**CS5542 – Big Data Analytics and Apps**

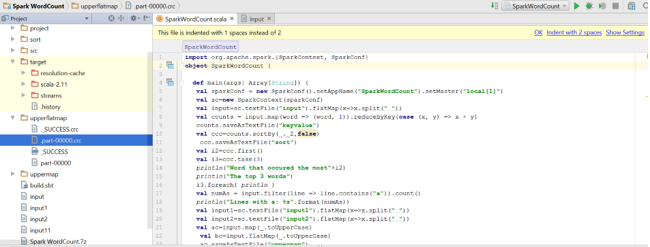
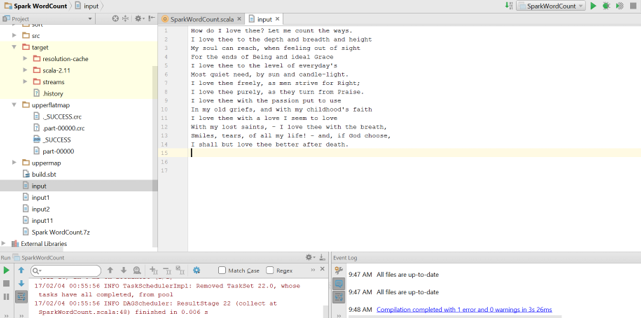
**Lab 2 – Assignment Submission**

