

LING115 Homework#2

Due: September 8, 2010

Instructions:

- Before you begin, create a directory named `hw2` under your home directory. Just to remind you, your home directory is `/home/students/<yourID>`. This is where you are when you first login.
- Write up your answers to questions 2 and 5 below using a text-editor (e.g. emacs, vim). Save the file as `<yourID>.hw2` (e.g. `hahnkoo.hw2`) and put it in your `hw2` directory.

1. Copy everything under `/home/ling115/hw2_out/` into your `hw2` directory.

2. If you solved problem 1 correctly, you should see a directory named `dummy_files` in your `hw2` directory. In that dummy directory, delete all files whose name consists of `dummy_` followed by one or two digits (e.g. `dummy_2`, `dummy_23`), but not the files whose name consists of `dummy_` followed by three digits (e.g. `dummy_123`). Using wildcards, you should be able to do this with a single command. What is that command?

3. Under the directory `/data/TREEBANK/RAW/WSJ/00/`, you should see that there are 99 files, whose names begin with `WSJ_00`. Concatenate `WSJ_0001` and `WSJ_0002` and save the output as a file named `WSJ_0001_plus_0002` in your `hw2` directory.

4. List all the **unique** words in the 99 files mentioned in problem 3 **ignoring case-distinction**. Save the output as a file named `WSJ_00.types` in your `hw2` directory.

5. The command `wc` is short for word count. When you run it with the name of input file as its argument, it prints out four types of information related to the input file: (1) number of lines in the input file, (2) number of words in the input file, (3) number of bytes in the input file, and (4) name of the input file. How many **unique** words are there in the `WSJ_00.types` file you created in problem 4?

6. Using the 99 files mentioned above, create a word frequency list sorted in descending order of token frequency like the one we created in section 4.4 of the lecture note titled 'More fun with shell commands'. Save the word frequency list as `WSJ_00.count` in your `hw2` directory.