

## CS 172 Project Part B - 6/8/2018

James Thi  
Qiwen Lyu  
Patrick Aben

**Twitter:** Write a program that parses the JSON objects of your big files from Part A and inserts them into Lucene. Handle the fields like username, location, and so on. Create a Web-based interface.

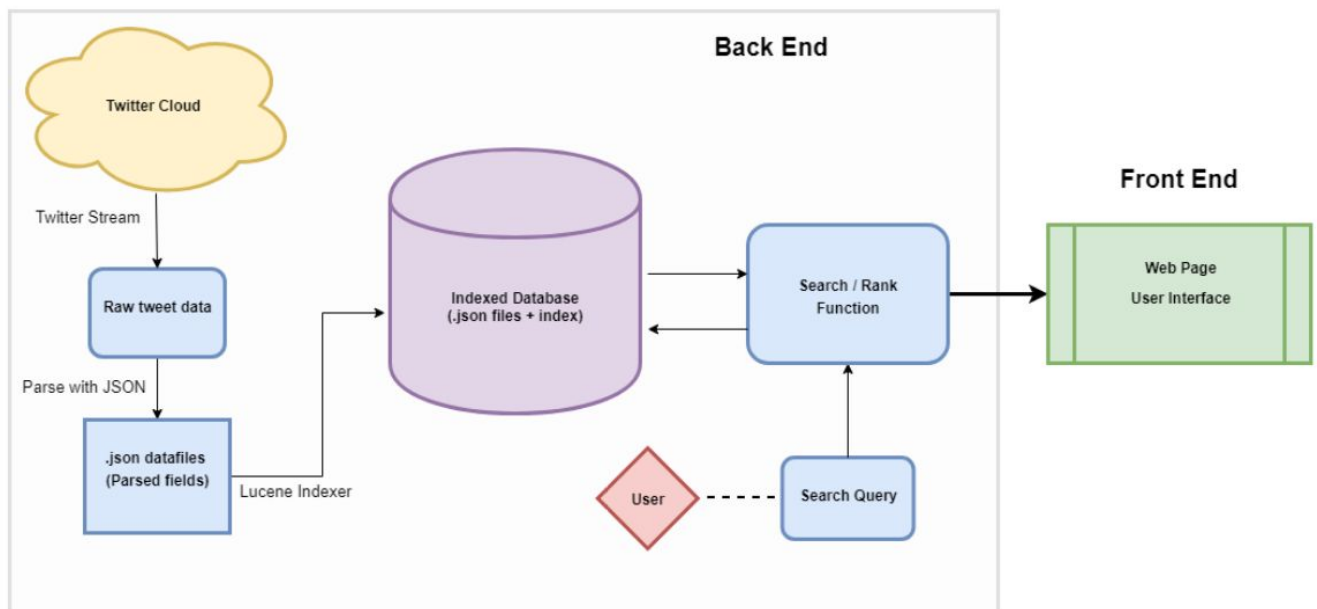
### Collaborations

Nelly (Qiwen) started off by creating the project and added all necessary libraries and dependencies. She built the index for the 5gb of datafiles using Lucene. Patrick then wrote the searching and ranking portion of our project, so that given a query and the number of results it would output, in decreasing score, the relevant results. James then wrote the separate front-end project for the Web-based UI, producing a well-formatted interface for the user to interact with and submit his own queries/number of results. **Blank** created the batch executable which allows the program to be run very easily.

### System Overview

In this project we successfully met all requirements given by the prompt. Using the datafiles (.json) containing tens of thousands of tweets, we indexed 5GB worth of tweets so they were properly searchable and rankable through Lucene. Then we created the web-based user interface to allow the user to easily enter a search query to view ranked results, in decreasing score. The user can also choose the number of results shown, or a default of 10 results is given. This whole part of the project was done using Lucene libraries and a separate Spring-boot project for the front-end Web UI.

### Architecture



## **Index Structures**

In the main app class we created a class “page” to store all important fields: title, latitude, longitude, source, date, tweet\_urls, hashtags and text. We also created a constructor to assign these values into the page. A new function called getDoc is then used to add all the related fields listed above into the document.

After all preparation is complete, the .json files can be parsed into the main function. The first line is empty so it is skipped. Then for each line, which represents one tweet, Json parse and Bufferreader are used to read in the different string object successfully. The constructor and function are then called to push it into the index.