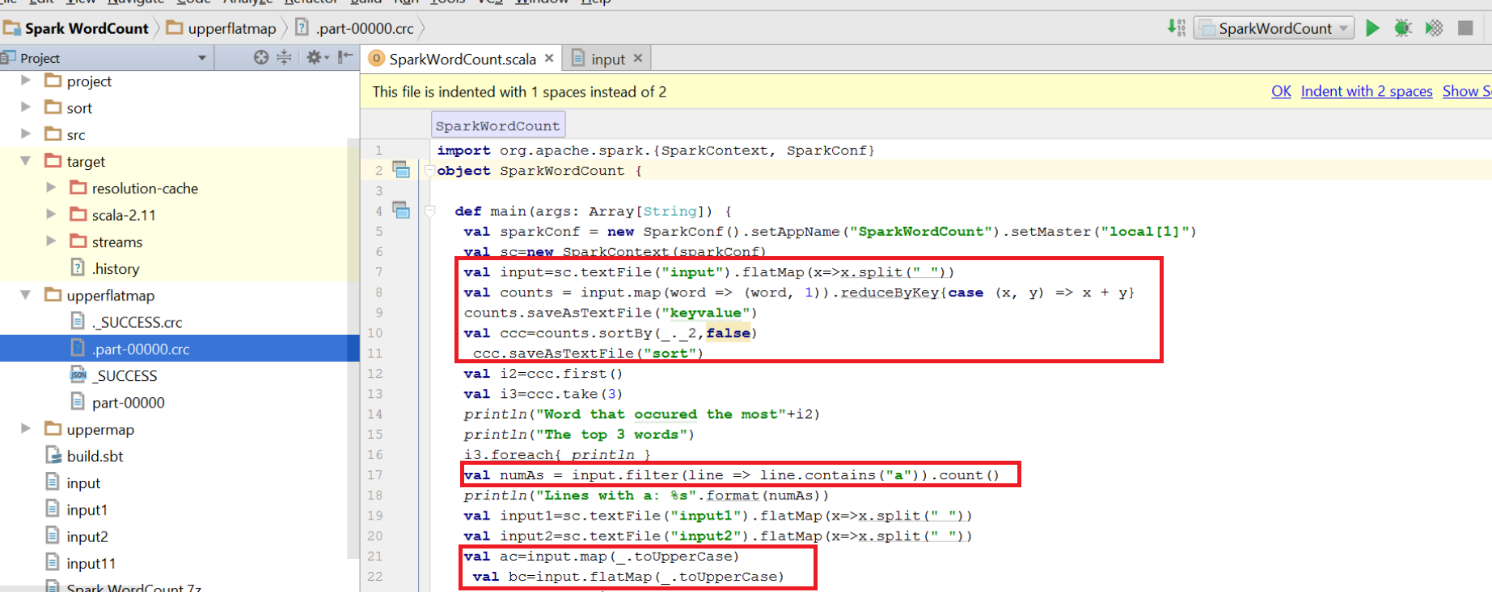
Name: Lava Kumar S

Class ID : 38

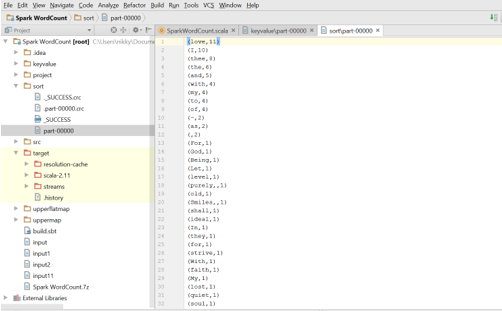
**Question**  
Write a spark program with an interesting use case using text data as the input and program should have at least Two Spark Transformations and Two Spark Actions.  
  
**Transformations & Actions**  
RDDs support two types of operations: transformations, which create a new dataset from an existing one, and actions, which return a value to the driver program after running a computation on the dataset. For example, map is a transformation that passes each dataset element through a function and returns a new RDD representing the results. On the other hand, reduce is an action that aggregates all the elements