50.007 Machine Learning

Design Project

Li Zi An
Liu Wei Long
Tan Enyi

Summary

For part 1, we calculate the emission parameters as specified in the project requirements. This is relatively straightforward but we made sure to use the csv.reader(csvfile, delimiter = "\t", quotechar = None, skipinitialspace = True) instead of the default cvs.reader as it could not parse the data correctly (specifically when a “ is encountered).

For part 2, the same caution was undertaken as previously mentioned. A Pocket Algorithm was used for the Viterbi Algorithm in order to reduce the memory load of the program and to increase the general efficiency. One note is that we integrated the start and the stop states into a single matrix with the rest of the transitions so that the coding in the later parts would be vastly simplified i.e. we do not need to have a base case or end case but rather one loop where special operations are invoked for start and stop states.

For part 3, we managed to improve the algorithm by converting all the alphabets to uppercase. This improved the accuracy by the greatest as the permutations of cases for a particular word has no impact on how the word is to be tagged. We also converted all numerals to the digit 0 for increased hit rate as all numbers have the same tag in the system. The same can be said of URLs starting with http:// and twitter users starting with @USER. We tried implementing custom dictionaries to increase the accuracy once again with databases of emoticons, companies and cities but it show very little or no improvement to the system as a whole.

Execution

For all 3 parts of the project, we have automated the code to take in the specified files and give a single output which is the accuracy. The user is only required to run the Python script and need not enter any arguments.