SURV 622/SURVMETH 622 Fundamentals of Data Collection

Assignment #3: Automated Coding of Tweets

Due date: Monday, April 6, 2019

This exercise is intended to expose you to the use of machine learning for interpreting tweets. It fits closely with Assignment #2 in that you are asked in the current assignment to conduct an analysis of the tweets you collected in Assignment #2 by listening to the Twitter API. Please download MLtexttutorial.R for example code. You will need to edit some code. Be sure to have the latest version of R installed.

You will work in the same group you worked in for Assignment #2 and will be provided with R code that will produce much but not all of the information you need. There are two parts to the assignment:

Part 1: Cleaning, preparing, and hand-coding the tweets to create a dataset

1. 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.
2. Each member of the group should read and assign sentiment to 200 tweets. Using categories like "Positive" and "Negative" is highly recommended, but the exact methodology is up to you. Be sure that all team members use the same set of coding categories.
3. Pool your hand-coded tweets to increase the size of your hand-coded corpus.
4. Create features based on cleaned text data.

Part 2: Sentiment analysis versus ML coding of sentiment

1. Use your hand-coded tweets as your data to build a machine learning model, with the hand-coded sentiments as your label. Try a series of different models such as K-Nearest Neighbors and Decision Trees, and decide on the final model based on performance metrics.
2. Use sentiment analysis (from Assignment #2) to automatically code sentiment in the testing set(s).
3. Describe the differences between the sentiment analysis and ML. Be clear about what performance metrics are being used, and how to interpret the results. Why do you think these differences are observed?