NB4_project_report

May 4, 2014

In this notebook, you'll sumarize your finding for the class presentation.

The class presentation will use this notebook in slide mode, as I do during lecture time.

You'll need to download the following zip file, unzip it, and run the following notebook install-support.ipynb to install the slide capabilities.

The notebook notebook-slideshow-example.ipynb will give you examples on how to use slides within the iPython notebook.

You should also write this notebook so that you can convert it nicely into a pdf document using the commands

```
ipython nbconvert NB4_report.pynb --to latex
pdflatex NB4_report.tex
```

See here for further references on how to do that.

In this notebook, you'll

- describe your problem as stated in the propectus
- comment on your data sources, on their format, on the difficulties to get them
- present the main challenge you encountered
- present your finding in the form of expresive graphics

Be sure to include an introduction section motivating your visualizations, with a description of the substantive context and why it is interesting.

Cite the source of the data and any other references that you used in carrying out your project.

0.1 Team members responsible for this notebook:

- team member 1 **Shadman Sadek**: Compiling and completing this notebook
- team member 2 **Elizabeth Sabiniano**: Compiling and completing this notebook

1 Objective:

The purpose of this project is to examine how sentiment towards college varies throughout the world - both between continents (general geographical and socioeconomic areas) (and between states specifically in the US?) We used Python and Twitter API to help gather a massive amount of data (~500,000 tweets) based on the keyword **college**. The data gathered was organized into tweets with location and without location. Furthermore, tweets with location were further categorized based on global location.

Our data was then analyzed into positive, negative, and neutral sentiment categories. We developed visualizations that displayed word frequencies and a map that displays sentiment based on global region.

1.1 Data Sources:

All of our data was obtained from Twitter. These were then saved in pickle files to accomodate for the large datastrings.

1.2 Challenges:

Originally, our team wanted to examine the sentiments towards higher education, however the keywords "higher education" did not yield enough data for us to work with. That and the following are the challenges we faced in this study:

- Finding the right program to gather enough tweets.
- Gathering enough data with location
- Machine learning: being able to train our program to categorize as much of the tweets as accurately as possible

We tried TwitterSearch, Tweepy, before finally finding Twython, which yielded enough data for us to work with.

1.3 Visualizations: Word Frequency

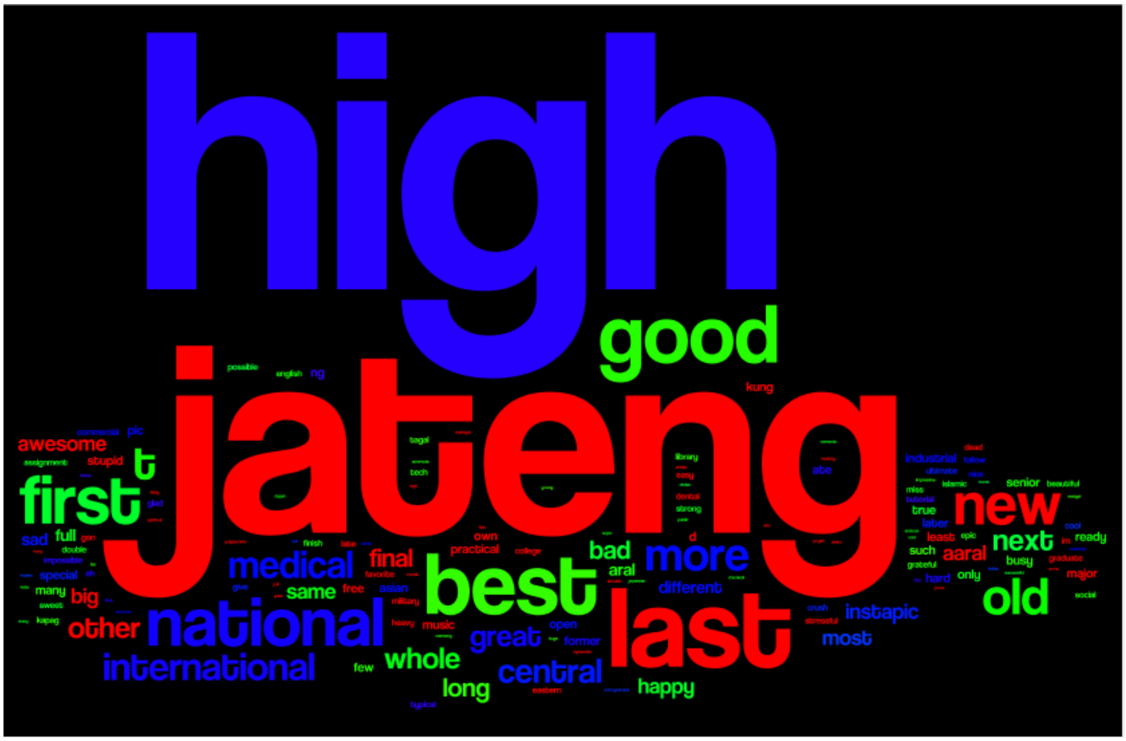
The following visualizations describe word frequencies in Asia, Africa, Europe, Latin America, United States, and the world. They were obtained through the word cloud program called Wordle. The most frequent words show up largest in the visualizations. Such vizualizations give us an idea of topics most commonly associated with our keyword **college**.

```
In [1]: from IPython.core.display import Image
```

