A Project Report On

"Analysis on Banking Industry"

Submitted in partial fulfillment of requirements of the course

"Principles of Big Data Management"

For the degree of

Master's in Computer Science

By

Nandanamudi, Jyothikiran	(jnyc8@mail.umkc.edu)	(16225795)
Swargam, Gopi Krishna	(gs446@mail.umkc.edu)	(16222026)
Junaid, Sidrah	(sjhv6@mail.umkc.edu)	(16225745)

Under the esteemed guidance of

Dr. Praveen Rao

Abstract:

The banking industry has been a revolution over a period of time. Banks today are growing swiftly and are being adaptable to inevitable changes being occurred across technology, customer nature and regulation. Nowadays, modern banking sector is shifting towards to technological trends to reach more number of customers from day to day. As banking system becomes more user friendly, customers still covet of human connection. Today, customers are able to quickly raise their issues through social media network which is acting as a social interaction medium" which puts greater pressure on the banks to adapt with the new solutions and new business models.

Even Banks are using social media to a greater extent to monitor their brand identity, testing new products like credit cards, debit cards, flexi loans, personal loans, vehicle loans and many other products are in line as well banks are using social media to get feedback about this products. Banks use social media for conversations and building trust. Even, employment opportunities are being provided using social media.

There are billions of active users of social media networks like Twitter, Facebook etc., worldwide, who are frequently active and connected by means of smartphones, computer and tablets. The posts and blogs which are being posted daily might be structured or unstructured data. These big data are waiting to be explored and as we see the social media embodies the leading and biggest source of consumer data. With the good management of Big Data it can lead to many great insights; we can find the root cause of the problems and failure that affects the revenues of the business.

Twitter is a fabulous source of information. Whenever something is happening people around the world start tweeting away. Twitter's kind of a feasible place to look for big data. It's an example of the ubiquitous services used by consumers and businesses alike that helps to generate this huge amount of data in the first place.

There are a number of types of software tools for analyzing unstructured data found in tweets. *Apache Spark™* is a fast and general engine for large-scale data processing. Spark offers a wide variety such as 80 high level operators that make it easy to build parallel operating apps. And you can use it interactively from the Scala, Python and R shells and you can write applications quickly in Java, Scala, Python, R.

It is an open source big data accessing framework built around speed, ease of use, and sophisticated analytics. Spark enables applications in Hadoop clusters to run up to 100 times faster in memory and 10 times speed even when running on disk. In addition to Map and Reduce functions, it supports queries written in SQL, streaming data, machine learning and graph data processing.

Table of Contents

I. Introduction:

Development of Analytical application for analyzing the twitter data collected for better understanding and for better decision making. We chose "Banking and other financial sectors" as our project concept for analyzing.

Banks find Twitter an indispensable tool for communications. Our project proposes eight different analytical queries on the Banking Industry on the data collected from twitter. One of the main feature of spark is speed and it has ability to run computations and operations in memory, but the system is also more efficient than MapReduce for complex applications running on disk.

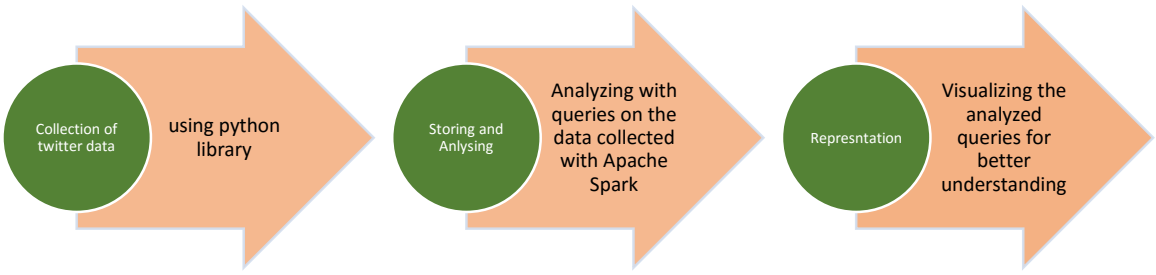
Spark SQL is Spark's package made available for working with structured data. Spark SQL allows developers to mix SQL queries with the programmatic data manipulations which are supported by by RDDs in Python, Java, and Scala, all within a single application, thus combining SQL with complex analytics. We used Spark SQL to write analytical queries in Java.

Big Data is something of a buzzword, but the data can be so complex and it is incomprehensible. Data visualization makes it easy to understand as it distill the lots of information into human understandable form typically focusing on the aggregates of the data. We made use of d3.js for visualization of our proposed analytical queries.

Our project draws few insights from the available data of the twitter based on the tweets posted by banks, their users and by common people who wants to share their opinion about the banking products and banking industry.

i. Analytical procedure

The sequence followed in analyzing the data is as shown below,



We used python streaming library to collect data being posted on banks in valid JSON format and is stored into a text file. We have collected almost 1200Mb of data for our analysis. We used spark locally in our Java work space for writing analytical queries on the collected data and has been visualized with d3.js.

ii. Software interface

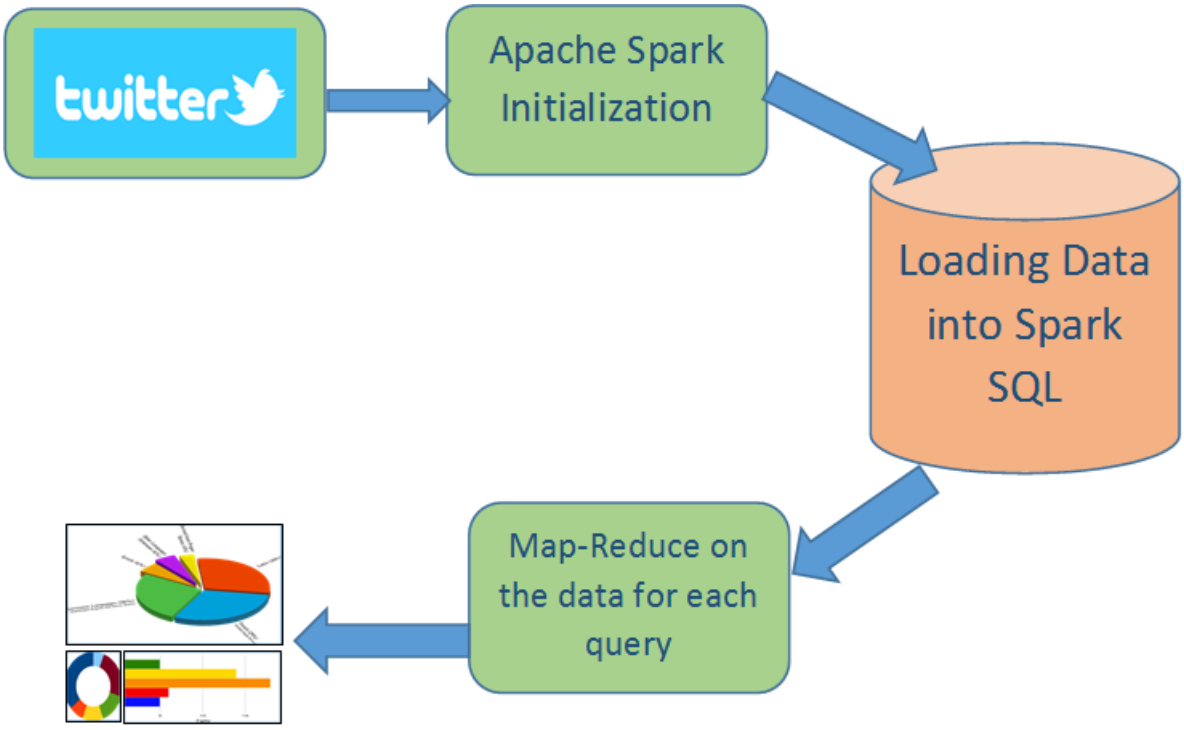
1. Client side: HTML, HTML5, JavaScript, d3.js, high charts
2. Framework: Apache Spark, Eclipse modelling framework.
3. Date base: Spark SQL

II. Architectural Design

The overall design of our project is shown in steps.

- Collection of data
- Load our input data.
- Parse our input into words.
- Reduce our words into a tuple pair that contains the word and the count of occurrences.
- Save our results.
- Visualizing the results saved with charts and graphs.

The above steps is implemented as in the below architecture diagram.



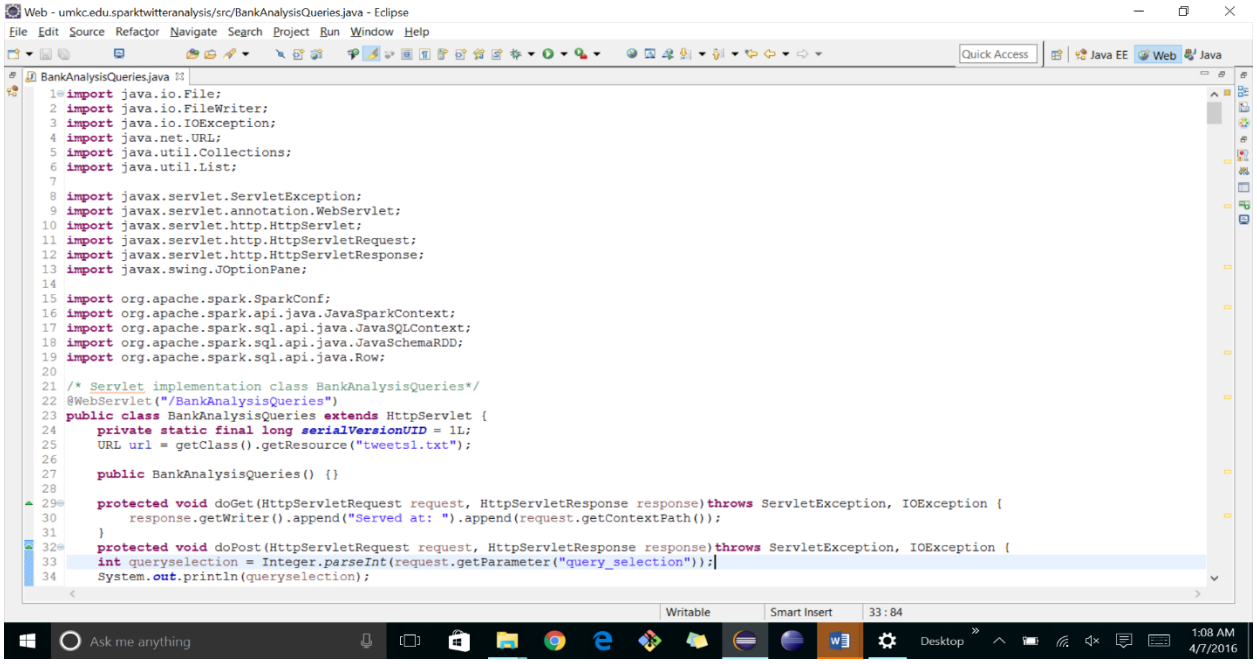
As represented above, tweets collected will be loaded into Apache Spark and then stored into Spark SQL. We will run queries which will execute map-reduce on the twitter data and will display the result.

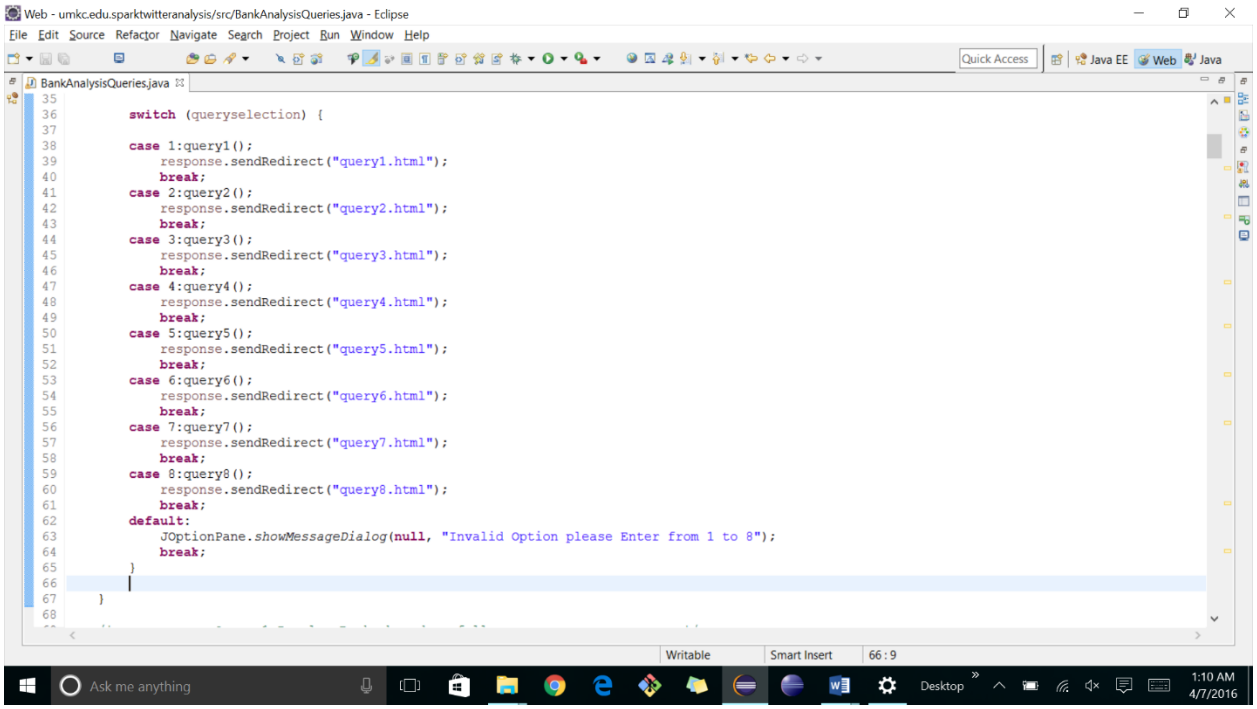
III. Project Implementation

As spark supports different programming languages Java is our choice. We used eclipse framework for implementing the Analytics using Apache Spark.

