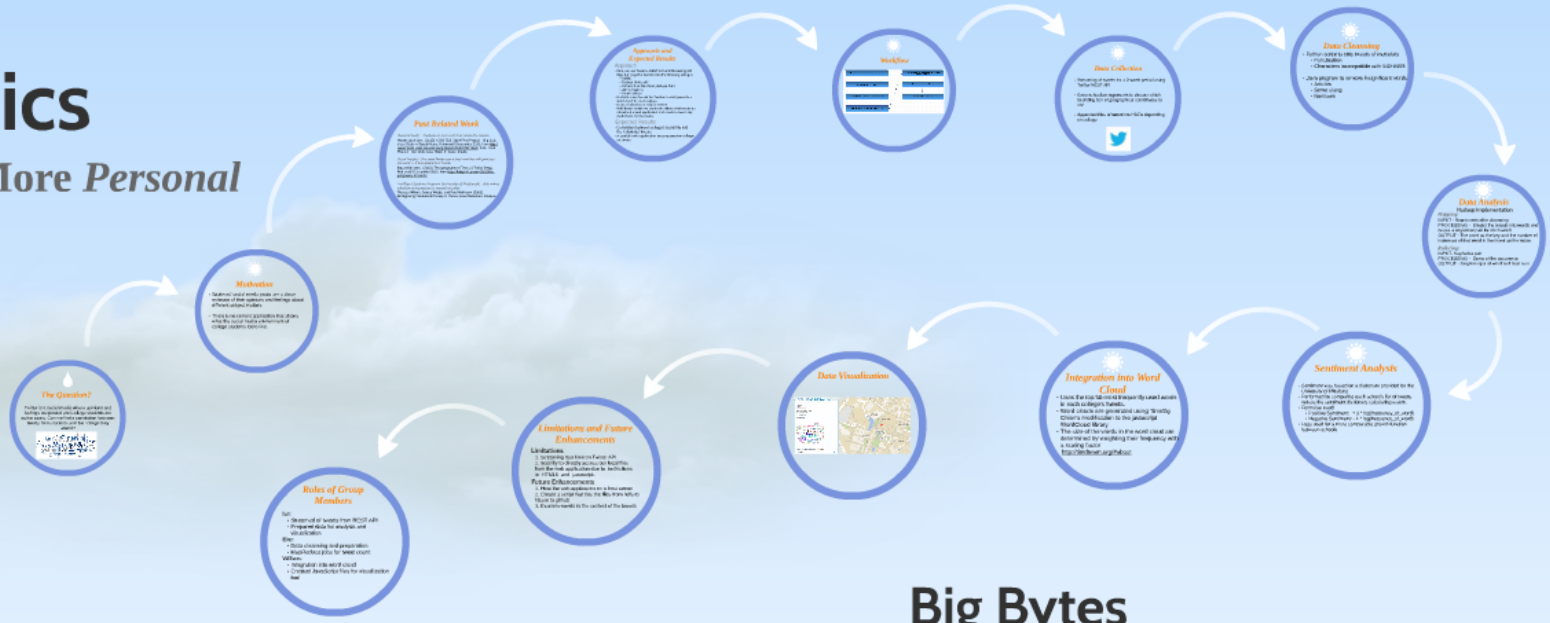
Twitterlytics

Making College More *Personal*

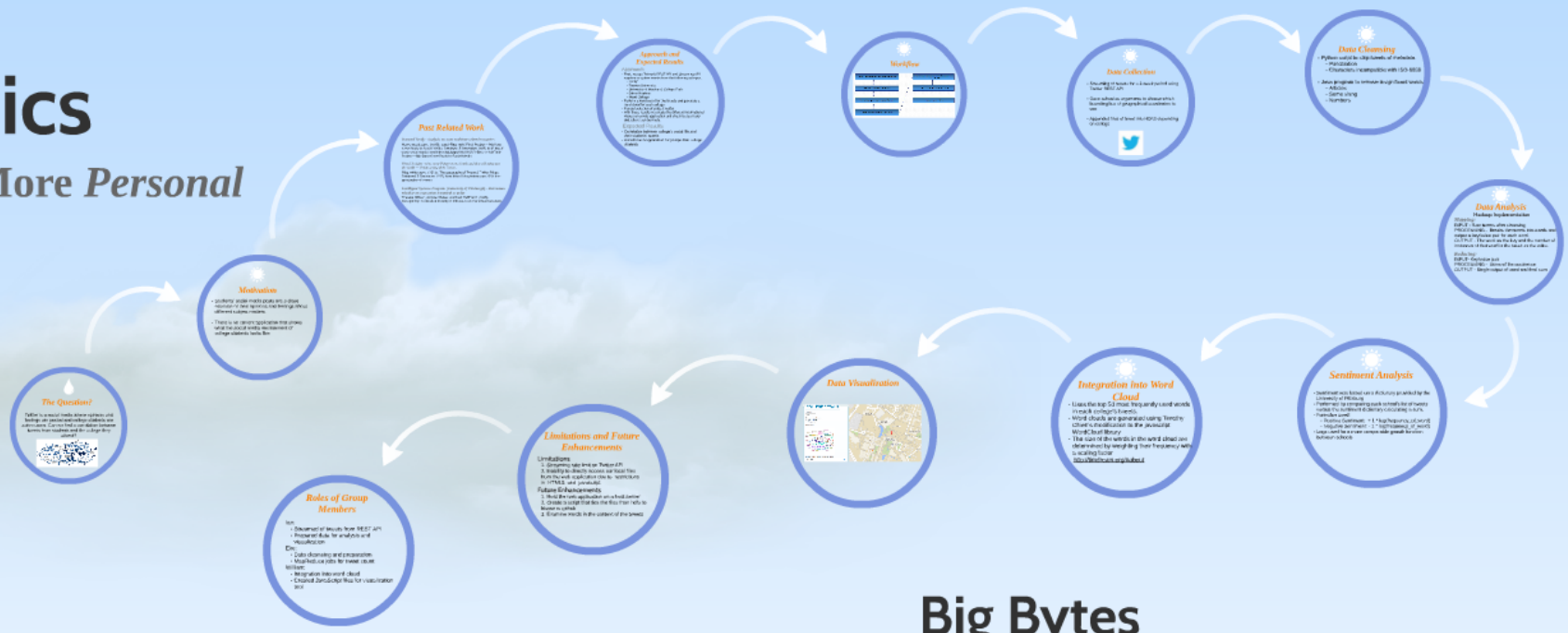


Big Bytes

Ian Kirk, Oladipupo Eke, William Trotman

Twitterlytics

Making College More *Personal*



Big Bytes

Ian Kirk, Oladipupo Eke, William Trotman

The Question?

Twitter is a social media where opinions and feelings are posted and college students are active users. Can we find a correlation between tweets from students and the college they attend?

Motivation

- Students' social media posts are a close estimate of their opinions and feelings about different subject matters.
- There is no current application that shows what the social media environment of college students looks like.

Past Related Work

Harvard Study - Analysis on near real-time related to movies

Howto-stack.com,. (2015). CSCI-E63, 2014 Final Project -- Big Data Case Study in Social Media. Retrieved 5 December 2015, from http://www.howto-stack.com/videos/g44yqpxVSuU/CSCI-E63,-2014-Final-Project----Big-Data-Case-Study-in-Social-Media

Visual Insights - Give every Twitter user a brush and they will paint you the world — if they geotag their Tweets.

Blog.twitter.com,. (2013). The geography of Tweets | Twitter Blogs. Retrieved 5 December 2015, from https://blog.twitter.com/2013/the-geography-of-tweets

Intelligent Systems Program (University of Pittsburgh) - determines whether an expression is neutral or polar

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis.

