SURV 622/SURVMETH 622 Fundamentals of Data Collection

Assignment #2: Using the Reddit API

Due date: March 25, 2024, 3:30 PM.

For this assignment, you will work with your pre-assigned group of three or four students, with at least one Michigan student in each group. Each group will 1) use the Reddit API to "listen" to and download a corpus of tweets using the RedditExtractoR package in R, 2) write a report that describes how the listening was done along with selected features of the corpus of posts created, 3) clean the corpus, and 4) hand-classify a subset of posts.

The posts of interest are to be determined by your group and should be on a polarizing topic with primarily two opposing viewpoints. Previous topics include the COVID-19 pandemic and vaccine-related issues, such as vaccine skepticism, vaccine mandates, boosters, and pandemic-related remote work. An important part of the exercise will be to choose appropriate topic, keywords, and subreddits for post selection.

You will want to begin listening as soon as possible. You will have to decide what Reddit content you select. Listening should continue for a week or longer. If after a day or two of listening it is evident that you are not selecting a sufficiently large corpus of posts to carry out the analyses described below (you want to aim for hundreds of posts at minimum so you can project over the time available for this assignment), you might want to broaden the scope of your keywords or expand the set of subreddits you are using.

The report each group is asked to turn in for this assignment should specify how the keywords and/or subreddits used for collecting the posts were selected. The keywords and/or subreddits should be listed, along with the dates when listening occurred. If keywords and/or subreddits 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 regularly, but if there are gaps, the time(s) of these gaps should be noted.

Next, clean and prepare the posts for analysis. Exclude posts that do not seem to concern your group's topic. You will need to manually inspect some posts in your corpus to get a sense of what irrelevant posts, if any, may be included in your corpus.

In addition to describing how the corpus of posts was created, the report also should contain some simple descriptive information about the posts. How many posts were collected in total and by day? How many posts after data cleaning? Is there a pattern with respect to the time of day or day of week when posts were created? Is there a relationship between events and frequency of tweets? What are the words in the set of posts 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, your group will hand-code a subset of your scraped Reddit posts. This hand-coded dataset will be used as a training set in Assignment #3. Using the cleaned corps of posts, randomly select

100*n posts for hand-coding, where n is the number of group members. Each member of the group should hand-code 100 different posts. Each member of the group should read and assign a stance to each post based on the chosen polarizing topic, e.g., favor, oppose, neutral, or irrelevant. Then, pool your 100*n hand-coded posts to create a corpus of labeled posts. Report on how many posts are in each category.

Send any questions to Robyn Ferg (fergr@umich.edu).