We used Spark locally in our Java process space. We created Sparkconf object which points to our Spark instance as in this case 'local' and in order to interact with Spark we need a SparkContext instance, so we created JavaSparkContext. We stored the tweets in Spark SQL database under tweetable and sql query is run on this table for selection of required data for analytics.

a. Source Code

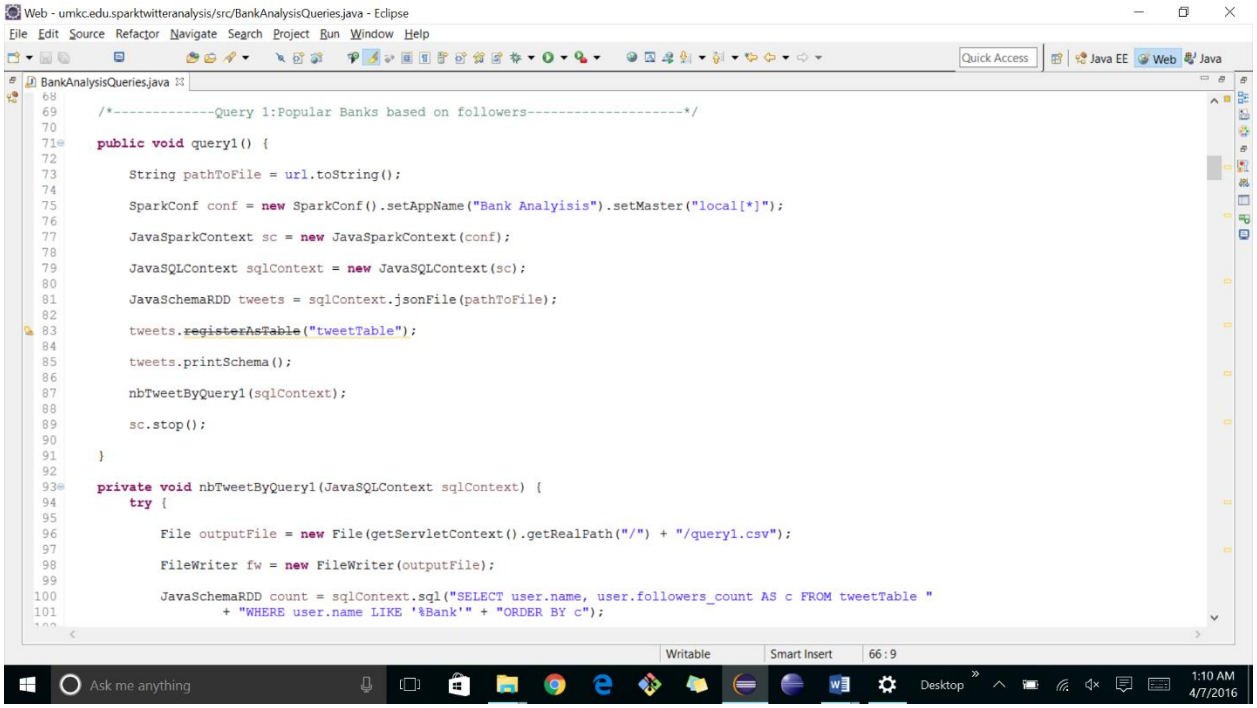


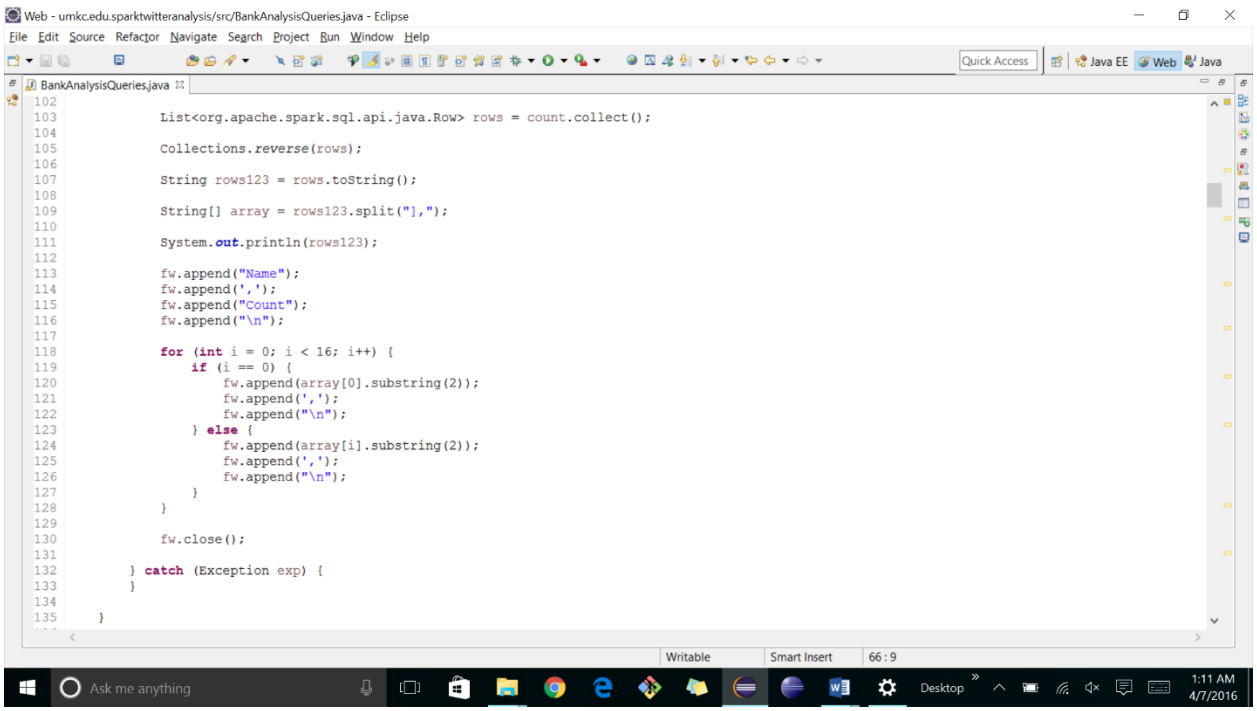


a. Analytical Queries:

Query 1: Popular Banks internationally based on their followers

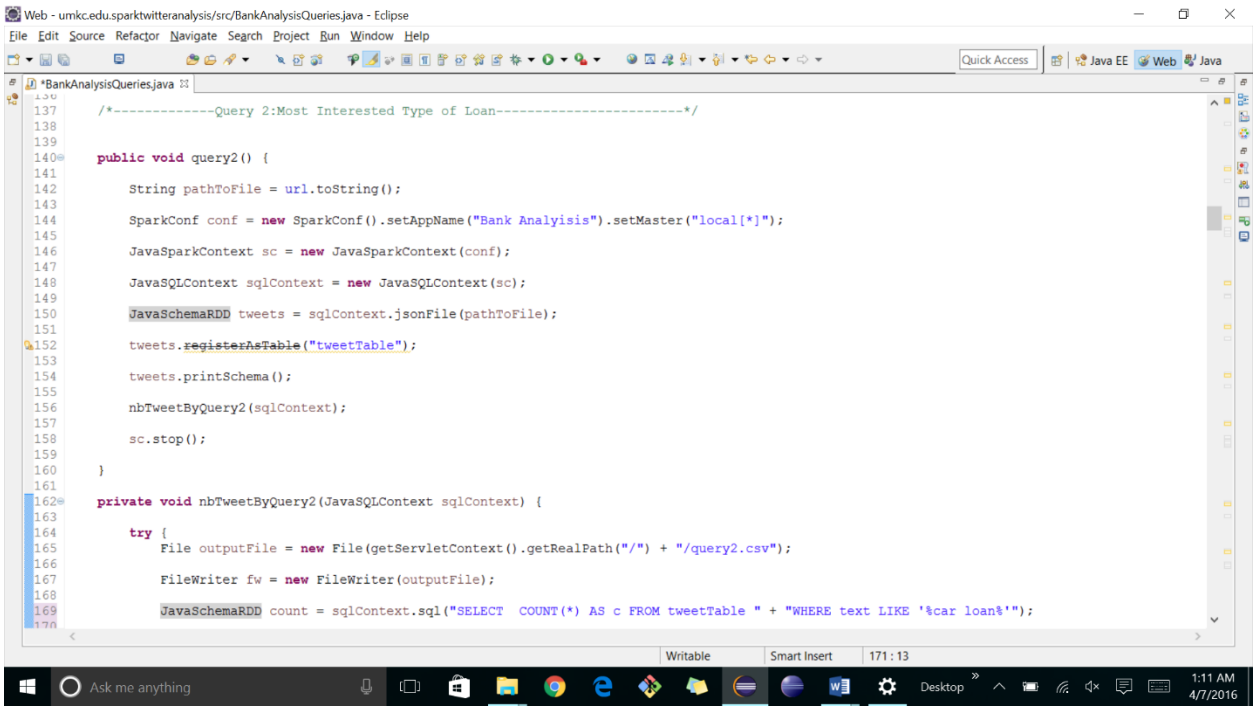
As there are lot of banks all over the world, people often would like to know the updates of some particular banks and often comment on them and usually follow some banks. So we did analysis based on this followers count and selected top banks.



