Author: Brian Pankau

Class: UWLAX DS710

Assignment: 13 Project Proposal

**What is your Question?** *Is Global Warming a Hoax?*

The questions that my research will attempt to answer are:

* Using sentiment analysis of the words used within the tweet, can it be determined whether or not the username of the tweet asserts or denies whether or not the concept of global warming is a hoax?
* What are the most frequently used words in the twitter texts that incorporate the hash tag #globalwarming?
* What are the words that are associated with each of the two viewpoint (supporter, denier)?
* Are there usernames that tweet more frequently than other tweeters?
* Do temporal factors (day of week, time of day) change the frequency distribution of the tweets?

As a stretch goal:

* What are the other top four hash tags that are referenced in the #globalwarming tweets?
* How do these other hashtags support the viewpoint of the tweeter?
* What is the frequency distribution of the words used within each of these five hashtags?

**What is your Audience?**

The audience of this research are lexical researchers who may be interested in using word frequency to categorize tweets without hashtag identifiers that discuss global warming concepts.

**What information will you gather about each tweet in order to address your question?**

This research study will gather

* Words used in tweets that contain the hash tag #globalwarming
* The username of the tweet
* The day of the week when the tweet was issued
* The time of the day when the twee was issued

**How will you collect the data?**

This research study will use a combination of REST and Streaming APIs and Python to collect tweets that contain the hash tag #globalwarming, and will use R to statistically analyze the usage of the words within the tweets to determine whether or not a username believes that global warming is a hoax.

**How will you analyze the data?**

The viewpoint for each tweet will be classified by hand as either a supporter or a denier for global warming. A list of stop words will be developed for exclusion.

The words within each tweet will be counted and associated with one of the two viewpoints. Frequency analysis will be used to determine:

* What is the most common word used?
* What is the least frequent word used?
* What is the average length of the tweet in the number of characters?
* What is the average length of a word within a tweet?

General Hypothesis: There is a set of words that can be used to categorize a tweeter’s viewpoint on global warming into two categories: a supporter of global warming, and a denier/non-support of global warming.