Centaurus: Twitter Sentiment Visualization

Team Members

Rohan Gokhale, Travis Robinson, Chase Hu

Introduction

Our team will create an website where a user will be able to search their favorite television show and then see how Twitter users feel about that show in real time. The user will be able to see how Twitter uses feel about their search by state. Behind the scenes, we will be ingesting tweets from the Twitter Stream API in real time into our Hadoop database, and will be analyzing that data to provide insights about how people feel about their favorite show. If not enough Twitter data can be collected on TV shows, we may choose to analyze other forms of media/companies (ex. books, news sources, airlines, grocery stores, etc.).

User Perspective

From the user perspective, a user will go to a website and be able to enter a word and will be shown a visualization of sentiment analysis (ex. 80% favorable) of that word for every state/county. Our visualization will be able to show details when the user hovers over the state/county.

Structure of Project

The functional structure of the project will likely be as follows

1. Ingestion - Getting streaming data from Twitter and loading it into our HDFS for processing
    a. Setup parameters for collecting Twitter data (ex. Location, keywords, sentiment dictionary, etc.)
2. Processing - Transforming ingested data into data set(s) that can be used for analysis and querying.

   a. Filtering - Covert Twitter data to columnar format, aggregation, indexing, etc
   b. Apply our Sentiment Model to the data to classify the tweets
3. Analyzing - Run analytical queries on the processed data sets to find (hopefully) interesting insights
4. Website - Display data for users to view and see how people in their state/county feel about a particular show. The website will have a search function so that users can search for a particular show, and see how those in their state/county feel about it. The website may also offer some form of visualization (such as a map of the US) that shows how that state/county compares to other states/counties.

Software, Languages, APIs, Development Tools

Amazon Web Services for spinning up a Hadoop cluster

HortonWorks Hadoop Distribution software for HDFS

Hive/Spark software for data analysis

Twitter Stream API for getting Twitter data

Scala to program sentiment analysis

Tableau for creating visualizations (maps) of the data the user requests

Task Breakdown

Task	**Who Will Perform**	**Time Estimate**
Setup the project environment : Week 3 Spinning up a Hadoop cluster on Amazon AWS with Spark. Download and install JVM, JDKs, Scala,	Rohan Gokhale	22

tools and software like python, maven, IDE (eclipse), github client, Scala, etc. Perform unit test on cluster.		
Database Design: Week 3 Design Tables and Schemas Ingest an existing twitter data set Perform unit test on database with existing twitter data set.	Chase Hu	20
Website Design: Week 3 Begin work on website using dummy images/data Testing of Website skeleton/layout using dummy data	Travis Robinson	18
Week 3 Progress Report	Chase Hu	1
Ingesting Data (Part 1): Week 4 Write a program that will continuously collect tweets using the Twitter API Perform unit test on data retrieval	Rohan Gokhale	30
Ingesting Data (Part 2): Week 4 Write a program that will parse data retrieved by program in (Part 1) and convert into our desired format, as well as retrieve user location (and any other information that may prove necessary) Perform unit testing on (Part 2) program	Travis Robinson	30
Analysis Prep: Week 4 Write (or otherwise acquire) a sentiment dictionary -----need to convert dictionary to something readable, such as database or array (I think database would be best for ease of changing, program could then refer to DB, though wouldn't be	Chase Hu	10

resource efficient)		
Week 4 Progress Report	Rohan Gokhale	1
Conduct Analysis: Week 5/6 Write a program(s) to perform a sentiment analysis on stored Twitter data. Analyze the data to see if the analysis provides meaningful results. Refine the program as needed.	Whole Team	75 (25 each)
Week 5 Progress Report	Rohan Gokhale	1
Data Visualization: Week 6 Use sentiment analysis data to create maps on Tableau Making sure the maps look presentable	Whole Team	21 (7 each)
Midterm Report: Week 6	Whole Team	3 (1 each)
Automate and Integrate: Week 7 Write a program to automate the previous task for all shows. Set up a Tableau account for the (hopefully) automated creation and storage of maps. Perform unit tests on the HDFS database and Tableau maps.	Rohan/Chase	30
Week 7 Progress Report	Chase Hu	1
Website Design: Week 7/8/9 Make an interactive web page that will allow a user to select the data the want to visualize and then see it from skeleton (Week 3). Integrate website with Tableau maps and database of positive and negative tweets and location. Perform usability tests on website.	Travis Robinson/Chase Hu	25
Week 8 Progress Report	Travis Robinson	1

Week 9 Progress Report	Travis Robinson	1
Testing: Week 9 Integration testing.	Chase Hu	10
Final Report: Week 10 Write the final report and prepare all documents/components for submission.	Travis Robinson	6

Total Time: 306

Time Per Member:

