CS172 Project Part 2 Documentation

Henry Garcia

Nicolas Lawler

Collaboration Details

- Henry
  - Built the UI
  - Refined the query processing code to work with the UI
  - Added to the documentation
- Nicolas
  - Built the index builder
  - Wrote the initial version of the query processing code
  - Wrote the foundation of the documentation

System Overview

Architecture

The System is split across two main parts, the IndexBuilder and a Web-Based UI

- IndexBuilder
  - Takes as input 2 directory paths, one for the location of the tweet files, and another for the desired output directory. (Note: uses JRE 1.8)
  - It iteratively builds a Document object for each tweet, then adds each document to the lucene index.
- User Interface
  - Contains a ListBox and TextBox in order to search tweets. The user can choose what they would like to search for (General, User, HashTags). The UI will search for the relevant tweets in the created index by searching hashtags, title, username, and body of the tweet. (hashtags and titles are boosted)
  - Once results are displayed the user can click on a tweet and the corresponding marker on the map will move. Similarly the user can click on the map point to highlight the tweet it corresponds to.
  - A user can also click on a Tweet's username or HashTag in order to add it to their searchbar.

Index Structure

The indexes are built using the "user", "text", "created_at", "geo_location", "linkTitle", "favorite_count", "retweet_count", “hashtags”, and "language" fields of the tweet. We only use the “screenName” property of the “user” field, and we separate the “geo_location” field into latitude and longitude. We store each of these fields within the index, so that the information can be presented to the user through the UI. “text”, “hashtags”, and “linkTitle” are stored as a TextField so that terms are tokenized, while all the others are stored as string or numeric fields. The “created_at” , “hashtags”, and “text” fields are boosted to give them more weight in searches.

Search Algorithm

- Empty Search
  - An empty search will return the most recent tweets ordered by time (newest tweets to oldest tweets)
- General Search
  - Takes the Text provided by the user and searches for matches in the “hashtags”, “text”, “user”, and “linkTitle” fields by separating the query by spaces. The “hashtag” and “linkTitle” fields are boosted by 2f. This is to allow “hashtags” and “linkTitle” field matches to get ranked higher, but also increase the likelihood that an exact match of pure text will still appear as a top result.
  - hashtags and usernames can be separated by “@” and “#” symbols respectively. However, text will be searched with spaces.
- HashTag Search
  - Takes the text provided by the user and searches for matches in the “hashtag” field. returning all results that contain specified hashtag
  - hashtags can be separated by spaces or by hashtag. (ex. searching “starwars startrek” is the same as “#starwars#startrek”
- User Search
  - Takes the text provided by the user and searches for matches in the “user” field. returning all results that contain the specified user.
  - usernames can be separated by spaces or by “@”. (ex. searching “anon123 cool4school” and “@anon123@cool4school” will return the same results)

Implementation Notes

- The index builder makes use of the Java 8 Streams API to process the incoming data in a declarative way.
- The index builder only accepts empty directories for the output directory, it will halt if the user attempts to pass in a non-empty directory.
- When a user submits a query the Title and hashtag fields are boosted by 2f (this allows fields with title and hashtag to be boosted, but also allows a user to find an exact match tweet that only contains text)