Asia

```
In [17]: Image(filename='../visualizations/AsiaWords.png')
```

```
Out[17]:
```

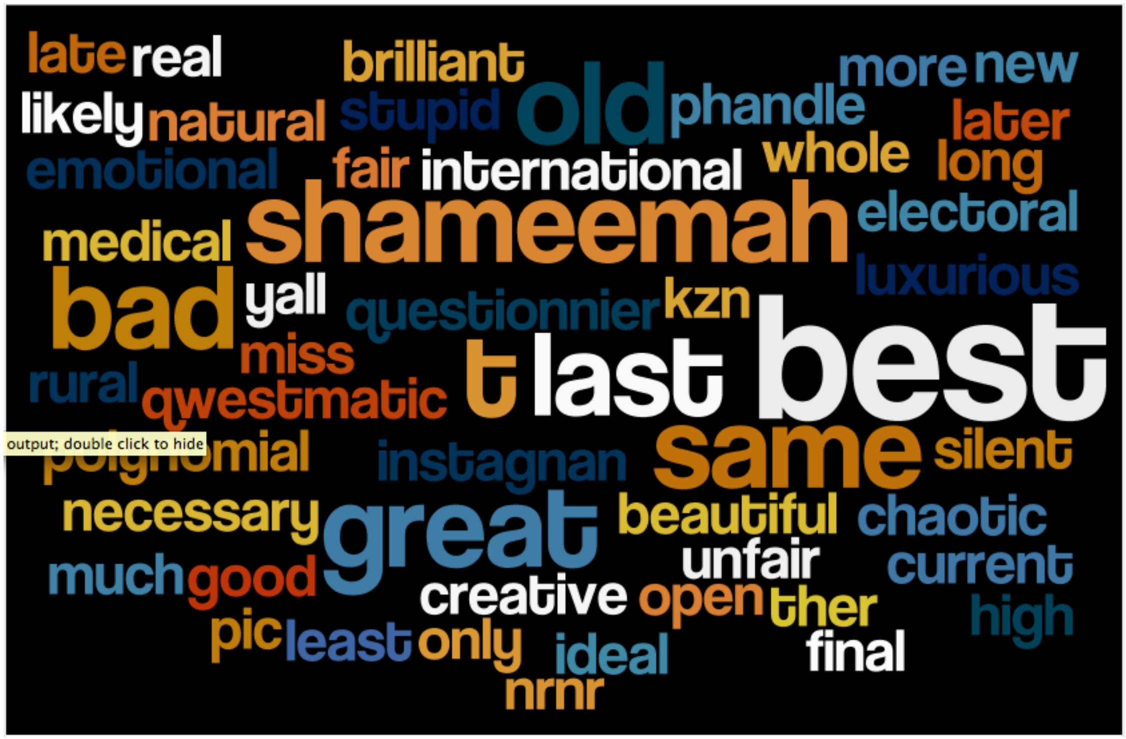


jateng stands for Jateng-DIY, a province in Indonesia. Over the period at which we gathered our data, the ESA Week 2014 (a scholastic competition) is taking place in this region.

Africa

```
In [4]: Image(filename='../visualizations/AfricaWord.png')
```

```
Out[4]:
```