The size of the index is a small fraction of the data itself, but still took up almost 2GB of memory. To work within the limitations of our computers, our index is instead written onto the computer’s physical disc rather than RAM. A directory containing the index must be referenced for the ranking system to later check and use to search with query terms, and is constantly referenced whenever new searches are made. The index is comprised of the pages constructed during the documenting portion of the JSON files.

## **Search and Rank Algorithm**

The search algorithm checks the indexed data and uses the index files to quickly find instances of the query terms and frequency per document. After finding the tweets with the highest frequency occasions of the query term appearing, the ranking algorithm lists the tweets from highest to lowest. In our searching system, we also count the number of times the query terms appear in the title of links, as well as hashtags. Because we added hashtag consideration in ranking, query terms that appear in a hashtag are counted twice, once for being in the text of the tweet, and a second for also being part of the tweet’s hashtags. This creates a bias in rank, where hashtagged terms are ranked much higher than those without. For example, a tweet that mentions a query term 5 times would be ranked lower than a tweet that simply hashtagged the same word 3 times. Or if there are two separate query terms, and a tweet that only mentions one term in a hashtag is higher ranked than a tweet that directly references both query terms. Lastly, it is also possible that the title of a link that the tweet references may have a high frequency of query terms, and not mention them at all in the text of the tweet to be higher ranked than a tweet that directly references all query terms, but fewer times. This means that our search algorithm considers all query terms to be independent of one another.

## **System Limitations**

Our Web UI is set up so it should not produce any errors. Search queries that only contained white space or only contained 2 or less characters produced a message telling the user to enter a valid query. Entering a non-positive integer into the “Number of results” textbox still produced search results with a valid query, but at a default “number of results” of 10. From our testing there was no other way to break the system.

Ideally we would have liked to provide additional details in the results section, such as the username of the tweet's poster, the profile picture of the user, or the link to the original tweet, but unfortunately in our datafiles we never recorded the username or the link to the original tweet. We did not see that option when we were recording data and it would be much too late to gather another 5gb of data. If we had that information though it would be simple to add that detail to each result using a couple html formatting commands.

We also would have done the extra credit and Nelly worked to create the .css and .js files needed to produce the google map api. However we tried multiple times to connect the .css and .js files to the html file, but were unable to make the two work together. The html file would not show anything and could not detect the .css and .js files regardless of what folder they were in. We think it is because we used spring-boot to build the Web UI, and the spring-boot template hid some connections and prevented the html from detecting the .css and .js files. If we recreated the project or did more research we might have been able to get the map working.