Limitations

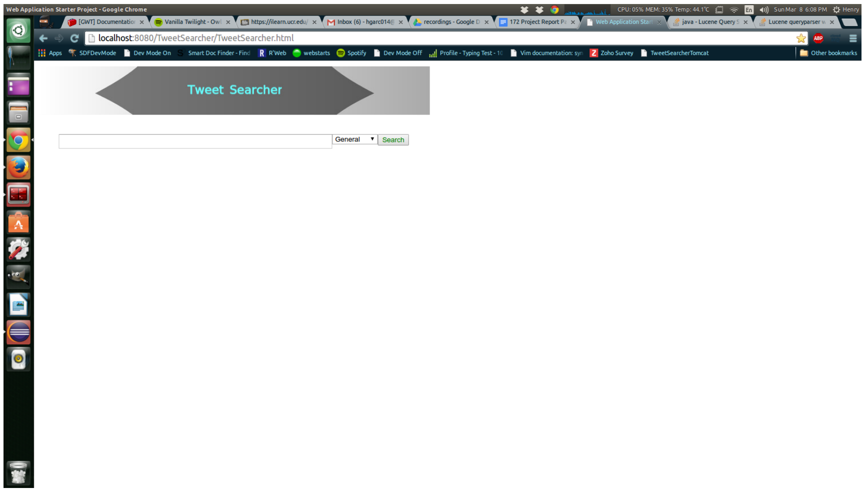
- The IndexBuilder relies on the directory full of tweet files created in the first part of the project. Because that directory is not being constantly updated, the tweets returned by searches may be out of date, depending on how long ago the TweetCrawler and IndexBuilder were run.
- Since the map uses Google’s API the user needs to have internet in order for the page to display correctly.

Deployment Instructions

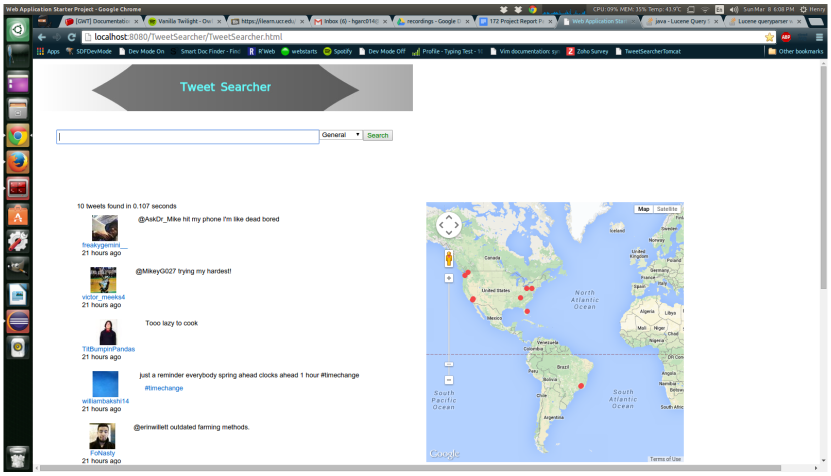
1. Index crawled tweets into a folder called index
2. place folder into TweetSearcher/war/indexes
3. GWT Compile the TweetSearcher project.
4. after compiling go to war folder (TweetSearcher/war)
5. open deploy script and change the location of the tomcat folder.
6. run deploy script (note: assumes you have a folder called indexes at location /var/lib/tomcat7/indexes) (WARNING: deploy script will remove the following files if they already exist TweetSearcher, TweetSearcher.war, and index in the tomcat folders)
7. restart tomcat and then you should be able to go to your application on a browser (http://localhost:8080/TweetSearcher/TweetSearcher.html)

Screenshots

Main Page


Web Application Starter Project - Google Chrome
localhost:8080/TweetSearcher/TweetSearcher.html
Tweet Searcher
General
Search

Blank Search



10 tweets found in 0.107 seconds

freakygemini__
21 hours ago
@AskDr_Mike hit my phone I'm like dead bored

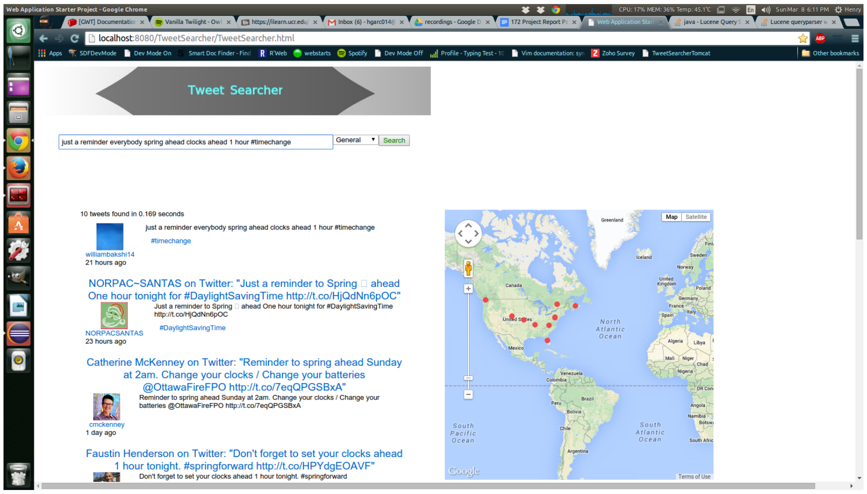
victor_meeks4
21 hours ago
@MikeyG027 trying my hardest!

