NLP Final Project
Fall 2011, Due December 15

For this final assignment, there are three types of projects to choose from, but you may also propose your own.

I. Annotation and Analysis of Data

For this task, you will annotate data in a corpus. Two corpora that are available are described here, but you may also collect your own data.

A. The Movie Review corpus sentence level

For this annotation, you will investigate the movie review corpus by analyzing the polarity of the opinions in the reviews at the sentence level. First, pick about 5 reviews (for each person) from the Movie Review corpus that you are comfortable working with. Produce a file that assigns a subjectivity/opinion label to each sentence.

First, consider whether the statement is subjective or objective. If it is objective, that is, states facts and expresses no opinion, label the sentence Objective. Next, for the subjective sentences, label them with an opinion label of Positive, Negative or Off-Topic. The positive and negative opinions should be that of the reviewer towards the movie itself or towards aspects of the movie, such as actors, director, or particular scenes. Opinions about anything else, like other movies, should be labeled Off-Topic.

For analysis purposes, in addition to the subjectivity/opinion label, you can give a second annotation label that indicates the level of difficulty of deciding the sentence. In this label, you can give an alternate label if you had trouble deciding between two specific labels, or you can just say Hard.

B. The Tweet corpus

For another research project, a number of tweets have been downloaded from Twitter that are on the topic of the Obama Health Care plan. For each tweet, we would like to have a label that says whether the opinion of the tweet is Positive, Negative or Neutral towards the plan. There are also some tweets not on the topic of health care reform and these can be labeled Irrelevant. For this project, I will draw a new set of tweets chosen randomly from the overall collection of ~1 million, separated into groups of approximately110 tweets, each in its own sheet of an excel file. I think that each person who chooses this task can annotate about 3 sheets of data. There is a short set of annotation guidelines to give more information about the task.

Again, in addition to the opinion labels, you can give a second annotation label that indicates the level of difficulty of deciding the sentence. In this label, you can give an alternate label if you had trouble deciding between two specific labels, or you can just say Hard.

C. Other Corpora

There are other corpora available to annotate, for example, we could use the Dialog corpus in which you would annotate the dialog acts. Or you may like to propose a corpus and task to annotate.

Semantic analysis of the Annotated data

The results of your project will be the annotated data file and a report that contains the following.

i. From the annotation, give a few examples each of tweets or sentences that typify each of the labels. Then give a few examples of the ones which were marked hard and give any comments that you have about why they were difficult.

ii. Analyze the sentences that you annotated and look for short phrase patterns or words that indicated to you the objectivity, subjectivity or opinion in that sentence. For each word or pattern, look them up in WordNet and see if the word or pattern can be generalized by adding hypernyms, synonyms, etc. Discuss whether you think that these patterns or words could be used to predict the opinions in future movie reviews.

iii. Analyze the sentences or words to design other clues that could be features for classification. For this, you can also consider if other types of NLP processing, such as POS tagging would be helpful in creating features.

II. Classification of Data

For this task, you should choose to work on classification of a data set.

Based on the data and the task, decide what level of NLP processing is desirable and carry out any NLP processing needed, e.g. you may want to run special purpose Tweet POS tagging. Read the data into the NLTK (either by reading the file, or using a PlainCorpusReader) and write Python/NLTK that defines features.

Produce the features in the notation of the NLTK and use one of their classifiers to train and test a classifier on the data, or produce the features as a csv file and use Weka to train and test a classifier.

Available Data:
Twitter data annotated with positive, negative and neutral opinions towards health care reform.
Subjectivity dataset from Pang and Lee contains 5000 each objective and subjective sentences.
Product debate corpus and Political debate corpus from Wiebe at the MPQA data web site.
Dialog Act data in the NLTK.
Senseval 2 data to train a Word Sense Disambiguation classifier for 4 words, in the NLTK.
Text Categorization data from the Reuters corpus.
(Others are available)

Available Resources:
Twitter tokenizer and POS tagger from Motif.
Lexical resources: the Subjectivity lexicon from Wiebe, the LIWD dictionary from Pennebaker and the ANEW dictionaries from Florida.
Stanford POS tagger, named entity recognizer and parser(s).
Mallet for finding topic models.

To complete this project, carry out at least one experiment where you use 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. Write a report that describes the data processing, the features and the classification experiment(s).

Earthcube Papers

This problem is similar to the classification task in that there is a set of documents and each document should be represented as a set of features. The difference is that this problem is more open-ended; the organizers want a visualization of clusters of the documents. One way is to define a set of features and use Weka to perform clustering. Another would be to investigate the use of topic modeling to either make clusters or features for clustering. We may consider that the visualization is out of scope for this course.

What to Hand In

If you are working in a group, you should choose a task for each person. If you do annotation, hand in the annotation data. Every group should hand in a report with the description of all that you did and the discussion of the results. As usual, submit these documents to the Blackboard system by the end of the day on the due date.