ICP\_2 Report

Jongkook Son(Kyle)

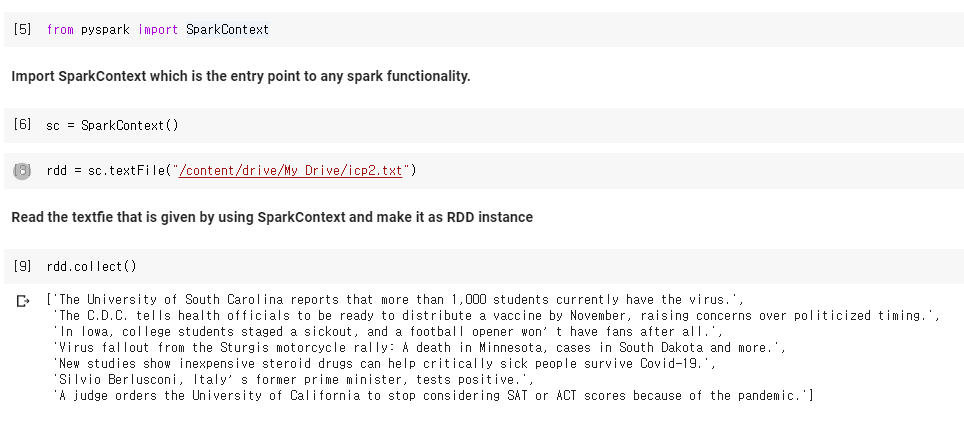
ICP Requirements:

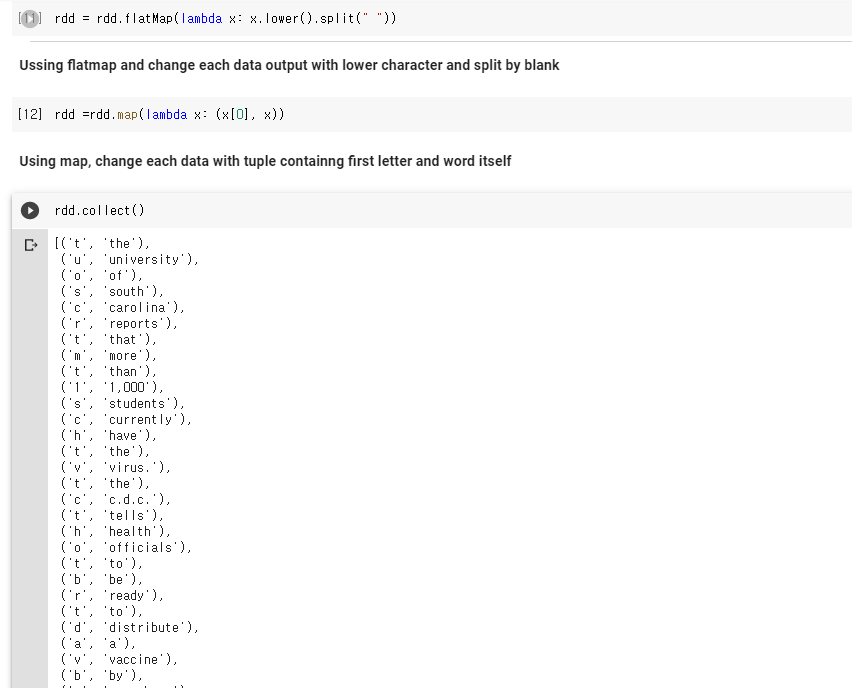
1. Spark Integration with Colab (or IDE that you are using) (50 points)
2. Creating a well commented Spark program and outputting the correct results and writing it to output file. (40 points)
3. Code quality, Pdf Report quality, video explanation (10 points)

a. What I learned in the ICP :

