Magazines Analysis
Africa Science and Society for Science

Executive Summary

In this exploratory analysis, I explored two sets of data sets pull from Twitter for the Africa Science and the Society for Science magazines. Both magazines mostly cover science, health, and environmental topics but from various places. Africa Science is created Kenya while the Society for Science is headquarters in Washington DC. The findings are not conclusive; however, they are informative of words patterns that are particular to one magazine than the other. Lexicon plots show words that are very prominent in one magazine than in the other.

Introduction

This analysis assumes that characteristics related to topics such as science, education, and environment will be widely covered in both magazines and appear frequently. I was curious to analyze words that followers use to describe these topics, being in different continents, yet speaking the same language: English. I used a Naïve Bayes Classifier to classify patterns of words. I was expecting to find pattern words of words such as development, health, education, agriculture, global, warming, … in the descriptions of twitter followers of Africa Science and Society for Science magazines. Unexpectedly, there are large patterns of followers of Africa Science magazine who did not seem to be real people but bots. Using a ratio of friends to followers, I looked at the patterns of subscribers whose accounts seemed dubious.

The Datasets

I downloaded twitter pulls of two magazines, Africa Sciences News and Society for Science. Africa Sciences News is Africa's leading website for science for development, health and environment. The magazine covers issues and informs of topics related to science, health, development, etc. in different African countries. It has 18,141 followers on twitter.

Founded in 1921 by journalist Edward W. Scripps and zoologist William Emerson Ritter, Society for Science & the Public is a non-profit membership organization dedicated to the public engagement in scientific research and education. It has 20,515 followers.

Methods

Naïve Bayes Classifiers

Naïve Bayes (NB) is a technique for constructing classifiers. The model assigns class labels to the followers' descriptions of their twit posts. Naïve Bayes is a conditional probability model. Given a set of words in the descriptions of twitter followers of Africa Science News or the Society for Science & the Republic magazine represented by a vector word = ($word_1$, …, $word_n$) representing some n features of words in the descriptions, the NB assigns to this occurrence of words the probabilities:

$$p(C_k \mid (word_1, \ldots, word_n) \text{ for each of K possible classes } C_k$$

where p is the probability that a given word/sentence is coming from the Africa Science or Society for Science twitter API data.

Word Classification

The NB classified whether a given word is from Africa Science Magazine or Society for Science based on the measured features. The measured features words are science, environment, health, … The features of classified results in this analysis show that word comedy is said 229 times in Africa Science magazine than in the Society for Science magazine. l'khbr is said about 149 times in Africa Science magazine than in the Society for Science magazine. ltknwlwjy is said about 125 times in Africa Science magazine than in the Society for Science magazine. Journalists is said about 104 times in Africa Science magazine than in the Society for Science magazine. The next five words are probably in African languages, which make sense if they are said that many times in Africa Science magazine than the USA magazine. I will ask my Kenyan friends what those words mean. I assume they are in Swahili. It is interesting to see that local languages are used with the English language, but it is natural to do so, assuming that that all readers understand it. Now, I want/am interested to also pull up the sentences where those words are used to see if I could guess/understand them in context.

Test of the Accuracy of the Model

I use random.shuffle and then split the features sets into a training set and test set to test for the accuracy of the NB classifier. The training model has correctly classified the word description for each magazine 79.54% of the time.