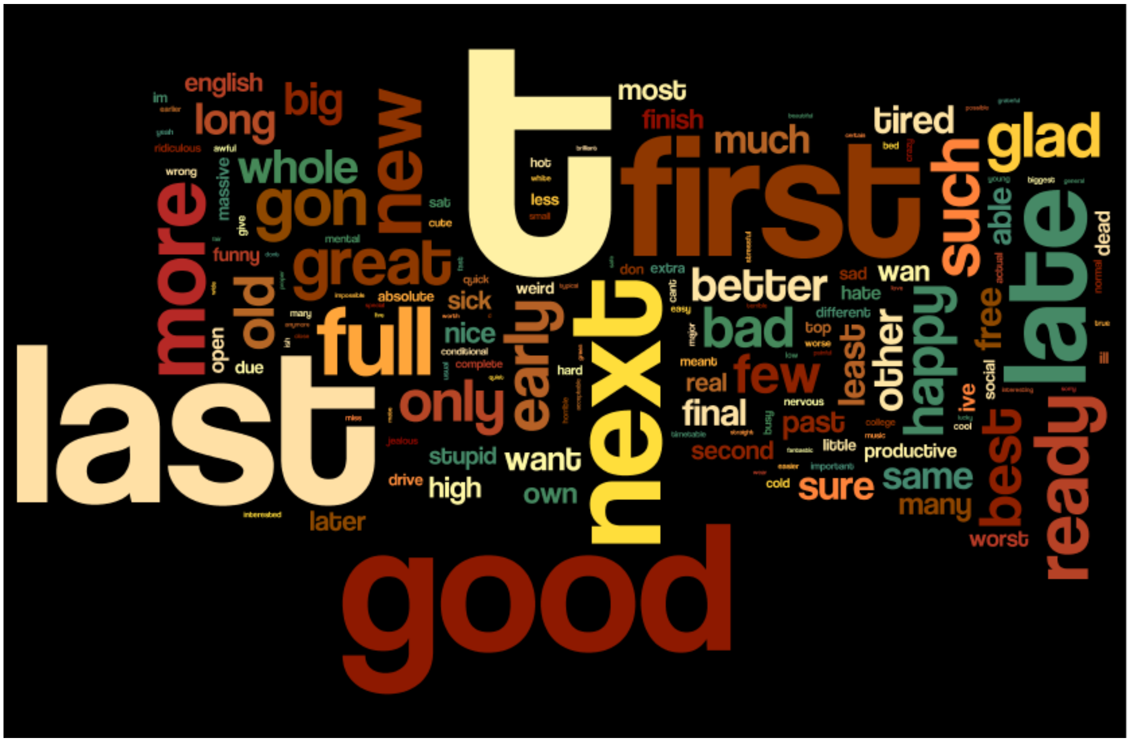


Each of the words above appeared with almost equal amount of frequency; thus, they are all close in size. *shameemah* is a user's name who has a decent amount of followers and obtained several retweets regarding college inquiries. *best* appears to be the most frequent word in this location.

Europe

```
In [8]: Image(filename='../visualizations/EuropeWords.png')
```

```
Out[8]:
```



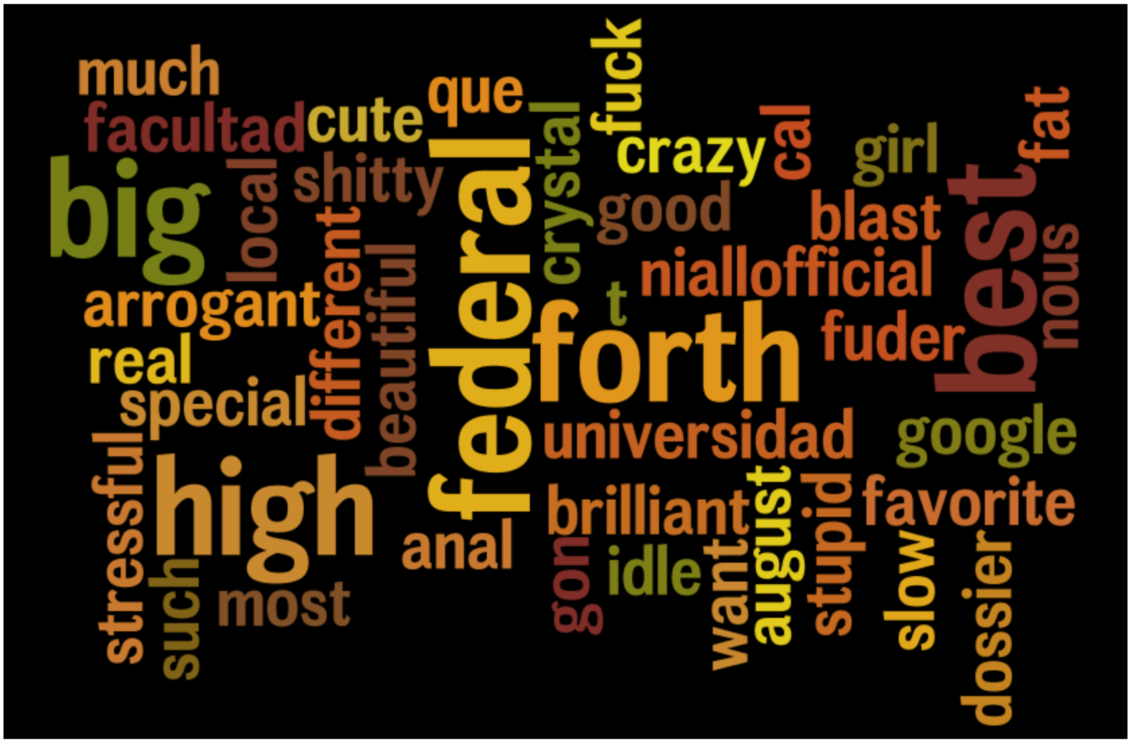
Although words such as *good* and *last* appear the most frequent in this region, the tweets at which they appear are mostly negations. For example:

