PYTHON FOR DATA SCIENCE

CSCI 5015

PROJECT PROPOSAL

Proposed title:

Applications of Python Programming in Twitter Mining

TEAM MEMBERS:

MAKANJUOLA OGUNLEYE

MIRACLE DICKSON

COURSE INSTRUCTOR:

DR JOHN NICHOLSON (ASSOCIATE PROFESSOR)

**INTRODUCTION**

**Hate Speech**

A hate speech is any speech made to arouse and propagate hate or prejudice against an Individual or a group. It is characterized by the use of insultive and derogatory words to stigmatize, minimize or

**STATISTICAL METHODS AND MODELS**

Python programming language has become a mainstream language and its applications cuts across several industries including, Information Technology, Finance, Government, Market research and many others. We are proposing to examine the application of Python in discovering information from text documents.

**MOTIVATION**

Our motivation is gotten from a Reality TV show competition in Nigeria called Big Brother Nigeria. This show has gathered public attraction and has generated a lot of controversies among Nigerians. Perhaps one thing that makes this show popular is its eviction system, it requires the public to vote through Text message or an online voting platform. Each contestant with the lowest vote during a voting period would be evicted. The interesting thing is that most Nigerians now take to twitter to share their opinions and perception about their favorite contestant. Thus, tons of thousands of tweets are added daily mostly from Nigerians about this reality show. We want to extract this data from twitter, analyze them and check for interesting information about them.

**DATA SET AND DATA COLLECTION:**

We are taking advantage of the fact that the show is currently ongoing, and we are trying to gather tweets about the show using the hashtag **#bbnaija.** We plan to extract between 500,000 to 1,000,000 tweets on this Hashtag. The good thing is that twitter has a feature that allows developers to download streams of tweets as they are being posted into a database. This feature is accessed through the twitter streaming API. We have already connected to twitter using twitter streaming API and we have been able to gather **262,801 tweets**. This number is increasing daily and we would stop when we reach the 500,000 tweets target or if we still have enough time, we would gather up to 1,000,000 tweets.

Attached is a link to our **262,801 tweets** in a csv format.

<https://goo.gl/WrMbt8>

**OBJECTIVES:**

We hope to answer the following questions at the end of the project:

1. How are tweeters distributed by location?

We would like to know the locations with the most tweets, and general online social engagements in Nigeria. Nigerians abroad are known to be very much engaged with politics and social activities in Nigeria. We would try to confirm this information.

1. Who are the top contestant by popularity on twitter?

Here we would count the number of times each contestant appears throughout the extracted tweets. Being able to discover this information can reveal to us the contestants that are mostly talked about.

1. Number of tweets that contains each contestant?

This question is a bit like the previous one, the difference is that in the previous question we are trying to get the total occurrence of each contestant name, but here we are trying to see the number of tweets about each contestant.

1. Words with highest correlation with each contestant and generating N-grams?

In text mining, an n-gram is a phrase or combination of words that may take on meaning that is different from, or greater than the meaning of each word individually.

We would generate n-grams using python to group words that have high correlation together.

1. Explore some Word Clustering algorithms such as (K – means clustering or Hierarchical Clustering)
2. Who are the Top tweeters and their location?
3. What are the sentiments around each contestant?

We would perform a sentiment analysis on the tweets that includes each contestant and graph these sentiments.

1. **WORD CLOUDS:** We would Generate a word cloud of the entire Text document.

References:

<https://marcobonzanini.com/2015/03/02/mining-twitter-data-with-python-part-1/>

<http://adilmoujahid.com/posts/2014/07/twitter-analytics/>

<https://www.toptal.com/python/twitter-data-mining-using-python>