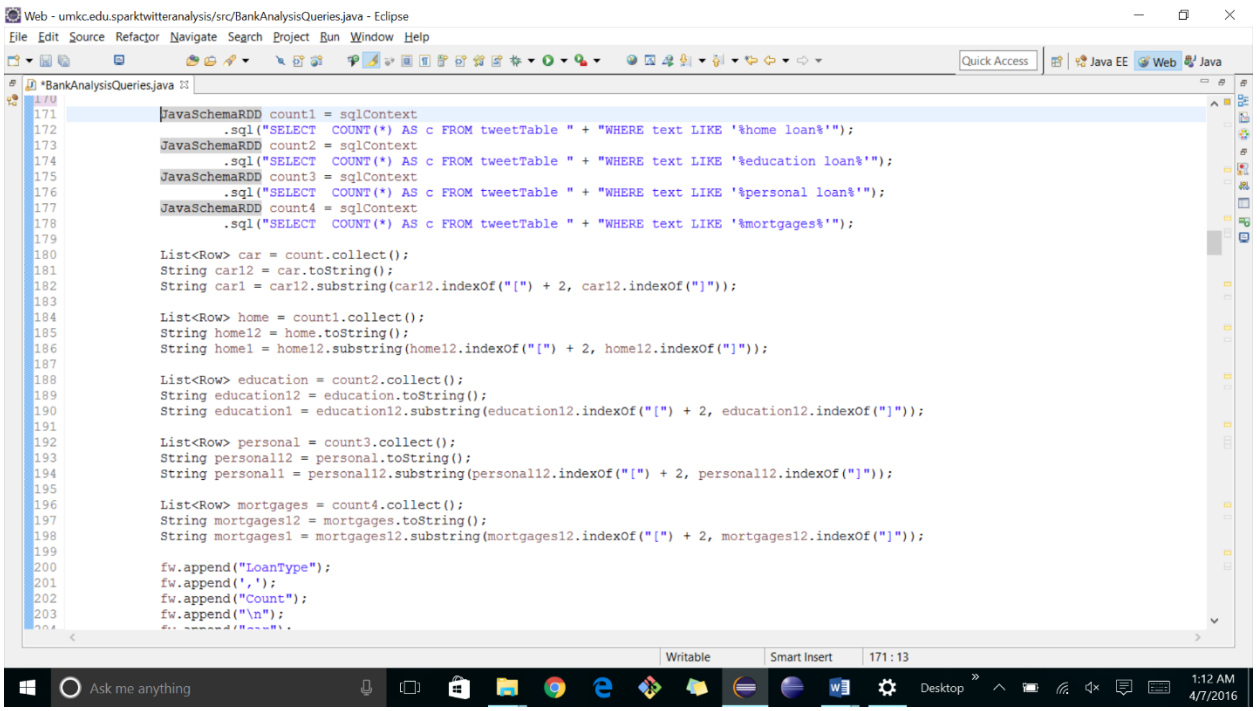


Query 2: Most interested type of loan

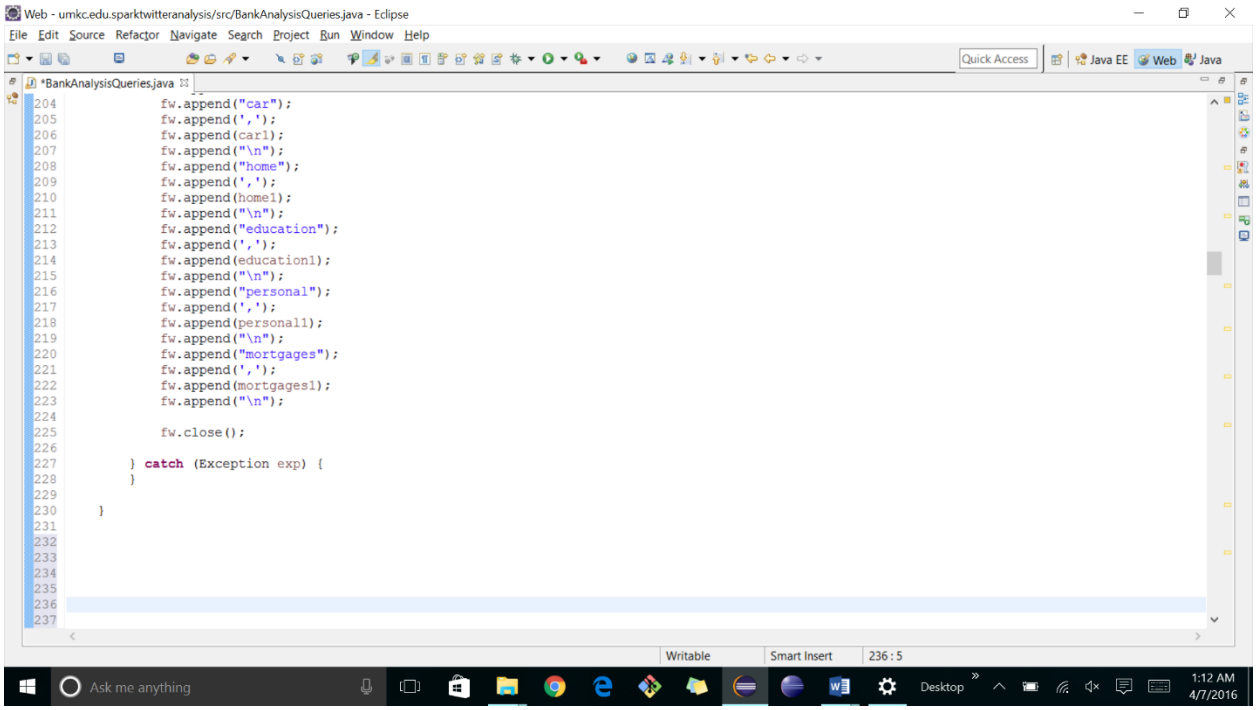
As we know banks are the sources of financial resources in terms of providing loans for the people when there is a need and often provides different offers on these loans. People do often tweet about this loans and wanted to know about these loans. So we did analysis on the most popular loans offered by the banks and what type of loans are most popular over the twitter.



```
/*-------------Query 2:Most Interested Type of Loan-----------------------*/


public void query2() {

    String pathToFile = url.toString();

    SparkConf conf = new SparkConf().setAppName("Bank Analyisis").setMaster("local[*]");

    JavaSparkContext sc = new JavaSparkContext(conf);

    JavaSQLContext sqlContext = new JavaSQLContext(sc);

    JavaSchemaRDD tweets = sqlContext.jsonFile(pathToFile);

    tweets.registerAsTable("tweetTable");

    tweets.printSchema();

    nbTweetByQuery2(sqlContext);

    sc.stop();

}

private void nbTweetByQuery2(JavaSQLContext sqlContext) {

    try {
        File outputFile = new File(getServletContext().getRealPath("/") + "/query2.csv");

        FileWriter fw = new FileWriter(outputFile);

        JavaSchemaRDD count = sqlContext.sql("SELECT  COUNT(*) AS c FROM tweetTable " + "WHERE text LIKE '%car loan%'");
```

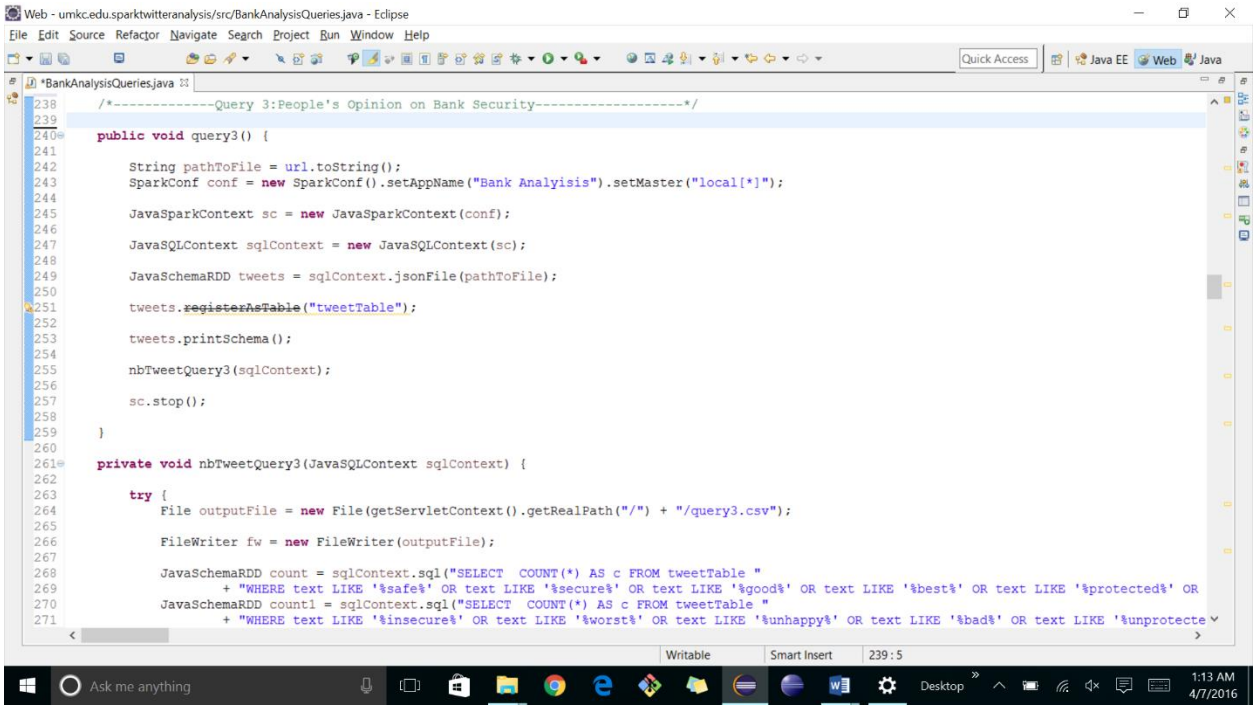


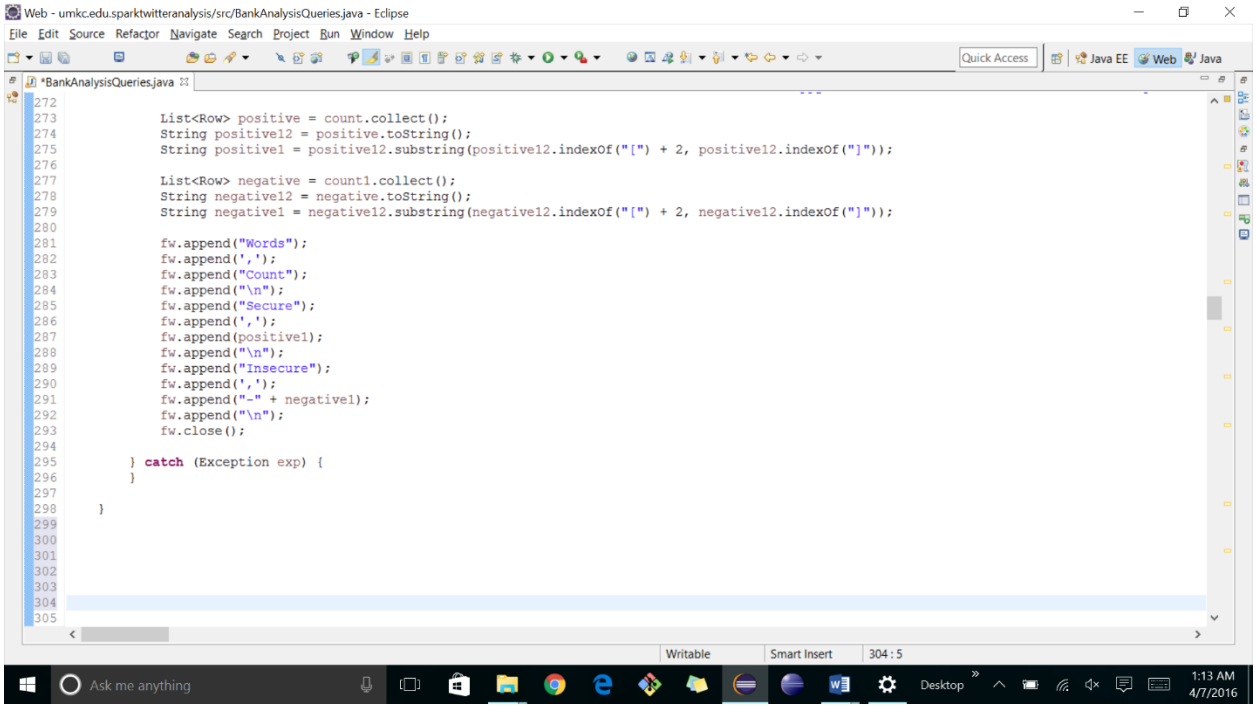
```
        JavaSchemaRDD count1 = sqlContext
                .sql("SELECT  COUNT(*) AS c FROM tweetTable " + "WHERE text LIKE '%home loan%'");
        JavaSchemaRDD count2 = sqlContext
                .sql("SELECT  COUNT(*) AS c FROM tweetTable " + "WHERE text LIKE '%education loan%'");
        JavaSchemaRDD count3 = sqlContext
                .sql("SELECT  COUNT(*) AS c FROM tweetTable " + "WHERE text LIKE '%personal loan%'");
        JavaSchemaRDD count4 = sqlContext
                .sql("SELECT  COUNT(*) AS c FROM tweetTable " + "WHERE text LIKE '%mortgages%'");

        List<Row> car = count.collect();
        String car12 = car.toString();
        String car1 = car12.substring(car12.indexOf("[") + 2, car12.indexOf("]"));

        List<Row> home = count1.collect();
        String home12 = home.toString();
        String home1 = home12.substring(home12.indexOf("[") + 2, home12.indexOf("]"));

        List<Row> education = count2.collect();
        String education12 = education.toString();
        String education1 = education12.substring(education12.indexOf("[") + 2, education12.indexOf("]"));

        List<Row> personal = count3.collect();
        String personal12 = personal.toString();
        String personal1 = personal12.substring(personal12.indexOf("[") + 2, personal12.indexOf("]"));

        List<Row> mortgages = count4.collect();
        String mortgages12 = mortgages.toString();
        String mortgages1 = mortgages12.substring(mortgages12.indexOf("[") + 2, mortgages12.indexOf("]"));

        fw.append("LoanType");
        fw.append(',');
        fw.append("Count");
        fw.append("\n");
```


Web - umkc.edu.sparktwitteranalysis/src/BankAnalysisQueries.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
Quick Access
Java EE
Web
Java
*BankAnalysisQueries.java
fw.append("car");
fw.append(',');
fw.append(car1);
fw.append("\n");
fw.append("home");
fw.append(',');
fw.append(home1);
fw.append("\n");
fw.append("education");
fw.append(',');
fw.append(education1);
fw.append("\n");
fw.append("personal");
fw.append(',');
fw.append(personal1);
fw.append("\n");
fw.append("mortgages");
fw.append(',');
fw.append(mortgages1);
fw.append("\n");
fw.close();
} catch (Exception exp) {
}
}
Writable
Smart Insert
236 : 5
Ask me anything
Desktop
1:12 AM
4/7/2016

Query 3: People's Opinion on Bank Security

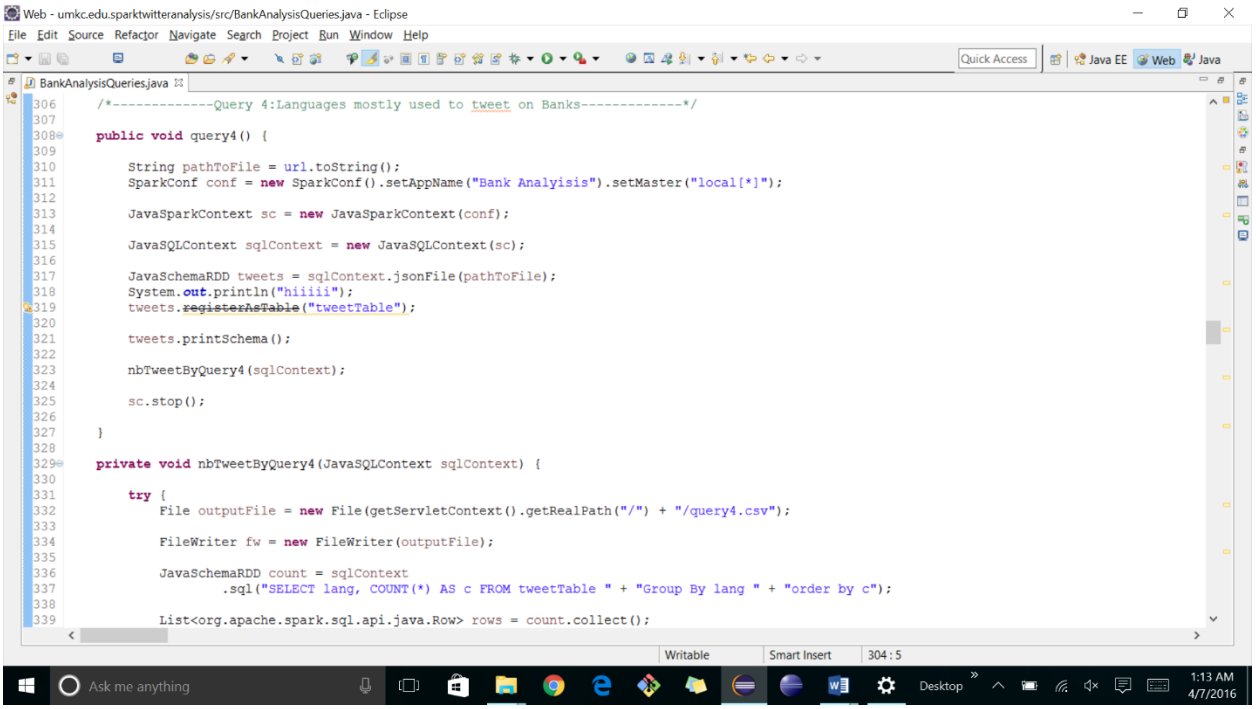
As there is a mixed opinion on banks when comes to security, service etc., from the customers as well as people. When we heard of robbery or fraud at a bank people would like to share and express their feelings. And banks do often reach their customers to ensure their relation. So, we did sentimental analysis based on the tweets collected i.e., what people think most about banks.