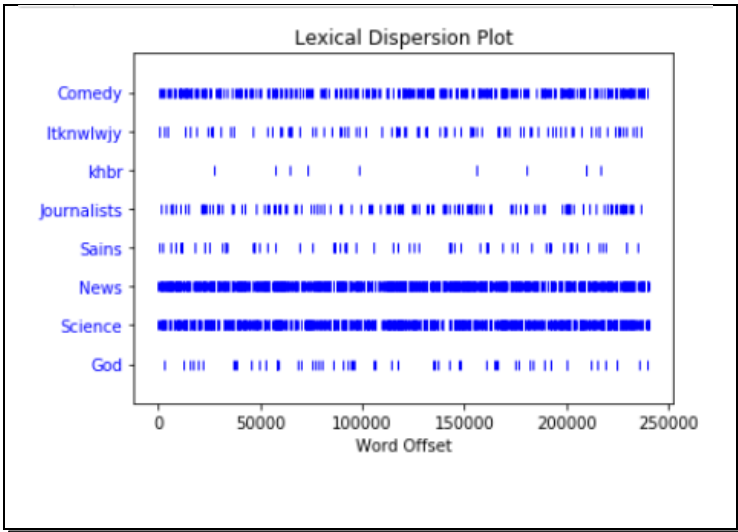
Lexicon

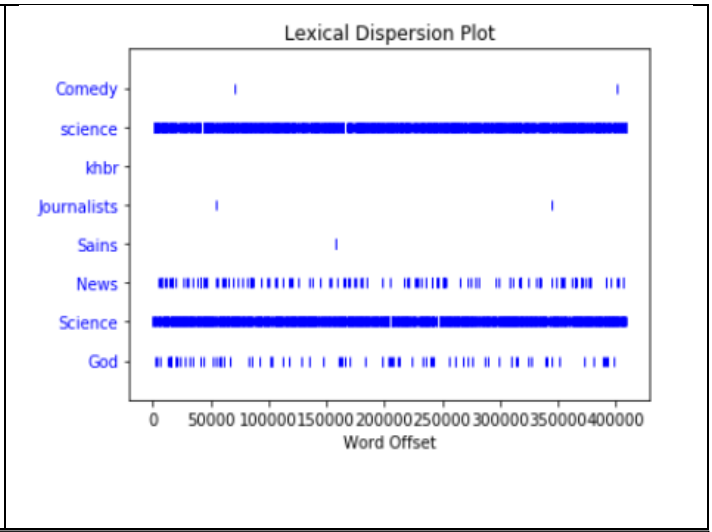
RESEARCH MORE ON WHAT DOES A LEXICON MEAN?

The classifier in the natural language tool kit help to visually look at the linguistic patterns of words that occur frequently in these magazines. Figure 1 and 2 respectively show a frequency of words from each of the descriptions of the magazines. Although the word *Comedy* is the top word used in the descriptions of Africa Science magazine, it is almost not used in the Society for Science magazine. The fact that this word is highly used by followers of Africa Science Magazine is very odd. A closer look at the sentence where those words are used reveal some suspicions of whether those words are used by real followers of by bots. For example, a sentence in which the word is used multiple times by the same follower.

Lexical Dispersion function in the NLTK automatically detects the frequency of how often a word occurs in a text and displays some words that appear in the same context. The lexical dispersion below shows the frequency of word types from the Africa Science magazine and the next one the frequency of words from the Society for Science & the Republic magazine

Figures 1 and 2 show changes in words in the two magazines overtime.





Left side: Figure 1: Lexical dispersion plot of Africa Science Magazine. Right side: Figure 2: Lexical dispersion plot of the Society for Science & the Republic.

Words in Context

Sentences Showing the Use of the Word "Comedy" in the Africa Science Twits.

Figure 3: Words in Context of Using the word Comedy in the Africa Science Twitter Page

```
050192901 Dungun , Terengganu 13 91 Comedy Filem Berita Dunia Sukan Tempatan B
4945105920 Gombak , Selangor 25 407 Comedy Music Movies Arts & Culture Photogr
l`rby @ lmtHd @ 30 90 Entertainment Comedy Photography Faith & Religion aaw711
192565815345152 Sibu , Sarawak 2 93 Comedy Filem Muzik Sains & Teknologi Hibur
 General Entertainment Music Movies Comedy Anime Arts & Culture Food Travel Ph
mpat , Kelantan 24 99 Business News Comedy General News Local Sports Local Spo
538852865 Malacca , Malaysia 10 104 Comedy Comedy Comedy Comedy Music Movies S
65 Malacca , Malaysia 10 104 Comedy Comedy Comedy Comedy Music Movies Science
cca , Malaysia 10 104 Comedy Comedy Comedy Comedy Music Movies Science & Tech
alaysia 10 104 Comedy Comedy Comedy Comedy Music Movies Science & Tech Science
uzik Sains & Teknologi Berita Sains Comedy Hiburan Am Filem Hiburan Muzik Tekn
ited Arab Emirates 6 101 Basketball Comedy General Entertainment Local Sports
0032 Muar , Johor 6 104 Muzik Anime Comedy Filem Pelancongan YongHanscom Yong
000102871040 London , England 9 285 Comedy Food Food Music Movies rihamakli Ri
ainment General Entertainment Music Comedy Science & Tech Science News Travel
```

On the other hand, the words *Science*, *News*, and *God* have high frequency in those magazines. The characters are different when the words are non-English words. For example, when looking at the words in context in the Africa Science magazine, the word "Journalists" is in sentences that are more formal and informative about different news channels in Africa and in the Emirates. However, the other top words "Comedy", "ltknwlwjy", "Sains" are in sentence that have unusual characters, misspelled words, and repetitive words in the same sentences. For example, technolgy is spelled as Teknologi and go along with the words Sains. For these reasons, some of the sentences are suspicious, leading the steps below

to know more about the followers of AfriScience magazine. The objective is to find which followers are real users and which ones are bots.

Exploratory Analysis of Africa Science Twitter Followers.

a) The Africa Science Magazine Dataset

A closer look at the twitter followers show that 350 followers have exactly 96 friends and mostly and less or equal to 68 followers. Most of them have 0 followers.

b) Bot accounts

Twitterbot is an automated software that acts as a real person twitter account by performing actions such as twitting, retweeting, liking, following, unfollowing or sending direct messages to other accounts. Some bots could have malicious contents, which are unwelcome news for users who might be following them back. There is some literature about these issues. Using a Depth-First Search (DSF) crawl and the public timeline API methods, Chu et al. 2012 defined characteristics that can be computed to detect bots. Those characteristics are: periodic and regular timing of tweets; whether the tweet content contains known spam; and the ratio of tweets from mobile versus desktop, as compared to an average human Twitter user. Bots are becoming smarter are there more methods are being defined to detect them. The dataset used on this analysis does not meet these characteristics. Moreover, due to time constraint and lack of in depth information, a simpler method is used to analyze the ratio of friend to follower.

