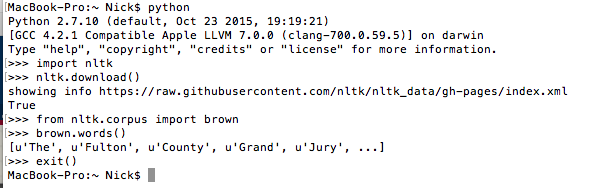
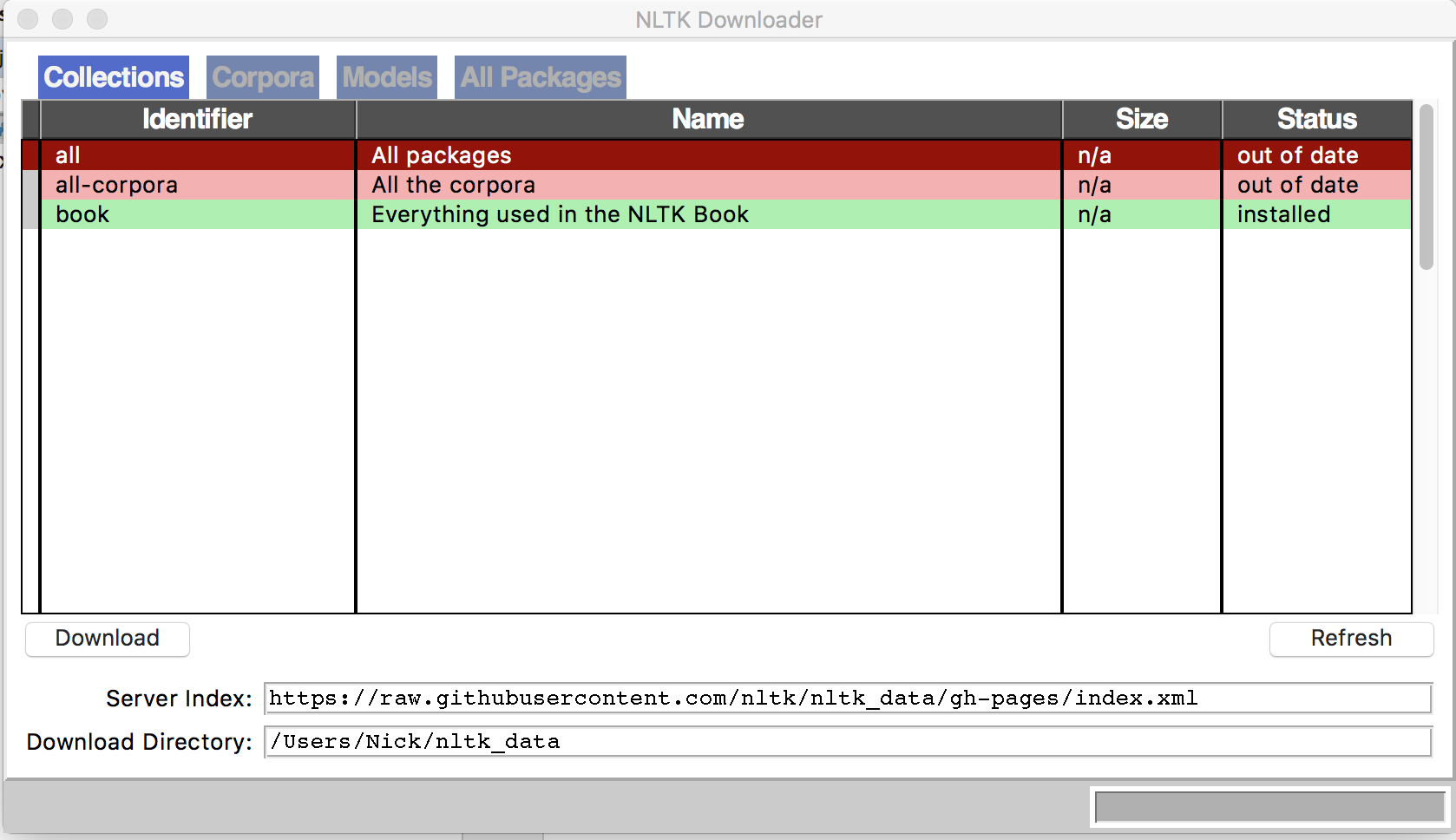
Nick Petty

