CS5542 Big Data Apps and Analytics

LAB ASSIGNMENT #2

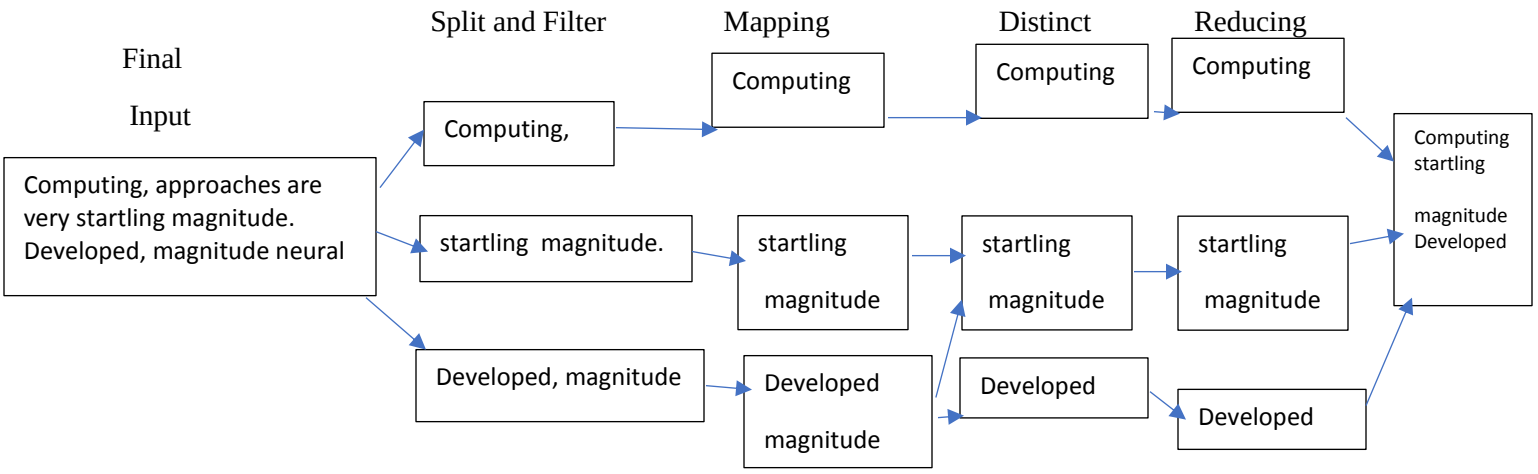
Name:Sidrah Junaid
Class Id:15

1. **Spark Programming:**

Write a spark program with an interesting use case using text data as the input and program should have at least Two Spark Transformations and Two Spark Actions.

Present your use case in map reduce paradigm as shown below (for word count).

USE CASE:



USE CASE IMPLEMENTATION:

The implemented use case shows the word having length of nine alphabets from the text present in input file.

1.The spark program is written in Scala using intelliJ IDE and it is showing the words having 9 alphabets. The provided input has whitespaces and contain punctuation marks.

2.Program goes line by line and split the words based on empty characters and punctuation marks.

3.Then it will filter the words whose length is nine. In the provided text there is some redundant words from which only distinct word selected.

4.Then, the whole output would be collected and select in a separate output text file.

TRANSFORMATIONS:

1.Flat map:

Each line in the input file will be checked and split based on whitespaces and punctuation mark.

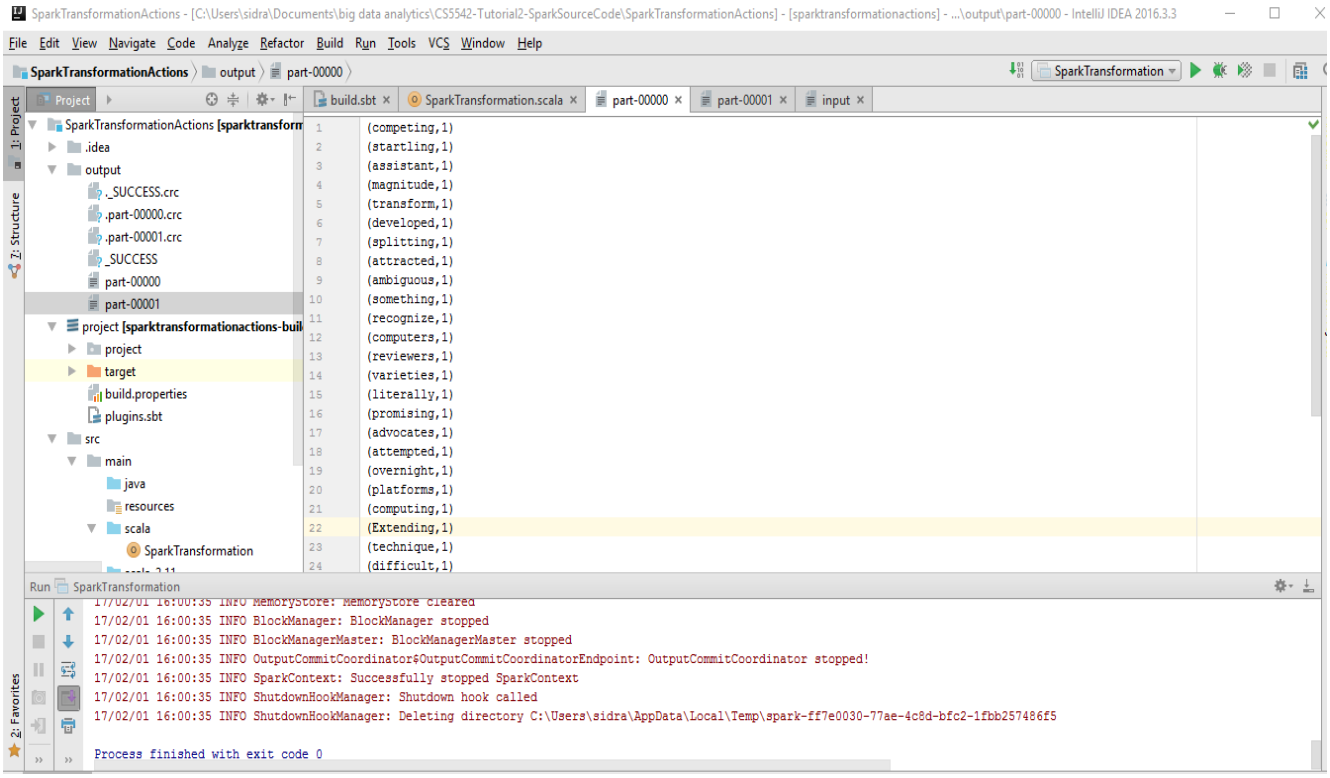
2.Distinct:
Here we select distinct words to avoid repetition.

