Program 6, CS 109, Fall 2022

Points

This is a 30-point assignment. It is due at 11pm on Thursday 1 December.

This is a group project. The groups are as follows.

1 Jeansonne and Weerasinghe
2 Pearce-Pearson and Rao
3 McGregor, Bach Nguyen, and Zhang
4 Buse, Magnus, and Vath
5 Blankenship, Kim, and Minh Nguyen
6 Cashman, Osmo, and Slabaugh
7 Robins, Shaw, and Tecson

Each of you will submit to me by email a summary of how things went in the group. This should not need to be longer than about half a page. Any time I do group projects, I want to know that all members of the group at least tried to do equal work, because I will normally assign the same grade to each member of the group.

The Assignment

You are to read a sample document from the Brown Corpus. This will be text from a news report, with each word tagged with part-of-speech (POS) tags.

You are to read the text and split into tokens to create a list of word/POS tokens. You are to create two dictionaries.

One will have the words as the key. The value for this will be a Python `set` of the POS tags for that word.

The other dictionary will be the reverse. It will have the POS tags as the key, and the value will be the words tagged with that POS.

Having split the original text into a list of word/POS tokens, you will then need to walk through that list, split on the slash, and store POS pointing to a set of words and word pointing to a set of POS in the two dictionaries.

You should use a Python `set` and not a list, so you don't get duplicates.

In the Canvas for this program is a sample text document, `ca01.txt`, and an 8-line header of that, so you can work on small data as you develop code.

I have included my output for both the entire document and the 8-line header.

Simplifications

You do NOT need to remove punctuation; indeed, punctuation has its own POS tags, as you will see.

You do NOT need to worry about lower-casing; you can just read and use the word/POS tokens as they are.