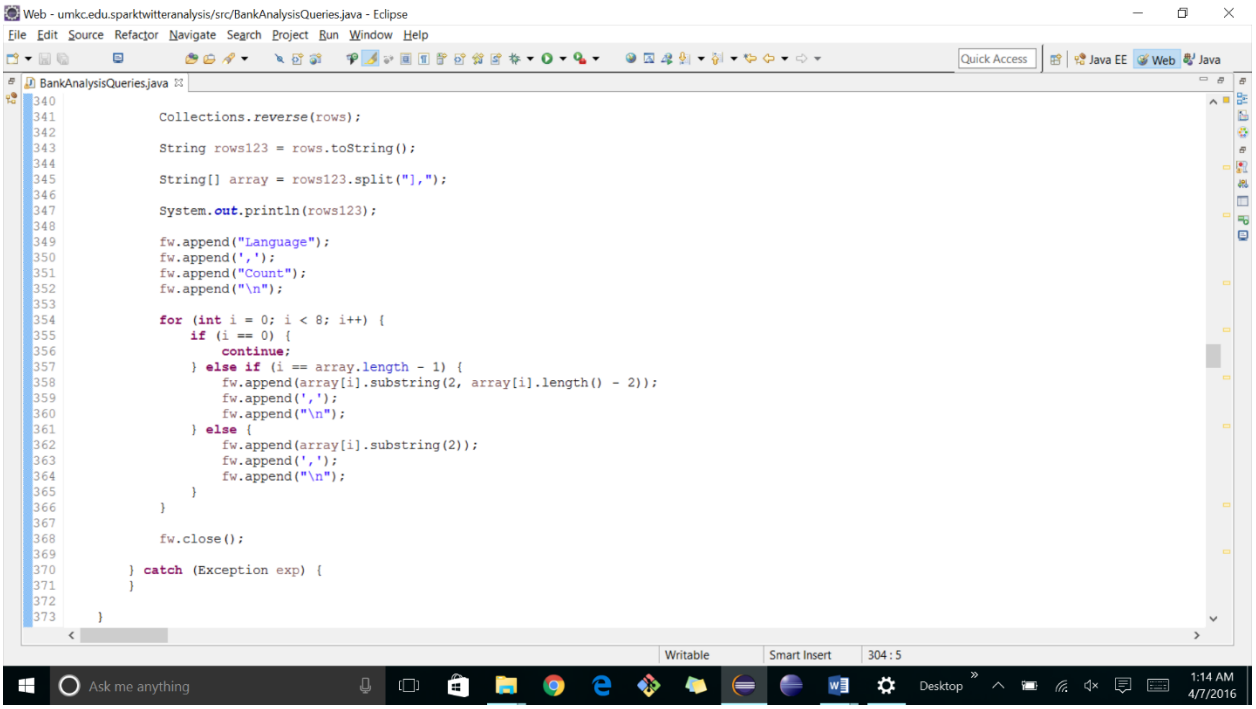




Query 4: Languages mostly used to tweet about banks

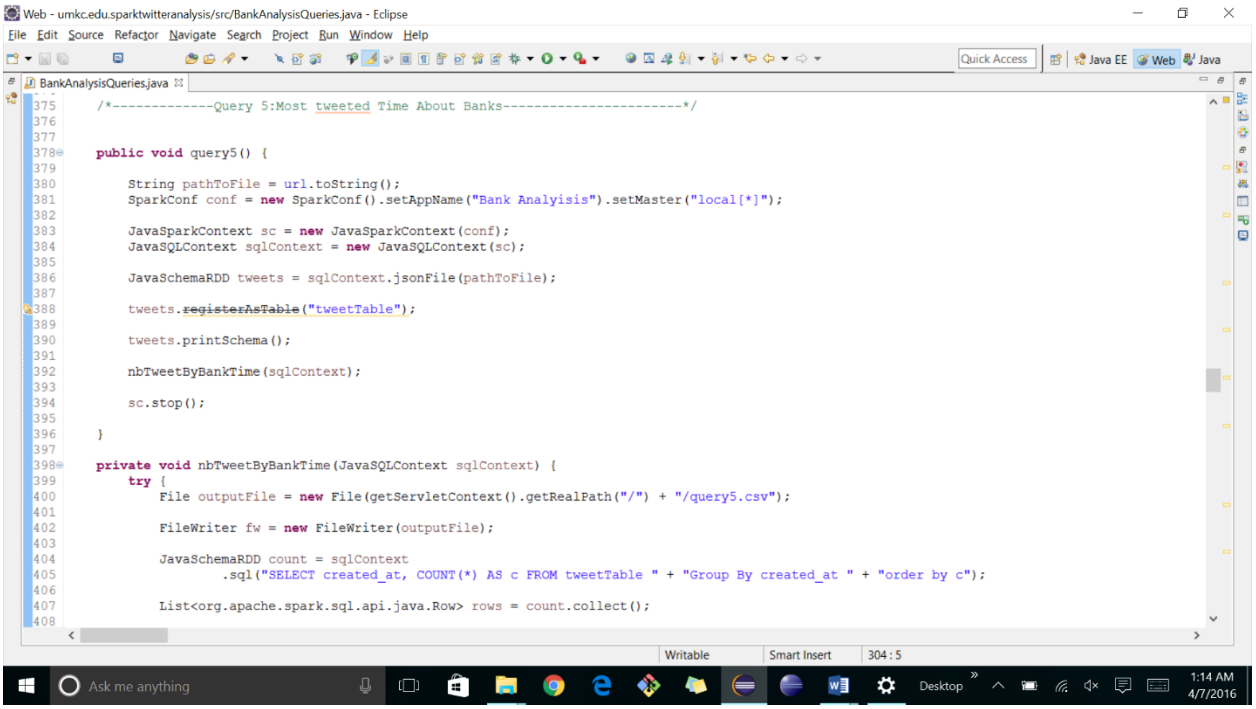
Twitter is social network which is being used almost worldwide. So, people of different areas uses different languages to tweet about banks. So we did analysis based on the language for finding the most used languages to tweet by the people all over the world.



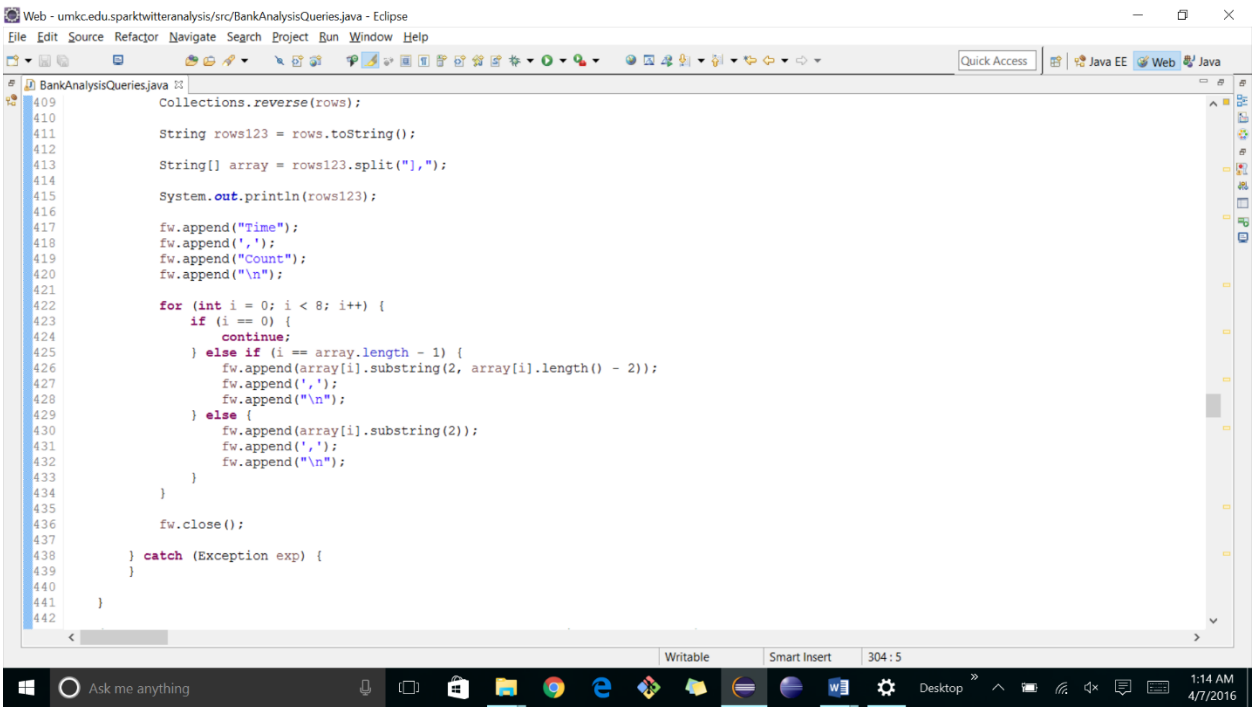


Query 5: Most tweeted specific time at a particular instant about banks

As we all see, when there is some interesting thing is happening regarding banks or well something unexpected happens then people wants to share with their friends and family. So, based on this we did analysis on at what specific time many tweets were recorded about banks.



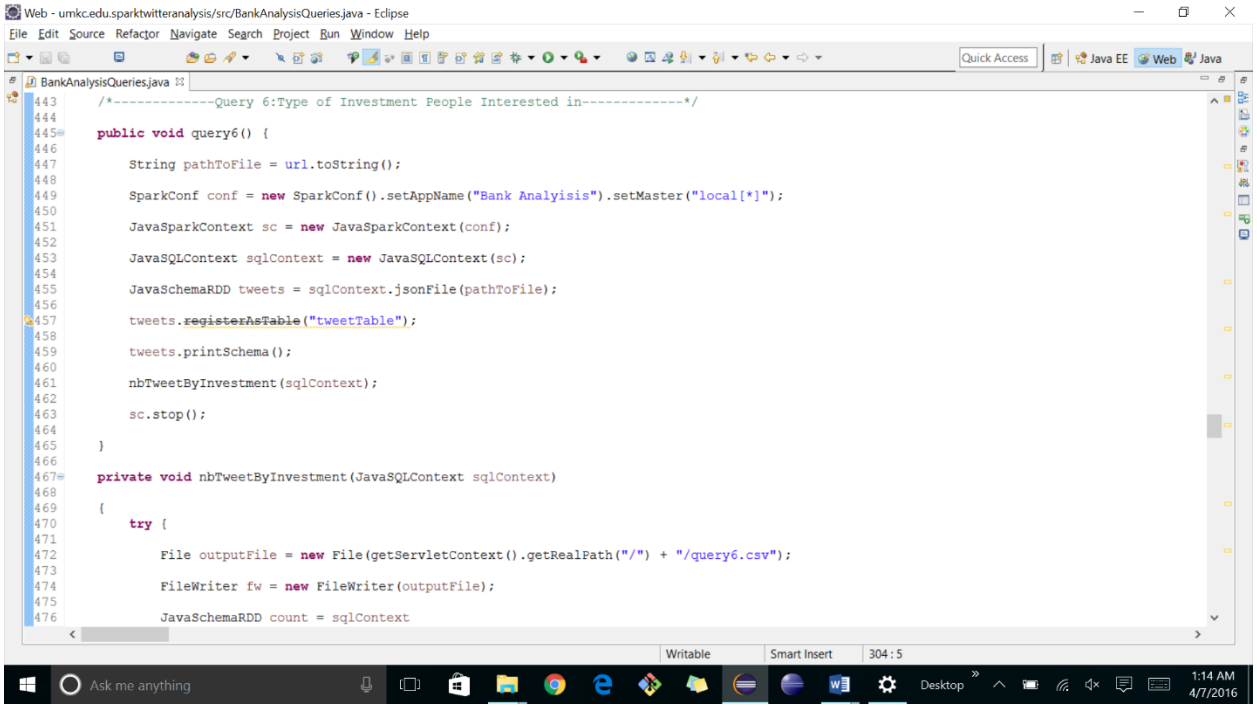
```
/*-------------Query 5:Most tweeted Time About Banks----------------------*/


public void query5() {

    String pathToFile = url.toString();
    SparkConf conf = new SparkConf().setAppName("Bank Analyisis").setMaster("local[*]");

    JavaSparkContext sc = new JavaSparkContext(conf);
    JavaSQLContext sqlContext = new JavaSQLContext(sc);

    JavaSchemaRDD tweets = sqlContext.jsonFile(pathToFile);

    tweets.registerAsTable("tweetTable");

    tweets.printSchema();

    nbTweetByBankTime(sqlContext);

    sc.stop();

}

private void nbTweetByBankTime(JavaSQLContext sqlContext) {
    try {
        File outputFile = new File(getServletContext().getRealPath("/") + "/query5.csv");

        FileWriter fw = new FileWriter(outputFile);

        JavaSchemaRDD count = sqlContext
                .sql("SELECT created_at, COUNT(*) AS c FROM tweetTable " + "Group By created_at " + "order by c");

        List<org.apache.spark.sql.api.java.Row> rows = count.collect();
```



