 | shall | 664 |
| 7 | sir | 629 |
| 8 | thee | 627 |
| 9 | good | 591 |
| 10 | henry | 587 |

Looks much better now!

2.6 Allow user to upload a zip file and automatically generate json file.

To prevent auto reload, disable use_reloader.

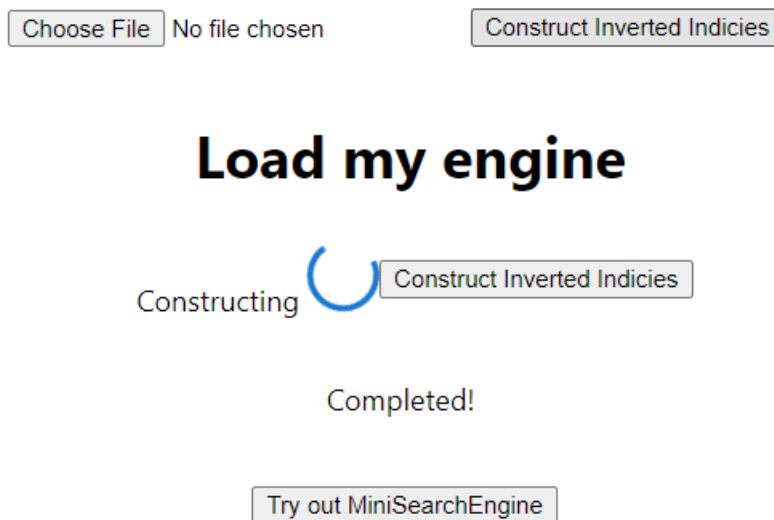
```
app.run(use_reloader=False, debug=True, host='0.0.0.0')
```



Delete 'input' and 'output' folder on HDFS to maintain a tidy environment.

```
2023-04-19 05:28:54,664 INFO streaming.StreamJob: Output directory: /haichuan0304/output
Deleted /haichuan0304/input
Deleted /haichuan0304/output
```

Rename the button:




2.7 Double check:

Delete inverted_index.json

```
haichuan0304@hw3-m:~/flask$ rm inverted_index.json
haichuan0304@hw3-m:~/flask$ ls
__pycache__  mapper_q.py  reducer_q.py  top_n_search.py
app.py       reducer.py.save  term_search.py  uploaded_files
```

localhost:3001

Problems - LeetCode CMU - Student Info... https://www.sma



Load my engine

Completed!

Try out MiniSearchEngine

SSH-in-browser

UPLOAD FILE DOWNLOAD

```
Reduce input records=93637
Reduce output records=7
Spilled Records=187274
Shuffled Maps =168
Failed Shuffles=0
Merged Map outputs=168
GC time elapsed (ms)=4695
CPU time spent (ms)=41080
Physical memory (bytes) snapshot=16665645056
Virtual memory (bytes) snapshot=144193667072
Total committed heap usage (bytes)=15605432320
Peak Map Physical memory (bytes)=639029248
Peak Map Virtual memory (bytes)=4655927296
Peak Reduce Physical memory (bytes)=345800704
Peak Reduce Virtual memory (bytes)=4670955520

Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0

File Input Format Counters
  Bytes Read=1035429
File Output Format Counters
  Bytes Written=1458601
2023-04-19 06:04:24,582 INFO streaming.StreamJob: Output directory: /haichuan0304/output
Deleted /haichuan0304/input
Deleted /haichuan0304/output
76.253.0.210 - - [19/Apr/2023 06:04:35] "POST /upload HTTP/1.1" 200 -
```

Works well:

localhost:3001/miniSearchIndex/topN

coding-interview-u... Problems - LeetCode CMU - Student Info... https://www.smar

SSH-in-browser

UPLOAD FILE

```
Reduce output records=7
Spilled Records=187274
Shuffled Maps =168
Failed Shuffles=0
Merged Map outputs=168
GC time elapsed (ms)=4574
CPU time spent (ms)=41100
Physical memory (bytes) snapshot=16623788032
Virtual memory (bytes) snapshot=144220295168
Total committed heap usage (bytes)=15588655104
Peak Map Physical memory (bytes)=636260352
Peak Map Virtual memory (bytes)=4675354624
Peak Reduce Physical memory (bytes)=347570176
Peak Reduce Virtual memory (bytes)=4663779328

Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0

File Input Format Counters
  Bytes Read=1035429
File Output Format Counters
  Bytes Written=1458601
2023-04-19 06:10:45,040 INFO streaming.StreamJob: Output directory: /haichuan0304/
Deleted /haichuan0304/input
Deleted /haichuan0304/output
76.253.0.210 - - [19/Apr/2023 06:10:55] "POST /upload HTTP/1.1" 200 -
76.253.0.210 - - [19/Apr/2023 06:11:41] "GET /top_n_search?n=10 HTTP/1.1" 200 -
```

Mini Search Engine: TOP-N Frequent Terms

Back to miniSearchIndex Choose your N value: 10 Search

| Id | Term | Frequency |
|----|------|-----------|
| 1 | thou | 1190 |
| 2 | d | 976 |
| 3 | king | 858 |
| 4 | lord | 788 |