TitBumpinPandas
21 hours ago
Tooo lazy to cook

williambakshi14
21 hours ago
just a reminder everybody spring ahead clocks ahead 1 hour #timechange
#timechange

FoNasty
21 hours ago
@erinwillett outdated farming methods.

General Search



localhost:8080/TweetSearcher/TweetSearcher.html

just a reminder everybody spring ahead clocks ahead 1 hour #timechange

10 tweets found in 0.169 seconds

williambakshi14
21 hours ago

just a reminder everybody spring ahead clocks ahead 1 hour #timechange

#timechange

NORPAC~SANTAS on Twitter: "Just a reminder to Spring □ ahead One hour tonight for #DaylightSavingTime http://t.co/HjQdNn6pOC"

NORPACSANTAS
23 hours ago

Just a reminder to Spring □ ahead One hour tonight for #DaylightSavingTime http://t.co/HjQdNn6pOC

#DaylightSavingTime

Catherine McKenney on Twitter: "Reminder to spring ahead Sunday at 2am. Change your clocks / Change your batteries @OttawaFireFPO http://t.co/7eqQPGSBxA"

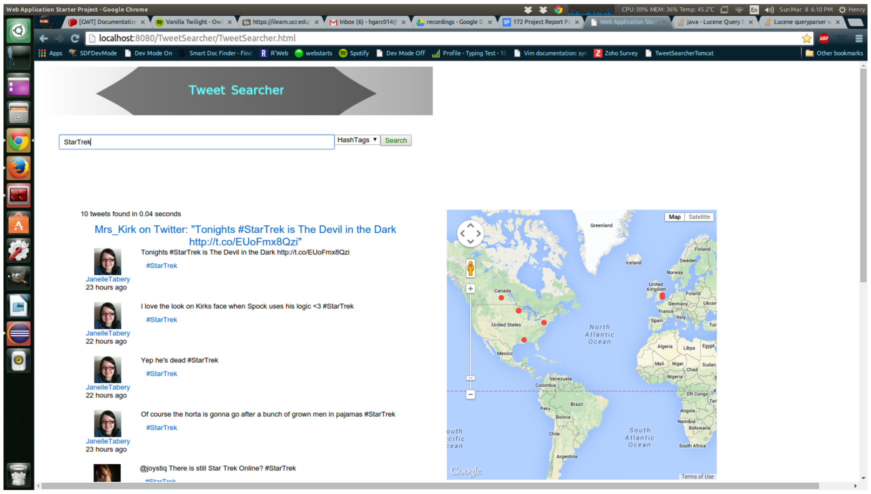
cmckenney
1 day ago

Reminder to spring ahead Sunday at 2am. Change your clocks / Change your batteries @OttawaFireFPO http://t.co/7eqQPGSBxA

Faustin Henderson on Twitter: "Don't forget to set your clocks ahead 1 hour tonight. #springforward http://t.co/HPYdgEOAVF"

Don't forget to set your clocks ahead 1 hour tonight. #springforward

HashTag Search



localhost:8080/TweetSearcher/TweetSearcher.html

StarTrek

10 tweets found in 0.04 seconds

Mrs_Kirk on Twitter: "Tonights #StarTrek is The Devil in the Dark http://t.co/EUoFmx8Qzi"

