Computer Science 4140
Project Part 1

Your assignment is to investigate three different ways to define the split between training and testing data when developing a tagger using the Brown Corpus: by genre (*category*), by source (*fileid*), and by sentence. Further requirements are noted below.

1. Name your source code *myTagger.py*
2. Your code must be executable at the command line (e.g., *python3 myTagger.py)*.
3. You must use n-fold cross validation to conduct the investigation. In the case of genre, each different genre must be a fold. In the case of source and sentence, divide the data into 10 equal folds based on a random shuffle of the data.
4. The program output should be three different tables, one table for each of the three ways to split the data. Each table should have two columns. The first column should have the header "Test Fold" and the second column should have the header "Result". For the genre table the "Test Fold" label should be the name of the genre. For the other tables the "Test Fold" label should be an integer from 0 to 9, inclusive which specifies the fold used for the test set for that experimental run. The "Result" label should be the experimental result expressed as a percentage with one number to the right of the decimal point. The tables should be neatly formatted and include a header line that says "Results by X" where X is the method for splitting the data (by genre, source, or sentence).
5. As the basis for your tagger, you may use the combination n-gram tagger approach described in Section 5.4 of Chapter 5. However, you should augment it with a Regular Expression tagger. Since comparative performance is one part of the evaluation, in order to keep a level playing field, your Regular Expression tagger may not have more than 20 patterns.
6. Besides the code, you will also submit a written report that compares the relative performance of your tagger on the different forms of data splitting and discusses which method is the most legitimate method of evaluation. Be sure to consider the difference in results depending on whether or not the original data is randomly shuffled. More specific details about the report will be forthcoming.
7. The submitted version of your program should be based on random shuffling of the data in constructing the folds for the source and sentence evaluations.

Note: For n-fold cross validation if the size of the data set is not divisible by "n", the folds will be of unequal size. However in this case the number of sentences and number of sources is divisible by 10 so no special coding is required.

Program Deadlines

April 12, 9:00 A.M.

Deadline for submitting completed program, including the report. The *<assignment_number>* should be *projp1*.

Project Evaluation

Your grade on this assignment will be weighted as follows:

Code Quality and Correctness	60%
Comparative Performance	10%
Report Quality	30%