First Steps: Assignment

Monday, 12 November 2018 16:21

The following is a quick guide on how to start the assignment.

It's expected that you play with the training data now to 1) further insight into the NLP algorithms given in the lectures 2) obtain feedback about your relative level of understanding. Not everything will work (or even be relevant) - learning what is and what isn't is an essential part of the course!

1. Get the Data

Training Data for the assessment is in the Data folder.

In addition, you now have a directory for Corpora which are pre-installed in NLTK. You can put your training data in this directory also. My directory is

/Users/mgl/nltk_data/corpora

Yours should be similar (obviously with your username and not mine). Create a directory and add the training data as txt files.

2. Read the corpus into Python.

For this we need a list of file-ids

>>> from os import listdir

>>> from os.path import isfile, join

>>> onlyfiles = [f for f in listdir(mypath) if isfile(join(mypath, f))]

(note - Mark isn't convinced there isn't an easier way to do this but the above works)

and we can set the address for where the corpus is by

>>> corpus_root = '/Users/mgl/nltk_data/corpora/assignment1/'

and then use a CorpusReader as previously done e.g.

corpus = nltk.corpus.reader.plaintext.PlaintextCorpusReader(corpus_root,

onlyfiles)

or read the text directly via

>>> import nltk.data

>>> nltk.data.load('corpora/assignment1/wsj_1947.txt')

3. Things to do

a. Tokenisation
b. Part of Speech Tagging
c. Named Entity Recognition.

4. A simple POS NER Tagger

I've uploaded a simple piece of code for using POS tagger as a NER tagger - you should use it as part of a larger backoff tagger (with different n-grams).

It's called name_tagger.py in the Code section of files. Does it work well enough for our purposes? Find out!

5. How close is your training data to the test data?

At the moment since we haven't done anything, it's quite far away. However you need to write some code to score your data (e.g. count how many times your system tags the test data correctly).