of the RDD using some function and returns the final result to the driver program (although there is also a parallel reduceByKey that returns a distributed dataset).

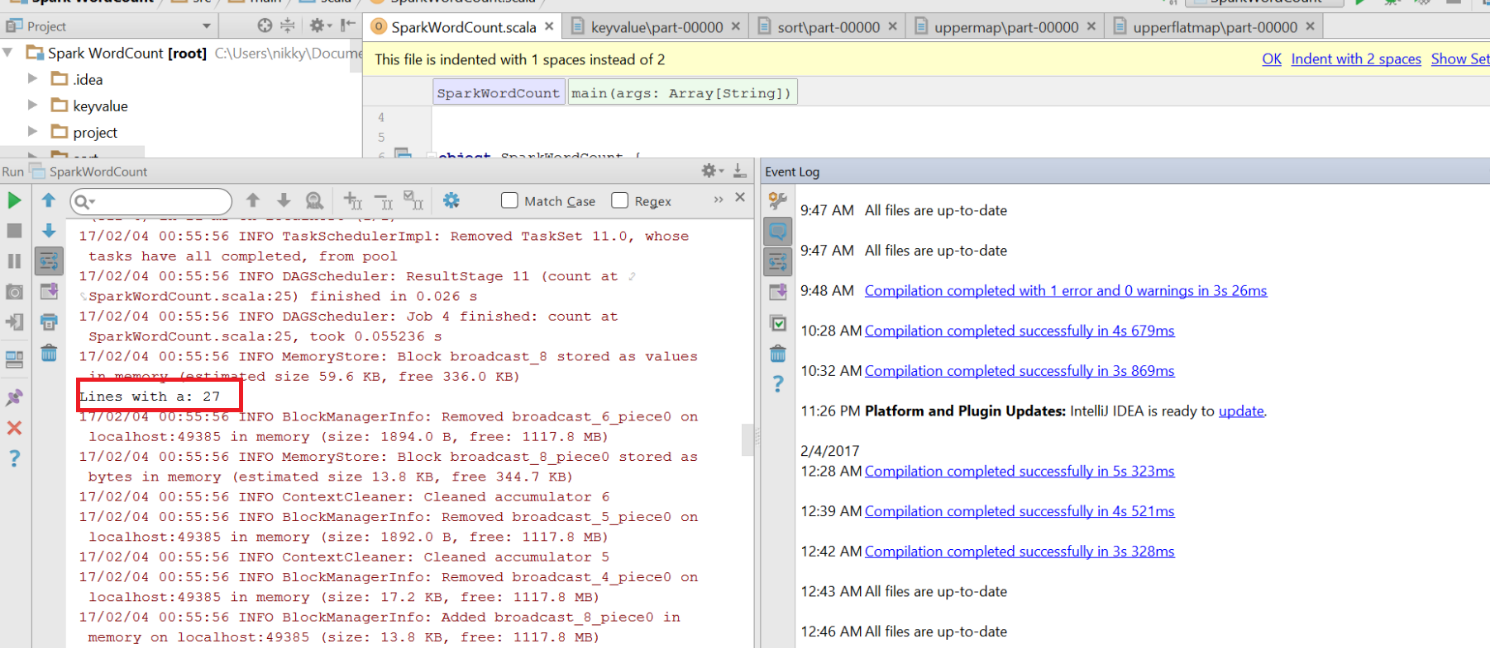
All transformations in Spark are lazy, in that they do not compute their results right away. Instead, they just remember the transformations applied to some base dataset (e.g. a file). The transformations are only computed when an action requires a result to be returned to the driver program. This design enables Spark to run more efficiently. For example, we can realize that a dataset created through map will be used in a reduce and return only the result of the reduce to the driver, rather than the larger mapped dataset.  
  
