COMP 598 Homework 1 – Mini Data Science Project

30 pts
Assigned Sept 9, 2021
Due Sept 17, 2021 @ 11:59 PM

This is an INDIVIDUAL Assignment – each student's work must be their own, each student completes this assignment, there are no teams for homework 1.

The goal of this assignment is to work through the data-handling phases of a mini data science project to put into practice the ideas we've discussed in the Unit 1 lecture. You are welcome to complete the exercises in this homework using whatever tools and programming languages you deem fit. In order to make ANY points, your assignment MUST pass the Homework 1 grading tests. Please watch the orientation video under Lecture Recordings in MyCourses for more information.

In this assignment, you will conduct an analysis of tweets produced by Russian trolls during the 2016 US election. These tweets were published by 538. You can read about them here.

In this mini-project, we'll be assessing the frequency with which troll tweets mention "Trump" by name.

1. Data Collection
    a. Download the raw tweet data. You will ONLY be using the data from the first file (IRAhandle_tweets_1.csv).
    b. Looking at only the first 10,000 tweets in the file, keep those that (1) are in English and (2) don't contain a question. This will be our dataset. To filter the right tweets out, take a look at the columns.
        i. There are specific columns that call our language. You can trust these.
        ii. Assume that a tweet which contains a question contains a "?" character.
    c. Create a new file (I would suggest in TSV – tab-separated-value - format) containing these tweets.
2. Data Annotation
    a. To do our analysis, we need to add one new feature: whether or not the tweet mentioned Trump. This feature "trump_mention" is Boolean (="T"/"F"). A tweet mentions Trump if and only if it contains the word "Trump" (case-sensitive) as a word. This means that it is separated from other alphanumeric letters by either whitespace OR non-alphanumeric characters (e.g., "anti-Trump protesters" contains "trump", but "I got trumped" does not).
    b. Create a new version of your dataset that contains this additional feature.
3. Analysis
    a. Using your newly annotated dataset, compute the statistic: % of tweets that mention Trump.
    b. It turns out that our approach isn't counting tweets properly … meaning that some tweets are getting counted more than once. Go through and look at your annotated data. Identify where the counting problem is coming from.

Submission Instructions

Download the template code from https://github.com/druths/comp598-2021 .

Your submission should pass the unit tests and contain – at minimum - the following:

- README.md (5 pts)
    - In 3 sentences or less, explain where the counting problem is coming from.
- dataset.tsv (20 pts)
    - This should be the output of your Data Annotation phase.
    - Format is tab-separated value, utf-8 (as long as you don't do anything fancy, it will be in utf-8) (5 pts)

    - The first line should be a header line (3 pts)
    - The file should contain the following columns, in this order: tweet_id, publish_date, content, and trump_mention. Tweets should appear in the same order they appeared in the original file from 538. (12 pts)
- results.tsv (5 pts)
    - Format is tab-separated value
    - The first line should be a header line, with headers “result” and “value”.
    - The second line should contain the result for “frac-trump-mentions”. If necessary, truncate your answer to three decimal places.

For partial credit purposes, you may also include the code that you used to do this work. It must be readable in a standard text editor. Remember that code readability and partial credit are correlated 😊