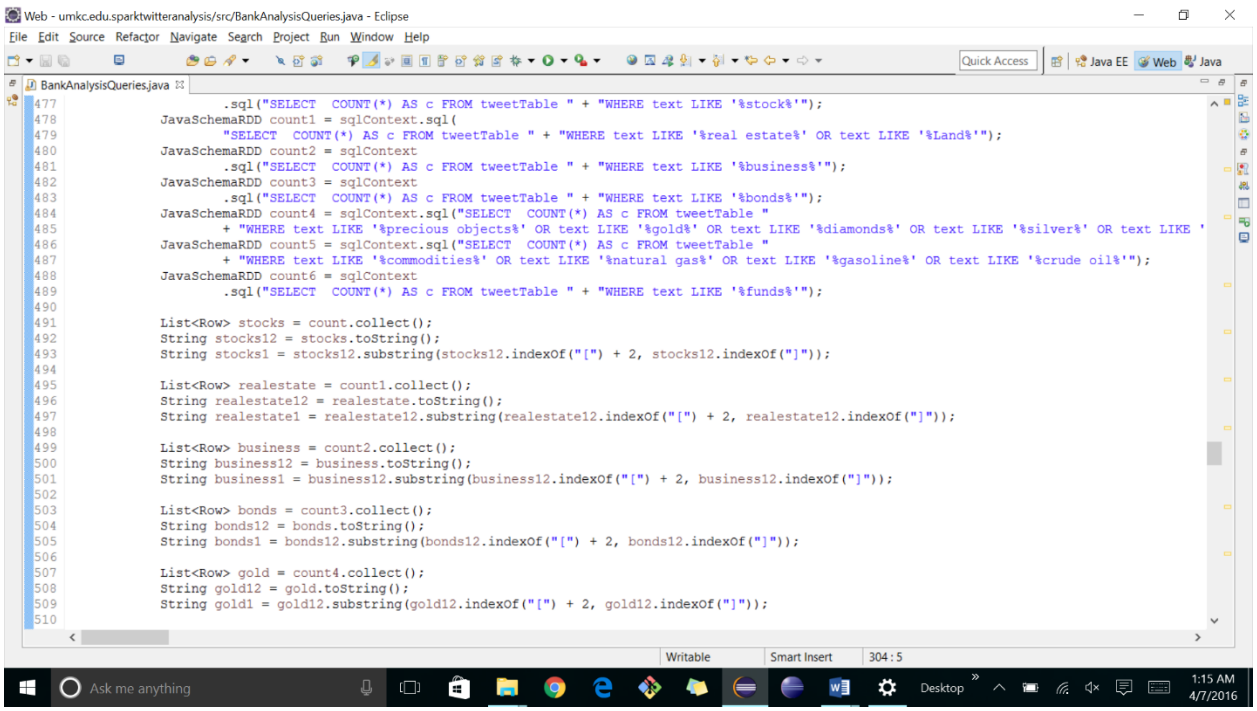
```
        Collections.reverse(rows);

        String rows123 = rows.toString();

        String[] array = rows123.split("],");

        System.out.println(rows123);

        fw.append("Time");
        fw.append(',');
        fw.append("Count");
        fw.append("\n");

        for (int i = 0; i < 8; i++) {
            if (i == 0) {
                continue;
            } else if (i == array.length - 1) {
                fw.append(array[i].substring(2, array[i].length() - 2));
                fw.append(',');
                fw.append("\n");
            } else {
                fw.append(array[i].substring(2));
                fw.append(',');
                fw.append("\n");
            }
        }

        fw.close();

    } catch (Exception exp) {
    }

}
```

Query 6: Type of investment people interested in most

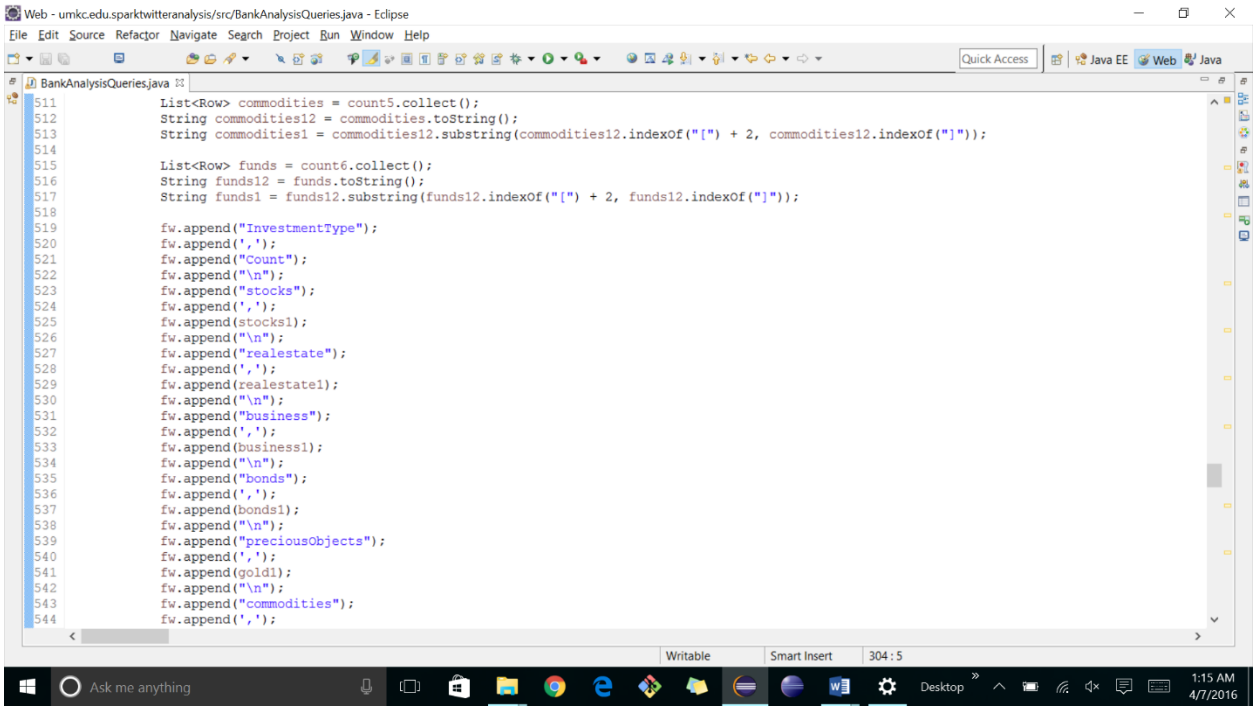
People do often saves money they earned in banks. Banks will invest these money in different investment sectors like real estate, stocks, business etc. and people do often talk about different type of investment and wants to gain some prior knowledge before investing. So, they do often tweet about these different investments. We did analysis based on this to find what type of investment people want to know about or what interests them most.

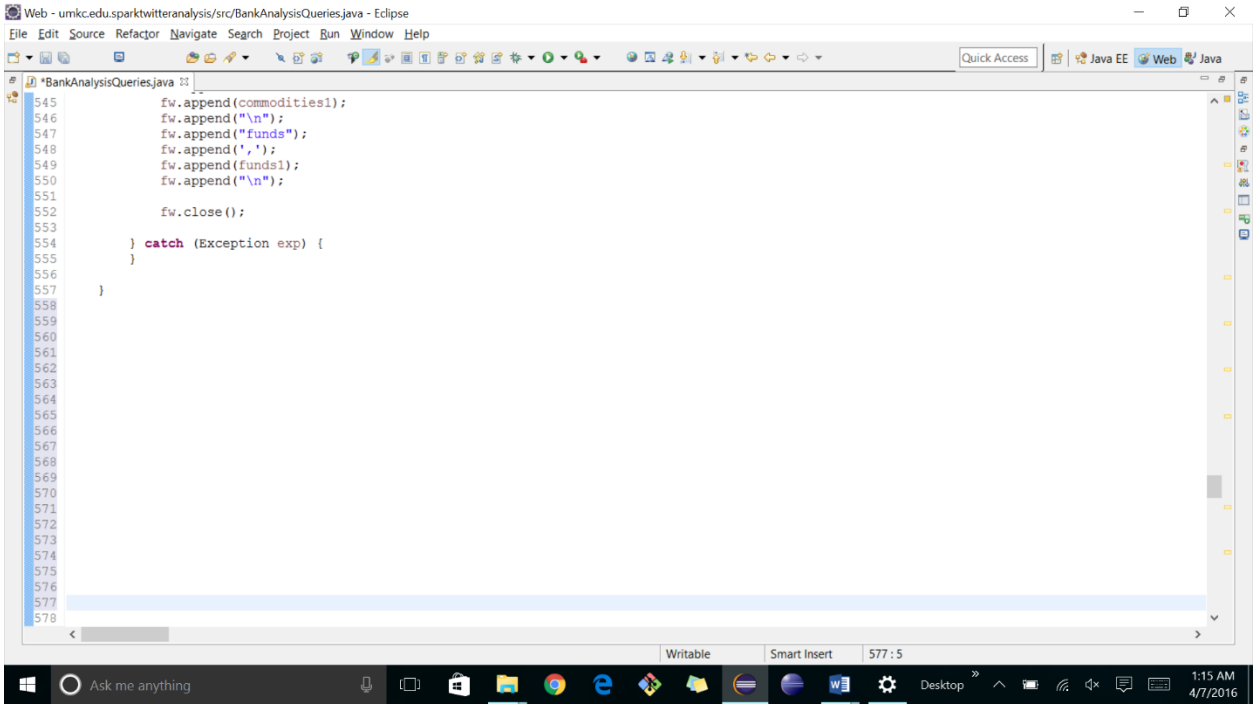


```
/*-------------Query 6:Type of Investment People Interested in-------------*/

public void query6() {

    String pathToFile = url.toString();

    SparkConf conf = new SparkConf().setAppName("Bank Analysis").setMaster("local[*]");

    JavaSparkContext sc = new JavaSparkContext(conf);

    JavaSQLContext sqlContext = new JavaSQLContext(sc);

    JavaSchemaRDD tweets = sqlContext.jsonFile(pathToFile);

    tweets.registerAsTable("tweetTable");

    tweets.printSchema();

    nbTweetByInvestment(sqlContext);

    sc.stop();

}

private void nbTweetByInvestment(JavaSQLContext sqlContext)

{
    try {

        File outputFile = new File(getServletContext().getRealPath("/") + "/query6.csv");

        FileWriter fw = new FileWriter(outputFile);

        JavaSchemaRDD count = sqlContext
```



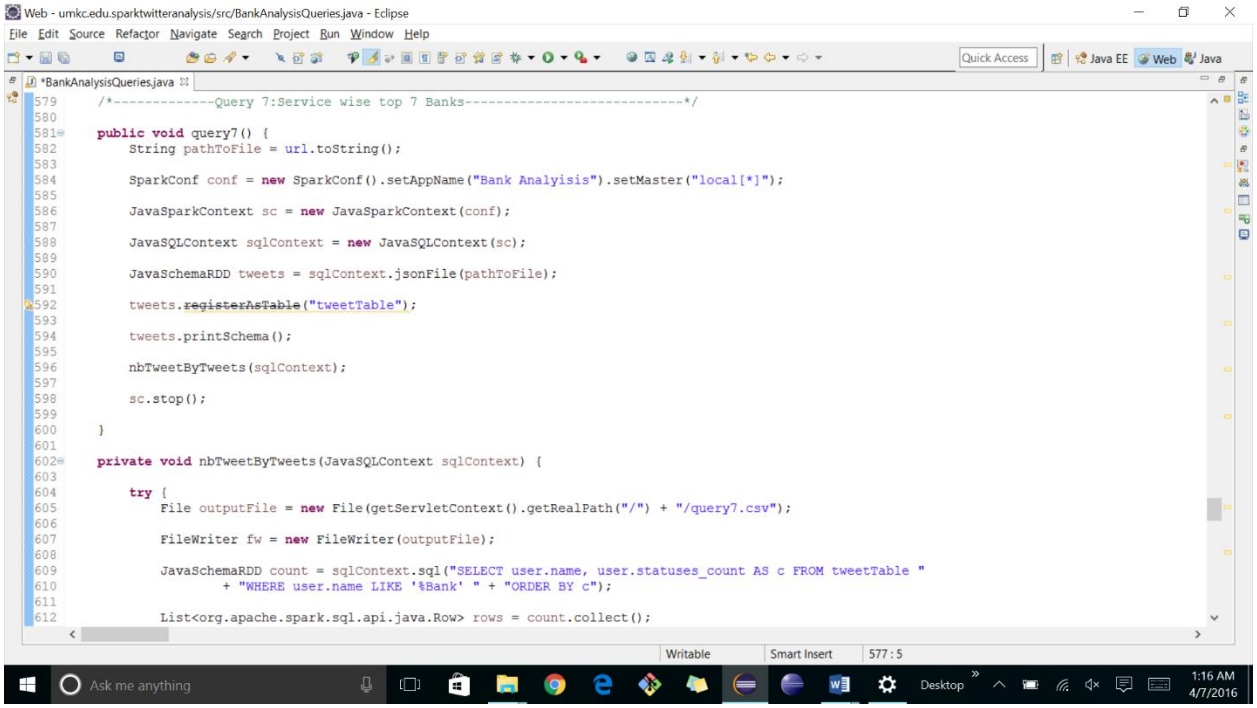
```
                .sql("SELECT  COUNT(*) AS c FROM tweetTable " + "WHERE text LIKE '%stock%'");
        JavaSchemaRDD count1 = sqlContext.sql(
                "SELECT  COUNT(*) AS c FROM tweetTable " + "WHERE text LIKE '%real estate%' OR text LIKE '%Land%'");
        JavaSchemaRDD count2 = sqlContext
                .sql("SELECT  COUNT(*) AS c FROM tweetTable " + "WHERE text LIKE '%business%'");
        JavaSchemaRDD count3 = sqlContext
                .sql("SELECT  COUNT(*) AS c FROM tweetTable " + "WHERE text LIKE '%bonds%'");
        JavaSchemaRDD count4 = sqlContext.sql("SELECT  COUNT(*) AS c FROM tweetTable "
                + "WHERE text LIKE '%precious objects%' OR text LIKE '%gold%' OR text LIKE '%diamonds%' OR text LIKE '%silver%' OR text LIKE '
        JavaSchemaRDD count5 = sqlContext.sql("SELECT  COUNT(*) AS c FROM tweetTable "
                + "WHERE text LIKE '%commodities%' OR text LIKE '%natural gas%' OR text LIKE '%gasoline%' OR text LIKE '%crude oil%'");
        JavaSchemaRDD count6 = sqlContext
                .sql("SELECT  COUNT(*) AS c FROM tweetTable " + "WHERE text LIKE '%funds%'");

        List<Row> stocks = count.collect();
        String stocks12 = stocks.toString();
        String stocks1 = stocks12.substring(stocks12.indexOf("[") + 2, stocks12.indexOf("]"));

        List<Row> realestate = count1.collect();
        String realestate12 = realestate.toString();
        String realestate1 = realestate12.substring(realestate12.indexOf("[") + 2, realestate12.indexOf("]"));

        List<Row> business = count2.collect();
        String business12 = business.toString();
        String business1 = business12.substring(business12.indexOf("[") + 2, business12.indexOf("]"));

        List<Row> bonds = count3.collect();
        String bonds12 = bonds.toString();
        String bonds1 = bonds12.substring(bonds12.indexOf("[") + 2, bonds12.indexOf("]"));

        List<Row> gold = count4.collect();
        String gold12 = gold.toString();
        String gold1 = gold12.substring(gold12.indexOf("[") + 2, gold12.indexOf("]"));
```

