The Kidneys Can Tweet?! An Exploratory Analysis of Twitter Hashtags Related to CKDu.

Daniel Smith {daniel.smith2@emory.edu}

2/6/2019

Overview and Motivation

Being an avid twitter user has showed me the value that global connections can have when discussing research ideas and connecting with other scientists. Applying data analytic techniques to tweets has proven to be a wealth of knowledge for other researchers. Due to this, I am eager to gain new skills that will allow me to function more proficiently in the world of nursing data science. Thru this project, I will gain insight and knowledge about the current conversations occurring between other scientist in regard to the research behind the current epidemic of chronic kidney disease of unknown etiology (CKDu), which is the focus of my dissertation project with Drs. Hertzberg and Mac. For my dissertation, I am studying the occupational histories and exposures of immigrant patients receiving emergent dialysis in Atlanta, GA, who are hypothesized to suffer from CKDu. By conducting an analysis of tweets related to CKDu, I will gain a more in-depth understanding of this global epidemic that will shape my understanding and thinking of my own dissertation project while, at the same time, helping me to gain new data science skills.

Project Objectives

This project will seek to characterize tweets about the CKDu epidemic. **Q1**. What are the frequencies of hashtags used when tweeting about the CKDu epidemic? **Q2**. What are the frequencies of keywords used when tweeting abotu the CKDu epidemic? **Q3**. Where are people from who are tweeting about the CKDu epidemic? **Q4**. What are the emtions of tweets about the CKDu epidemic?

By exploring the four questions above, I hope to learn methods for accessing and searching twitter data, cleaning and processing extracted tweets, and analyzing tweets for frequency, location of origin, and sentiment.

Data

Using R package 'rtweet' data will be extracted from Twitter.com a social media website that allows users to tweet short messages of 280 characters or less. Common data that can be acquired via tweet analysis includes textual/sentiment analysis, analysis of the social connections between users, and geolocation analysis of where a tweet originated. This project will be focused on analyzing tweets with keywords and hashtags related to the global CKDu epidemic. Currently, candidate key words include 'chronic kidney disease of unknown etiology', 'enfermedad renal crónica no tradicional', 'mesoamerican nephropathy', and 'chronic interstitial nephritis in agricultural communities'. Candidate hastags are #CKDu #CINAC #MeN and #ERCnt.

Data Wrangling:

From what I have seen in the literature thus far, there will be preprocessing of the tweet data required before I can analyze my data. Tweets will need to be converted to lower case letters; usernames, links, punctuation, digits, and stopwords will need to be removed; and stemming of words (i.e. making "walking""","walk", and"walks"" all appear in the base form of "walk") will have to occur. The R package 'tm' will be used for preprocessing of the data.

Exploratory Analysis:

Exploratory Analysis will be conducted first with a subset of the total tweets acquired from Twitter. The 'tm' package in R will be used to summarize and give frequencies for the keywords and hashtags of the subset of downloaded tweets. Location of origin for the subset will be analyzed using both the 'ggplot2' and 'dplyr' packages. Simple sentiment analysis (positive vs. negative) will be conducted on this subset using the 'syuzhet' package in R.

Analysis

Once exploratory analysis is completed, the totality of the extracted tweets will be analyzed using similar methods as described above. However, in addition to the simple sentiment analysis (positive vs. negative), eight emotion sentiment analysis (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust) will be conducted to further understand the emotions behind CKDu tweets.

Sample Plots & Tables

Sample Plot of Number of CKDu Tweets per Country

```
tweets <- c(100,305,420,334,278,130)
country <- c("Costa Rica", "Nicaragua", "El Salvador", "Mexico", "Guatemala", "USA")
df <- data.frame(country, tweets)
ggplot(data =df, aes(x = country, y = tweets)) + geom_bar(stat = "identity")
```

Schedule

Completed Activities

-January 28, 2019 Meeting with Steve Pittard to discuss rTweet package and potential difficulties with analyzing tweets.

-January 28, 2019 Apply for Twitter Developer Account. A developer account is needed for access to Twitter's API that will allow for tweet extraction and data analysis in R. Developer Account Applicaiton Status: **Approved!**

Future Activities/Deadlines

-Week of **Feb 11**: Finalize milestone 1 and submit by February 13.

-Week of **Feb 18**: Extract Tweets from Twitter

-Week of **Feb 25**: No activities scheduled due to SNRS Conference Attendance.

-Week of **March 4**: Preprocess the subset of tweets to be used in the exploratory analysis.

-Week of **March 11**: Analyze subset of tweets using the 'tm', 'ggplot2', and 'dplyr' packages to calculate frequencies of hashtags & keywords and to calculate location of origin.

-Week of **March 18**: Analyze subset of tweets for simple sentiment analysis using the 'syuzhet' package.

-Week of **March 25**: Finalize milestone 2 and submit by March 27.

-Week of **April 1**: Preprocess totality of the retrieved tweets.

-Week of **April 8**: Analyze totality of tweets using the 'tm', 'ggplot2', and 'dplyr' packages to calculate frequencies of hashtags & keywords and to calculate location of origin. Additionally, analyze totality of tweets for simple sentiment analysis using the 'syuzhet' package.

-Week of **April 15**: Finalize milestone 3 and prepare manuscript plus powerpoint for project.

-Week of **April 22**: Extra week built in as a buffer.

-Week of **April 29**: Present project on May 1.