Rohan: 102

Chase: 102.5

Travis: 101.5

Diagrams

Flow chart of project operations

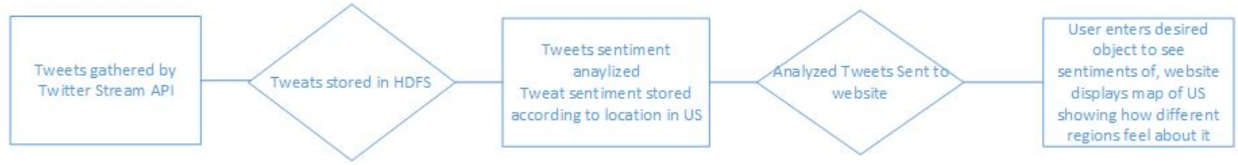
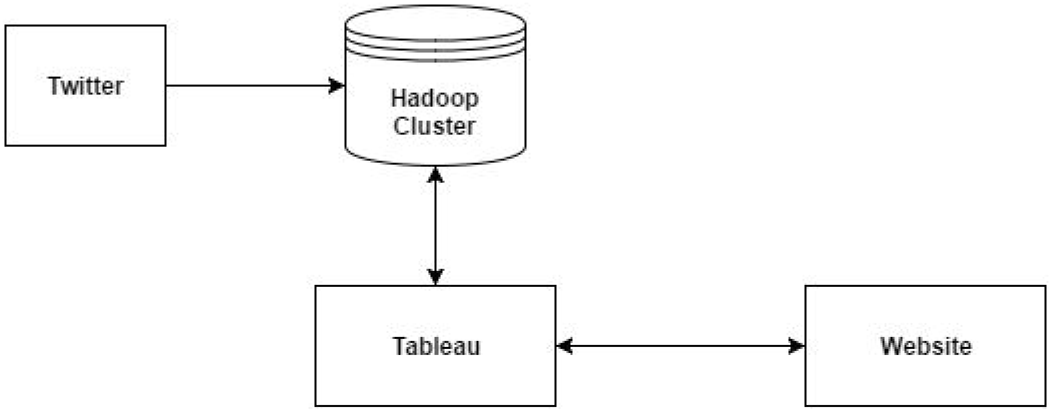

Tweets gathered by Twitter Stream API
Tweats stored in HDFS
Tweets sentiment anaylized Tweat sentiment stored according to location in US
Analyzed Tweets Sent to website
User enters desired object to see sentiments of, website displays map of US showing how different regions feel about it

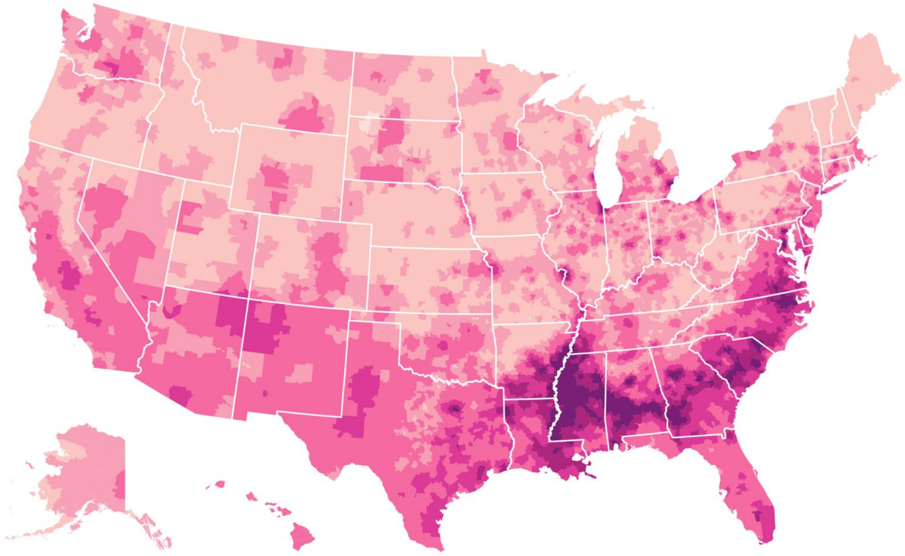
Diagram outlining interaction between the Twitter API, our HDFS, Tableau, our website, and the user.



Example visualization of TV Show keyword

TV Show: Duck Dynasty

*Darker areas show more favorable Twitter data. In this regional map, Duck Dynasty is most popular in the southeastern portion of the country.



Conclusion:

Our team hopes to give people a good resource to review tv shows (or, if not enough data can be gathered on TV shows, then on some other interesting entity) and see how those in their state/county feel about it. Tastes vary widely between groups of people (for example Oregonians and Minnesotans), and it would be fun for users to be able to see how people in other areas feel about their favorite or least favorite things.

This project will be completed by surveying Twitter data pulled from the stream API, and analyzed via Hadoop and associated analysis tools. It should take roughly 300 hours to complete and should be completed by the deadline.

Sources:

1. http://www.nytimes.com/interactive/2016/12/26/upshot/duck-dynasty-vs-modern-family-television-maps.html?WT.mc_id=2016-KWP-AUD_DEV&WT.mc_ev=click&ad-keywords=AUDDEVREMARK&kwp_0=302296&kwp_4=1169511&kwp_1=532257
2. http://blog.cloudera.com/blog/2012/09/analyzing-twitter-data-with-hadoop/
3. https://dev.twitter.com/streaming/public
4. https://dev.twitter.com/streaming/overview
5. https://dev.twitter.com/streaming/overview/connecting
6. https://www.youtube.com/watch?v=ziqx2hJY8Hg