JanelleTabery
23 hours ago

Tonights #StarTrek is The Devil in the Dark http://t.co/EUoFmx8Qzi

#StarTrek

JanelleTabery
22 hours ago

I love the look on Kirks face when Spock uses his logic <3 #StarTrek

#StarTrek

JanelleTabery
22 hours ago

Yep he's dead #StarTrek

#StarTrek

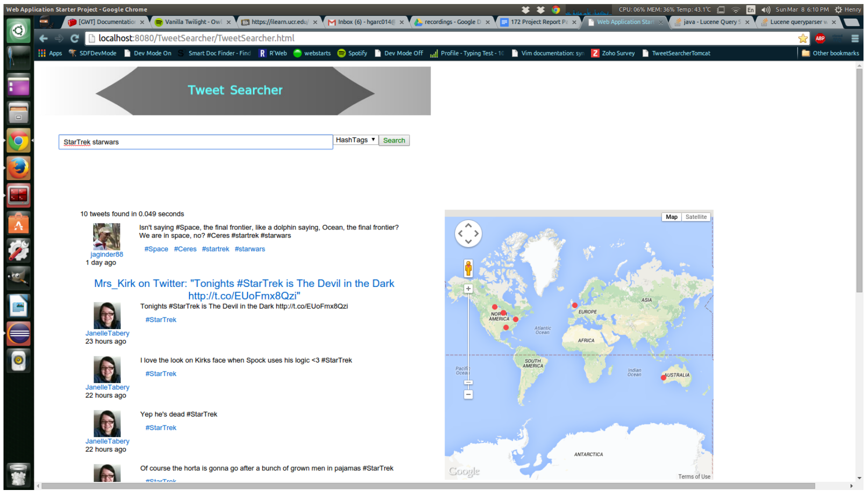
JanelleTabery
23 hours ago

Of course the horta is gonna go after a bunch of grown men in pajamas #StarTrek

#StarTrek

@joystiq There is still Star Trek Online? #StarTrek

Multi HashTag Search



localhost:8080/TweetSearcher/TweetSearcher.html

StarTrek starwars

10 tweets found in 0.049 seconds

jaginder88
1 day ago

Isn't saying #Space, the final frontier, like a dolphin saying, Ocean, the final frontier? We are in space, no? #Ceres #startrek #starwars

#Space #Ceres #startrek #starwars

Mrs_Kirk on Twitter: "Tonights #StarTrek is The Devil in the Dark http://t.co/EUoFmx8Qzi"

JanelleTabery
23 hours ago

Tonights #StarTrek is The Devil in the Dark http://t.co/EUoFmx8Qzi

#StarTrek

JanelleTabery
22 hours ago

I love the look on Kirks face when Spock uses his logic <3 #StarTrek

#StarTrek

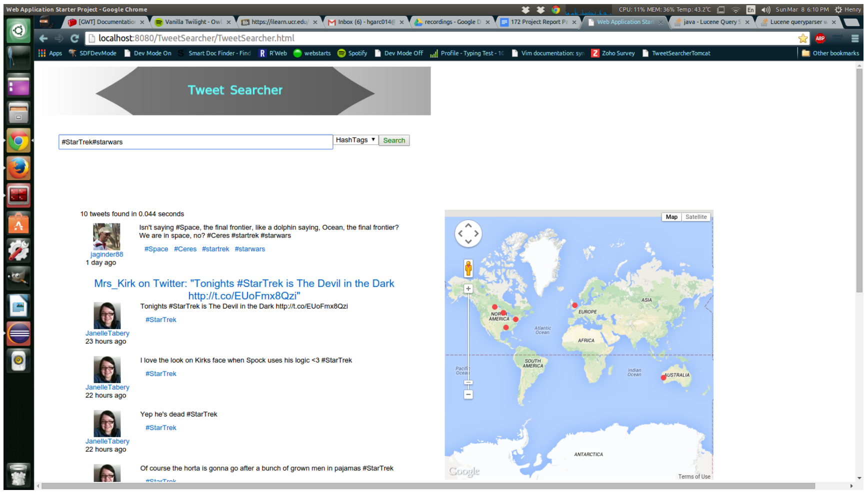
JanelleTabery
22 hours ago

Yep he's dead #StarTrek

#StarTrek

Of course the horta is gonna go after a bunch of grown men in pajamas #StarTrek

Multi HashTag Search (with #)



