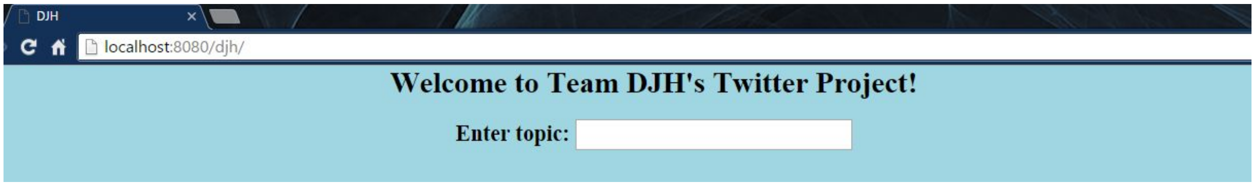
Project Report

Topic: Tweet Clustering System

Link: https://github.com/TeamDJH/DJH

Team DJH Members: David Gwin, Jacob Reid, Helam Franco

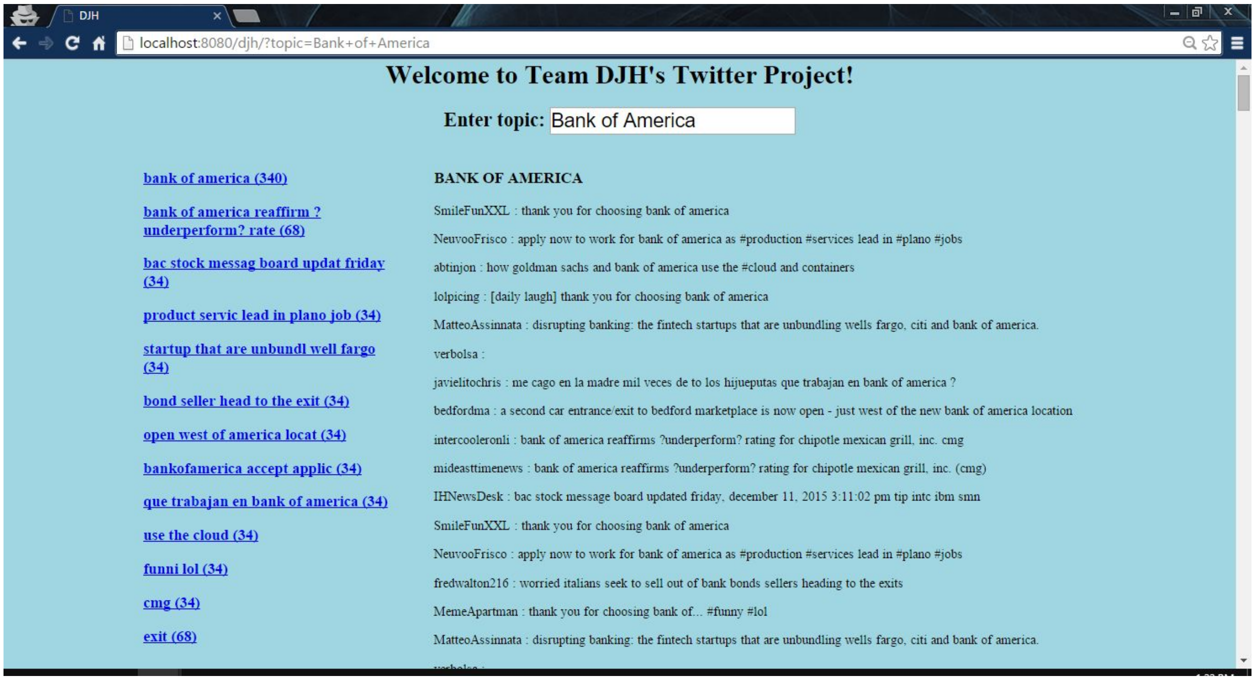


Welcome to Team DJH's Twitter Project!

Enter topic:

Overview:

This program allows you to search for tweets relating to a specific topic. It accesses the Twitter database and returns the results clustered into relevant topics. In the example below we searched for Bank of America.



Welcome to Team DJH's Twitter Project!

