Report

1. Project Description

Our search engine is based on the theme of science, technology, engineering, and mathematics, also known as "STEM." All of the data for our search engine is crawled from Reddit's subreddits, hence the name of our search engine is the *STEM Subreddit Search.* The STEM Subreddit Search Engine displays the top 20 results of any query. Each query contains the ranking, index score, type, identifier, title, text, subreddit, reddit score, and the file path of each result. Below you will find a high-level overview of our project.

2. **Contribution**
    a. **Brian**
        - Project Documentation
        - Data Crawled: 231MB
        - Research
    b. **Edward**
        - Data Crawled: 140MB
        - Debugging
        - Research
    c. **Hugo**
        - Created Crawling Script
        - Project Documentation
        - Data Crawled: 414MB
        - Research
    d. **Kenny**
        - Data Crawled: 60MB
        - Research
        - Project Documentation
        - Video Demo
        - Dark Mode
    e. **Yongfeng**
        - Data Crawled: 304MB
        - Research
        - Project Documentation
        - Frontend
        - Backend

3. Overview

a. Architecture

Crawler

The crawler is a simple Python script which uses the PRAW library in order to retrieve information from Reddit. At the core of the script is a loop over the list of subreddits to be crawled. Using PRAW, inside of this loop, the script retrieves posts from the current subreddit and scrapes their ontents and that of their comments. In order to scrape the comments, there is an inner loop which is used to implement a breadth-first search.

Search Engine

For our web architecture we are using the spring boot initializer with Spring Web dependency. For our APIs, we are using thymeleaf and the lucene query search. You may refer to the dependencies in pom.xml in the project demo folder.
For our lucene parser, we use MultiFieldQueryParser and set the weight of the title field to 0.65 and weight of the text field to 0.35. We use Filereader to read the file, and CSVReader to read and parse our data. This allows us to connect our local file system from the lucene backend and retrieve data to show the results.

b. Index Structure

Using Lucene, we implemented the StandardAnalyzer() function because it is sufficient for the size and type of our data. We index and analyze the title and text fields of our data; however, we do not store them because they are large. We do not index or analyze the type, score, id, or file paths fields; however, these fields are stored. We do not store, index, or analyze subreddit field. The total collected data folder at 1.15 GB is compressed to 445 MB for the query to search from.

c. Search Algorithm

We are using Lucene's default ranking algorithm which uses a combination of the Vector Space Model and the Boolean Model. The VSM ranks documents by the frequency that a query term appears in a document relative to the number of times the term appears in all the documents in the collection. The Boolean model is used to isolate relevant documents from the non-relevant documents using boolean logic. In addition, Lucene adds capabilities and refinements onto this model to support boolean and approximate string match searching.

d. Crawling Strategy

The crawler takes as input a list of subreddits that should be scraped. It then requests as many recent posts as the Reddit API will allow, around a 1000. The crawler then iterates over those posts, scrapes them, and then scrapes their comments. Since comments in Reddit are organized in a hierarchical fashion, there is a need to recursively ask for child comments in order to scrape all of them. Once all comments on all of the retrieved posts have been scraped for a particular subreddit, it then moves onto the next subreddit in the list. Once a subreddit has been crawled, it is uploaded to a shared drive. A total of 1.15GB of data was crawled by our team.

e. Data Structures

i. A Python list is used as a sort of queue to store which subreddits should be scraped.

ii. A stack data structure is used to scrape the comments as part of a breadth-first search approach.

iii. A Python list of Python lists is used as a temporary data structure for storing the scraped data before being put into a Pandas data frame.

iv. A Pandas data frame data structure is used to organize the collected data and to easily dump the results to a CSV file.

v. Storing acceptable csv files into lucene (index folder). Only some information is taken from the csv file; specifically, the type, score, id, and file path.

vi. List of matches that are stored in data type article which is passed into index html to be shown on the web interface via thymeleaf and bootstrap tables.

4. Limitations

a. The Reddit API's rate limit is up to 60 requests per minute.
b. The Reddit API's crawling limit is up to the 1000 most recent posts.
c. The LuceneProject takes 4 arguments: file directory, index directory, query, and number of hits to display. Default values will be used if they are not specified. However, these default values do not apply to different machines/environments.
d. Two arguments in the demo: index directory and the number of hits to display are fixed, meaning that these default values do not apply to different machines/environments.
e. The data is stored in a local file system instead of a Database System. Therefore, it may be less efficient at reading and writing.
f. There are dozens of corrupted(incomplete) records across data files. Since there are millions of different records, the Lucene project will only filter out corrupted records instead of fixing and storing them.