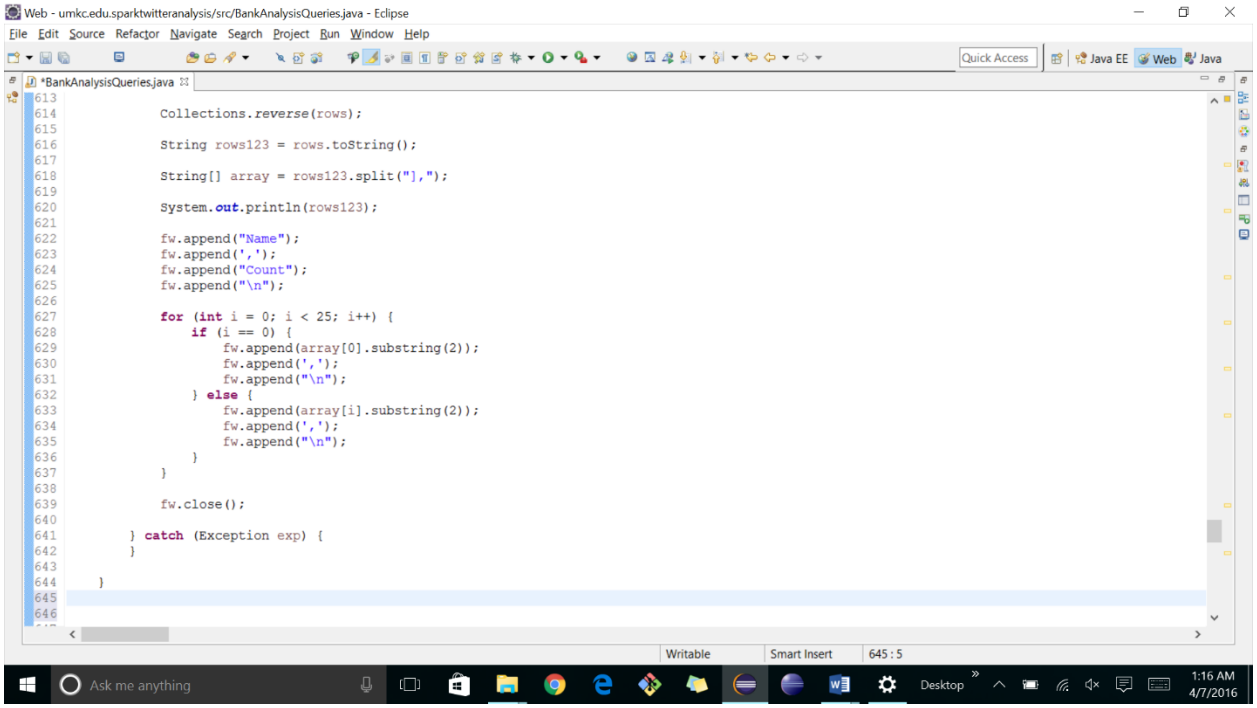

Web - umkc.edu.sparktwitteranalysis/src/BankAnalysisQueries.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
Quick Access
Java EE
Web
Java
BankAnalysisQueries.java
List<Row> commodities = count5.collect();
String commodities12 = commodities.toString();
String commodities1 = commodities12.substring(commodities12.indexOf("[") + 2, commodities12.indexOf("]"));
List<Row> funds = count6.collect();
String funds12 = funds.toString();
String funds1 = funds12.substring(funds12.indexOf("[") + 2, funds12.indexOf("]"));
fw.append("InvestmentType");
fw.append(',');
fw.append("Count");
fw.append("\n");
fw.append("stocks");
fw.append(',');
fw.append(stocks1);
fw.append("\n");
fw.append("realestate");
fw.append(',');
fw.append(realestate1);
fw.append("\n");
fw.append("business");
fw.append(',');
fw.append(business1);
fw.append("\n");
fw.append("bonds");
fw.append(',');
fw.append(bonds1);
fw.append("\n");
fw.append("preciousObjects");
fw.append(',');
fw.append(gold1);
fw.append("\n");
fw.append("commodities");
fw.append(',');
Writable
Smart Insert
304 : 5
Ask me anything
Desktop
1:15 AM
4/7/2016


Web - umkc.edu.sparktwitteranalysis/src/BankAnalysisQueries.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
Quick Access
Java EE
Web
Java
*BankAnalysisQueries.java
fw.append(commodities1);
fw.append("\n");
fw.append("funds");
fw.append(',');
fw.append(funds1);
fw.append("\n");
fw.close();
} catch (Exception exp) {
}
}
Writable
Smart Insert
577 : 5
Ask me anything
Desktop
1:15 AM
4/7/2016

Query 7: Service wise top banks

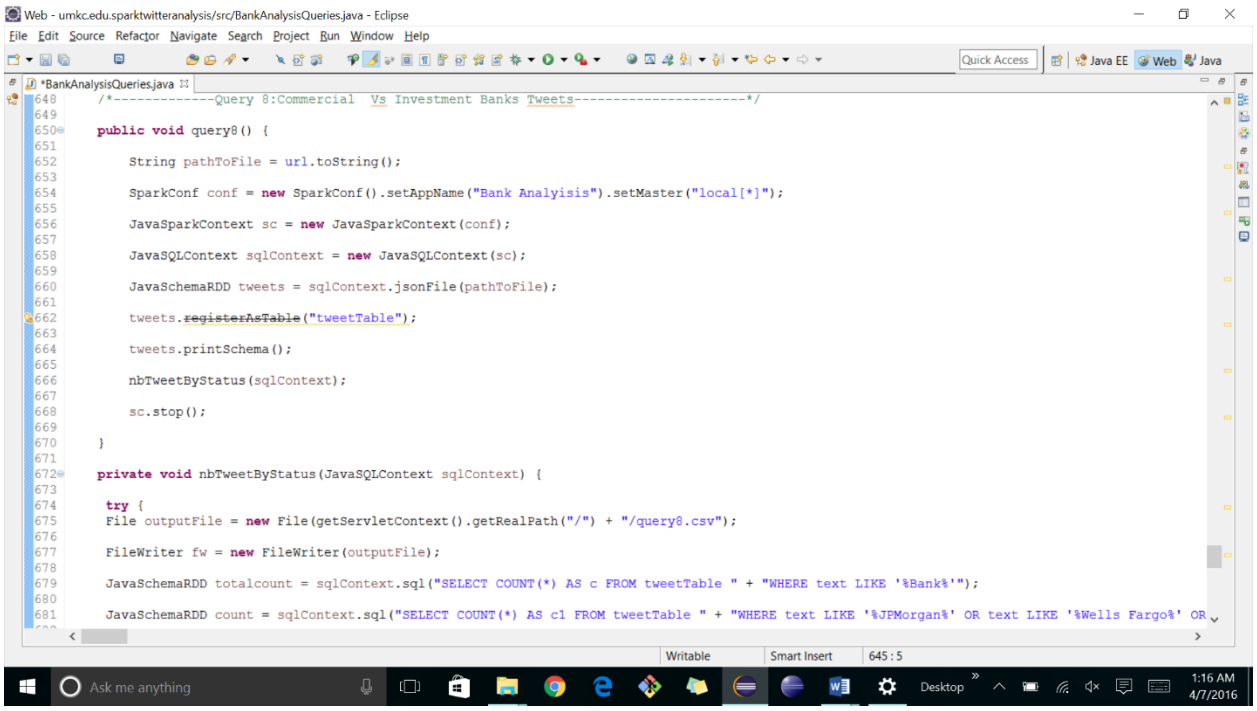
Nowadays banks are using twitter as source for reaching out people and wants to know the opinion about their products to ensure quality. They even provide customer support through this social networking site. So, we did analysis on banks to find out which banks is in touch with their customers the most i.e., which banks is providing more service to their customers through twitter.

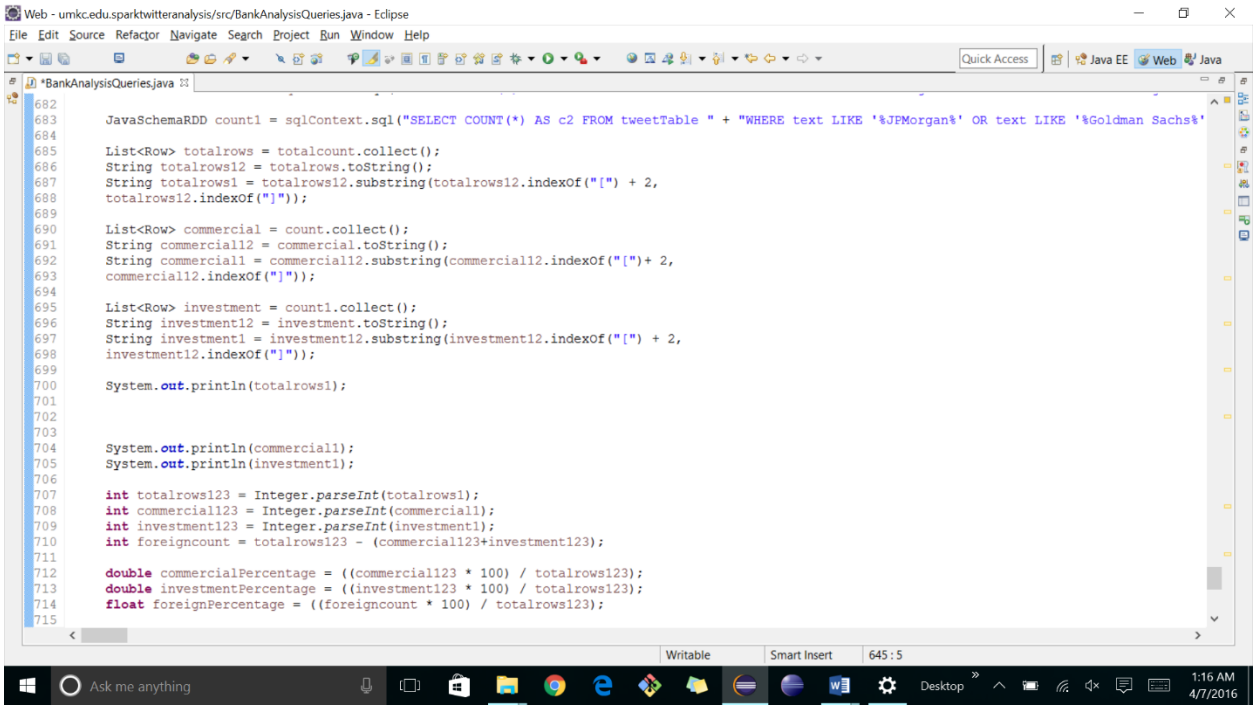


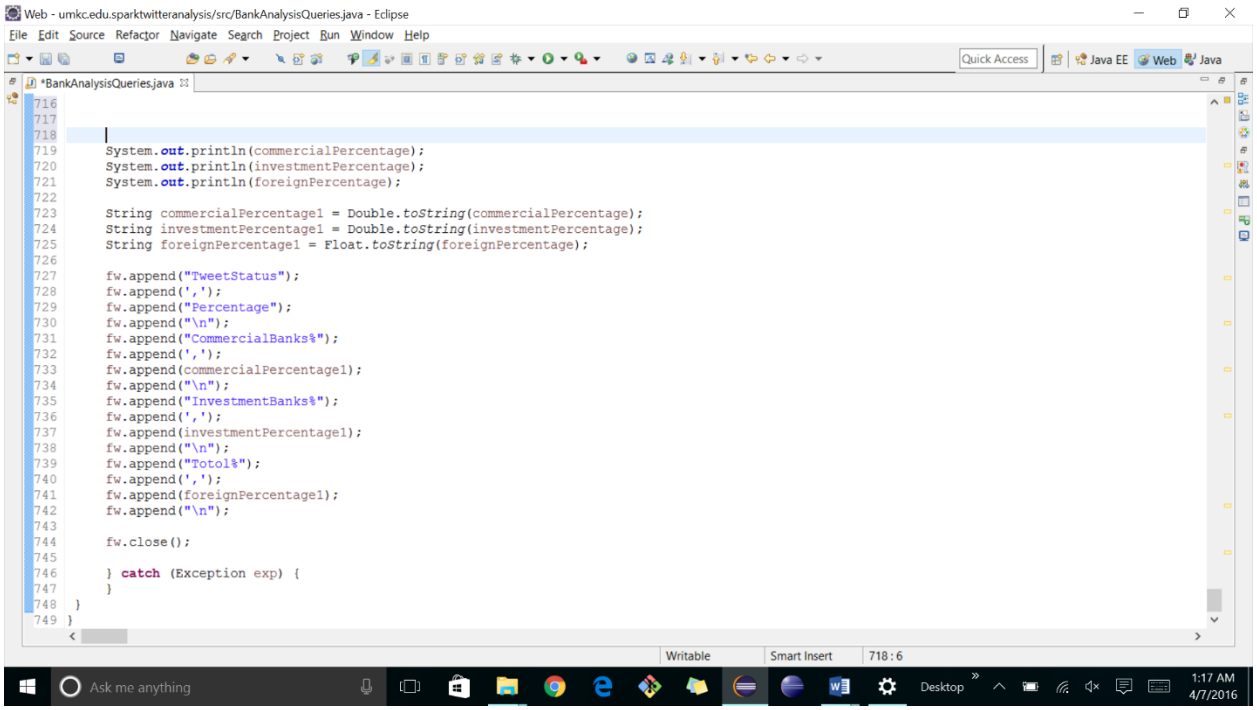


Query 8: Commercial Vs investment banks of USA

Banks are often categorized into Commercial and Investment Banks. The U.S finance industry has grown about 50% to 60% more compared to other countries over the past 2 decades. Bank regulation is very fragmented in USA when compared to other Group of Ten (economics) countries. So, we did analysis on finding out the percentage of tweets often tweeted about commercial banks and investment banks and tweets regarding to banks of all other countries.