**Spark Program**  
A scala program has been written to demonstrated spark actions and transformations.  
  
**Source Code:**

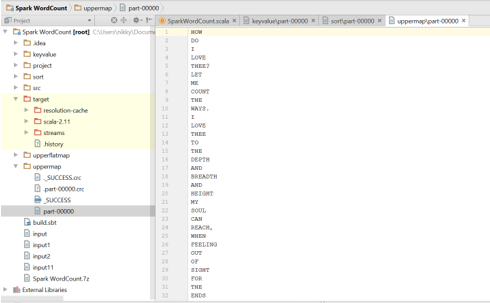
  
   
  
**Input1**  


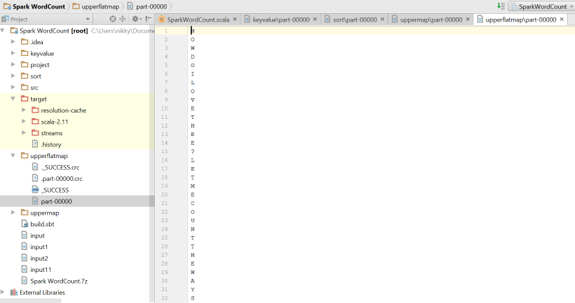
**Transformations**  
The input has been read and a **flat map** operation is performed to divide the input files to strings. Then a word count has been performed using **map** and **reduceByKey** to divide the input file to (key,value) pairs.Then A sort operation is performed using **sortby** to find the most occured words.Also a **filter** operation is performed to find the number of lines that contains charachter'a'. 

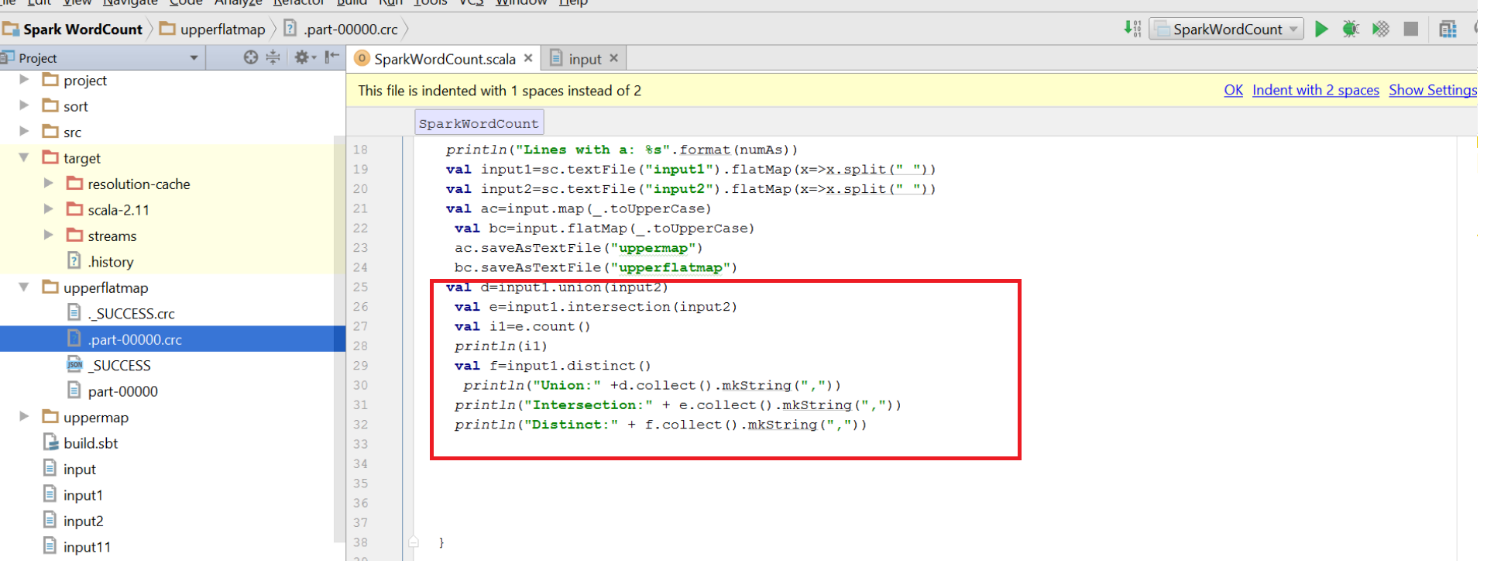
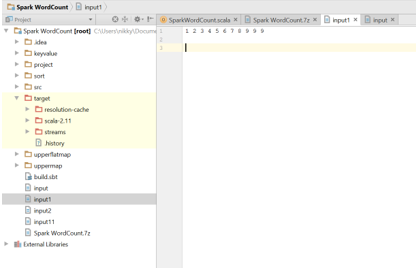
**Output after performing a series of Transformation:**