5. **Deployment Instructions**

Crawler Deployment

a. install python libraries: *python -m pip install praw pandas*
b. go to https://www.reddit.com/prefs/apps/ and create a new app, a personal script
c. name the project, assign a description, and use localhost for the **redirect uri** on the reddit app: *http://127.0.0.1/*
d. copy the app id and the secret id into the script file to the praw.Reddit() constructor
e. enter in your OS and your Reddit username in the user agent also located in the to the praw.Reddit() constructor
f. run the script

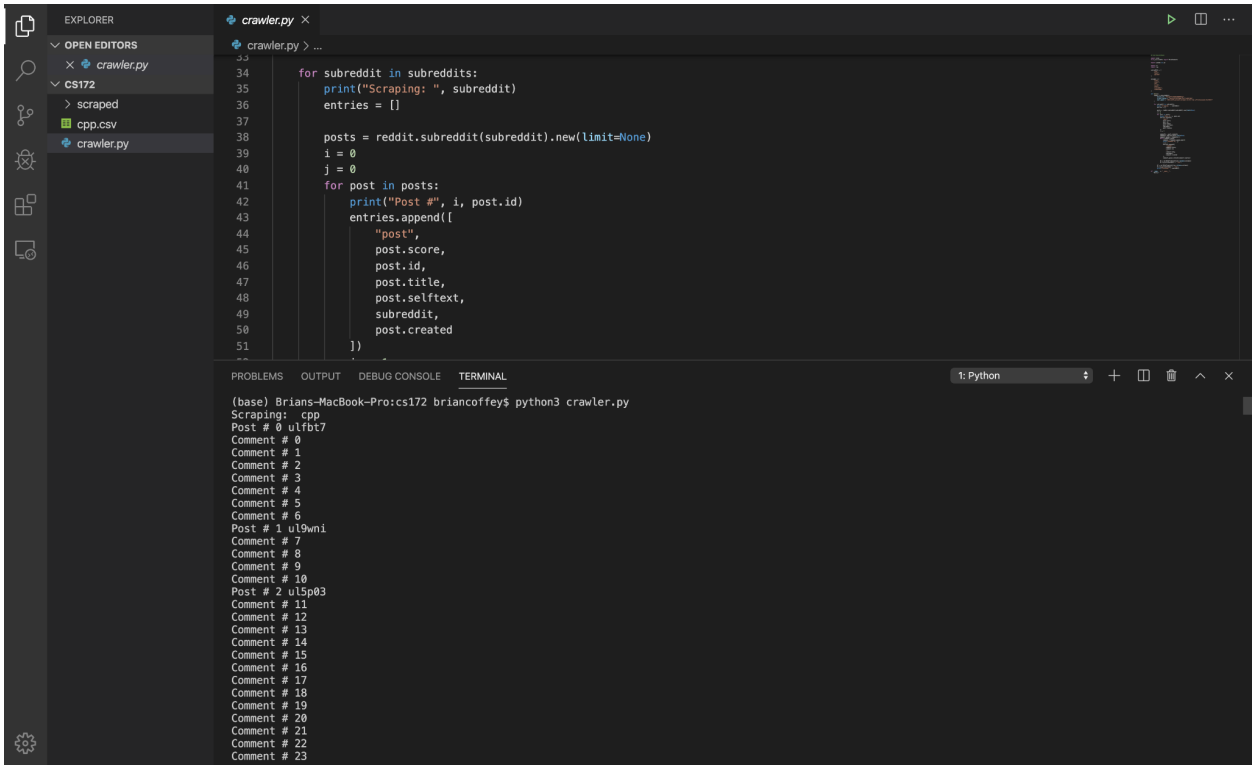
Index Deployment

a. Download IntelliJ IDEA.
b. Once downloaded, specify the indexdir location and filedir location. Filedir should be the folder where the csv files are.
c. Running, it will create an index folder from the scanned csv files. The hard coded string, "computer science" is the test query.
d. After running the code, the index folder will appear and the test query results will show in the terminal.

