Comp814 Text Mining

POS tagging Lab

Objective

The objective is of this lab is to be able train, store, retrieve a pos tagger and evaluate their accuracies.

Task

Use the sample Code from lectures to do the following.

1. Using Brown corpus from NLTK split the corpus into 80% training and 20% for testing for all of the following exercises.
2. Instantiate the following taggers from NLTK.
   a. Unigram tagger
   b. TnT tagger
   c. Perceptron tagger
   d. CRF tagger
3. Train all of the taggers and store the trained models as a pickle file.
4. Retrieve the pickle file and test them on the testing data.
5. Tabulate and compare the accuracies and choose the best one out of the lot. You can base this choice on the F1 value.
6. Use the best tagger to do the following.
   a. Download 10 news articles from 10 different news sites on a dominant topic of the day.
   b. By reading the articles determine at least 3 nouns that best represents the chosen topic. Lets call this set $\Tau$
   c. Your task is to determine the percentage of nouns in the set $\Tau$ compared to all nouns in the 10 articles under study.
7. Upload your python code file to Blackboard by 6pm Friday this week. You can upload a zip file containing the python code and a separate one containing the table of comparisons. Else you can paste the table in the python file at the end as comments.
8. **Further study for the following week.**
   a. Download articles on the same topic to Expand the data set to 20 articles.

b. Formulate a way to computationally determine the dominant topic in the 20 articles instead of doing it by reading it as you did as part of the lab.

c. Implement the strategy to determine 5 words that represent the dominant topic.

d. Extract 5 context words that appear with the topic words.

e. Determine the most common context words on the topic that you have chosen. (hint: we normally disregard common words such as “a” “the” etc).