SURV 622/SURVMETH 622 Fundamentals of Data Collection

Assignment #2: Using the Twitter API and Sentiment Analysis of Tweets

Due date: March 23

For this assignment, you will work with your group of four or five students. Using the rtweet package in R, each group will 1) use the Twitter API to "listen" to and download a corpus of tweets, 2) write a report that describes how the listening was done along with selected features of the corpus of tweets created, 3) clean the corpus, and 4) analyze the sentiment of tweets using dictionary-based methods in R.

The tweets of interest are tweets that discuss either the University of Maryland men's basketball team (for the Maryland students) or the University of Michigan men's basketball team (for the Michigan students). An important part of the exercise will be to choose appropriate keywords and hashtags for tweet selection.

You will want to begin listening as soon as possible. There are several upcoming events that might be expected to affect the volume of tweets about these teams and the sentiments expressed in them. These include the regular season game between Michigan and Maryland on March 8; for Michigan, the game against Wisconsin on February 27 (if possible), Ohio State on March 1, Nebraska on March 5; for Maryland, the game against Minnesota on February 26 (if possible), Michigan State on February 29, and Rutgers on March 3. Because you will be conducting some longitudinal analyses for which change in Twitter sentiment will be of interest, try to collect tweets over a period that spans at least one event, e.g., a very close game, a blow-out, something controversial or very positive – ideally multiple events – relevant to the team you are studying. You will have to decide what Listening – what Twitter content you select – should continue for a week or longer.

The report each group turns in for this assignment should specify how the keywords and/or hashtags used for collecting the tweets were selected. The keywords and/or hashtags should be listed, along with the times when listening started and ended. If keywords and/or hashtags were changed during the data collection process, please note the changes that were made, the reason(s) for them and when they occurred. Listening ideally will occur continuously, but if there are gaps, the time(s) of these gaps should be noted.

Next, clean and prepare the tweets for analysis. Exclude tweets that do not seem to concern the men's basketball team at either the University of Michigan or the University of Maryland, for example that concern football or women's basketball.

In addition to describing how the corpus of tweets was created, the report also should contain some simple descriptive information about the tweets. How many tweets were collected in total and by day? Is there a pattern with respect to the time of day or day of week when tweets were created? Is there a relationship between events and frequency of tweets? What are the words in the set of tweets you have assembled that appear most frequently? How does this change if you exclude “stop words” such as “a,” “an,” “the,” “is,” and others that are common in English sentences but are generally not informative?

Next, describe the sentiment in the corpus. Visually examine the dictionaries so you understand what kind of words they contain and how they differ from each other (you will be provided with code that uses five different sentiment dictionaries). Then analyze the sentiment of the cleaned corpus. Pick two dictionaries and describe how the results differ across the dictionaries for your team. Using one dictionary, how does sentiment change over time for the team? Using that same dictionary, how does the sentiment of original tweets and re-tweets differ? Why do you suppose this is the case?

Send any questions to Fred Conrad (fconrad@umich.edu) and Robyn Ferg (fergr@umich.edu).