Homework 1

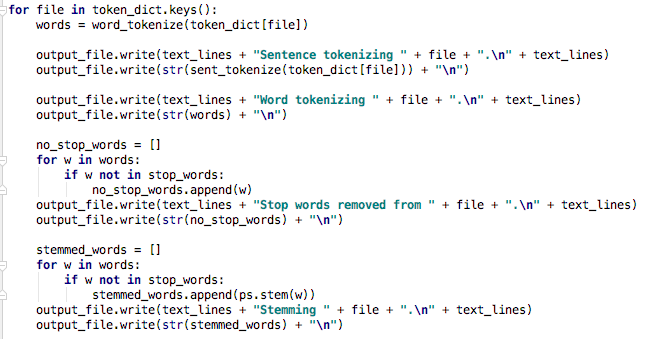
Source code is attached in the file HW1\_Nick\_Petty\_src.zip and program output is attached in the file hw1output.txt.

1. Screen shots are attached of Python and NLTK in the Mac Terminal.





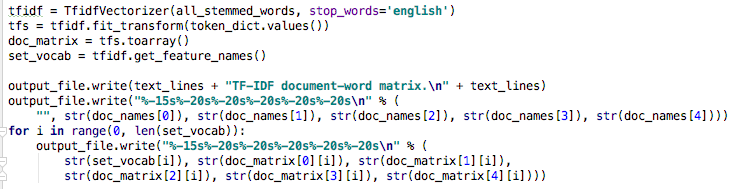
2. This code section tokenizes each document into sentences and words, removes stop words, and stems the words:



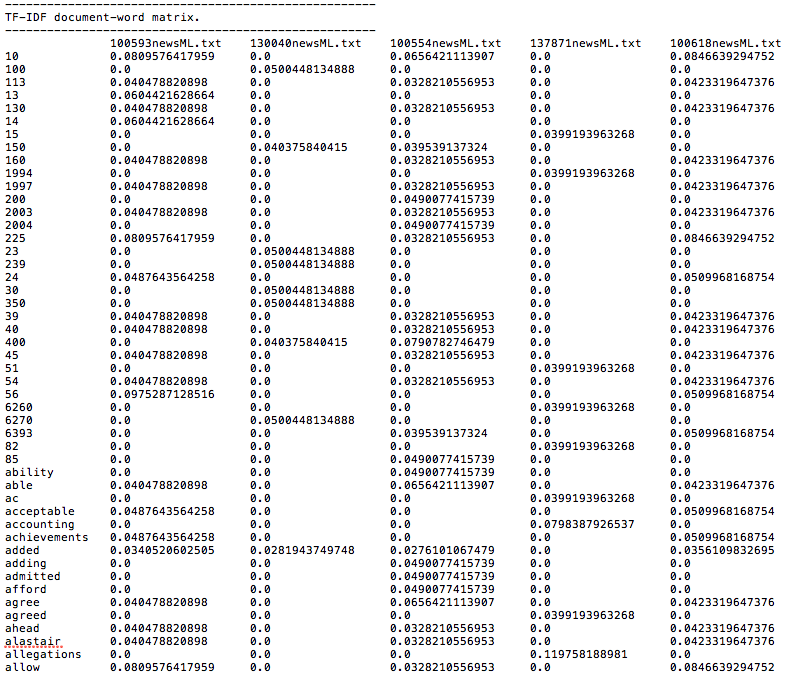
This is the output for 1 document:



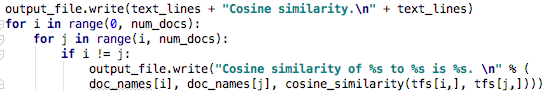
3. This code section calculates the tf-idf for each word and generates the document-word matrix:



This is the top few rows of the matrix:



4. This section of code calculates pairwise cosine similarity for the documents:

This is the calculated result:

