# COMP 370 Homework 9 – Data Annotation

Assigned Nov 1, 2024  
Due Nov 8, 2024 @ 11:59 PM

In this assignment, we’re interested in manually annotating the main topics discussed on the /r/mcgill subreddit vs. the /r/concordia subreddit – and we’ll be using other people as annotators. 😊

## Task 1: Prepare a taxonomy (coding) guide

Using your taxonomy from HW8, prepare a taxonomy (coding) guide. It should have the following sections:

1. Motivation – why this taxonomy was built
2. Each taxonomy category
   1. Definition
   2. Examples (what’s in)
   3. Edge cases (what’s in, but might not look like it; what’s out, but might look like it)

This document will likely be 2-3 pages long.

## Task 2: Prep data coding

Use the extract\_to\_tsv.py script from last homework to extract 20 posts from concordia and 20 posts from McGill.

## Task 3: Collect annotations

Find two people who are willing to be an annotator for you. This could be a classmate in COMP 370, a friend, significant other, parent, etc… Run a coding orientation for each (or both at the same time):

1. Explain your study to them
2. Walk them through your taxonomy guide
3. Answer any questions

Give them a copy of the Concordia and McGill tsvs that you made in Task 2. Make sure to provide these in a format that will work for them (Excel, Google Sheets, etc…).

Perform the annotation exercise yourself as well.

Once they’ve returned the completed annotation documents, you should have three annotations from three people (you and your two annotators). Proceed to Task 4.

## Task 3: Compute annotator agreement

1. Obtain all posts you have majority agreement on (at least 2 annotators agreed on the same label)
2. Obtain all the posts you have complete agreement on (all 3 annotators agreed on the same label)

## Task 4: Assess errors

Look at the places where there were annotator disagreements. Where did annotators completely disagree? Where did they not completely agree? The disagreements are due to SOMETHING. What do you think it is? (It could be more than one thing)

## Task 6: Derive your labeled data

For all posts with majority agreement, give that post the annotation that was assigned by at least two annotators.

For those that did not have a majority annotation, act as the arbitrator and decide which of the categories named should be assigned.

This produces the final labeled dataset.

## Task 5: Plot results

Using only your final labeled dataset, make a bar chart showing the relative abundance of each topic, contrasting Concordia and McGill. What differences do you observe?

## Submission Instructions

Submit a zip containing:

* taxonomy\_guide.md
* mcgill.json
* concordia.json
* final\_labeled\_dataset\_mcgill.tsv
* final\_labeled\_dataset\_concordia.tsv
* results.png
* error\_assessment.md
  + Give your error assessment in this file.