localhost:8080/TweetSearcher/TweetSearcher.html

Tweet Searcher

#StarTrek#starwars | HashTags ▾ | Search

10 tweets found in 0.044 seconds

jaginder88
1 day ago
Isn't saying #Space, the final frontier, like a dolphin saying, Ocean, the final frontier? We are in space, no? #Ceres #startrek #starwars
#Space #Ceres #startrek #starwars

Mrs_Kirk on Twitter: "Tonights #StarTrek is The Devil in the Dark http://t.co/EUoFmx8Qzi"

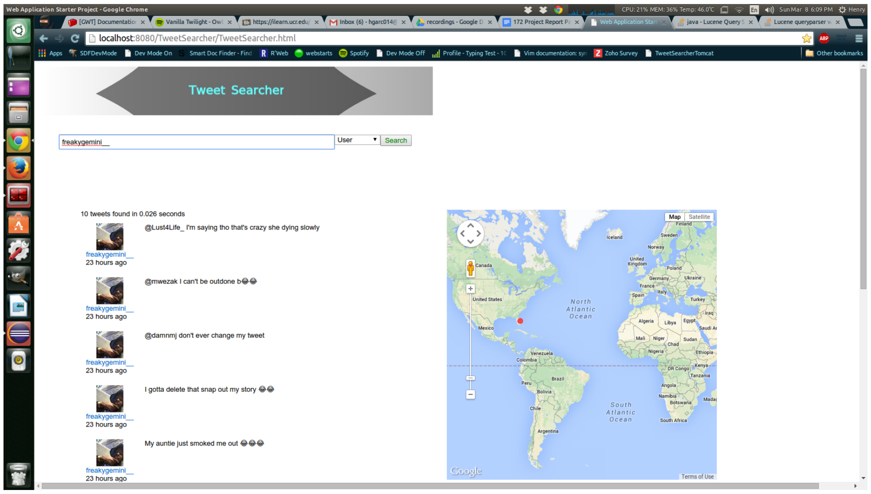
JanelleTabery
23 hours ago
Tonights #StarTrek is The Devil in the Dark http://t.co/EUoFmx8Qzi
#StarTrek

JanelleTabery
22 hours ago
I love the look on Kirks face when Spock uses his logic <3 #StarTrek
#StarTrek

JanelleTabery
22 hours ago
Yep he's dead #StarTrek
#StarTrek

Of course the horta is gonna go after a bunch of grown men in pajamas #StarTrek

User Search



localhost:8080/TweetSearcher/TweetSearcher.html

Tweet Searcher

freakygemini__ | User ▾ | Search

10 tweets found in 0.026 seconds

freakygemini__
23 hours ago
@Lust4Life_ I'm saying tho that's crazy she dying slowly

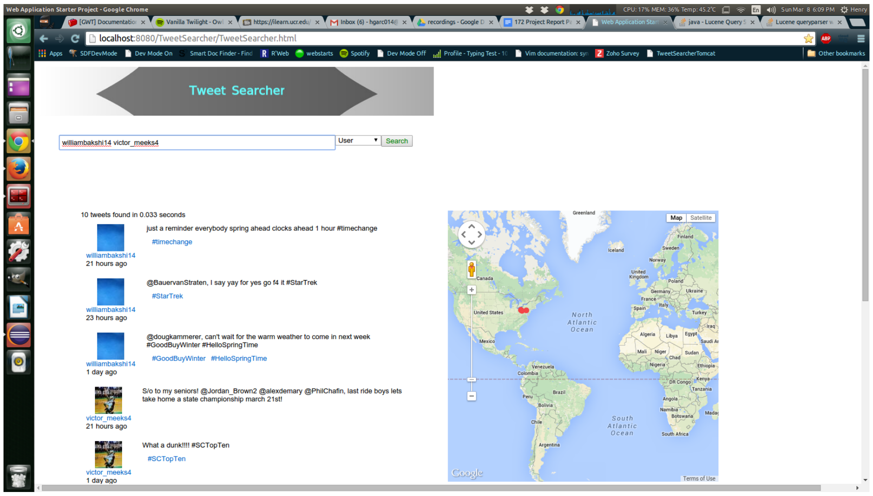
freakygemini__
23 hours ago
@mwezak I can't be outdone b😂😂

