NLP Homework 2 - Monday

Due Date: Tuesday, 10/24, 11:59 pm.

In this assignment, you will continue to work on the "text" in Fake.csv and True.csv.

Sentiment Polarity - Classifier Building

You will explore the sentiment polarity of each sentence in the text, based on what we have learned from lab 5. You will build a classifier to classify the sentiment polarity of a given sentence as *positive* or *negative*. You are expected to 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

You are encouraged to explore other features in your experiments (e.g., the existence or number of adjective words, verbs, noun phrase, etc.). 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

Sentiment Polarity – Analysis of Fake and Real "text"

You will use the classifier you have built to identify if a sentence in Fake news article or Real news article is positive or negative. You will only analyze the sentences in the first 50 fake news articles and first 50 real news articles in the Fake.csv and Real.csv. For analysis of Fake news articles, you will then provide a csv file that contains the following fields: "text", "the number of positive sentences in text", "the number of negative sentences in text". Each “text” corresponds to a “text” entry in Fake.csv and there will be 50 rows in this csv file. You will do the same for True.csv data.

Next, you will examine whether the Fake content tends to contain more positive or negative sentences.

How to Submit Homework: **<u>Due Date: Oct. 24, 11:59 pm.</u>**

Go to the Blackboard system and the Assignment for Homework 2 and submit your notebook file and the CSV files mentioned above. The file needs to be run already and results are displayed. Please provide enough comments in the code and name the variables mindfully to make it more intuitive for people to understand the logic behind your code. Additionally, in this notebook file, please:

1. display the first 20 rows of the two CSV files
2. explain the features you used and your experiments
3. explain how you examine whether the Fake content tends to contain more positive or negative sentences.
4. discuss what you have learned from this sentiment analysis assignment – it can be reflections on your methodology and process of doing this homework; or on the results you have obtained

NOTE: Please name your submission files such that they contain your first and last name.