Approach and Expected Results

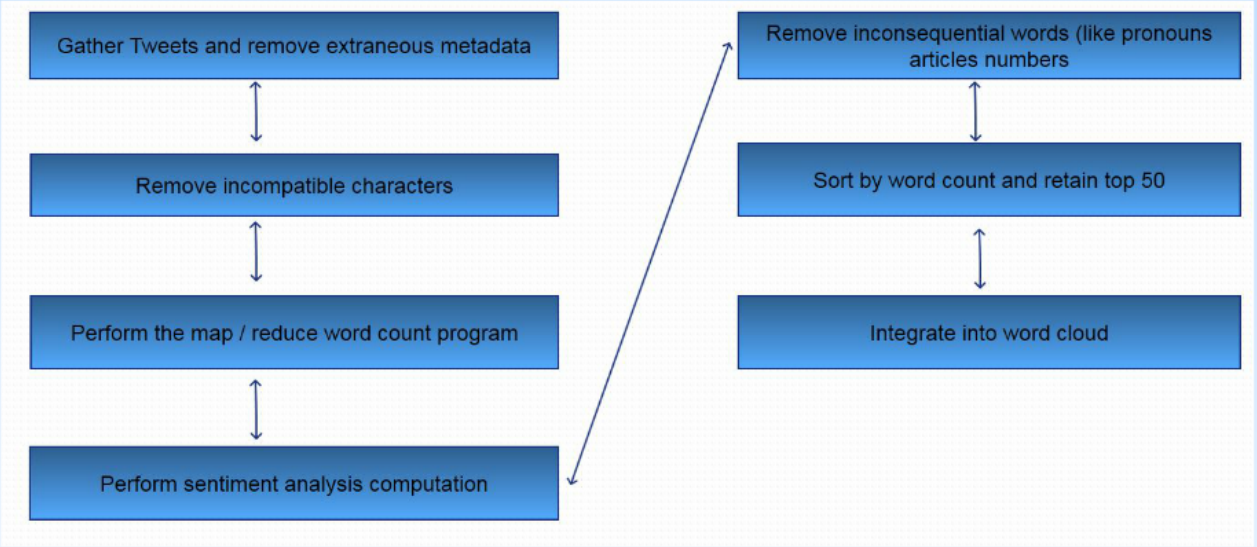
Approach

- First, we use Twitter's REST API and Streaming API together to gather tweets from the following colleges.
  - UMBC
  - Towson University
  - University of Maryland, College Park
  - Johns Hopkins
  - Hood College
- Perform a word count for the tweets and generate a word cloud for each college
- Manual selection of subject matter
- With those results we create the different informational visuals on a web application and check to see if any deductions can be made.

Expected Results

- Correlation between college's social life and their students' tweets
- A useful web application for prospective college students

Workflow



Data Collection

- Streaming of tweets for a 2-week period using Twitter REST API
- Gave school as arguments to choose which bounding box of geographical coordinates to use
- Appended files of tweet into HDFS depending on college

Data Cleansing

- Python script to strip tweets of metadata
  - Punctuation
  - Characters incompatible with ISO-8859

- Java program to remove insignificant words
  - Articles
  - Some slang
  - Numbers

Data Analysis

Hadoop Implementation

Mapping:

INPUT - Raw tweets after cleansing
PROCESSING - Breaks the tweets into words and output a key/value pair for each word.
OUTPUT - The word as the key and the number of instances of that word in the tweet as the value.

Reducing:

INPUT- Key/value pair
PROCESSING - Sums of the occurence
OUTPUT - Single output of word and final sum

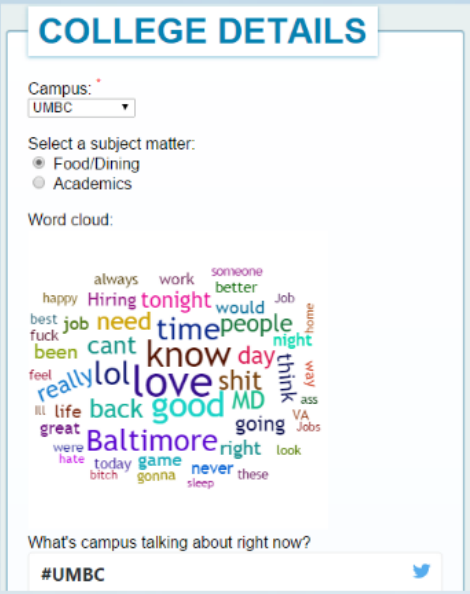
Sentiment Analysis

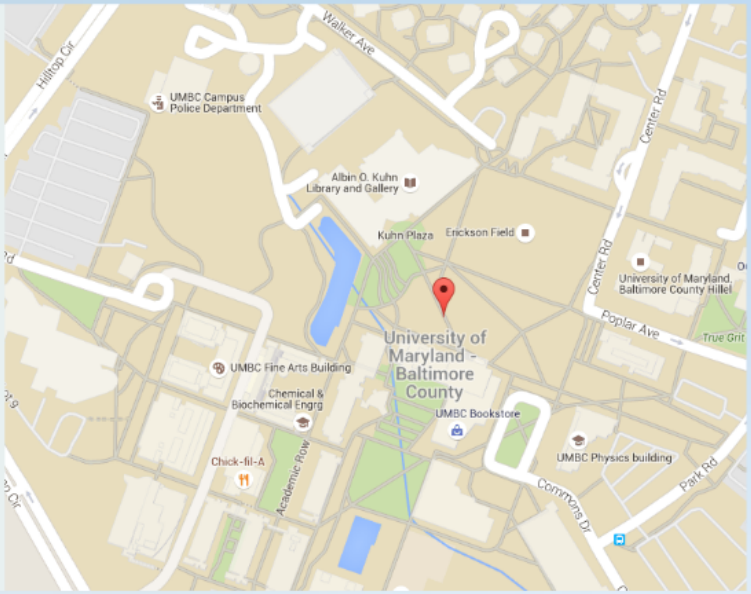
- Sentiment was based on a dictionary provided by the University of Pittsburg
- Performed by comparing each school's list of tweets versus the sentiment dictionary calculating a sum.
- Formulae used:
  - Positive Sentiment: + 1 * log(frequency_of_word)
  - Negative Sentiment: - 1 * log(frequency_of_word)
- Logs used for a more comparable growth function between schools

Integration into Word Cloud

- Uses the top 50 most frequently used words in each college's tweets.
- Word clouds are generated using Timothy Chien's modification to the javascript WordCloud library
- The size of the words in the word cloud are determined by weighting their frequency with a scaling factor
  http://timdream.org/#about

Data Visualization





Limitations and Future Enhancements

Limitations

1. Streaming rate limit on Twitter API
2. Inability to directly access our local files from the web application due to restrictions in HTML5 and javascript.