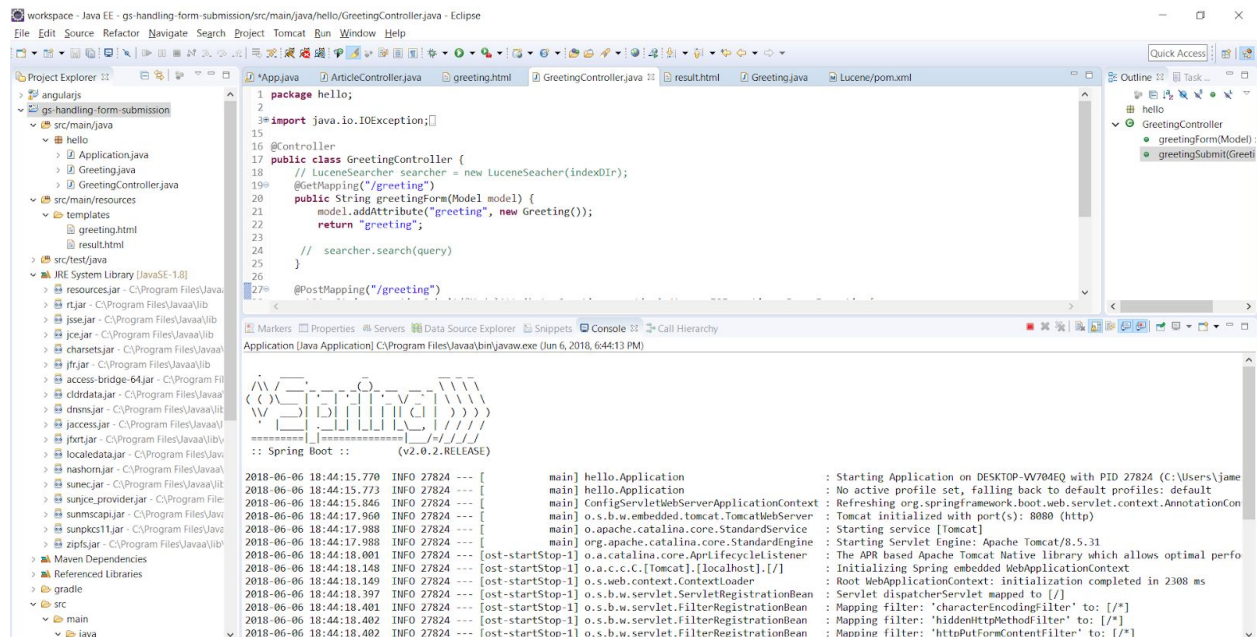
## **How to Deploy System**

### *How to launch the spring-boot project*

Once the Spring-boot project has been launched, the local webpage can be accessed. In a web browser head to *localhost:8080/greeting* . A welcome page will appear with two text boxes: one to enter a search query, and one to enter the number of results desired. Any search query of 3 or more non-whitespace characters should process correctly. If a non-positive integer is entered under the number of results, the system will default to a maximum of 10 outputted results. Simply fill in the two text boxes and click Search to run the search function, or Clear to clear both text boxes. From the results page there is also a link to Submit another query.

## Screenshots

### Spring-boot launched and running in Eclipse:



### Welcome screen of the Web UI:



## Results page with Search Query: “fortnite” and Number of Results: “30”:

Twitter Search Engine

localhost:8080/greeting

Twitter Search Engine v.1.0.0.0.0

Created by Patrick, Nelly, James

[Submit another query](#)

SEARCHING FOR: fortnite

Results shown: 30

1 (score:27.610804) -> Tweet: Fortnite Crashes all the time... @FortniteGame Help! #Fortnite https://t.co/FbnMSdDhYg - Date: Sat May 12 20:42:05 PDT 2018 - Linked Tweet: ["Classy on Twitter: \"Fortnite Crashes all the time... @FortniteGame Help! #Fortnite\""] - Tweet Source: [Twitter Web Client](#)

2 (score:27.28858) -> Tweet: Our #Fortnite music video is back! Which Fortnite emote is your fave and did you spot it in our video? https://t.co/RkaALSPFq9 - Date: Wed May 16 13:29:02 PDT 2018 - Linked Tweet: ["FORTNITE RAP BATTLE SONG - Fortnite in Real Life - Fortnite the Movie - YouTube"] - Tweet Source: [Twitter for iPhone](#)

3 (score:27.079807) -> Tweet: RT @FortniteClub: How the most enlightened people play Fortnite (Credit: u/ Horsechub) #Fortnite https://t.co/8vp5RDLQ8b - Date: Sun May 20 15:16:34 PDT 2018 - Linked Tweet: ["Fortnite Club on Twitter: \"How the most enlightened people play Fortnite (Credit: u/ Horsechub) #Fortnite\""] - Tweet Source: [Twitter for iPhone](#)

4 (score:27.051188) -> Tweet: RT @SilentDroidd: Have you guys seen the new JetPack in Fortnite?? lulul #Fortnite https://t.co/tBKICrFzg - Date: Tue May 22 07:29:00 PDT 2018 - Linked Tweet: ["Droidd on Twitter: \"Have you guys seen the new JetPack in Fortnite?? lulul #Fortnite\""] - Tweet Source: [Twitter for iPhone](#)

5 (score:26.802498) -> Tweet: Diving out the Fortnite partybus like #Fortnite @FortniteGame @VancityReynolds https://t.co/AujgkmKnpY - Date: Fri May 18 14:35:19 PDT 2018 - Linked Tweet: ["Aaron Bailey on Twitter: \"Diving out the Fortnite partybus like #Fortnite @FortniteGame @VancityReynolds\""] - Tweet Source: [Twitter for iPhone](#)