Enter topic: Bank of America

bank of america (340)

bank of america reaffirm ? underperform? rate (68)

bac stock messag board updat friday (34)

product servic lead in plano job (34)

startup that are unbundl well fargo (34)

bond seller head to the exit (34)

open west of america locat (34)

bankofamerica accept applic (34)

que trabajan en bank of america (34)

use the cloud (34)

funni lol (34)

cmg (34)

exit (68)

BANK OF AMERICA

SmileFunXXL : thank you for choosing bank of america

NeuvooFrisco : apply now to work for bank of america as #production #services lead in #plano #jobs

abtinjon : how goldman sachs and bank of america use the #cloud and containers

lolpicing : [daily laugh] thank you for choosing bank of america

MatteoAssinnata : disrupting banking: the fintech startups that are unbundling wells fargo, citi and bank of america.

verbolsa :

javielitochris : me cago en la madre mil veces de to los hijueputas que trabajan en bank of america ?

bedfordma : a second car entrance/exit to bedford marketplace is now open - just west of the new bank of america location

intercooleronli : bank of america reaffirms ?underperform? rating for chipotle mexican grill, inc. cmg

mideasttimenews : bank of america reaffirms ?underperform? rating for chipotle mexican grill, inc. (cmg)

IHNewsDesk : bac stock message board updated friday, december 11, 2015 3:11:02 pm tip intc ibm smn

SmileFunXXL : thank you for choosing bank of america

NeuvooFrisco : apply now to work for bank of america as #production #services lead in #plano #jobs

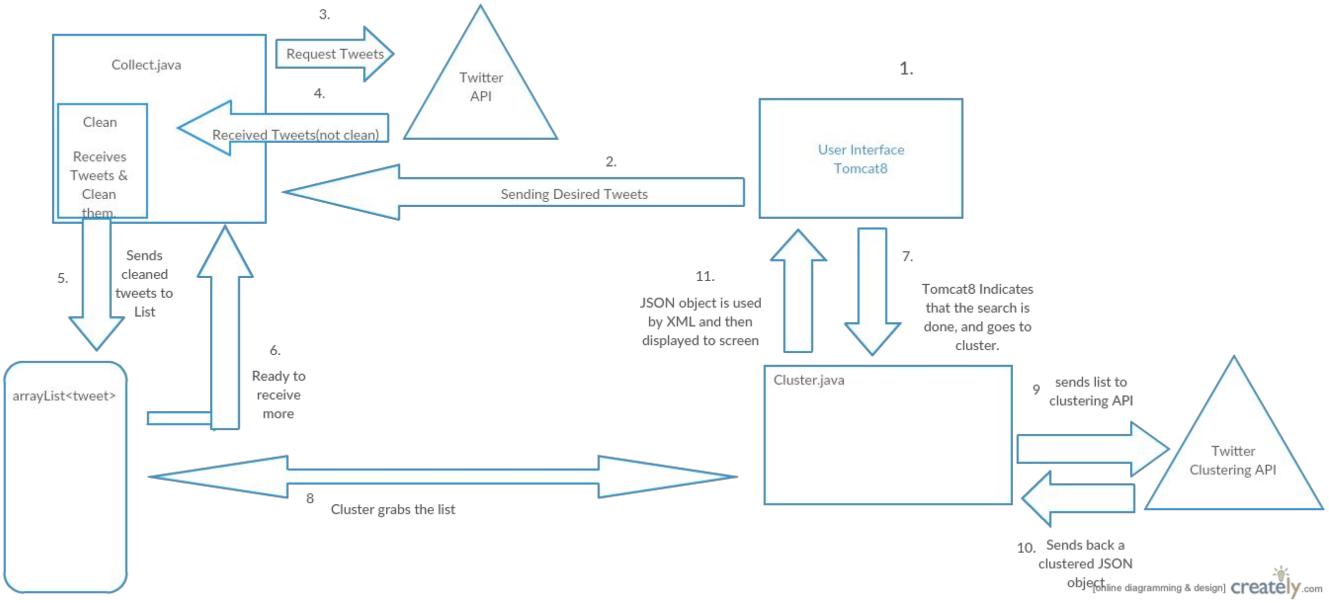
fredwalton216 : worried italians seek to sell out of bank bonds sellers heading to the exits

MemeApartman : thank you for choosing bank of... #funny #lol

MatteoAssinnata : disrupting banking: the fintech startups that are unbundling wells fargo, citi and bank of america.

Program Architecture:

We split our project into three main classes: Collect, Cluster and Tweet. The Tweet class we wrote to be able to use tweet objects. In the Collect class we connected to the Twitter API using the Twitter4j library, and collected and cleaned the tweets. The Cluster class takes the cleaned tweets and clustered them using the RxNLP's Sentence Clustering API. We then built a servlet to give our program a user interface using Apache TomCat.