Future Enhancements

1. Host the web application on a host server
2. Create a script that ties the files from hdfs to hbase to github
3. Examine words in the context of the tweets

Roles of Group Members

Ian:

- Streamed of tweets from REST API
- Prepared data for analysis and visualization

Eke:

- Data cleansing and preparation
- MapReduce jobs for tweet count

William:

- Integration into word cloud
- Created JavaScript files for visualization tool

Sources and References

The Best 376 Colleges 2012. Robert Franek, Princeton Review (Firm), Laura Braswell, Seamus Mullarkey.

Dev.twitter.com,. (2015). The Streaming APIs | Twitter Developers. Retrieved 8 December 2015, from https://dev.twitter.com/streaming/overview

Howto-stack.com,. (2015). CSCI-E63, 2014 Final Project -- Big Data Case Study in Social Media. Retrieved 5 December 2015, from http://www.howto-stack.com/videos/g44yqpxVSuU/CSCI-E63,-2014-Final-Project----Big-Data-Case-Study-in-Social-Media

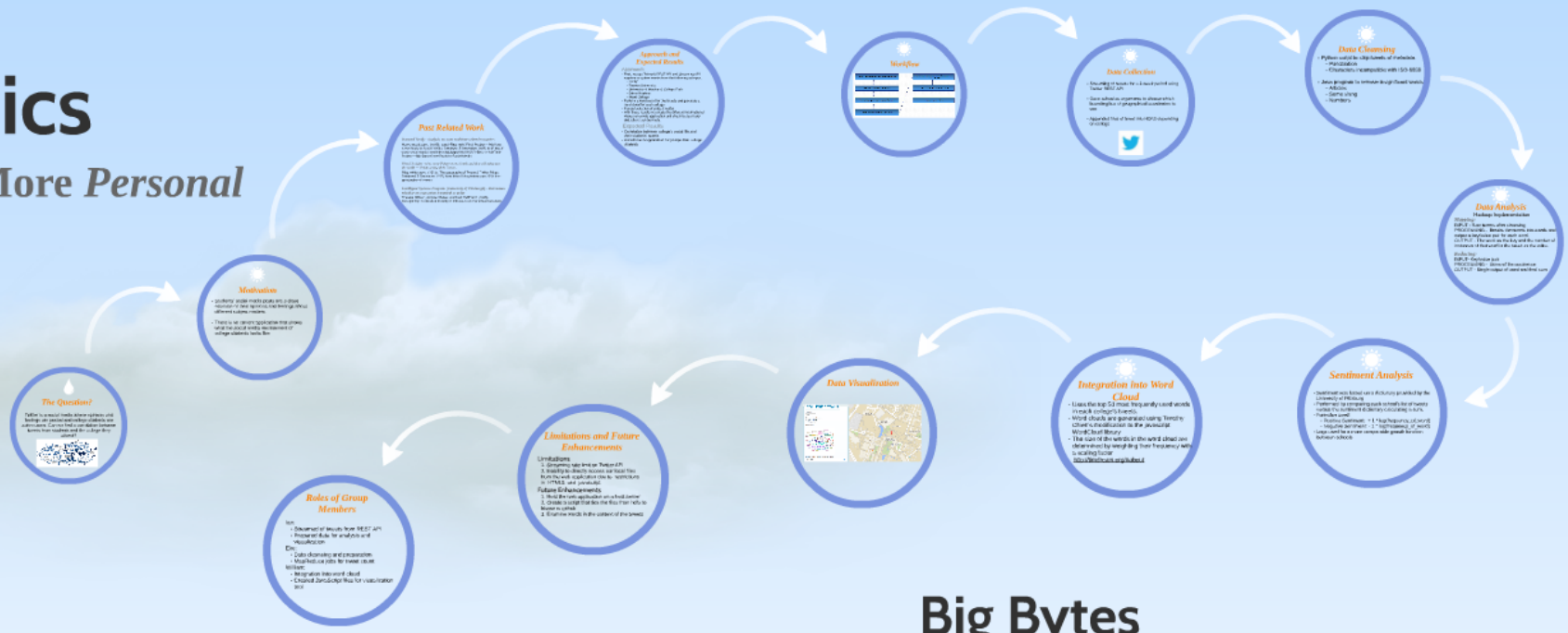
Blog.twitter.com,. (2013). The geography of Tweets | Twitter Blogs. Retrieved 5 December 2015, from https://blog.twitter.com/2013/the-geography-of-tweets

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. Proc. of HLT-EMNLP-2005.

Questions?

Twitterlytics

Making College More *Personal*



Big Bytes

Ian Kirk, Oladipupo Eke, William Trotman