- "Can't wait to finish college for good. 6 weeks left"
- 'Not sure college is a good idea today'
- "It's safe to say I haven't missed college one bit these last 2 weeks don't wannnna go back"
- 'so much to do, zero motivation for the last week of college, this is not good..'

Latin America

```
In [10]: Image(filename='../visualizations/LatinAmericaWords.png')
```

```
Out[10]:
```

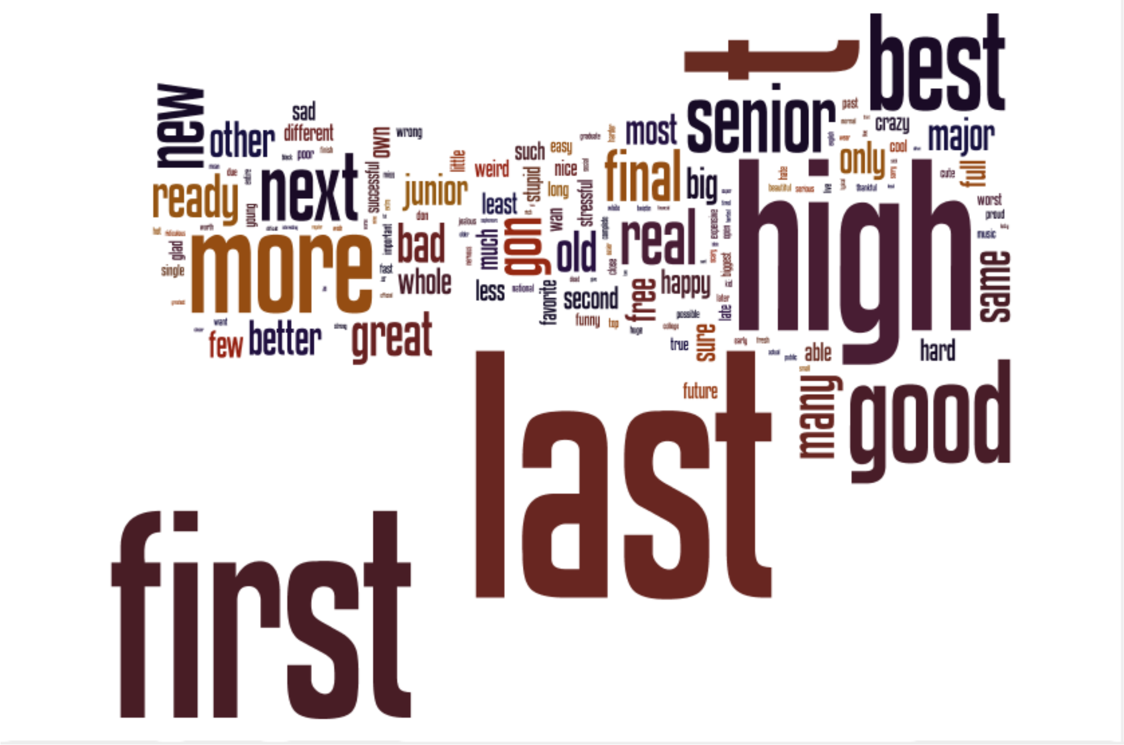


Similar to Africa, most the words listed for Latin America have the same frequency. The bigger words are only one or two counts away from the smaller ones.

