COMP 598 Homework 8 – Data Annotation

30 pts
Assigned Nov 12, 2020
Due Nov 20, 2020 @ 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 8.

Non-standard (i.e., built-in) python libraries you can use:

- pandas
- requests

In this assignment, we're interested in how many posts were made concerning vote counts in the US 2020 election over the three days of data you collected from Reddit for homework 6. To do this, you will be manually coding posts you collected in homework 6.

Task 1: Prep for coding (15 pts)

Write a script extract_to_tsv.py that accepts one of the files you collected from Reddit and outputs a random selection of posts from that file to a tsv (tab separated value) file. It should function like this:

> extract_to_tsv.py -o <out_file> <json_file> <num_posts_to_output>

If there aren't num_posts_to_output posts in the file, then the script should just output all those it finds. If there are more than num_posts_to_output (which is likely the case), then it should randomly select num_posts_to_output of them and just output those.

The output format (written to out_file) is:

> Name <tab> title <tab> coding
> <name of first post chosen> <tab> <title of first post chosen> <tab>
> <name of second post chosen> <tab> <title of the second post chosen> <tab>
> …
> <name of the n'th post chosen> <tab> <title of the nth post chosen> <tab>

Note that:

- we're including the "name" field because it uniquely identifies the post, in case you ever need to go back and check something in the original data
- whitespace between column value and the tab is optional
- the third column "coding" is intentionally blank. We'll be completing that in the next task.

Run your script on your three files you created (one for each day) to extract 75 posts for each day. You should have one file for each day.

Task 2: Code posts (15 pts)

Our typology in this assignment has two categories:

- vote count (v): this post title directly refers to the tallying of votes or the number of votes in the US 2020 election. Note that just mentioning the election (even though an election involves the # of votes cast) isn't enough. Examples:
  - Nevada begins counting votes.
  - Trump disputes the latest numbers reported by Georgia.
  - The recount will require election clerks to work overnight

- non-vote count (n): the post title doesn't directly refer to vote counting – the process or the tallies themselves.
    - Biden looks poised to win the election.
    - The situation is looking grim for democrats in Pennsylvania.
    - Trump claims voter fraud in multiple states.

Code all the posts that were extracted from your files by putting a "v" or "n" in the coding column next to each post. To do this, you can use a text file or, another option, would be to use a spreadsheet application – just make sure you export your results in tsv format.

Submission Instructions

Your MyCourses submission must be a single zip file entiled HW7_<studentid>.zip. It should contain the following items:

- scripts/
    - extract_to_tsv.py – script for Task 1
- data/
    - <date1>_posts.tsv
    - <date2>_posts.tsv
    - <date3>_posts.tsv

Notes on Grading

Your assignment will be graded

1. By the functionality of your script (in task 1)
2. Every sampled post was coded (in task 2)
3. The correctness of your codings (in task 2). As with all coding, there is a grey area and you and the TA won't agree on all the codings. To avoid these tricky cases, the TAs will take a selection of your codings (at random) and grade agreement on only the "obvious" cases – posts that are quite obviously about (or not about) voting. The idea is that this is more of a sanity check to ensure you took the coding seriously. So don't sweat to much the posts that are in a grey area.