CS 565: Assignment-2

Hidden Markov Models (HMMs), Part-of-Speech (PoS) Tagging

Due Date: 14th March, 2016

Data

The training dataset for this assignment is part of the 'Brown Corpus' (Description: https://en.wikipedia.org/wiki/Brown_Corpus). This dataset contains the following data files.

Filename	**Content**
Brown_train.txt	Untagged training data
Brown_tagged_train.txt	Tagged training data

Download data from:
https://github.com/sap01/IITG_CS565_Spring2016/tree/master/Assignment2/Data

Data Files Format

- The untagged data files have one sentence per line, and the tokens are separated by blank spaces. For example:

```
At that time highway engineers traveled rough and dirty roads to accomplish their duties.
```

- The tagged data files have one sentence per line as well, but the 'token/tag' pairs (instead of just the tokens) are separated by blank spaces. For example:

```
At/ADP that/DET time/NOUN highway/NOUN engineers/NOUN traveled/VERB rough/ADJ and/CONJ dirty/ADJ roads/NOUN to/PRT accomplish/VERB their/DET duties/NOUN ./.
```

To Do

The objective of this assignment is to implement Part-of-Speech (PoS) tagger algorithms for learning trigram Hidden Markov Model (HMM) from the given corpus. A PoS tagger's job is to find out the most probable PoS tag sequence for each sentence in the given data. We are going to use trigram tagger because it provides an excellent balance between performance and computational complexity. The task is divided into five deliverables.

Deliverable 1

Write a program to implement vanilla trigram HMM tagger i.e. plain HMM tagger with Laplace Smoothing and mapping of out of vocabulary (oov) words (i.e. words with count < 6) to rare word.

Deliverable 2

Write another program to improve the previous model with better smoothing technique of your choosing and optionally better grouping of words (as discussed in the class).

Deliverable 3

Write an evaluation program that takes input from two tagged data files with same word tokens but potentially different tags. For example:

Input tagged data file 1.txt

At/ADP that/DET time/NOUN highway/NOUN engineers/NOUN traveled/VERB rough/ADJ and/CONJ dirty/ADJ roads/NOUN to/PRT accomplish/VERB their/DET duties/NOUN ./. Using/VERB privately-owned/ADJ vehicles/NOUN was/VERB a/DET personal/ADJ hardship/NOUN for/ADP such/ADJ employees/NOUN ,/. and/CONJ the/DET matter/NOUN of/ADP providing/VERB state/NOUN transportation/NOUN was/VERB felt/VERB perfectly/ADV justifiable/ADJ ./.

Input tagged data file 2.txt

At/ADP that/DET time/NOUN highway/VERB engineers/NOUN traveled/VERB rough/ADJ and/CONJ dirty/ADJ roads/NOUN to/PRT accomplish/VERB their/DET duties/NOUN ./. Using/DET privately-owned/ADJ vehicles/NOUN was/VERB a/DET personal/ADJ hardship/NOUN for/ADP such/ADJ employees/NOUN ,/. and/CONJ the/DET matter/NOUN of/ADP providing/VERB state/NOUN transportation/NOUN was/VERB felt/VERB perfectly/ADV justifiable/ADJ ./.

The first input file corresponds to the desired output. Whereas the second input file is generated by the developed model. The next step is to compare the second file with the first file and calculate the precision, recall, and F1-measure. The output should be as follows:

Precision: 0.573245
Recall: 0.809938
F1-measure: 0.675858

Note: The values in the above example are imaginary.

Deliverable 4

Compose a readme file (.txt or .pdf) describing how to execute your programs.

Deliverable 5

- Create a folder named ‘<First Roll No.>_<Second Roll No.>’.
  E.g. ‘120101007_120101018’.
- Put your readme file inside that folder.
- Create a subfolder named ‘Supplementary Material’ inside the aforementioned folder.
  Put all your supplementary files, like - source code files inside this subfolder.
- Compress the main folder as ‘.tar.gz’. E.g. ‘120101007_120101018.tar.gz’.
- Submit the ‘.tar.gz’ file to Canvas Assignment 2 portal via your Dropbox account by 11:59 PM, March 14, 2016.