freakygemini__
23 hours ago
@damnmj don't ever change my tweet

freakygemini__
23 hours ago
I gotta delete that snap out my story 😂😂

freakygemini__
23 hours ago
My auntie just smoked me out 😂😂😂

Multi User Search



localhost:8080/TweetSearcher/TweetSearcher.html

Tweet Searcher

williambakshi14 victor_meeks4 | User ▾ | Search

10 tweets found in 0.033 seconds

williambakshi14
21 hours ago
just a reminder everybody spring ahead clocks ahead 1 hour #timechange
#timechange

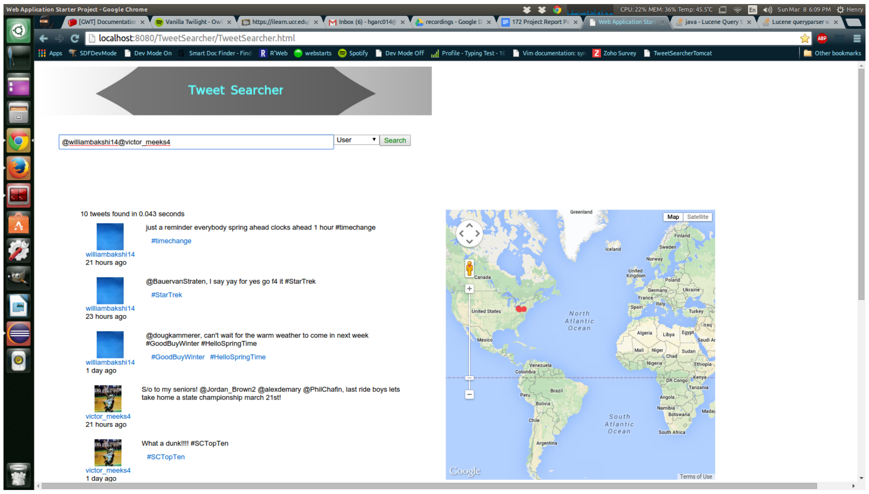
williambakshi14
23 hours ago
@BauervanStraten, I say yay for yes go f4 it #StarTrek
#StarTrek

williambakshi14
1 day ago
@dougkammerer, can't wait for the warm weather to come in next week #GoodBuyWinter #HelloSpringTime
#GoodBuyWinter #HelloSpringTime

victor_meeks4
21 hours ago
S/o to my seniors! @Jordan_Brown2 @alexdemary @PhilChafin, last ride boys lets take home a state championship march 21st!

victor_meeks4
1 day ago
What a dunk!!!! #SCTopTen
#SCTopTen

Multi User Search (with @)



@williambakshi14@victor_meeks4

User ▾

10 tweets found in 0.043 seconds

williambakshi14
21 hours ago

just a reminder everybody spring ahead clocks ahead 1 hour #timechange

#timechange

williambakshi14
23 hours ago

@BauervanStraten, I say yay for yes go f4 it #StarTrek

#StarTrek

williambakshi14
1 day ago

@dougkammerer, can't wait for the warm weather to come in next week #GoodBuyWinter #HelloSpringTime

#GoodBuyWinter #HelloSpringTime

victor_meeks4
21 hours ago

S/o to my seniors! @Jordan_Brown2 @alexdemary @PhilChafin, last ride boys lets take home a state championship march 21st!

victor_meeks4
1 day ago

What a dunk!!!! #SCTopTen

#SCTopTen