Assignment 3

L545/B659

Due: Friday, October 22, 2021, 23:59

You are allowed to work in groups of 2 members each. We recommend that you form groups such that one member has a strong coding background and the other has a strong linguistics background. This will foster good learning and better exchange of ideas. You ARE NOT allowed to discuss solutions with another groups. You MUST NOT copy or search for answers anywhere and MUST SUBMIT work that belongs to your group only.

You are given three parts of data `part1.txt`, `part2.txt`, `part3.txt`.

1. Annotate them by using `TreeTagger` (use the model trained on Penn Treebank), save the output to three single files. (**30'**)
2. Based on the TreeTagger outputs, manually analyze the results to discover ambiguities and errors, and save three tag-corrected files corresponding to the three output files in question 1. Your annotation decision should be made based on the *Part-of-Speech Tagging Guidelines for the Penn Treebank Project* (you can find it on canvas). (**45'**)
3. For each part of data, if we consider your corrected data as *gold standard*, write a script to calculate an agreement score between the TreeTagger output and the corrected gold standard, that is, the percentage of words that are correctly tagged. Then write a brief report including scores for the three parts of data and explanation for each. (**25'**)
4. **Bonus**: As for the ambiguities and errors you found in Question 2, can you list the most frequent three error or ambiguity types you discovered? Please explain each type and give examples. (**10'**)

What you should submit:

1. 3 TreeTagger output files and 3 corrected files.
2. Your script for calculating the agreement score and a brief report of results.
3. If you do the bonus question, you should submit another file.