United States

```
In [2]: Image(filename=’../visualizations/USWords.png’)
```

```
Out[2]:
```

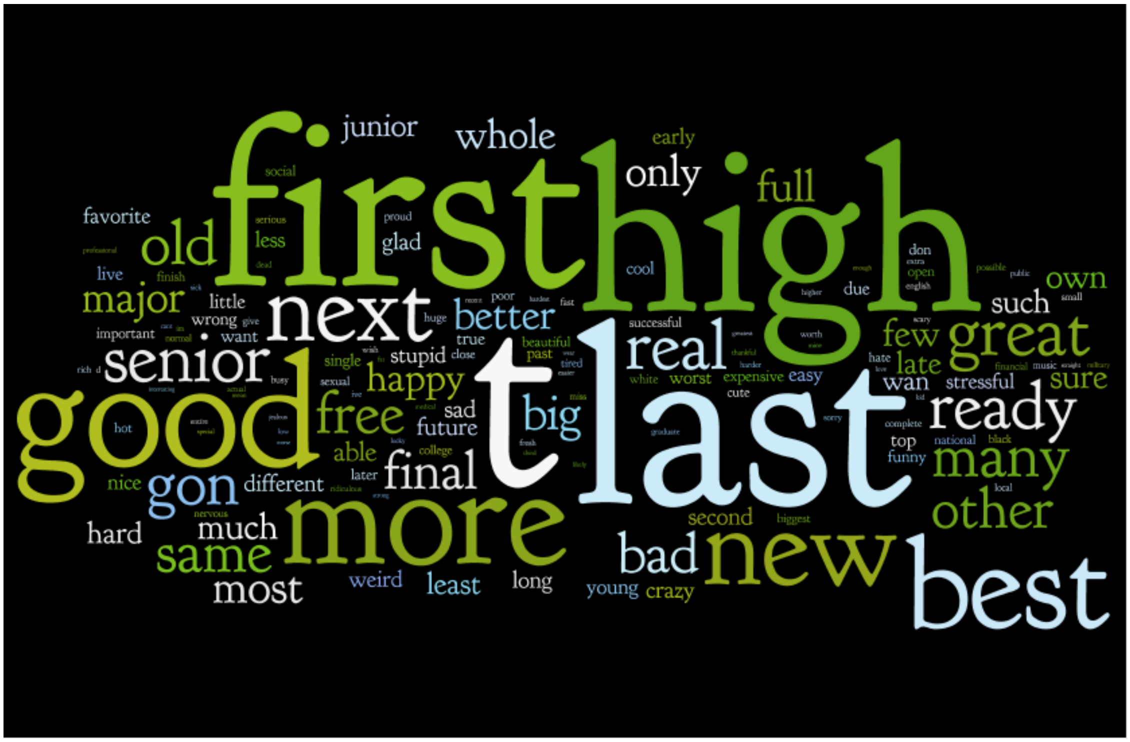


Most of the tweets obtained from the US either talk about the user's first year of college or the last days they will spend as a college student.

World

```
In [3]: Image(filename='../visualizations/WorldWords.png')
```

```
Out[3]:
```



This shows all the frequent words across the locations from which we gathered our data.

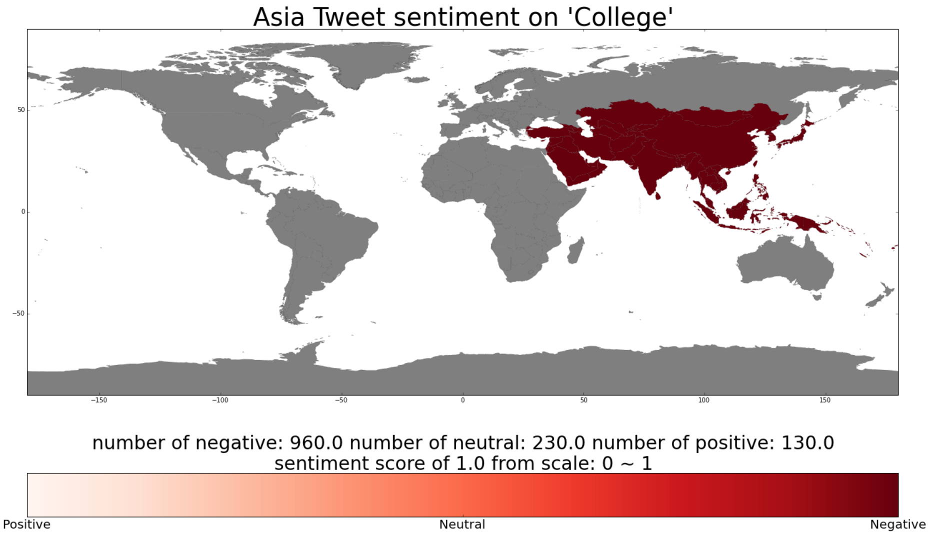
1.3.1 Visualizations: Global Sentiment

The following visualizations show differences in sentiment based on the keyword **college** by global region. These visualizations give us an idea of differing sentiment based on location and can help in making determinations on why there is differing sentiment if any (demographics? political?, socioeconomics?).