c) Analysis Africa Science Magazine Using the Follower-Friend Ratio

The expectation in twitter account is that the more people you follow, more likely friends, then they are more likely to follow you back. Even though this is not the case with some celebrities, because they usually have large numbers of followers who would like to keep up on their activities. Each twitter user account has on their page:

1. The number of follower of the user
2. The number of tweets that they've posted
3. The number of people that that user is following

These information about the user can be used to compute the ratio. If 1 is greater than 3 then, this user has a positive ratio, it could mean that this user is worth looking a; If 3 is greater than 1, then the likelihood that this user is a spammer or marketer is very high (CrunchBase, 2009). Table 1 present a descriptive statistic of the doubtful 350 accounts of Africa Science followers, also known as bot.

d) Visualizing the Africa Magazines Twitter Followers Account

The graph shows that there are many accounts with this ratio close to 1. 3 is greater than one providing these ratios in Figure 1.

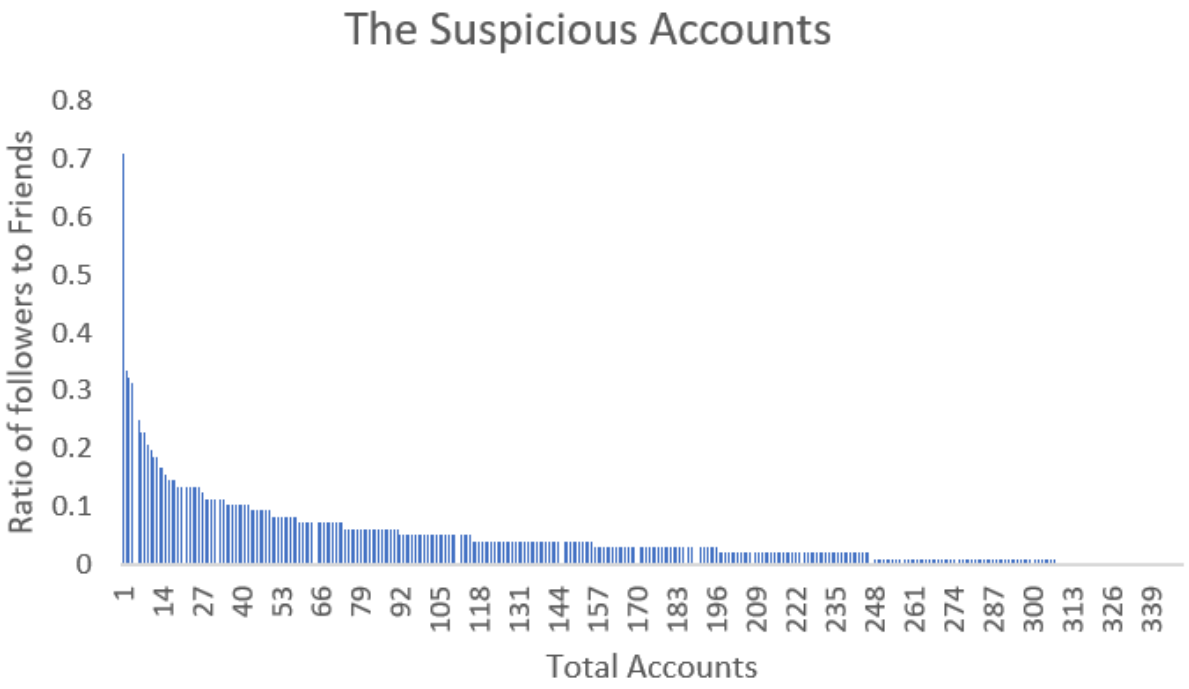


Figure 3: Ratio of the suspicious accounts

Conclusion

An exploratory analysis of the Twitter API data of these two magazines is very informative of words choices and patterns. Word choices can be helpful to know the nature and behaviors of subscribers.

Sources:

Datasets:

Africa Sciences News: twitter: @Afriscience. Link: www.africasciencenews.org. Society for Science & the Public: twitter: @Society4Science, link: www.societyforscience.org

Course Material:

Prof: John Chandler. Naïve Bayes Classifier: ADA material, www.nltk.org/book/ch06.html ,

Sentiment Analysis

Articles and other:

Chu, Z. Gianvecchio, S. Wang, H. Jajodia, S. 2012. "Detecting automation of twitter accounts: are you a human, bot, or cyborg?" IEEE Transactions on Dependable and Secure Computing, Vol 9. No. X http://www.cs.wm.edu/~hnw/paper/tdsc12b.pdf

CrunchBase: "Twitter's Golden Ratio (That No Ones Likes To Talk About)" https://techcrunch.com/2009/08/26/twitters-golden-ratio-that-no-one-likes-to-talk-about/

Brandom, R. 2017. "How to spot a Twitter bot". https://www.theverge.com/2017/8/13/16125852/identify-twitter-bot-botometer-spambot-program