

LING115 Homework#8

Due: 10/24

Instructions:

- Use `/home/ling115/wsj.t2c` as the corpus.
- By default, `TGrep2` returns the first subtree in each sentence that matches the specified pattern. To identify *all* subtrees that match the pattern, use `TGrep2` with `-a` option.
- Create a `hw8` directory in your home directory and save your work in that directory.

1. [1 point x 4 = 4 points] For each subtree matching the descriptions below, answer the following:

- 1) What is the pattern you have to specify in order to extract the subtree using `TGrep2`?
- 2) How many matching subtrees can you find in the `TREEBANK` corpus? Remember to run `TGrep2` with `-a` option.

Save your answers as `hw8.q1`.

1.1. Either `NN` (=singular or mass noun) or `NNS` (=plural noun)

1.2. `NP` (=noun phrase) that immediately dominates a `DT` (=determiner) and an `NN`

1.3. `NP` that follows `VBD` (=verb in past tense) with an optional `DT` immediately before the `NP`

1.4. `ADJP` (=adjective phrase) that immediately dominates `RB` (=adverb) and `JJR` (=comparative adjective), where `RB` immediately precedes `JJR`

2. [4 points] Python provides a function called `getoutput`, which is defined in a module called `commands`. You can use this function inside your Python program to run an external program (what you would normally run directly in the shell) and get the standard output from the program as a string. For example, the following lines of code will allow you to save the output of running `ls -la` as a string variable named `ls_out` in Python:

```
>>> import commands
>>> ls_out=commands.getoutput('ls -la')
```

Write a Python program that prints out the average number of adjectives (`JJ`, `JJR`, or `JJS`) a noun phrase (`NP`) dominates in the `TREEBANK` corpus. Save your program as `hw8.q2`.