6 (score:26.802498) -> Tweet: Diving out the Fortnite partybus like #Fortnite @FortniteGame @VancityReynolds https://t.co/AujgkmKnpY - Date: Fri May 18 14:35:19 PDT 2018 - Linked Tweet: ["Aaron Bailey on Twitter: \"Diving out the Fortnite partybus like #Fortnite @FortniteGame @VancityReynolds\""] - Tweet Source: [Twitter for iPhone](#)

7 (score:26.452526) -> Tweet: 🙌 YASSS It's time for a great show GS🔴 GILLYGILL🔴 #Fortnite fortnite https://t.co/m5yulWZgU8 https://t.co/bVJjxMIi8I - Date: Sat May 19 11:21:29 PDT 2018 - Linked Tweet: ["#Fortnite fortnite gang joining up", "R F P on Twitter: \"🙌 YASSS It's time for a great show GS🔴 GILLYGILL🔴 #Fortnite fortnite https://t.co/m5yulWZgU8\""] - Tweet Source: [Live.me](#)

8 (score:26.452526) -> Tweet: 🙌 YASSS It's time for a great show GS🔴 GILLYGILL🔴 #Fortnite fortnite https://t.co/m5yulWZgU8 https://t.co/bVJjxMIi8I - Date: Sat May 19 11:21:29 PDT 2018 - Linked Tweet: ["#Fortnite fortnite gang joining up", "R F P on Twitter: \"🙌 YASSS It's time for a great show GS🔴 GILLYGILL🔴 #Fortnite fortnite https://t.co/m5yulWZgU8\""] - Tweet Source: [Live.me](#)

9 (score:26.034601) -> Tweet: RT @krypto9095xbox: When your fortnite character dance better then you 🤖🤖🤖🤖🤖🤖 #fortnite @FortniteGame @Ninja https://t.co/TeXTCiTk5 - Date: Mon May 21 12:27:20 PDT 2018 - Linked Tweet: ["krypto9095 849k subscribers on Twitter: \"When your fortnite character dance better then you 🤖🤖🤖🤖🤖🤖 #fortnite @FortniteGame @Ninja\""] - Tweet Source: [Twitter for Android](#)

## Results page with Search Query: “UCR” and Number of Results: “invalidnumber30asdf”:

Twitter Search Engine

localhost:8080/greeting

Twitter Search Engine v.1.0.0.0.0

Created by Patrick, Nelly, James

[Submit another query](#)

SEARCHING FOR: UCR

Invalid input for number of results; defaulted to max 10 results.

Results shown: 10

1 (score:25.587397) -> Tweet: Little surprise for our moms today in the UCR. #mothersSunday... https://t.co/Y61KOCeM2 - Date: Sun May 13 03:56:43 PDT 2018 - Linked Tweet: ["Nathan Davis on Instagram: \"u201CLittle surprise for our moms today in the UCR. #mothersSunday #weLoveMom. #SeeyouSunday.u201D\""] - Tweet Source: [Instagram](#)

2 (score:24.357569) -> Tweet: RT @Festival\_Cannes: #Photocall GIRL by LUKAS DHONT 📺 #Cannes2018 #UCR https://t.co/q1MAV6H7GR - Date: Wed May 16 22:09:45 PDT 2018 - Linked Tweet: ["Festival de Cannes on Twitter: \"#Photocall GIRL by LUKAS DHONT 📺 #Cannes2018 #UCR\""] - Tweet Source: [Twitter for Android](#)

3 (score:21.811665) -> Tweet: RT @Yhillar\_: Found someone's tassel at the UCR botanic gardens HMU if it's yours @ canyon c/o '18 https://t.co/KkwReFPkCC - Date: Thu May 17 16:28:37 PDT 2018 - Linked Tweet: ["Hill on Twitter: \"Found someone's tassel at the UCR botanic gardens HMU if it's yours @ canyon c/o '18\""] - Tweet Source: [Twitter for iPhone](#)

4 (score:19.864231) -> Tweet: UCR Report on Inland Empire's Housing Market Cycle https://t.co/IrvpCFdYz @ucriverside @ucr\_business #Riverside #InlandEmpire #housing #sanbernardino https://t.co/rmWBMchuD5 - Date: Tue May 22 16:49:01 PDT 2018 - Linked Tweet: ["UCR Report on Inland Empire's Housing Market Cycle 'u2013 InlandEmpire.us\", \"InlandEmpire.US on Twitter: \"UCR Report on Inland Empire's Housing Market Cycle https://t.co/IrvpCFdYz @ucriverside @ucr\_business #Riverside #InlandEmpire #housing #sanbernardino\"\""] - Tweet Source: [HubSpot](#)