Collect.java
3.
Request Tweets
Twitter API
4.
Received Tweets(not clean)
Clean
Receives Tweets & Clean them.
1.
User Interface Tomcat8
2.
Sending Desired Tweets
5.
Sends cleaned tweets to List
6.
Ready to receive more
arrayList<tweet>
11.
JSON object is used by XML and then displayed to screen
7.
Tomcat8 Indicates that the search is done, and goes to cluster.
Cluster.java
9
sends list to clustering API
Twitter Clustering API
8
Cluster grabs the list
10.
Sends back a clustered JSON object

Modules used:

Apache TomCat: used to make servlet to be able to make a user interface with a front-end and back-end.

GitHub: used to manage code changes and keep everyone's code up to date.

Slack: used to manage communication between team members.

RxNLP Sentence Clustering API: used to cluster tweet text into topics.

Twitter Streaming API: used to collect tweets from Twitter data base.

Twitter4j: used to make using Twitter Streaming API easier to use.

Mashape: used to get a key to access the RxNLP Sentence Clustering API.

XML: used to display clustered tweets as per instructors request.

JSON: used to pass data to the RxNLP Sentence Clustering and also was the format that the data was in upon return from the RxNLP Sentence Clustering.

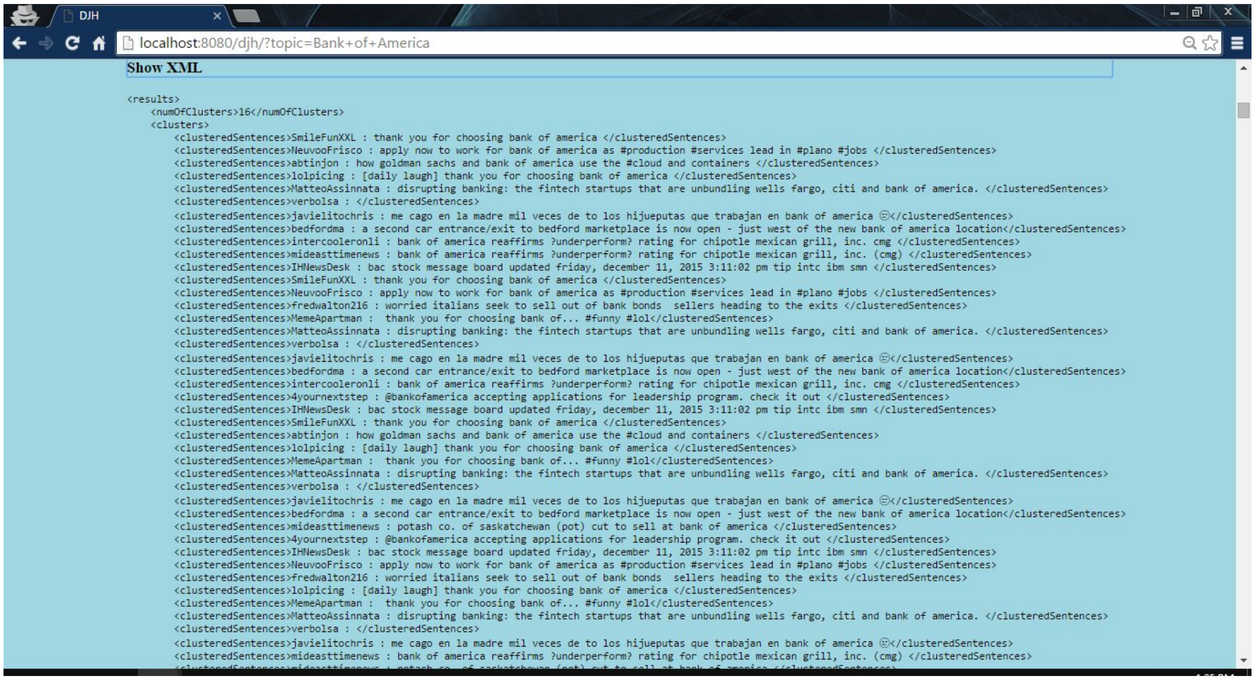
HTML: used to format and display the clustered tweets.

CSS: used to style the display of the clustered tweets.

JSP: used to call and run the program and to get the results and pass it to the JavaScript and HTML.

JAVA: used to write the core of the program; Collect, Cluster and Tweet classes.