Asian Sentiment

```
In [6]: Image(filename='../visualizations/Asia.png')
```

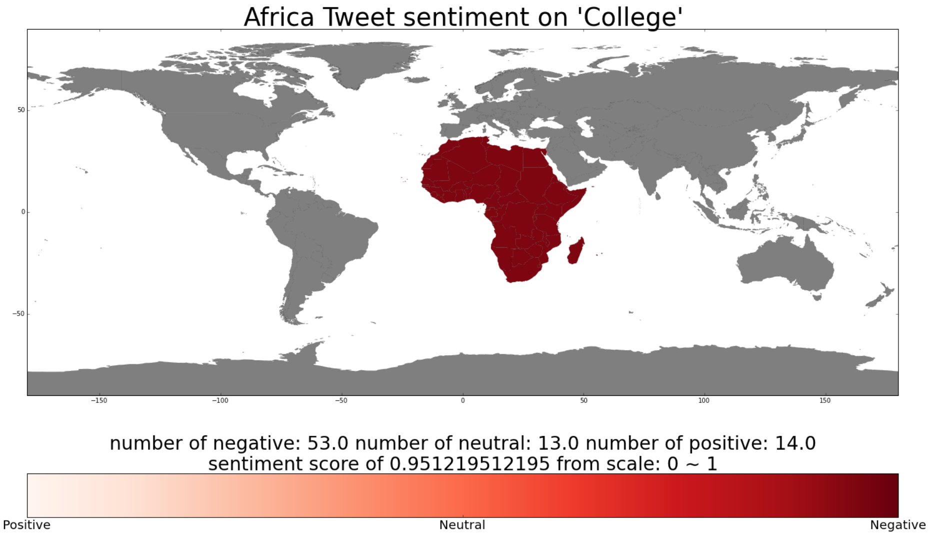
```
Out[6]:
```


Asia Tweet sentiment on 'College'
50
0
−50
−150
−100
−50
0
50
100
150
number of negative: 960.0 number of neutral: 230.0 number of positive: 130.0
sentiment score of 1.0 from scale: 0 ~ 1
Positive
Neutral
Negative

African Sentiment

```
In [19]: Image(filename=’../visualizations/Africa.png’)
```

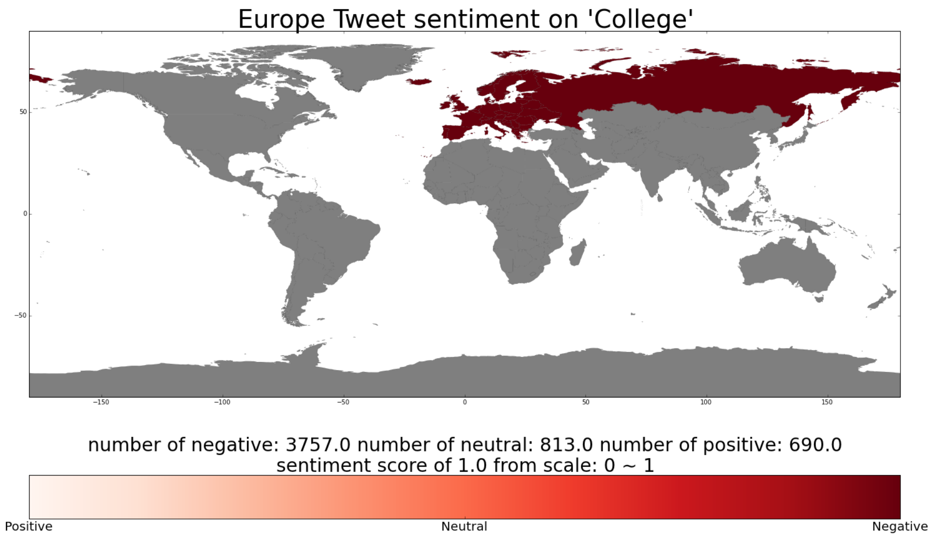
```
Out[19]:
```


Africa Tweet sentiment on 'College'
50
0
−50
−150
−100
−50
0
50
100
150
number of negative: 53.0 number of neutral: 13.0 number of positive: 14.0
sentiment score of 0.951219512195 from scale: 0 ~ 1
Positive
Neutral
Negative

European Sentiment

```
In [20]: Image(filename='../visualizations/Europe.png')
```

