

Homework 1¹

Due by 11:59pm 2/7/2019

Objective

- Gain hands-on experiences with some of the concepts we've covered in class thus far. In particular, this assignment focuses on:
 - text pre-processing
 - baseline machine learning methods for text classification tasks using bag-of-words representation
 - applications of experimental methodology
- Practice posing research questions and setting up experiments to address them. Communicate findings in a clear and organized manner to others.

Project Requirement

The corpus we will be using for this assignment is the [SFU Opinion and Comments Corpus](#). In particular, we will focus only on the annotated constructiveness and toxicity corpus (you can also just get the CSV file [here](#)). You will develop a classifier for *toxicity* of user comments to news articles (given a comment from Column F, predict the level of toxicity (the left-most number in Column I of the corresponding row)). The required components for this assignment are:

1. Set up your global experimental framework:
 - Make appropriate cross-validation splits. [NB: you should think about how you want to randomize the data and be able to justify your choice in the write-up. Note that in the CSV file, the instances are sorted by the the original articles and then by comment orders.]
2. Build a baseline linear regression classifier:
 - You need to extract and preprocess the comment text so as to determine your vocabulary set for this task.
 - Train a logistic regression classifier where the features are just the vocabulary set. You may use standard off-the-shelf packages for training the classifier.
 - Record the performance of the classification. How does it compare against a majority-vote baseline?
3. Make some simple improvements to the baseline. Now that you have gone through one iteration of classifier design, what are some processing decisions that might be improved? Perhaps you need to adjust your initial preprocessing procedures. Perhaps you need to reduce your number of features. Perhaps you want to slightly increase the

¹ This assignment is also available on Google Drive:
<https://drive.google.com/drive/folders/1KPhPFPTfsmITLCX4NQdqkNdKlfdYGO5Q?usp=sharing>

complexity of your features (allow for some common phrases, for example). For this step, develop an “improved” classifier. Describe what has changed from Step 2.

4. Perform a rigorous comparison between your Step 2 classifiers and Step 3 classifiers. Apply appropriate statistical tests. (It’s OK if the “improved” version didn’t really improve.)
5. Pose a question based on this classification task; conduct an experiment to answer the question; discuss the outcomes of the experiment and draw some conclusions.
 - This portion is intended to be more open-ended and exploratory. You are encouraged to not pose the exact same question as someone else.
 - You may frame your question to explore any of the issues that we have discussed thus far (e.g., stemming, sentiment analysis, word embeddings). This is a good opportunity to further explore something that we didn’t go into a great deal of details in the lectures, but is in the readings.
 - If you are a little unsure as to what a valid question is, note that you probably have already implicitly asked a question in Step 3. Perhaps you had two (or more) different ideas about how to reduce the number of features; then, the question you implicitly asked was: “Which feature reduction method is better suited for this task?” One way to complete this step is to rigorously compare the options and discuss the outcomes.
 - Your question does not necessarily have to uncover some deep insight about NLP, but it should be at least somewhat “interesting.”
 - Prior to running the experiment, you ought to be *unsure* about what the outcome will be. For example, if you were asking the feature reduction question, you should pick two methods such that they both sound plausible on paper.
 - Some other question examples:
 - You think some commonly accepted belief is not quite right (e.g., stemming is unnecessary for sentiment-related tasks in English), so you pose some diagnostic questions to see what’s going on.
 - You think this classification problem is related to some other problem, and you want to explore the connections.
 - Note that the complexity of the techniques used in-and-of itself does not constitute “interestingness.”
 - The outcome to your question does not have to be “good.” For example, suppose you are trying to compare feature reduction method, it’s OK if the outcome does not clearly shows a winner. (You do need to discuss the outcome and what it might suggest, however.)
 - To address your question, you are allowed to use external resources, including:
 - Standard off-the-shelf packages and resources such as: NLTK, Stanford CoreNLP, SciKit.
 - Other parts of the SFU Opinion and Comments Corpus
 - Pre-trained word embeddings
 - Make sure your question’s scope is narrow enough that you can finish the experiment in a week or so. Also, please refrain from developing something that

requires a great deal of computational resources (storage, computing cycles) because the TA needs to be able to evaluate your work. In general, when you're in doubt, simplify.

What to commit

- Your code and data files
 - Please document enough of your program to help the TA grade your work.
- A README file that addresses the following:
 - Describe the computing environment you used, especially if you used some off-the-shelf modules. (Do not use unusual packages. If you're not sure, please ask us.)
 - List any additional resources, references, or web pages you've consulted.
 - List any person with whom you've discussed the assignment and describe the nature of your discussions.
 - Discuss any unresolved issues or problems.
- A REPORT document that discusses the following:
 - Describe what you did for Step 2 and report the baseline performances and compare it against majority voting.
 - Describe your model for Step 3 and report its performances. Compare this model against the previous baselines.
 - Pose your question (Step 5). Provide some motivation or explanation for why you asked this question; or you may offer some hypotheses (what you think the outcome will be). Describe how you want to set up the experiment to answer your question. Present the experimental results. Discuss the outcomes and draw some conclusions.

Grading Guideline

Assignments are graded qualitatively on a non-linear five point scale. Below is a rough guideline:

- 1 (40%): A serious attempt at the assignment. The README clearly describes the problems encountered in detail.
- 2 (60%): Correctly completed the assignment through Step 2, but encountered significant problems with later steps. Submitted a README documenting the problems and a REPORT for the outcomes of Step 2.
- 3 (80%): Correctly completed the assignment through Step 4, but has a significantly flawed Step 5. Submitted a README and a REPORT.
- 4 (93%): Correctly completed the assignment through Step 4. For step 5, the question posed is clear and rigorously answered through experimentation. The REPORT content is solid.

- 5 (100%): Correctly completed the assignment through Step 4. For step 5, the question posed is clear and interesting; it is rigorously answered through experimentation. The REPORT content is well-written and insightful.