JavaScript: used to show selected topic in the user interface.



```
<results>
    <numOfClusters>16</numOfClusters>
    <clusters>
        <clusteredSentences>SmileFunXXL : thank you for choosing bank of america </clusteredSentences>
        <clusteredSentences>NeuvooFrisco : apply now to work for bank of america as #production #services lead in #plano #jobs </clusteredSentences>
        <clusteredSentences>abtinjon : how goldman sachs and bank of america use the #cloud and containers </clusteredSentences>
        <clusteredSentences>lolpicing : [daily laugh] thank you for choosing bank of america </clusteredSentences>
        <clusteredSentences>MatteoAssinnata : disrupting banking: the fintech startups that are unbundling wells fargo, citi and bank of america. </clusteredSentences>
        <clusteredSentences>verbolsa : </clusteredSentences>
        <clusteredSentences>javielitochris : me cago en la madre mil veces de to los hijueputas que trabajan en bank of america 😑</clusteredSentences>
        <clusteredSentences>bedfordma : a second car entrance/exit to bedford marketplace is now open - just west of the new bank of america location</clusteredSentences>
        <clusteredSentences>intercooleronli : bank of america reaffirms ?underperform? rating for chipotle mexican grill, inc. cmg </clusteredSentences>
        <clusteredSentences>mideasttimenews : bank of america reaffirms ?underperform? rating for chipotle mexican grill, inc. (cmg) </clusteredSentences>
        <clusteredSentences>IHNewsDesk : bac stock message board updated friday, december 11, 2015 3:11:02 pm tip intc ibm smn </clusteredSentences>
        <clusteredSentences>SmileFunXXL : thank you for choosing bank of america </clusteredSentences>
        <clusteredSentences>NeuvooFrisco : apply now to work for bank of america as #production #services lead in #plano #jobs </clusteredSentences>
        <clusteredSentences>fredwalton216 : worried italians seek to sell out of bank bonds  sellers heading to the exits </clusteredSentences>
        <clusteredSentences>MemeApartman :  thank you for choosing bank of... #funny #lol</clusteredSentences>
        <clusteredSentences>MatteoAssinnata : disrupting banking: the fintech startups that are unbundling wells fargo, citi and bank of america. </clusteredSentences>
        <clusteredSentences>verbolsa : </clusteredSentences>
        <clusteredSentences>javielitochris : me cago en la madre mil veces de to los hijueputas que trabajan en bank of america 😑</clusteredSentences>
        <clusteredSentences>bedfordma : a second car entrance/exit to bedford marketplace is now open - just west of the new bank of america location</clusteredSentences>
        <clusteredSentences>intercooleronli : bank of america reaffirms ?underperform? rating for chipotle mexican grill, inc. cmg </clusteredSentences>
        <clusteredSentences>4yournextstep : @bankofamerica accepting applications for leadership program. check it out </clusteredSentences>
        <clusteredSentences>IHNewsDesk : bac stock message board updated friday, december 11, 2015 3:11:02 pm tip intc ibm smn </clusteredSentences>
        <clusteredSentences>SmileFunXXL : thank you for choosing bank of america </clusteredSentences>
        <clusteredSentences>abtinjon : how goldman sachs and bank of america use the #cloud and containers </clusteredSentences>
        <clusteredSentences>lolpicing : [daily laugh] thank you for choosing bank of america </clusteredSentences>
        <clusteredSentences>MemeApartman :  thank you for choosing bank of... #funny #lol</clusteredSentences>
        <clusteredSentences>MatteoAssinnata : disrupting banking: the fintech startups that are unbundling wells fargo, citi and bank of america. </clusteredSentences>
        <clusteredSentences>verbolsa : </clusteredSentences>
        <clusteredSentences>javielitochris : me cago en la madre mil veces de to los hijueputas que trabajan en bank of america 😑</clusteredSentences>
        <clusteredSentences>bedfordma : a second car entrance/exit to bedford marketplace is now open - just west of the new bank of america location</clusteredSentences>
        <clusteredSentences>mideasttimenews : potash co. of saskatchewan (pot) cut to sell at bank of america </clusteredSentences>
        <clusteredSentences>4yournextstep : @bankofamerica accepting applications for leadership program. check it out </clusteredSentences>
        <clusteredSentences>IHNewsDesk : bac stock message board updated friday, december 11, 2015 3:11:02 pm tip intc ibm smn </clusteredSentences>
        <clusteredSentences>NeuvooFrisco : apply now to work for bank of america as #production #services lead in #plano #jobs </clusteredSentences>
        <clusteredSentences>fredwalton216 : worried italians seek to sell out of bank bonds  sellers heading to the exits </clusteredSentences>
        <clusteredSentences>lolpicing : [daily laugh] thank you for choosing bank of america </clusteredSentences>
        <clusteredSentences>MemeApartman :  thank you for choosing bank of... #funny #lol</clusteredSentences>
        <clusteredSentences>MatteoAssinnata : disrupting banking: the fintech startups that are unbundling wells fargo, citi and bank of america. </clusteredSentences>
        <clusteredSentences>verbolsa : </clusteredSentences>
        <clusteredSentences>javielitochris : me cago en la madre mil veces de to los hijueputas que trabajan en bank of america 😑</clusteredSentences>
        <clusteredSentences>mideasttimenews : bank of america reaffirms ?underperform? rating for chipotle mexican grill, inc. (cmg) </clusteredSentences>
        <clusteredSentences>mideasttimenews : potash co. of saskatchewan (pot) cut to sell at bank of america </clusteredSentences>
```

