Assignment P2b: Crowdsourcing and Argumentation

Deadline: Nov 12, 5 pm

In this assignment, you will be combining labels from P2a, analyzing them, and building arguments from them.

Part 1: Detecting Fake Labels in Crowdsourcing

Crowdsourcing involves multiple steps, each of which is posed as a question below.

1. *Combining data:* The attached folder contains P2a submission files with consent labels. Consider each student as a crowd worker and write a code to combine all submission files into a single dataset. The dataset must have 2,500 app reviews and columns containing consent labels. *Include the code snippet in your report.*

   A review is labeled by two or more crowd workers. Your dataset must include labels from all workers (as different columns) who were assigned that review.

2. *Reliability of labels:* The dataset comprises all labels received from students PLUS some fake labels from three fake files created by the TA. Your task is to computationally identify the top ten candidate files for being fake. You can use statistical measures such as interrater reliability (Cohen's kappa score or Kripendoff's alpha) to solve this task. Or, you can propose a new computational algorithm for this task. Do explain your steps and output in the report.

3. *Combining labels and analyzing data:* Remove the top 5 fake candidates (according to your approach) and use majority voting (discussed in class) to combine labels from multiple crowd workers. Your final dataset should have exactly three consent columns for each review.

   Pick the three most prominent consent categories in the final dataset. Computationally analyze what type of words are used by multiple parties (abuser and victim) to address each other. Is there a difference in the choice of words used across the three consent categories? You can use the NLTK library for this task and word clouds for visualization.

Part 2: Argumentation

1. From the final dataset, build an argument in favor of the abuser, illustrating why they like to track or harass the victim. Validate the argument by listing at least 20 reviews that satisfy it. Visualize your argument using the lucid chart.

2. From the final dataset, build an argument in favor of the victim, showing why they don't like to be tracked or harassed by the abuser. List at least 20 reviews from the dataset that satisfy it. Visualize your argument using the lucid chart.

3. Show how one of the two arguments above attacks the other argument. Visualize this attack using the lucid chart. Explain which part of the argument (premise, claim, warrant, or fact) is attacked and how.

   Describe your approach and all your findings in the report.

Please submit a .zip file, including the following items:

1) Your report in PDF format, including results and a description of your approach in each part.
2) Your program, along with your source codes, and all necessary libraries.
3) Instructions or a Readme file for compiling and executing your program.