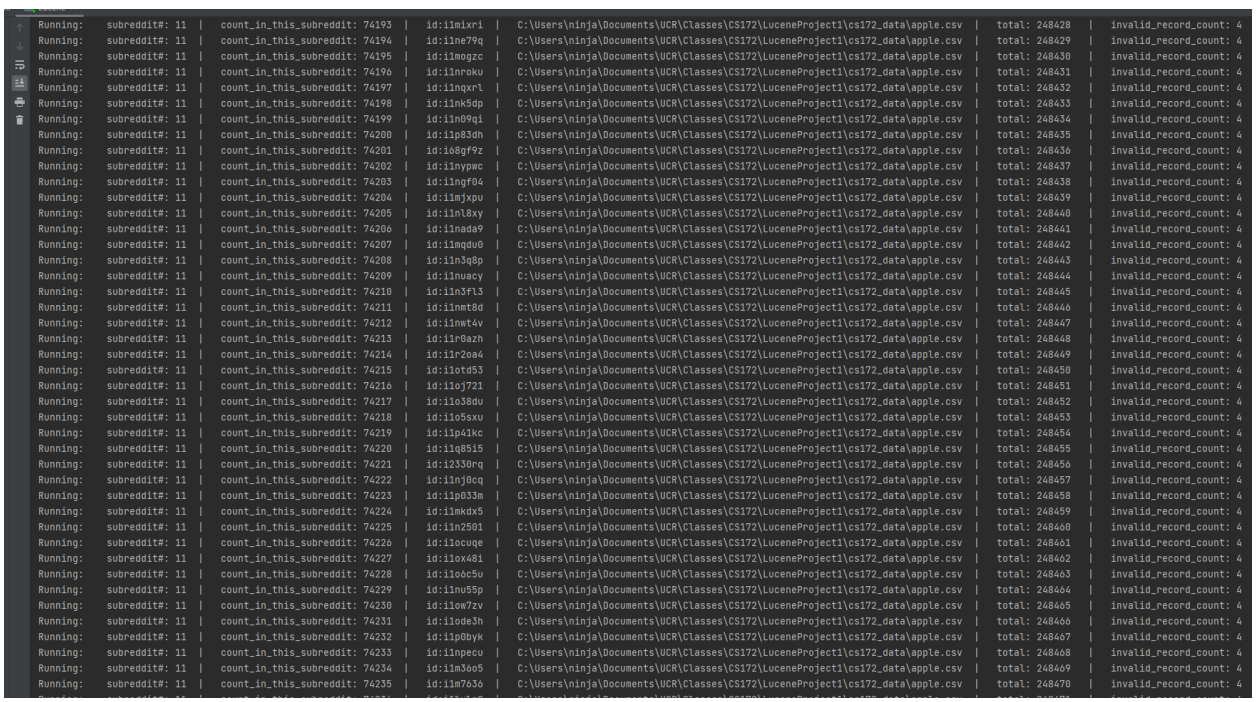
Web Application Deployment

a. Download IntelliJ IDEA.
b. Once downloaded, specify the indexdir for the dataset. Identify the path leading to the Index folder created from the index application.
c. Build and run the application. Access on localhost:8080.
d. On the query. You can put in any sort of query and the first 20 results will show. You can continue to show queries by putting them in the textbox and pressing enter. Searching "*computer science*", you will see the same results pop up but just in a web interface instead.

