NLP Homework 4

Due Monday, November 2, 11:59 pm.

In this homework, you will conduct sentiment analysis to gain some understanding about the emotions reflected from the texts that you've worked on in the previous two homework (in the field "text").

Based on what we have learned from this class, you will explore the sentiment polarity of the comments at the sentence level. This includes how to process the words and how to conduct the sentiment polarity analysis using classifiers. You will build a classifier to classifies the sentiment polarity of a given sentence as as *positive, negative,* or *neutral.* You will provide a CSV file that contains the following fields: "title", "author", "country", "the number of positive sentences", "the number of negative sentences", "the number of neutral sentences". You will analyze all the *positive* sentences to identify top 50 adjective phrases, adverb phrases, and verb phrases, and do the same with all the *negative* sentences. You will present these results in tables.

In your report, please explain in detail the processing techniques that you have applied, the features you used for the classification task, and your experiments. For the data preprocessing/cleaning task, we have learned about several techniques such as tokenization, sentence creation, regular expression processing, stop word filtering, etc. You should describe the techniques you used in this assignment.

For the classification task and the experiments, you should start with the "bag-of-words" features where you collect all the words in the training corpus and select some number of most frequent words to be the word features. You should use at least NaiveBayes classifier and multi-fold cross-validation. You need to obtain precision, recall, and F-measure scores. In your experiments, you should use at least two different sets of features and compare the results. For example, you may take the unigram word features as a baseline and see if the features you designed improve the accuracy of the classification. Here are some of the types of experiments that we have done so far:

- Filter by stop words or other pre-processing methods
- Representing negation
- Using a sentiment lexicon with scores or counts: Subjectivity

There are many datasets available for training on sentiment polarity. Below are some examples. Please choose one dataset for the training purpose and briefly explain why you choose it:

- The sentence_polarity corpus introduced in class
- http://www.cs.jhu.edu/~mdredze/datasets/sentiment/
- http://help.sentiment140.com/for-students

- https://www.kaggle.com/crowdflower/twitter-airline-sentiment

How to Submit Homework:

Go to the Blackboard system and the Assignment for Homework 3. Attach your report file and submit. Your submission should include:

1) your report in a PDF format
2) Table including two lists of sentences: negative vs positive (Please include in your report)
3) Your Python code and the processing screenshots (Please submit in one separate folder zipped)