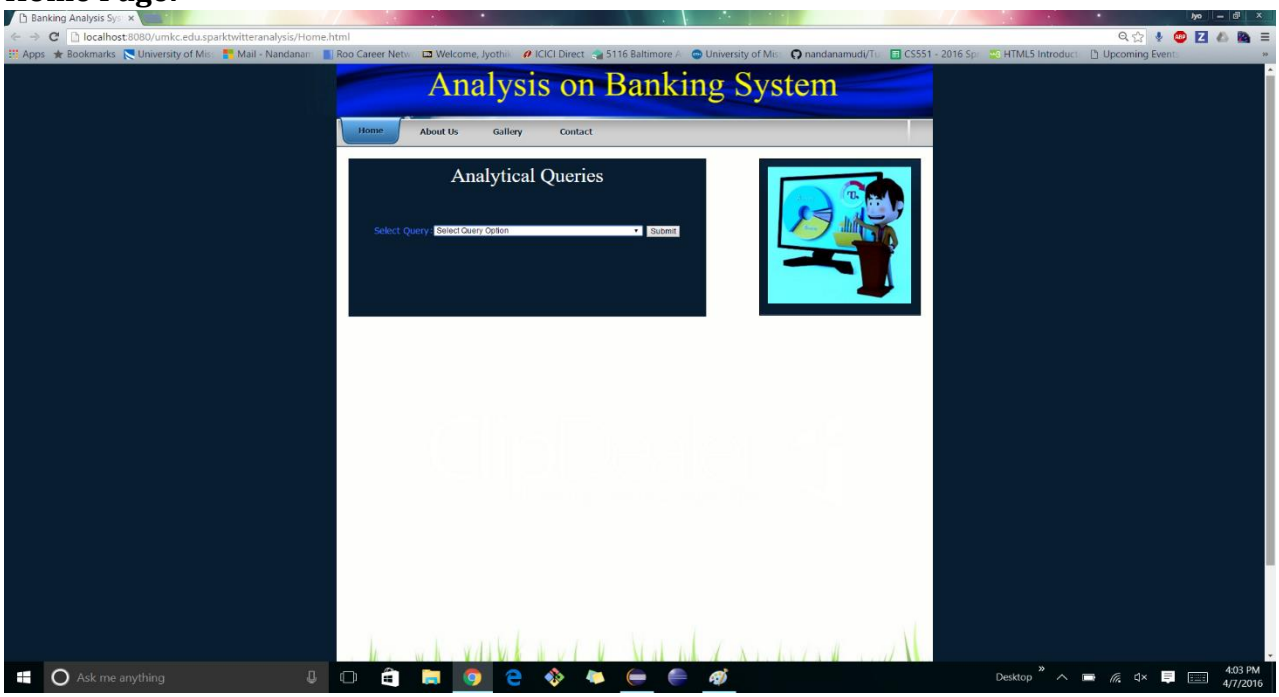



Web - umkc.edu.sparktwitteranalysis/src/BankAnalysisQueries.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
Quick Access
Java EE
Web
Java
*BankAnalysisQueries.java
System.out.println(commercialPercentage);
System.out.println(investmentPercentage);
System.out.println(foreignPercentage);
String commercialPercentage1 = Double.toString(commercialPercentage);
String investmentPercentage1 = Double.toString(investmentPercentage);
String foreignPercentage1 = Float.toString(foreignPercentage);
fw.append("TweetStatus");
fw.append(',');
fw.append("Percentage");
fw.append("\n");
fw.append("CommercialBanks%");
fw.append(',');
fw.append(commercialPercentage1);
fw.append("\n");
fw.append("InvestmentBanks%");
fw.append(',');
fw.append(investmentPercentage1);
fw.append("\n");
fw.append("Totol%");
fw.append(',');
fw.append(foreignPercentage1);
fw.append("\n");
fw.close();
} catch (Exception exp) {
}
}
}
Writable
Smart Insert
718 : 6
Ask me anything
Desktop
1:17 AM
4/7/2016

b. Visualization results

On client HTML, HTML5, CSS and Javascript were used to represent the analytical queries in attracting way. For pictorial representation we chose d3.js library.

Home Page:

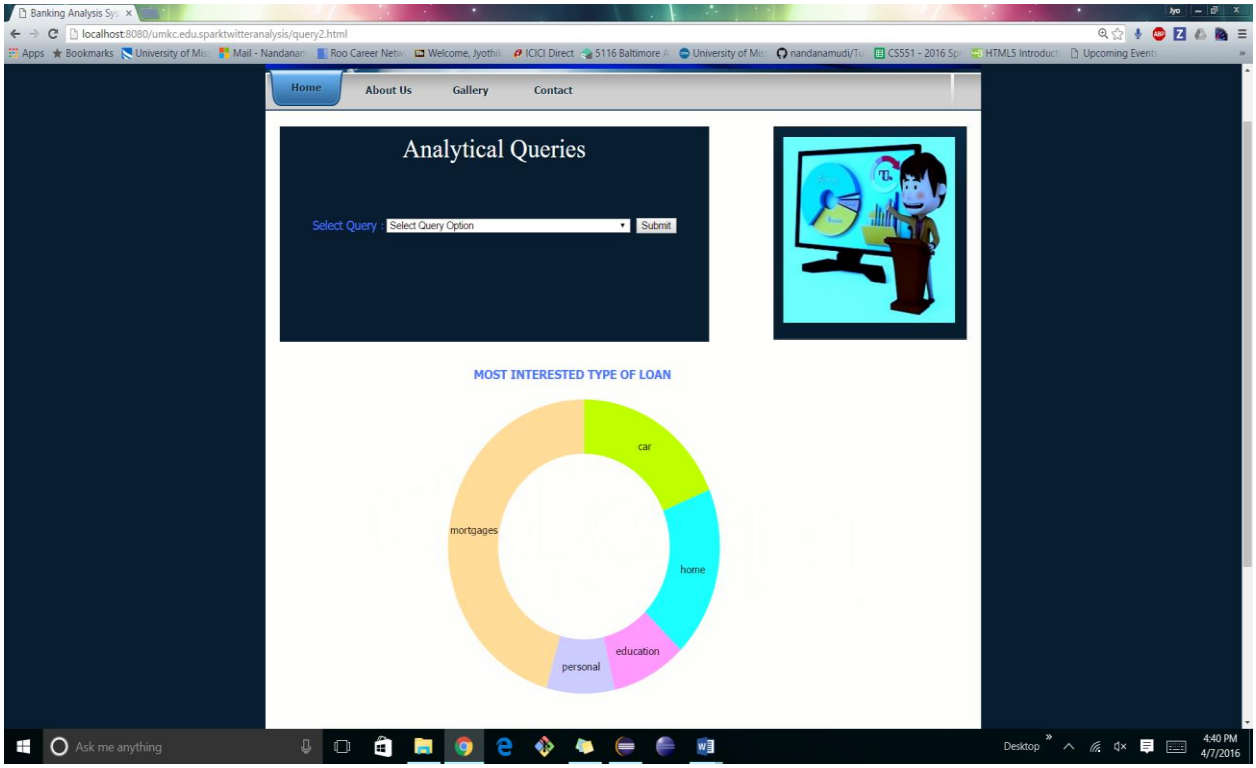


Query 1 Output:

The query 1 is about finding the popular bank across the social networking site: twitter

Our analysis shows that World Bank being most popular as it is being followed by 1712829 followers all over the world. Guaranty Trust Bank in the second, Jodrell Bank in third place, Kotak Mahindra Bank in fourth place, Diamond bank in fifth place and ICICI Bank in sixth place respectively with 527567, 230042, 113469, 87193, 83472 respectively.

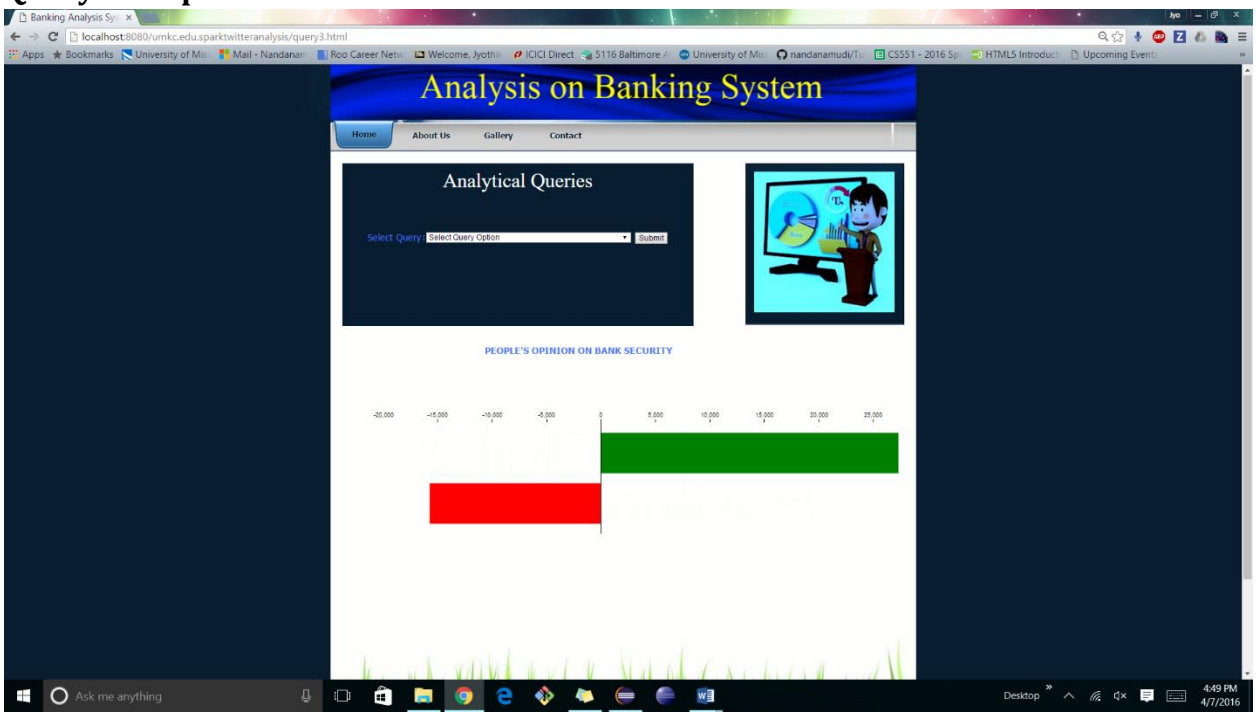
Query 2 Output:



The query 2 is about finding the most interested type of loan.

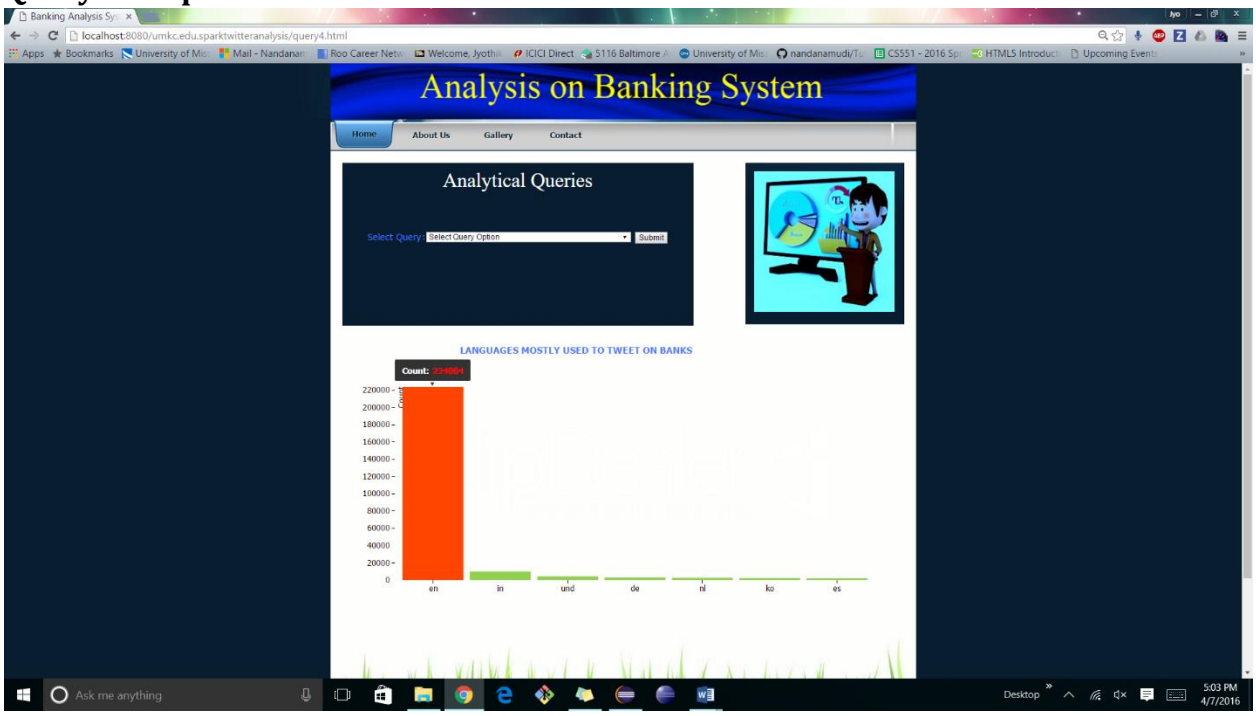
Our analysis shows that Mortgages is the type of loan people talk about loan and vehicle, home, education and personal takes the next level of interest in order respectively.