6. Screenshots



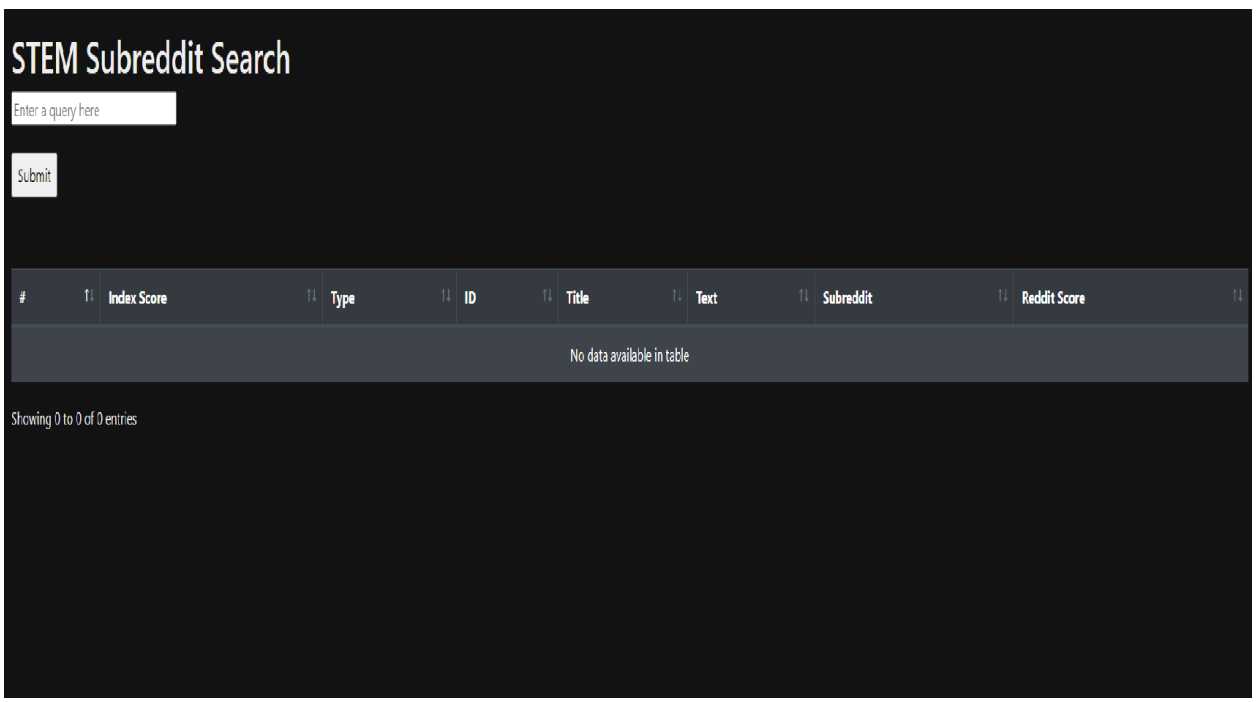
Script in Action



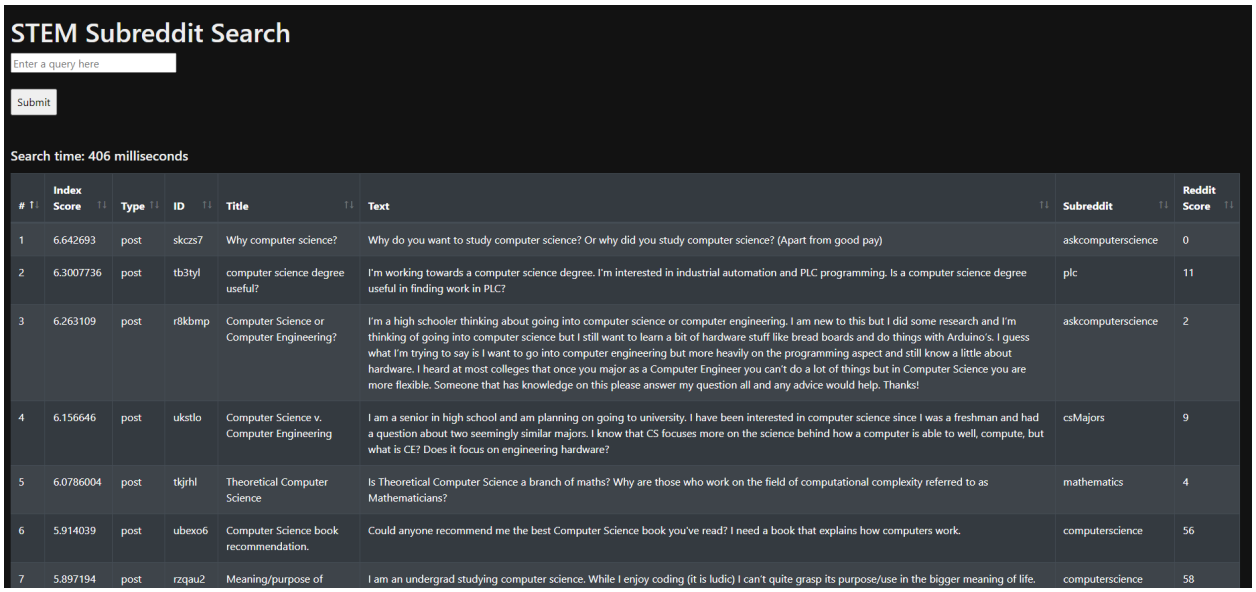
Lucene in Action / Ranking

```
Found 1655+ hits!
Displaying top 20 results:

Result#: 1
Score: 6.042693 |   Reddit Score: 0 |   Subreddit: askcomputerscience   |   ID: skczs7
Type: post
Title: Why computer science?
Text: Why do you want to study computer science? Or why did you study computer science? (Apart from good pay)



Result#: 2
Score: 6.3007736    |   Reddit Score: 11    |   Subreddit: plc  |   ID: tb3tyl
Type: post
Title: computer science degree useful?
Text: I'm working towards a computer science degree. I'm interested in industrial automation and PLC programming. Is a computer science degree useful in finding work in PLC?



Result#: 3
Score: 6.263109 |   Reddit Score: 2 |   Subreddit: askcomputerscience   |   ID: r8kbmp
Type: post
Title: Computer Science or Computer Engineering?
Text:     I'm a high schooler thinking about going into computer science or computer engineering. I am new to this but I did some research and I'm thinking of going into computer science but I still want to lear



Result#: 4
Score: 6.156646 |   Reddit Score: 9 |   Subreddit: csMajors |   ID: ukstlo
Type: post
Title: Computer Science v. Computer Engineering
Text: I am a senior in high school and am planning on going to university. I have been interested in computer science since I was a freshman and had a question about two seemingly similar majors. I know that CS f



Result#: 5
Score: 6.0786004    |   Reddit Score: 4 |   Subreddit: mathematics  |   ID: tkjrhl
Type: post
Title: Theoretical Computer Science
Text: Is Theoretical Computer Science a branch of maths? Why are those who work on the field of computational complexity referred to as Mathematicians?
```