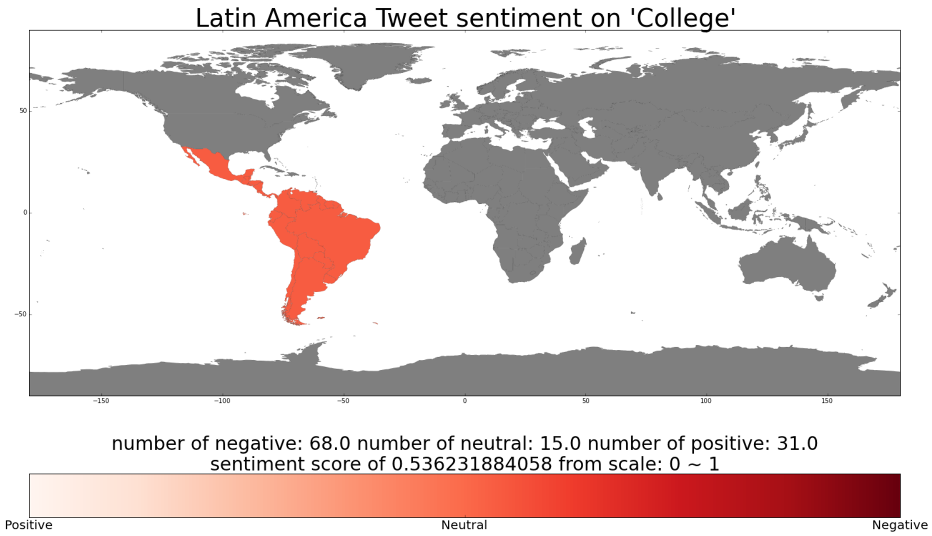
```
Out[20]:
```


Europe Tweet sentiment on 'College'
50
0
−50
−150
−100
−50
0
50
100
150
number of negative: 3757.0 number of neutral: 813.0 number of positive: 690.0
sentiment score of 1.0 from scale: 0 ~ 1
Positive
Neutral
Negative

Latin American Sentiment

```
In [21]: Image(filename='../visualizations/Latin America.png')
```

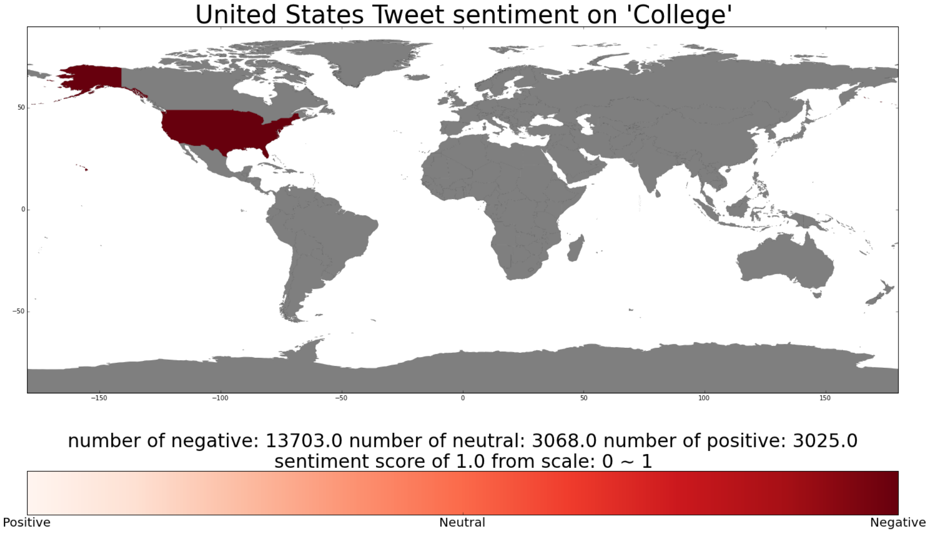
```
Out[21]:
```


Latin America Tweet sentiment on 'College'
50
0
−50
−150
−100
−50
0
50
100
150
number of negative: 68.0 number of neutral: 15.0 number of positive: 31.0
sentiment score of 0.536231884058 from scale: 0 ~ 1
Positive
Neutral
Negative

American Sentiment

```
In [22]: Image(filename='../visualizations/United States.png')
```

