IST736 Text Mining
HW5 (10 points)

Training/Evaluation Data Acquisition Through AMT

Use the documents that you gathered in HW1, and use AMT platform to request five sentiment labels per document. Note that you can set up your tasks in different ways, such as hiring five workers, each annotating all comments, or hiring 10 workers, each annotating half of the comments, etc.

After obtaining the labels, calculate pair-wise Kappa values among the AMT workers, and then calculate the average Kappa value as the overall agreement among the AMT workers. Also calculate the Kappa agreement between each AMT worker's annotations and your manual annotations. Based on these calculations, discuss the AMT workers' annotation reliability.

Describe in your report:

a. The experiment design: How many turkers do you aim to hire? What is the workload and payment for each turker? What is your requirement for the turkers (language proficiency, geographical location, past work performance, etc.)? Please also explain why you think this is the best choice for your experiment design to obtain the best-quality data in the most efficient way, e.g., your spam-control strategy.
b. The experiment outcome: How long did it take to obtain all labels? How much did you pay in total? Did any unexpected events occur during the process? Did you find any spammers? If yes, how did you find out and remove spam data? What is the average Kappa agreement among the workers? What are their levels of agreement with your ground truth? Do all AMT workers share similar marginal distributions?
c. Conclusion: Do you think AMT is a viable approach for obtaining training labels? What lessons did you learn in this experiment?

Attach a spreadsheet with the following seven columns: original comment, your label (ground truth, “positive,” “negative,” or “neutral”), and AMT response #1, #2, #3, #4, #5.

Submission format: Same as HW1 (up to 4 pages, at least 12-point font, 1" margin on all sides)