COMP 598 Homework 7 – Data Annotation

35 pts
Assigned Nov 4, 2021
Due Nov 12, 2021 @ 11:59 PM

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

- **pandas**
- **requests**

In this assignment, we're interested in the main topics discussed on the /r/mcgill subreddit vs. the /r/concordia subreddit. We'll do this using human annotation … and you're the annotator 😊

Task 1: Data collection (10 pts)

First, let's collect the latest 100 posts (using the **/new** endpoint (do not use the /hot endpoint)).

Write a script "`collect_newest.py`" that collects the **100** newest posts in the subreddit specified. It should run as follows:

```
python3 collect_newest.py -o <output_file> -s <subreddit>
```

Collect two data files - one for mcgill and one for concordia subreddits. This involves running your script two times. Note that in the output data files, **you should have exactly one post (in JSON format) per line**. Do not indent the JSON output. The files should be named **concordia.json and mcgill.json**. Place them in the root folder of the submission template. Please read the README.md file in the repository for further instructions.

Task 2: Prep for coding (10 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:

```
python3 extract_to_tsv.py -o <out_file> <json_file> <num_posts_to_output>
```

If `<num_posts_to_output>` is greater than the file length, then the script should just output all lines. If there are more than `<num_posts_to_output>` (which is likely the case), then it should randomly select *num_posts_to_output* (the parameter you passed to the script) 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>
```

Here is an example:

```
Name  title coding
t3_jmmrja   "Easy Computer Science classes"
t3_jmm91k   "Cloudberry (+ Tri-pawed squirrel) Appreciation Post"
t3_jmg17h   "Breaking a lease over a persistent cockroach infestation?"
t3_jmfc0t   "Don't know how to cook"
t3_jmfj91   "everything is falling apart"
```

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.

We also need a specific output for this exercise (which will be completed on task 3). Run the following:

```
python3 extract_to_tsv.py -o annotated_mcgill.tsv mcgill.json 50
python3 extract_to_tsv.py -o annotated_concordia.tsv concordia.json 50
```

That means, run your script on your McGill and Concordia files you created, 50 lines in each. The output files, **annotated_mcgill.tsv and annotated_concordia.tsv,** should be submitted in the submission_template. Please check the README.md for further information.

Task 3: Code posts (10 pts)

Our typology in this assignment has three categories:

- **course-related (c)**
- **food-related (f)**
- **residence-related (r)**
- **other (o)**

Of course, there's a lot that will go into the other category. Using the files you produced in Task 2, **annotated_mcgill.tsv and annotated_concordia.tsv,** code all the posts that were extracted from your files by putting the appropriate category ("c", "f", "r", or "o") capturing what the post is MOSTLY about. In other words, you'll edit the files that you produced in the last task, so you're completing the third column.

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. **We won**'t be able to grade your assignment if you don't provide a .tsv file!

Here is an example:

```
Name  title coding
t3_jmmrja   "Easy Computer Science classes"     c
t3_jmm91k   "Cloudberry (+ Tri-pawed squirrel) Appreciation Post" o
t3_jmg17h   "Breaking a lease over a persistent cockroach infestation?" r
```

```
t3_jmfc0t   "Don't know how to cook"      f

t3_jmfj91   "everything is falling apart" o
```

Task 4: Analyze (5 pts)

Write a script called “analyze.py” which outputs the number of each category that appears in your annotated files. The script should run like this:

```
python3 analyze.py -i <coded_file.tsv> [-o <output_file>]
```

The “-o …” argument is optional. If omitted, print the result to stdout. In either case, the output should be written in JSON format like this:

```
{

“course-related”: 70,

“food-related”: 30,

“residence-related”: 20,

“other”: 80

}
```

Once you’ve run this, you can see how differently the university student communities use their subreddit.

Submission Instructions

Please check hw7 README.md and its template directory.