Computer Science 4140
Project Part 2

Given your assigned chunk type from the CoNLL corpus, write functions to do the following tasks.

- List all the tag sequences that occur with each instance of the given chunk type.
- Count the frequency of each tag sequence, and produce a ranked list in order of decreasing frequency. Each line should consist of an integer (the frequency) and the tag sequence.

In addition, you should inspect the high-frequency tag sequences. Use this study as the basis for developing a better chunker.

Deliverables

1. Source code for the two functions. Name this file *analysis.py*. It should be executable at the command line. The structure of the file should be to specify all needed imports, then specify the definitions of the functions, and then specify calls to the functions.
2. Source code for your new chunker. Name this file *mychunker.py*. It should be executable at the command line.
3. A brief report (no more than one page) that discusses your analysis of the high frequency tag sequences, your proposed approach for developing a better chunker, and an analysis of the results of your approach, comparing it to baseline results. Name this file *reportpart2.pdf*.

Assigned Chunk Type

NP	PP	VP
Bair	Barnes	Backshall
Cahoon	Jones	Graves
Colglazier	Kurdewan	Hotalen
Pena	Magsino	Jackson
Suncin	Mills	Weeks

Program Deadlines

April 25, 9:00 A.M.

Deadline for submitting all project deliverables. The <*assignment_number*> should be *projp2XX* where *XX* will be your assigned chunk type (*NP*, *PP*, or *VP*).

Project Evaluation

Note that for the evaluation of your proposal for improving chunker performance, the majority of the weighting will be on the depth and thoughtfulness of your analysis rather than the actual performance (though actual performance does matter). Your grade on this assignment will be weighted as follows:

Code Quality and Correctness	70%
Comparative Performance	10%
Report Quality	20%