5 (score:17.223783) -> Tweet: @UCRSoftball way to go UCR !!!! Great team !!! - Date: Fri May 18 20:00:12 PDT 2018 - Tweet Source: [Twitter for Android](#)

6 (score:17.223783) -> Tweet: @UCRSoftball way to go UCR !!!! Great team !!! - Date: Fri May 18 20:00:12 PDT 2018 - Tweet Source: [Twitter for Android](#)

7 (score:16.475857) -> Tweet: @KareemGongora The UC system and UCR specifically have a budget ask this year for things like deferred maintenance and to build a new building at UCR. - Date: Wed May 16 06:59:56 PDT 2018 - Tweet Source: [Twitter for iPhone](#)

8 (score:16.058872) -> Tweet: RT @taylorchloeoe: Students: we need more parking UCR: MSE2? Students: no more parking.... UCR: Spend \$30,000 on ugly sweaters? Studen... - Date: Thu May 17 19:04:07 PDT 2018 - Tweet Source: [Twitter for iPhone](#)

9 (score:16.058872) -> Tweet: RT @taylorchloeoe: Students: we need more parking UCR: MSE2? Students: no more parking.... UCR: Spend \$30,000 on ugly sweaters? Studen... - Date: Thu May 17 19:16:40 PDT 2018 - Tweet Source: [Twitter for iPhone](#)



## Results page with Search Query: “embedded engineering” and Number of Results: “”:

Twitter Search Engine

localhost:8080/greeting

Twitter Search Engine v.1.0.0.0.0

Created by Patrick, Nelly, James

[Submit another query](#)

SEARCHING FOR: embedded engineering

Invalid input for number of results; defaulted to max 10 results.

Results shown: 10

1 (score:31.237547) --> Tweet: plates plates and more plates #engineering https://t.co/pA7ygLwgLD - Date: Tue May 22 07:49:16 PDT 2018 - Linked Tweet: ["rhehamsnow on Twitter: \"plates plates and more plates #engineering u2026 \""] - Tweet Source: [Twitter for Android](#)

2 (score:31.014107) --> Tweet: RT @IntEngineering: Drones pull off first transmission line cable construction.. #engineering https://t.co/aYzjKeYFka - Date: Tue May 22 19:25:36 PDT 2018 - Linked Tweet: ["Interesting Engineering on Twitter: \"Drones pull off first transmission line cable construction.. #engineering u2026 \""] - Tweet Source: [Twitter for Android](#)

3 (score:30.01082) --> Tweet: Graduation 🎓 @nyuniversity #engineering @nyutandon @barclayscenter https://t.co/a8TrHymypD - Date: Tue May 15 05:44:23 PDT 2018 - Linked Tweet: ["Michael Herzenberg on Twitter: \"Graduation 🎓 @nyuniversity #engineering @nyutandon @barclayscenter u2026 \""] - Tweet Source: [Twitter for iPhone](#)

4 (score:28.238129) --> Tweet: Ground #engineering expert provides advanced structural solutions for commercial businesses https://t.co/iFc7kGjz7 - Date: Wed May 23 01:46:00 PDT 2018 - Linked Tweet: ["Ground engineering expert provides advanced structural solutions for commercial businesses | Warehouse & Logistics News"] - Tweet Source: [Buffer](#)

5 (score:27.37644) --> Tweet: Embedded Software Developer - £45,000 - £50,000, Swindon, £45k-50k, Engineering https://t.co/EbksuKVY5a #jobs #hiring - Date: Tue May 22 04:34:05 PDT 2018 - Linked Tweet: ["Embedded Software Developer - £45,000 - £50,000"] - Tweet Source: [jdibu Adpost](#)

6 (score:26.062891) --> Tweet: #Fargo #ND #USA - Senior Embedded Software Engineer - Product Engineering Title Senior Embedded Software Engineer - 32319 N https://t.co/i2NdYj6Zgl #CAREER #JOB - Date: Wed May 23 02:48:58 PDT 2018 - Linked Tweet: ["Builders Yellow Pages JOBS: SENIOR EMBEDDED SOFTWARE ENGINEER FARGO ND USA"] - Tweet Source: [Twibble.io](#)