3.Filter:
Here filter is applied so the words only having length 9 will be saved in main_filter.

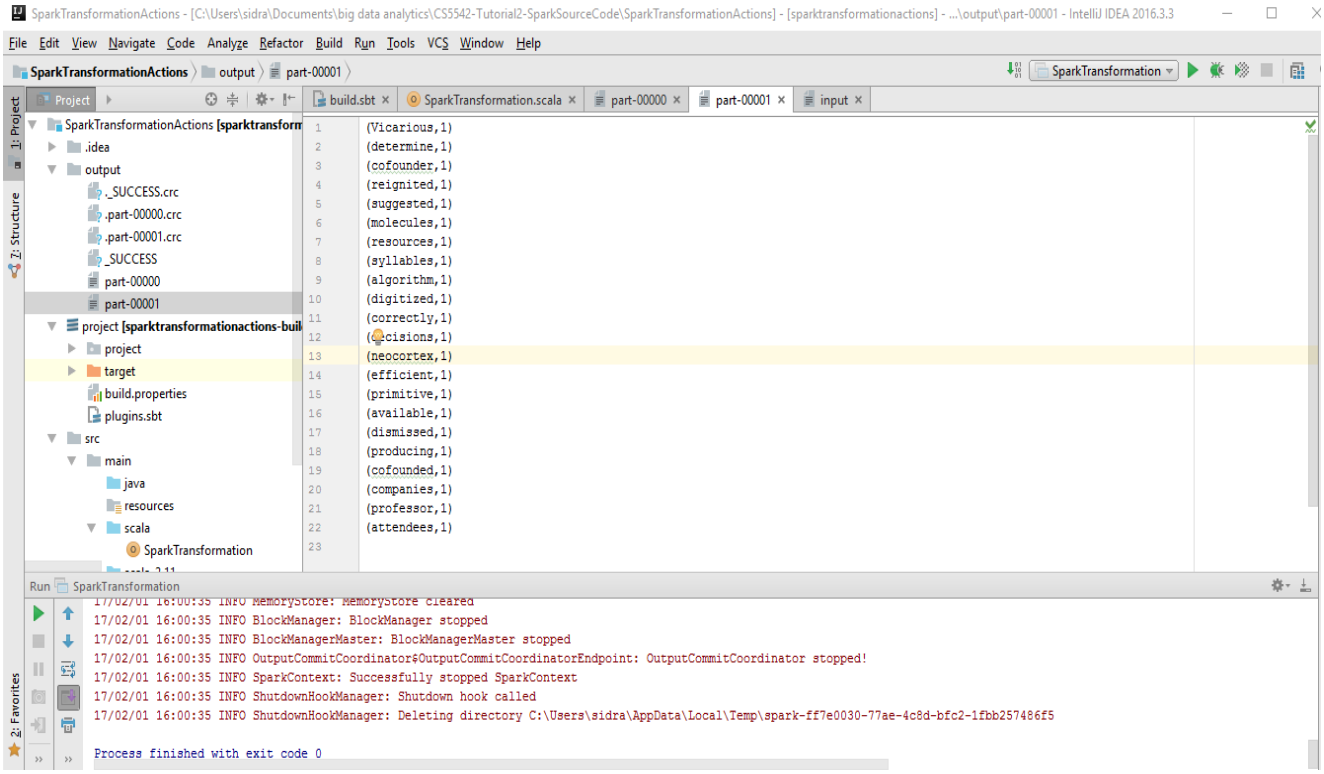
Actions:
1.Collect:
Return all the elements of the input file as an array at the driver program.

2.saveAsTextFile:
Write the elements of the input file as a text file (or set of text files) in a given directory in the local filesystem

Output:
File one:



File two:



Input file: