COMP 370 Homework 8 – Data Annotation

Assigned Nov 7, 2023
Due Nov 15, 2021 @ 11:59 PM

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

First, let's collect some reddit posts (using the ***/new.json*** endpoint – details here). We'll collect two data files. One from the McGill subreddit and one from the Concordia subreddit.

For the purpose of this assignment, collect them manually. Meaning, in a web browser, get the json dump and download it to a file. You should have a a mcgill.json file and a concordia.json file.

Task 2: Prep for coding

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: Develop an ontology

We’re analyzing what these posts are about. You need to develop an ontology. We’re aiming for 5-8 categories (this tells you the level of resolution).

Using the files you produced in Task 2, **`annotated_mcgill.tsv and annotated_concordia.tsv,`** conduct an open-coding of those posts. (It’s easiest to use something like Excel for this part of the process).

1. Start by going through all the posts and putting down the category that “seems right to you”
2. Review all the “categories” that you put down. You might have a lot. Consolidate them down to 5-15 categories.
3. Go back through an annotate posts again using just those ones. Keep track of any issues you encountered.
4. Based on your experience with the second round, tighten up the categories one more time so that you have the 5-8 in your typology. (if you’re not there yet, just repeat steps 3 & 4 again, possibly a couple times)

Submission Instructions

Submit a zip containing:

- mcgill.json
- concordia.json
- annotated_mcgill.tsv
- annotated_concordia.tsv
- extract_to_tsv.py
- README.md
    - In this file, list the final categories you came up with.