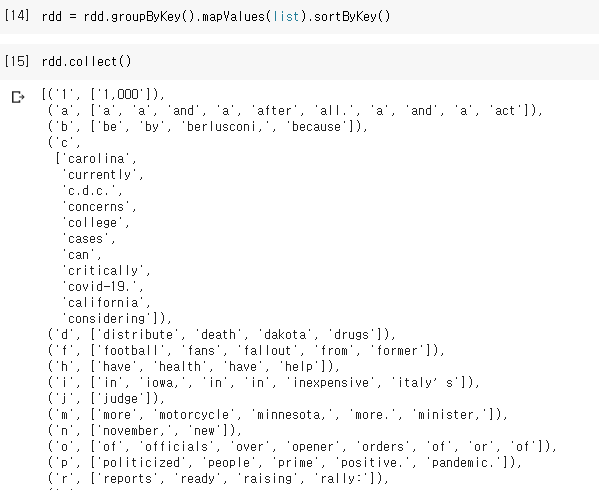
=> I could have learned the basics of pyspark. First of all, I learned how to make RDD data by Using SparkSession and SparkContext And after making RDD data, I was also able to read it make it a Data frame. After that, I learned basic data analysis skills by using spark. By using the ml module in pyspark, implementing some simple model for ml analysis.

b. ICP description what was the task you were performing and Screen shots that shows the successful execution of each required step of your code

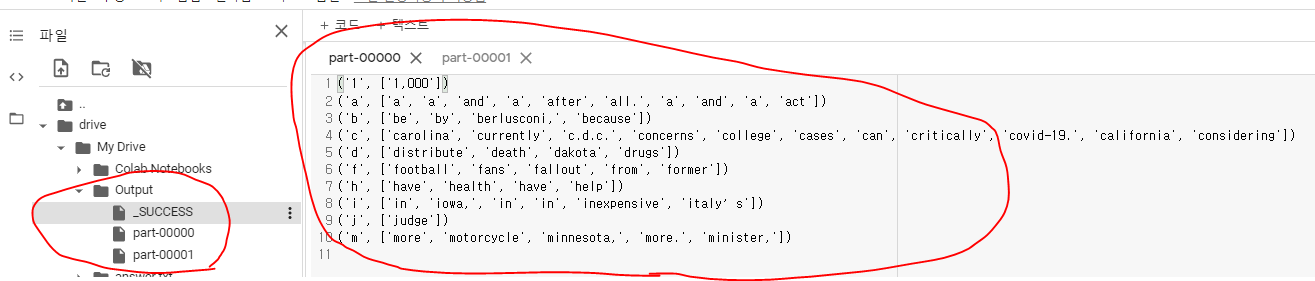
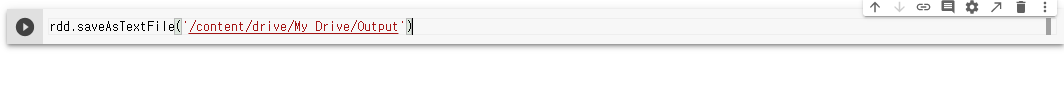
1. Thanks to the starter code that was given, I could easily integrate spark into google colab
2. After installing pyspark in Google Colab, I import sparkcontext which is the point to any spark functionality. Then I read textfile and made it as an RDD. Then as You can see below, it is read by a line.
3. Then I used a flatmap and make each character of the data to be lower case letter and also make each line to be separate with blank by using split method. And after that by using map, make each word to be a tuple containing first letter of word and word itself.



1. Since RDD has Key-Value format as an tuple now, We can use group By key method and also sort it by key value to be arranged. After this it is clear that I group the words in a given text file based on starting letters successfully.



1. Finally, I saved it as textfile and saved it on google drive and uploaded it to submission repository.



c. Challenges that I faced

=> I was very new to Pyspark and bigdata. So I should have researched from basic operation. It takes some time to explore but it makes me more knowledgeable and used to it.

d. Out put file link if applicable

=> I uploaded it on my github repository submission folder

e. Video link (YouTube or any other publicly available video platform) =><https://www.youtube.com/watch?v=14N-vGMXSrE>

f. Any inside about the data or the ICP in general

=> I was able to learn how to make RDD with Spark Context and also could get used to grouping RDD with Key and value which was very new to me.