Christopher Lang (EMPL ID: 12091305)

Homework 3, STA9792

Writeup

**Question 5.1**

**I/PRP need/VBP a/DT flight/NN from/IN Atlanta/NN**

The word “Atlanta” is improperly tagged. It is a proper noun (NNP), not simply a noun (NN)

**Does/VBZ this/DT flight/NN serve/VB dinner/NNS**

The word “this” is not a determiner but an adjective, as it refers to the “flight”, a noun (NN)

**I/PRP have/VB a/DT friend/NN living/VBG in/IN Denver/NNP**

I can’t find an error

**Can/VBP you/PRP list/VB the/DT nonstop/JJ afternoon/NN flights/NNS**

I can’t find an error

**Question 5.8**

A Python program is written that implements a bigram Hidden Markov Model with Viterbi algorithm for token POS classification

There are two files. One more important one is **HMMTagger.py** that defines a python class that does the actual model training and POS classification. When tagging a new sentence using the method **tag\_sentence**, a Viterbi algorithm is executed to tag the new sentence

Training and especially POS classification is a bit slow due to inefficiencies in how data from the transition matrices are queried, beware

Another file called **train\_hmmtagger.py** is the entry point into training the HMM tagger. This was initially tested with **nltk**’s built in **treebank** corpus. The script should automatically download this corpus for the tagged sentences. I also attempted to test with **nltk**’s brown corpus but was returning weird results, hence that was skipped

The program is developed in Python 3.5.3 Anaconda distribution, 64 bit Windows 10. Important dependencies include:

* **numpy** – for array use
* **scipy** – for sparse matrices
* **nltk** – for labeled sentences (POS)
* **tqdm** – for progress bar

**Question 5.9**

An Excel spreadsheet is also provided called **in-sample confusion matrix.xlsx**, showing the in-sample confusion matrix for a random subset of 100 sentences from the **treebank** corpus. A sample was used due to the shear slowness of tagging