Project 2

Vinoth Rajasekar – vinoth@uw.edu

I was able to successfully complete project 2 using Python. Below are the sample results.

the	4398064
a	2032214
to	1893205
of	1888409
and	1759680
in	1486132
that	814646
for	793617
is	712518
on	564762
by	559404
with	512398
he	494957
it	484405
at	463595
said	442322

I used regex and Counter from collections library in python to clean and filter the data and produce the total count of the unigram words. Further, the results are sorted in the descending order as above based on the total count. Below is the procedure I followed to clean the data and produce the final output.

- Read AQUAINT corpus of English newswire files from /corpora/LDC/LDC02T31/nyt/2000
- Then, Process the AQUAINT corpus of English newswire to remove SGML tags.
- Next, filter only words and apostrophes using regex as we need to retain only letters and straight apostrophe.
- Next, trim any occurrence(s) of the straight apostrophe from the beginning and end of the word using regex
- Next, convert all the words to lowercase letters.
- Now we have parsed and cleaned the English newswire corpus. Now, split the data by space and count the word occurrence using the Counter method from collections library.

- Repeat the above steps for all the files in the corpora.

- At the end of reading all the files, we have the total count occurrence for each unigram word.

- At last, using sort method, I sorted the results in descending order, such that the word with highest count appears first.

Challenges:

I was referring to the project2 discussion in the canvas. The discussions helped me to understand the edge cases and helped to implement the solution around it in my code.

Secondly, I was running into memory error, when I executed the coding using run.sh. Because, I parsed the results and kept them in memory and at the end I did count on all the word occurrences. This was not effective and efficient. I made an implementation to make the count on the words as I read each file and cumulatively add the new file results with the old file using the Counter method in python. This helped to get around the memory error and was able to successfully execute the code using both run.sh and condor and produce the results successfully.