7 (score:25.782915) --> Tweet: "87% of women engineers... would recommend #engineering as a great career" - @TheManufacturer's interview with @RAEngNews's new chief executive Hayaatun Sillem is a great weekend read. https://t.co/0u7Laz2XdV - Date: Fri May 18 08:58:17 PDT 2018 - Linked Tweet: ["The Manufacturer sits down with Hayaatun Sillem - newly appointed chief exec of the Royal Academy of Engineering - The Manufacturer"] - Tweet Source: [Twitter Web Client](#)

8 (score:25.782915) --> Tweet: "87% of women engineers... would recommend #engineering as a great career" - @TheManufacturer's interview with @RAEngNews's new chief executive Hayaatun Sillem is a great weekend read. https://t.co/0u7Laz2XdV - Date: Fri May 18 08:58:17 PDT 2018 - Linked Tweet: ["The Manufacturer sits down with Hayaatun Sillem - newly appointed chief exec of the Royal Academy of Engineering - The Manufacturer"] - Tweet Source: [Twitter Web Client](#)

## Results page with Search Query: “history of the civil war” and Number of Results: “70”:

Twitter Search Engine

localhost:8080/greeting

Twitter Search Engine v.1.0.0.0.0

Created by Patrick, Nelly, James

[Submit another query](#)

SEARCHING FOR: history of the civil war

Results shown: 70

1 (score:44.420517) --> Tweet: \*CIVIL WAR\* History of the War for the Union Civil, Military & Naval Part 7 Hurry #militaryhistory #historymilitary #civilmilitary https://t.co/410dJWOXE - Date: Sun May 20 20:05:00 PDT 2018 - Linked Tweet: ["\*CIVIL WAR\* History of the War for the Union Civil, Military & Naval Part 7 | eBay"] - Tweet Source: [Tweet-Eye Web Application](#)

2 (score:33.741936) --> Tweet: RT @juegofthrones: civil war // infinity war https://t.co/EnuxAbPqNm - Date: Wed May 16 18:43:49 PDT 2018 - Linked Tweet: ["mr. stark? on Twitter: \"civil war // infinity war u2026 \""] - Tweet Source: [Twitter for Android](#)

3 (score:33.741936) --> Tweet: RT @juegofthrones: civil war // infinity war https://t.co/EnuxAbPqNm - Date: Thu May 17 07:05:05 PDT 2018 - Linked Tweet: ["mr. stark? on Twitter: \"civil war // infinity war u2026 \""] - Tweet Source: [Twitter for Android](#)

4 (score:32.19607) --> Tweet: The Spanish Civil War | Documentary Vine https://t.co/YOaFAN2BAk - Date: Mon May 21 12:01:20 PDT 2018 - Linked Tweet: ["The Spanish Civil War | Documentary Vine"] - Tweet Source: [Twitter Web Client](#)

5 (score:32.119553) --> Tweet: Civil War 2 https://t.co/dxdNWkMaMy - Date: Mon May 14 14:24:37 PDT 2018 - Linked Tweet: ["Getsuga Tenshō on Twitter: \"Civil War 2 u2026 \""] - Tweet Source: [Twitter for Android](#)

6 (score:31.551271) --> Tweet: I liked a @YouTube video https://t.co/ez75vpVtL The Battle of Winchester -Civil War- - Date: Thu May 17 19:39:46 PDT 2018 - Linked Tweet: ["The Battle of Winchester -Civil War- - YouTube"] - Tweet Source: [Google](#)