Search Results / Total Hits



Default Search Engine / No Query



STEM Subreddit Search

Enter a query here

Submit

Search time: 406 milliseconds

#	Index Score	Type	ID	Title	Text	Subreddit	Reddit Score
1	6.642693	post	skczs7	Why computer science?	Why do you want to study computer science? Or why did you study computer science? (Apart from good pay)	askcomputerscience	0
2	6.3007736	post	tb3tyl	computer science degree useful?	I'm working towards a computer science degree. I'm interested in industrial automation and PLC programming. Is a computer science degree useful in finding work in PLC?	plc	11
3	6.263109	post	r8kbmp	Computer Science or Computer Engineering?	I'm a high schooler thinking about going into computer science or computer engineering. I am new to this but I did some research and I'm thinking of going into computer science but I still want to learn a bit of hardware stuff like bread boards and do things with Arduino's. I guess what I'm trying to say is I want to go into computer engineering but more heavily on the programming aspect and still know a little about hardware. I heard at most colleges that once you major as a Computer Engineer you can't do a lot of things but in Computer Science you are more flexible. Someone that has knowledge on this please answer my question all and any advice would help. Thanks!	askcomputerscience	2
4	6.156646	post	ukstlo	Computer Science v. Computer Engineering	I am a senior in high school and am planning on going to university. I have been interested in computer science since I was a freshman and had a question about two seemingly similar majors. I know that CS focuses more on the science behind how a computer is able to well, compute, but what is CE? Does it focus on engineering hardware?	csMajors	9
5	6.0786004	post	tkjrhl	Theoretical Computer Science	Is Theoretical Computer Science a branch of maths? Why are those who work on the field of computational complexity referred to as Mathematicians?	mathematics	4
6	5.914039	post	ubexo6	Computer Science book recommendation.	Could anyone recommend me the best Computer Science book you've read? I need a book that explains how computers work.	computerscience	56
7	5.897194	post	rzqau2	Meaning/purpose of	I am an undergrad studying computer science. While I enjoy coding (it is ludic) I can't quite grasp its purpose/use in the bigger meaning of life.	computerscience	58

Search Engine Displaying Top 20 Results of the Query "computer science"

7. Video Demo

Video