Classes:

Collect class accesses the Twitter streaming API using Twitter4j. We registered our application with Twitter and received the proper tokens and keys. The collect class takes a string parameter of topic. it passes that topic to the Twitter streaming API through Twitter4j. It then loops through requesting tweets till it reaches 500. Each tweet that is read is stored in a an array list of tweets. The list is then passed to the clean function to remove duplicates and clean up the text and beep out profanity.

Tweet is a class we write to make a tweet object. Our tweet object has two String fields: 'username' and 'text'.

Cluster is passed the ArrayList of tweets from the Collect class. It takes objects from the arraylist and formats them into a JSONObject and sends it to the clustering API. The clustering API returns a JSONObject of all the cluster topics and clusters. The cluster class then maps the clustered tweets back to the original tweets in the arraylist and adds the username to the text. We found a function online that then formats the JSONObject into XML to be displayed in the browser.(http://stackoverflow.com/questions/25864316/pretty-print-xml-in-java-8/33541820#33541820)

Extra Credit:

We were able to implement a profanity filter that blocks a lot of those swear words. We also used TomCat make a servlet to be able to make a user interface with a front-end and back-end. We used JSP, HTML, CSS and JavaScript to display the user interface.

Division of Labor:

Jacob: Made Cluster and Tweet classes and setup RxNLP's Sentence Clustering. Setup servlet to display frontend using TomCat, JSP, JavaScript, HTML and CSS. Created mashape account to get API key. Also found a XML formatter function to turn the JSONObject into XML. (http://stackoverflow.com/questions/25864316/pretty-print-XML-in-java-8/33541820#33541820) Used CSS and javascript to display different tweets based on topics along with showing the XML upon click. Set up Git repository on GitHub for code collaboration. Figured out how to import all the external libraries via adding the jar files to build path. Wrote documentation and setup instructions using GitHub Wiki. Set up Slack communication tool to improve communication. Kept everyone informed and on the same page by communicating with team members often.

Helam: Implemented the cleaning part of the Twitter program. This included removing duplicate tweets, lowercase text, remove unwanted text (such as @,!,hyperlinks, etc). This was achieved by using regular expressions. Regex was a challenge to implement particularly the URL Regex and figuring out how to remove them from any string. The initial approach to solve the Regex url problem was making a huge ‘if, else’ statement that would convert the string to a char array, and search for the http, but after some help from other friends, it was finalized that this regex worked on a URL :
“(https?|http?|ftp|file)://[a-zA-Z0-9+&@#/%?=~_!:,.;]*[-a-zA-Z0-9+&@#/%=~_|]”
Allowed us to remove the substring of the url and everything after it until a whitespace. Also implemented a profanity filter.

David: Wrote getProperties, getAccess and writeProperties classes in the collect class. Created Twitter account, registered the application with Twitter. Read through the Twitter API. Helped implement the Twitter4j. Maintained communication though slack communication tool.

Issues:

We struggled adding the jar files for libraries we needed to use into our project. We also struggled working on the same project together, we fixed that problem by using GitHub to manage our code changes. We struggled with communication but we fixed that by setting up Slack communication tool. We also struggled because a member of our group (Jorge) dropped the class and did not tell us. He had been assigned to work on the clustering class and we thought he was working on it but we found out a week before the presentation that he had dropped the class and had not done any work on the project. This sent us into panic mode as we tried to catch up and get the project together.