Thomas Munoz Vasquez

Big Data Programming

Due March 3

Assignment 2

The objective of this code is to remove stopwords, count the frequency of words, and sort the words based on their frequency in descending order. It was achieved by using python and spark in a virtual machine.

To achieve the removal of stopwords, we input the file name and define the stopwords in a set of unique characters. We use flatmap to split each line of text into a list of words based on spaces, we follow with filter to filter out the stopwords from the file and finally we use map to ensure all words are converted to lowercase.

To count the frequency of words, we use map to transform each RRD to a key-value pair with the starting value of 1, we use reduceByKey to count the frequency (times it appears in file).

Finally, we use sortby to sort their frequency in descending order, it is sort in ascending order by default, so we set ascending to false.

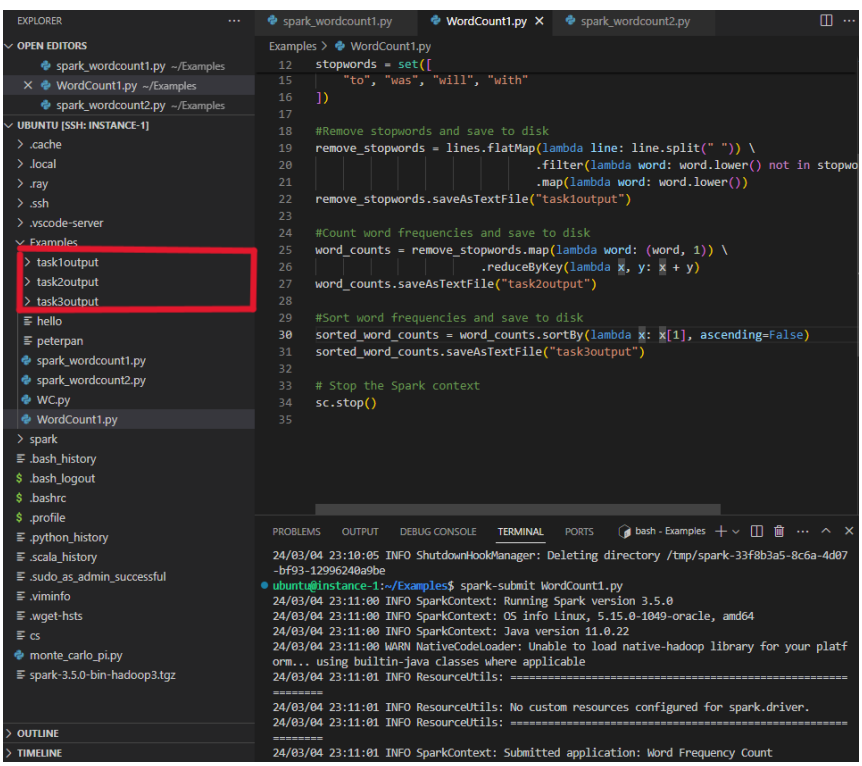


*Figure 1. Running command and creation of tasks outputs.*

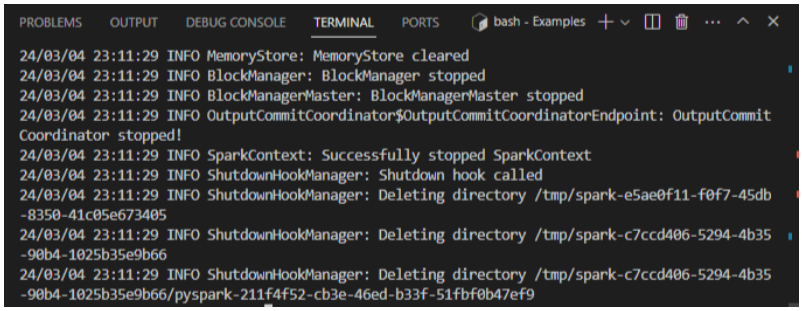


*Figure 2. Terminal output (runtime 29 secs)*

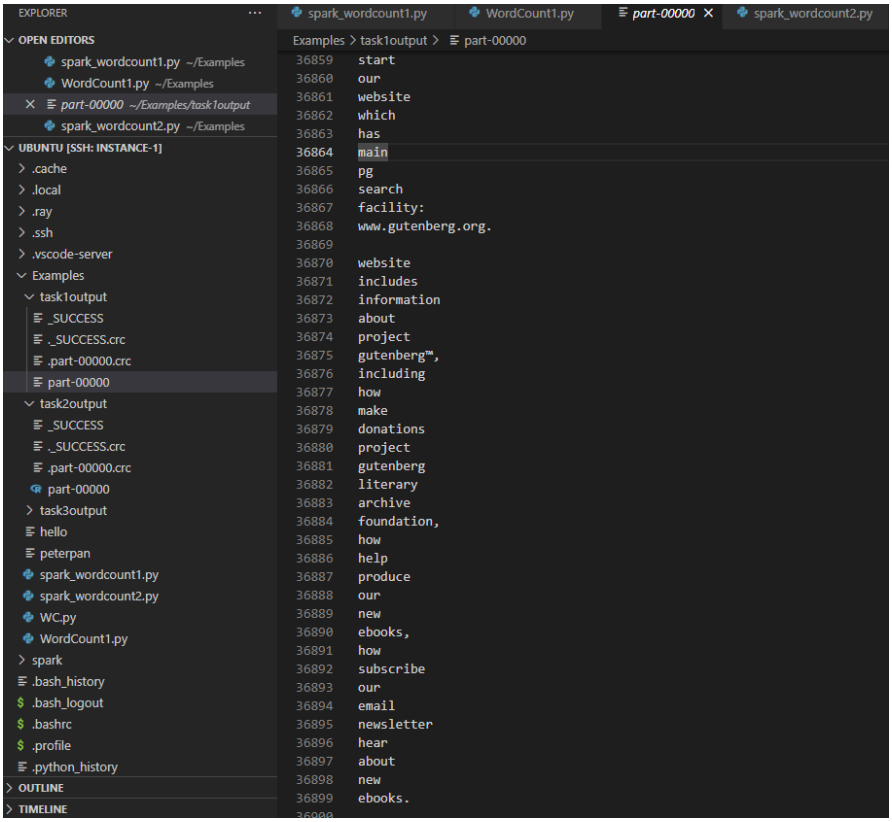


*Figure 3. Task1Output (remove stopwords).*

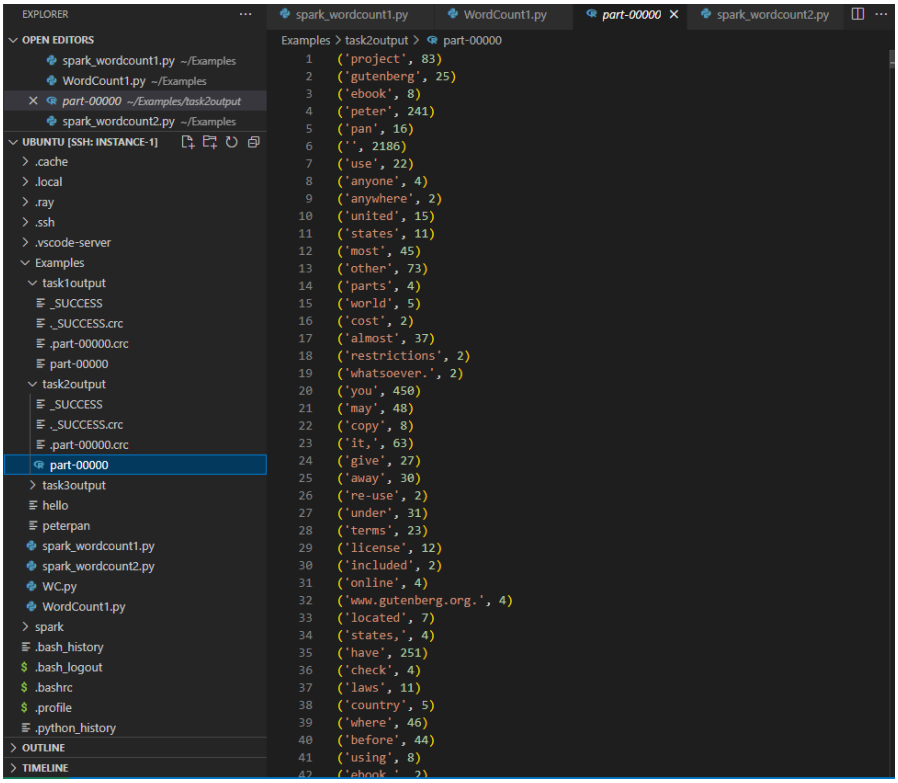


*Figure 4. Task2output ( count word frequencies<key/value>.)*

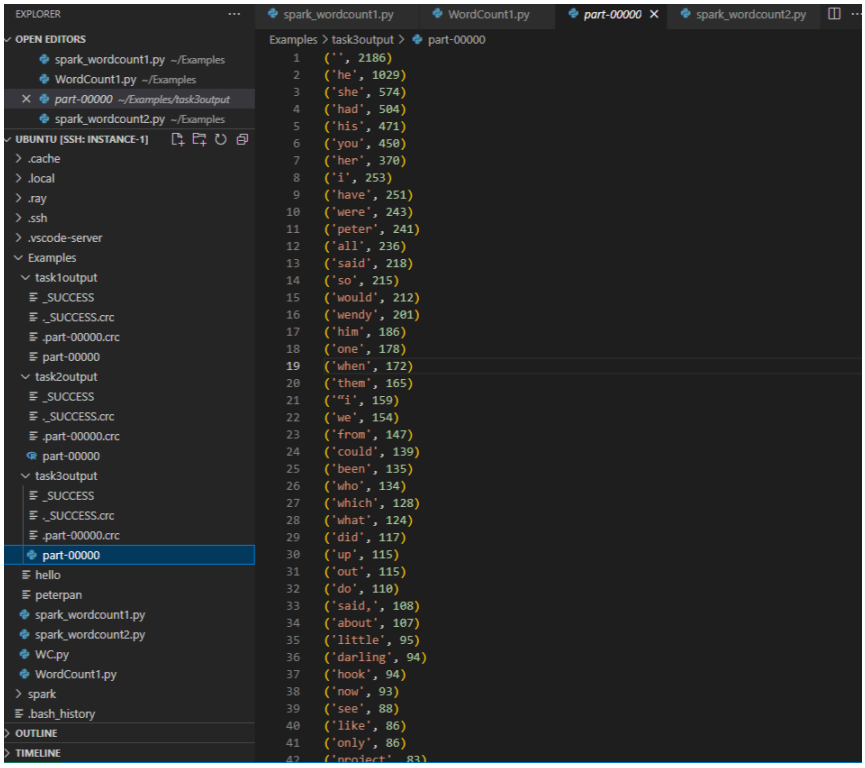


*Figure 5. Task3output (Sorted word frequencies).*