Query 3 Output:



The query 3 is about analyzing the data to find the opinion of people on bank security. We come to conclusion people are with good opinion about banks as you can see in the snapshot the green color indicates positive side opinion and the red colored bar indicates the negative opinion.

Query 4 Output:



The query 4 is finding the most communicative language about banks across twitter. The most used language is English as we all know with a total count of 224000 tweets of all the tweets collected.

Query 5 Output:

The query 5 is about finding the time at which there are many tweets about banks. For our analysis on the collected data the output of this query is on Mar 26, 2016 Saturday at 20:57:29, 51 tweets has been recognized. And in the second place on Mar 17, 2016 Thursday at 20:40:00, 50 tweets has been recognized.

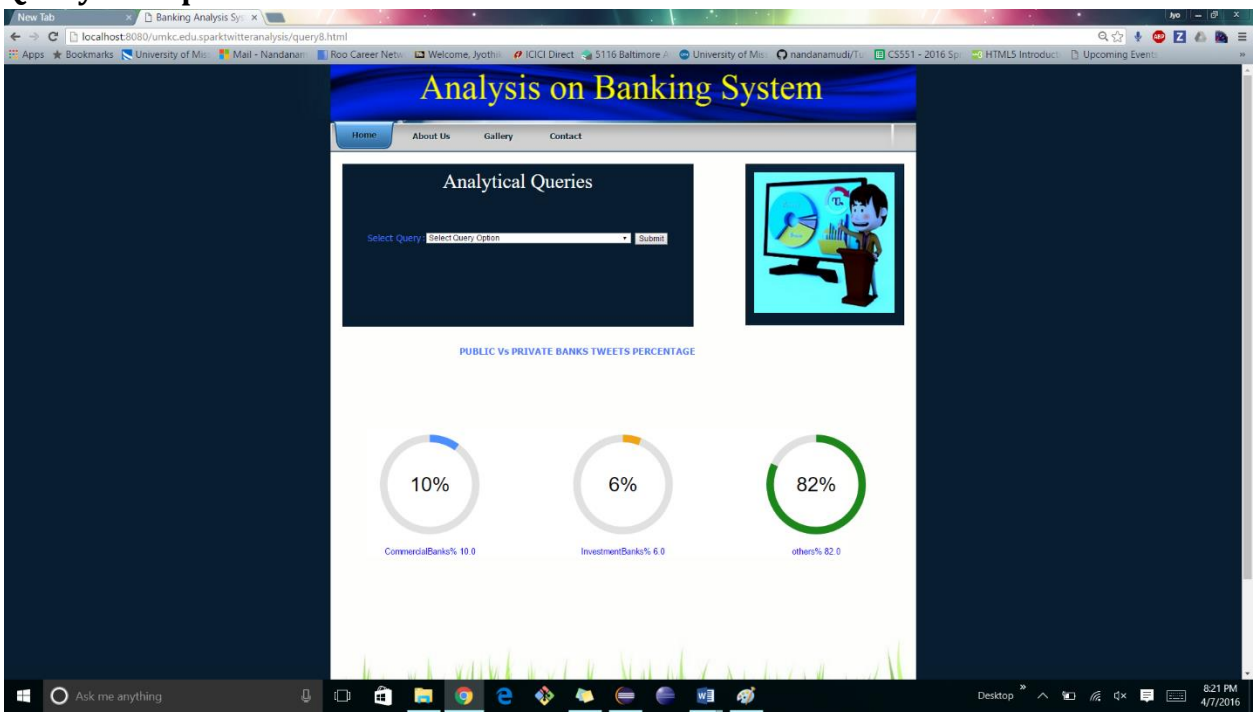
Query 6 Output:

The query 6 is about finding the most interested type of investment people interested in most. Our analysis shows that Business takes the most interested type of investment when compared to others. In the second place precious objects like gold, silver, diamonds, platinum etc., takes place and people are interested in stocks and real estate almost equally.

Query 7 Output:

The query 7 is about finding the service oriented banks that are using twitter the most to reach out the people and their customers. Here we chose the status count of bank for comparison. Our analytical results shows that Lloyds Banks stays in top position in service wise that uses twitter has medium for reaching the people. Guaranty Bank, HDFC Bank, TD Bank and Next Bank occupies the next positions respectively.

Query 8 Output:



The query 8 is to find the percentage of commercial banks, investment banks of USA of all the tweets available. Our analysis has shown that commercial banks occupy about 10% and investment banks is of 6% and tweets related to all other banks worldwide is of 82%.

iii. Other Pages:

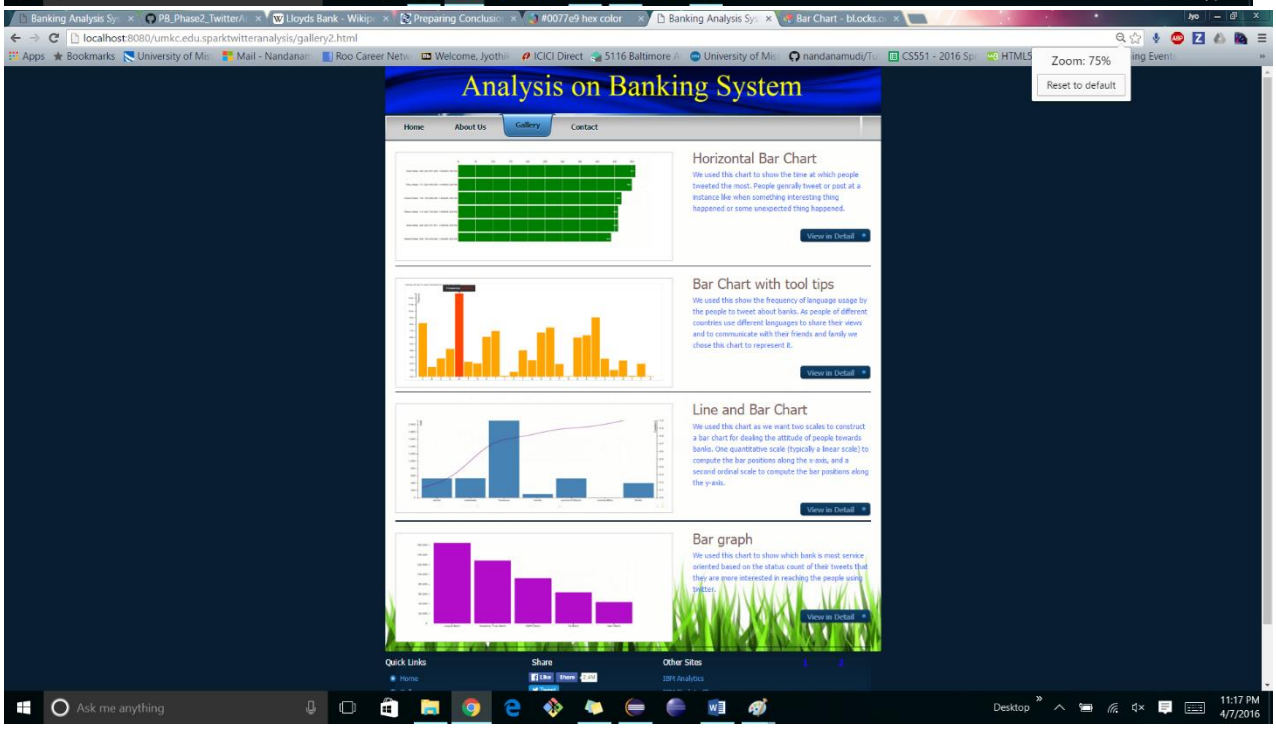
About Us Page:

Here we mentioned what the project is about and what all the steps involved in designing the project. We also mentioned few extra links like Facebook share, like buttons, twitter share tag, google following and share tags etc.

Recently IBM has done analysis on Banking Industry and you can find the links below at the bottom of the page which is an added reference to our project.

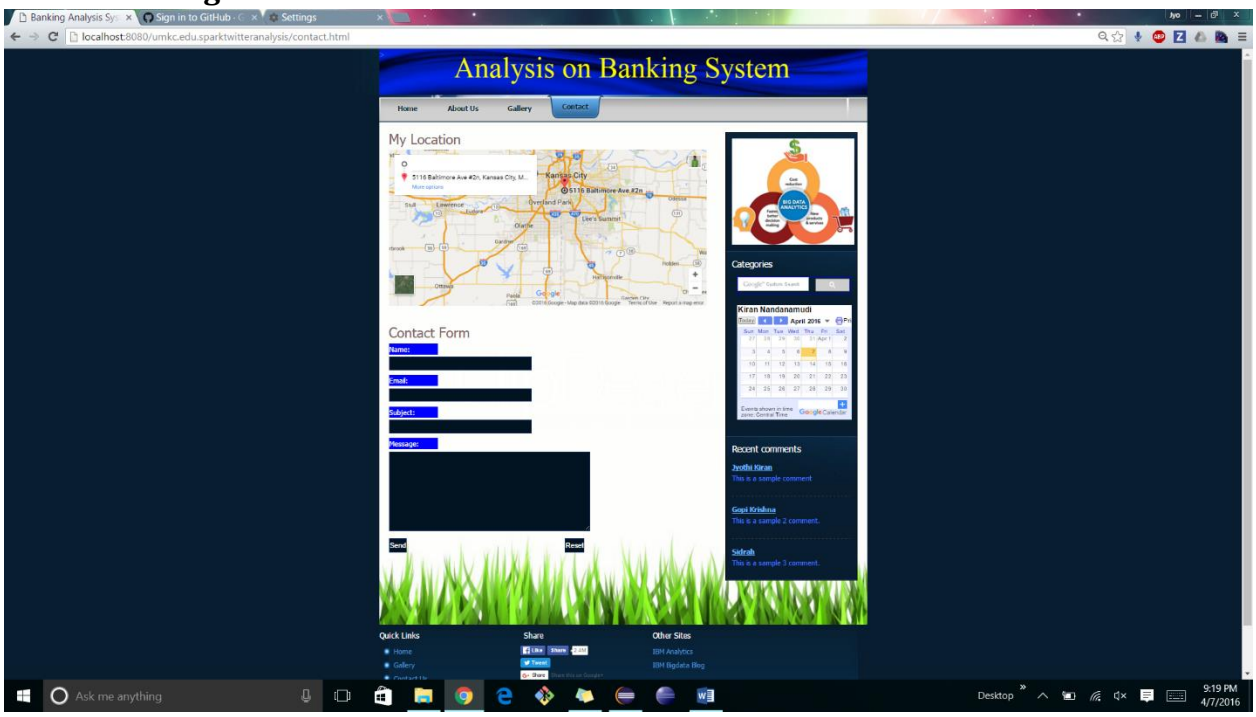
On the right side bar there is a search engine, where you can search all the information you want to and calendar plugin is an extra added quality of this page.

Gallery Page:



This page shows all the visualization charts and graphs we used in this project for analyzing. A brief information is shown explaining why we used that graph and links has been provided to shown the complete details of the graphs and charts for complete information of the visualization charts.

Contact Us Page:



This page just shows our location to contact us as well as contact from to post if there is anything there is need for clarification.

IV. Summary of the Project

The summary of our analysis on the banking sector is available in the below table to give outlook of whole project.

Query	Result
1. Most Popular Bank	World Bank
2. Most Interested type of loan	Mortgages
3. People Opinion on Bank Security	Secure
4. Most used language to tweet about banks	English (en)
5. Time at which people tweeted mostly on banks	March 26, 2016 at 20:57:29
6. Type of investment people often talk about	Business
7. Service wise top bank which uses twitter	Lloyd Bank of London

V. Conclusion and links:

Our analysis on banks has given some insights on the data collected from the twitter and gives some idea upon the banks. In overall this project can be extended to a greater extents for analyzing huge amount of data.

Link to the Project

https://github.com/nandanamudi/PB_Phase_2-Project

VI. References:

http://stackoverflow.com/questions/34236112/java-lang-noclassdeffounderror-org-apache-spark-sparkconf

https://m.oschina.net/blog/467423

https://www.mapr.com/ebooks/spark/03-apache-spark-architecture-overview.html

http://www.ibm.com/analytics/us/en/industry/banking/

http://www.ibmbigdatahub.com/blog/analytics-banking-services

http://www.taywils.me/2013/11/05/javasparkframeworktutorial.html

https://github.com/mbostock/d3/wiki/Gallery