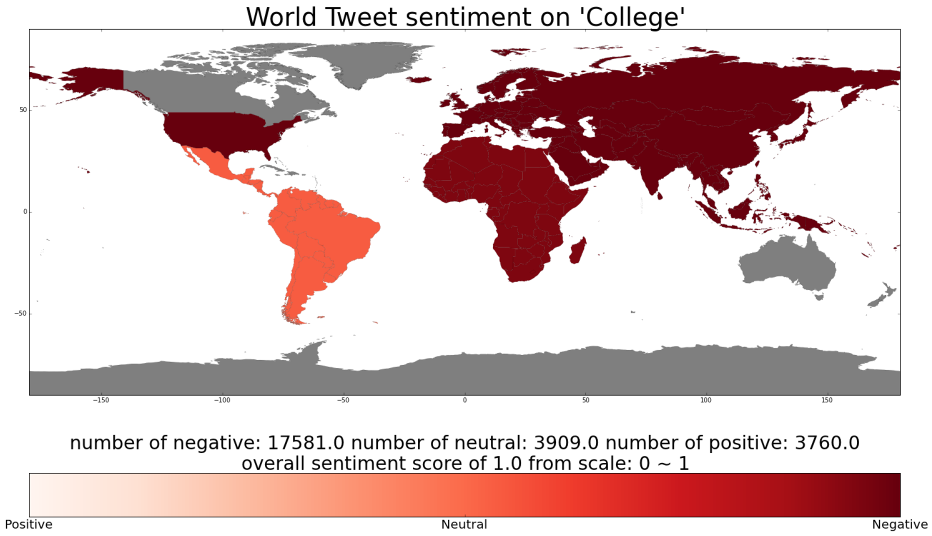
```
Out[22]:
```


United States Tweet sentiment on 'College'
50
0
−50
−150
−100
−50
0
50
100
150
number of negative: 13703.0 number of neutral: 3068.0 number of positive: 3025.0
sentiment score of 1.0 from scale: 0 ~ 1
Positive
Neutral
Negative

World Sentiment

```
In [24]: Image(filename='../visualizations/World.png')
```

```
Out[24]:
```


World Tweet sentiment on 'College'
50
0
−50
−150
−100
−50
0
50
100
150
number of negative: 17581.0 number of neutral: 3909.0 number of positive: 3760.0
overall sentiment score of 1.0 from scale: 0 ~ 1
Positive
Neutral
Negative

1.4 Conclusions:

Overall, we see a global sentiment of negativity towards **college**. All of the global regions besides Latin America vastly display a negative sentiment. Although in Latin America, the sample size is much smaller so it is difficult to make conclusive remarks on the level of negativity. These results suggest that there is a universal negative feeling towards college regardless of location, socioeconomic status, and culture.

According to this year's Twitter study of Beevolve, 73.7% of Twitter users are of age 15-25, where more than half of its total users live in the United States (with 50.99%). Twitter has not yet reached rural areas, thus we could assume that the areas identified in this research are mainly from highly industrialized areas of the world. Due to the dominance of 15-25 year-old users, we can also presume that most are in or on their way to college. There are many factors as to why our results mainly lead to negative sentiments towards college. Given that these data were obtained in a 2-week period (last two weeks of April), the tweets could have been highly influenced by finals/graduation/college application/acceptance, etc. During this period, the other side of the world, such as Asia, is undergoing final examinations and extreme pressure to do well.

Additionally, the words **first**, **high**, and **last** appeared the most in the respective regions analyzed above. These words may well explain the sentiment most users (gathered in this research) feel towards college. Examples from various locations include:

- "@EssentialFact: A study has found that the first two years of college are basically useless."really sums up everything.' *(Asia)*
- 'Sila ayos na college papers and files tas June-ish pa classes nila, ako last week of May na waley pa rin na aaccomplish. DLSU pls. Huhuhu' *(Asia)*
- "First I don't get to move Jaki into college and now I find out that we don't have the same spring break. FUCK EVERYTHING" *(US)*
- 'In an email to extended family my dad just referred to my school as "clown college" and discredited everything I've done over last 4 years.' *(US)*
- 'first day back at college tomorrow how am i going to do this help me.' *(Europe)*
- 'Last week in college poses a big struggle.. So much work to do but so much going out to do at the same time. *(Europe)*

- ‘Things i never learned in high school: how to: pay bills buy a house apply for college but thank god i can graph a polynomial function’ *(Africa)*

Seeing that the sample size for some of the sentiment analysis are very low, users in this region can cause some bias in the analysis. Users could also tweet as often as they want about the same topic for a certain length of time, such as college. The sentiments, which we have gathered, may be highly obtained from a few sample of users of whom expressed the same sentiment about college for some amount of time. We also have to consider that some of the tweets obtained are mixed with English and some other languages, including but not limited to Spanish, Indonesian, Filipino. Words written in these languages will not be classified by the Classifier (refer to ../NB3_data_analysis for the sentiment analysis). The foreign words may have been suggesting positive sentiments, but our Classifier is limited to the English language.

Therefore, we cannot fully discern whether other factors, such as socioeconomic status, cultural tradition of the country, or their current standards of education, play the most significant roles in the data gathered. However, we could conclude that based on these recent tweets, users feel an overarching negativity towards college due to a some of the following reasons: high amount of work, acclamation from family, college culture shock after Sprin Break, etc.