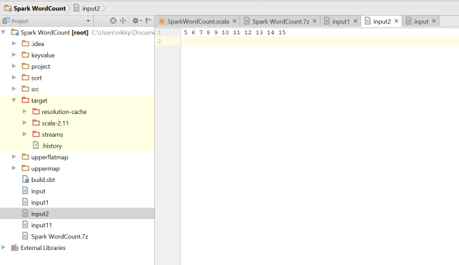


  
Also map and flat map were used with **upper case** functions to convert the files to upper case.

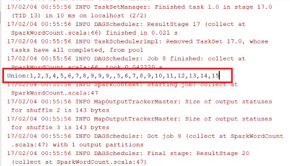


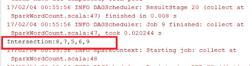


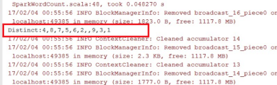
Also two input files are read to perform **Intersection**,**Union**,**Join** operations.  
  
**Program**  
  
**Input1**  


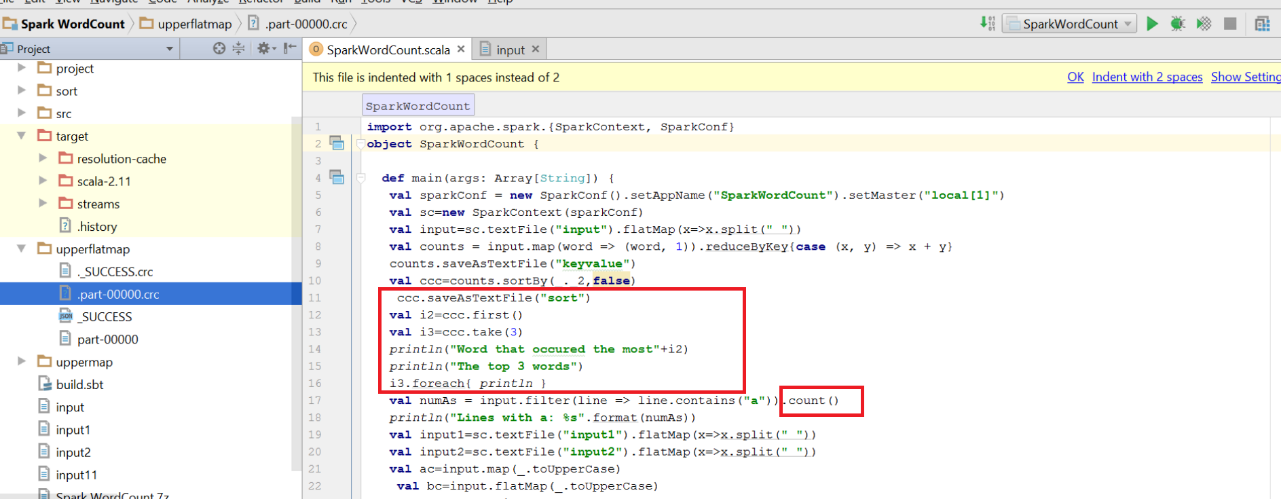
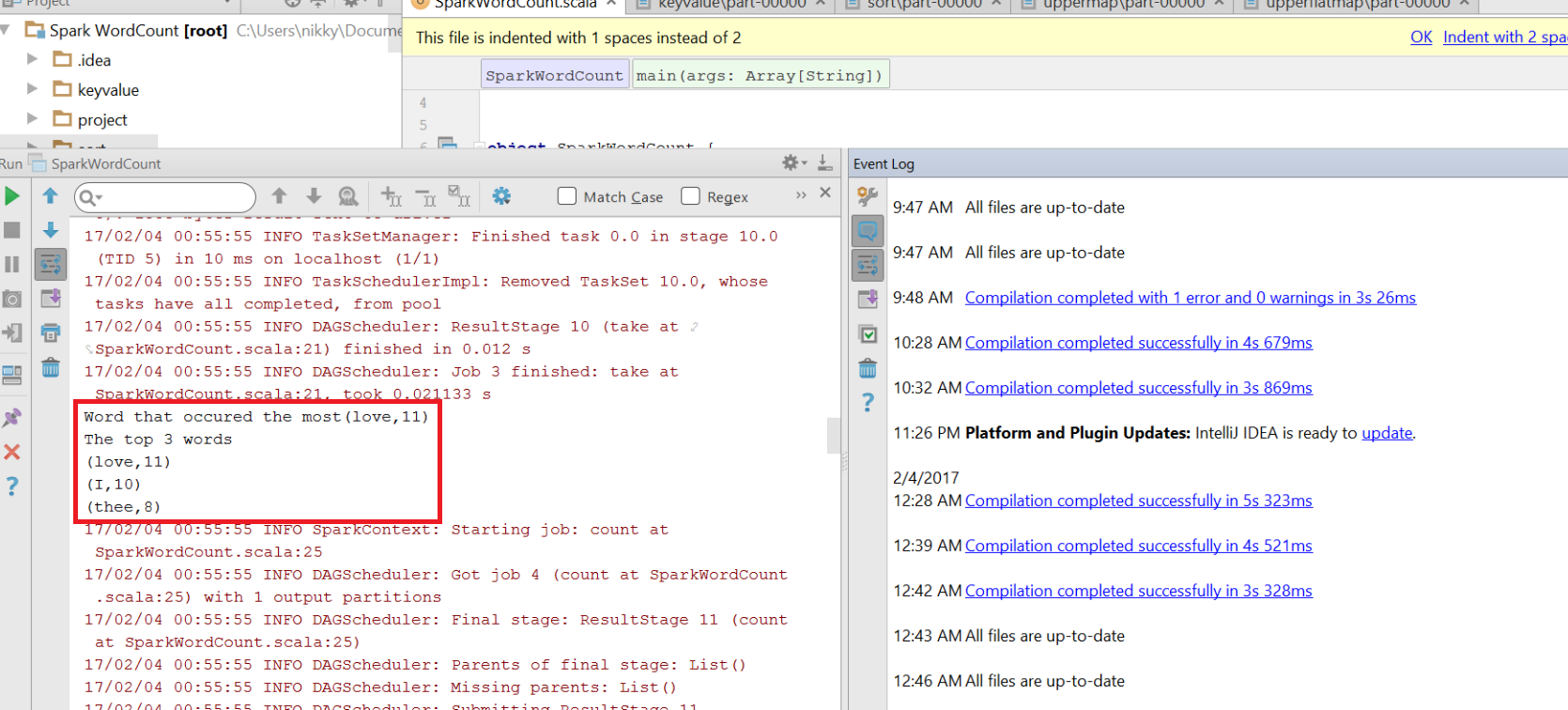


**Output:**







**Actions** After the sorting is done the most occured word using **first**,The top 3 words using **take**and in the filter operation**count** is used,also**saveAsTextFile** is used.  
.   
Hence 8 transformations and 4 actions have been performed.  
  
**Map Reduce Paradigm**  
s