

LING115 Homework #4
Due: September 22, 2010

Instructions:

All files I mention below are under `/home/ling115/hw4_out/`. Create a `hw4` directory under your home directory. Store all the files that you have created while answering the questions below in your `hw4` directory.

1. [1 point] Copy `bigram_v1.py` to your home directory. Line 19 of the file should read:

```
bigram='_' .join(words[i:i+2])
```

Explain what this line does by adding comments below line 21. Your answer may span two or more lines. Remember to begin every line of your answer with `#` to make sure that the line is meant to be a comment. After you have typed your answer, save the file as `q1.py`.

2. [1 point] You should be able to see that `bigram_v1.py` extracts all unique bigrams from the standard input, where a bigram is a sequence of two words. Use `bigram_v1.py` to list all unique bigrams in files under `/data/TREEBANK/RAW/WSJ/00/`. Save the output as `wsj.bigrams`.

3. [1 point] Modify `bigram_v1.py` so that the program extracts all unique trigrams in standard input instead of bigrams, where a trigram is a sequence of three words. Save the modified code as `trigram_v1.py`.

4. [1 point] Modify `bigram_v1.py` so that the program extracts all unique n -grams in standard input, where an n -gram is a sequence of n words. Let the user specify the n of n -grams as the argument to your program. Save the modified code as `ngram_v1.py`.

For example, entering the following should list all unique bigrams in `foo.txt`.

```
$ python ngram_v1.py 2 < /home/ling115/hw4_out/foo.txt
```

5. [2 points] Write a Python program that prints out how often the definite article 'the' appears in the standard input, regardless of its case: treat 'the', 'The', 'THE', etc. to be the same. Save your program as `count_the.py`.