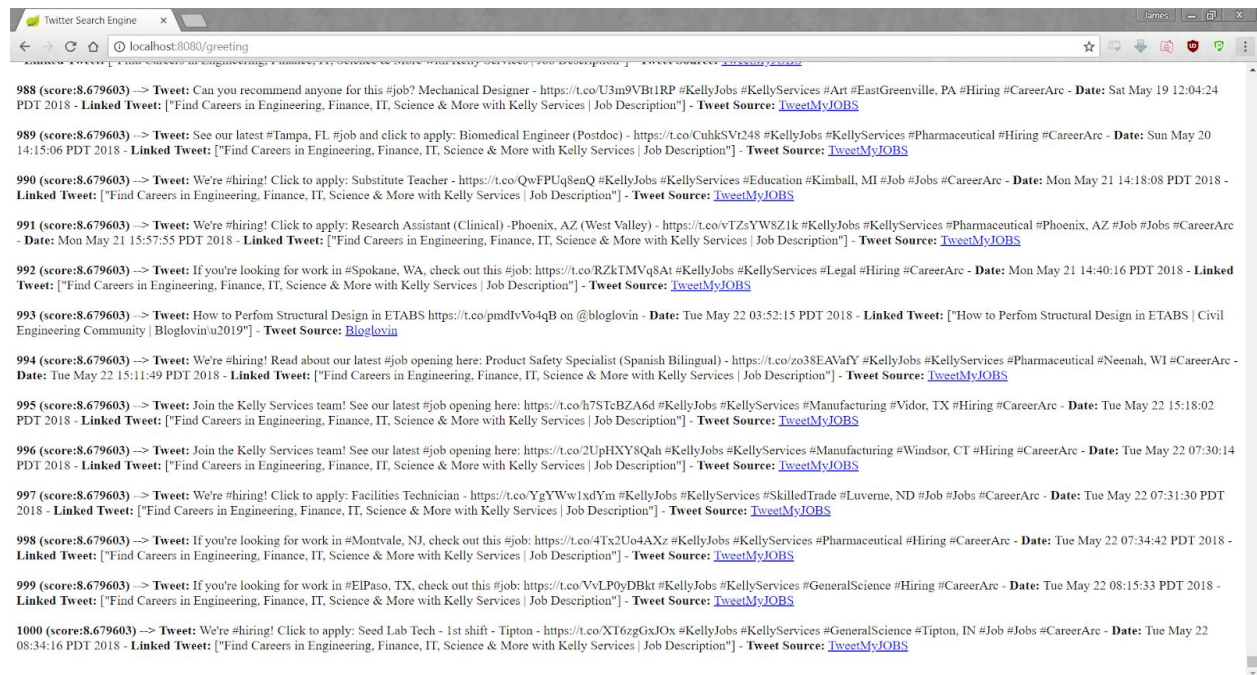
7 (score:31.381207) --> Tweet: RT @KannasEgo: A true civil war https://t.co/6eTjwWpOem - Date: Mon May 21 18:41:18 PDT 2018 - Linked Tweet: ["Ego on Twitter: \"A true civil war u2026 \""] - Tweet Source: [Twitter Web Client](#)

8 (score:31.345493) --> Tweet: Another civil war for you guys.. Zero Two or Ichigo? ♥ https://t.co/4dRyTKikH1 - Date: Thu May 17 05:59:14 PDT 2018 - Linked Tweet: ["DANTE ♥ on Twitter: \"Another civil war for you guys.. Zero Two or Ichigo? ♥ u2026 \""] - Tweet Source: [Twitter for Android](#)

9 (score:31.345493) --> Tweet: Who's winning the Democrats' civil war? https://t.co/30fajAdvz - Date: Tue May 22 18:44:20 PDT 2018 - Linked Tweet: ["The Democrats' Civil War is on Display in Georgia Primary | Time"] - Tweet Source: [Twitter Web Client](#)

10 (score:30.973259) --> Tweet: If this law passes the next world war or Civil War will happen not good at all 🚫 https://t.co/seZh4go331 - Date: Fri May 18 14:15:51 PDT 2018 - Linked Tweet: ["IG:JORDANMAKEITREAL on Twitter: \"If this law passes the next world war or Civil War will happen not good at all 🚫 \""] - Tweet Source: [Twitter for Android](#)

## Last part of Results page with Search Query: “engineering” and Number of Results: “1000d”:



## Results page with Search Query: "" and Number of Results: "10":



### Twitter Search Engine v.1.0.0.0.0

Created by Patrick, Nelly, James

[Submit another query](#)

SEARCHING FOR:

Please enter a valid query of 3 or more letters.

```
james@DESKTOP-2Q6MS8K MINGW64 ~/Desktop
$ java -jar twittermission.jar 100 C:/Users/james/Desktop
```