**CS5560 Knowledge Discovery Management**

**In Class Assignment #4**

**13th Sept 2018**

**Feedback Link :** <https://goo.gl/forms/MSfIsJPvsuNLXbkA2>

1. **Create a Spark project, to perform the data statistics about 10 abstracts.**

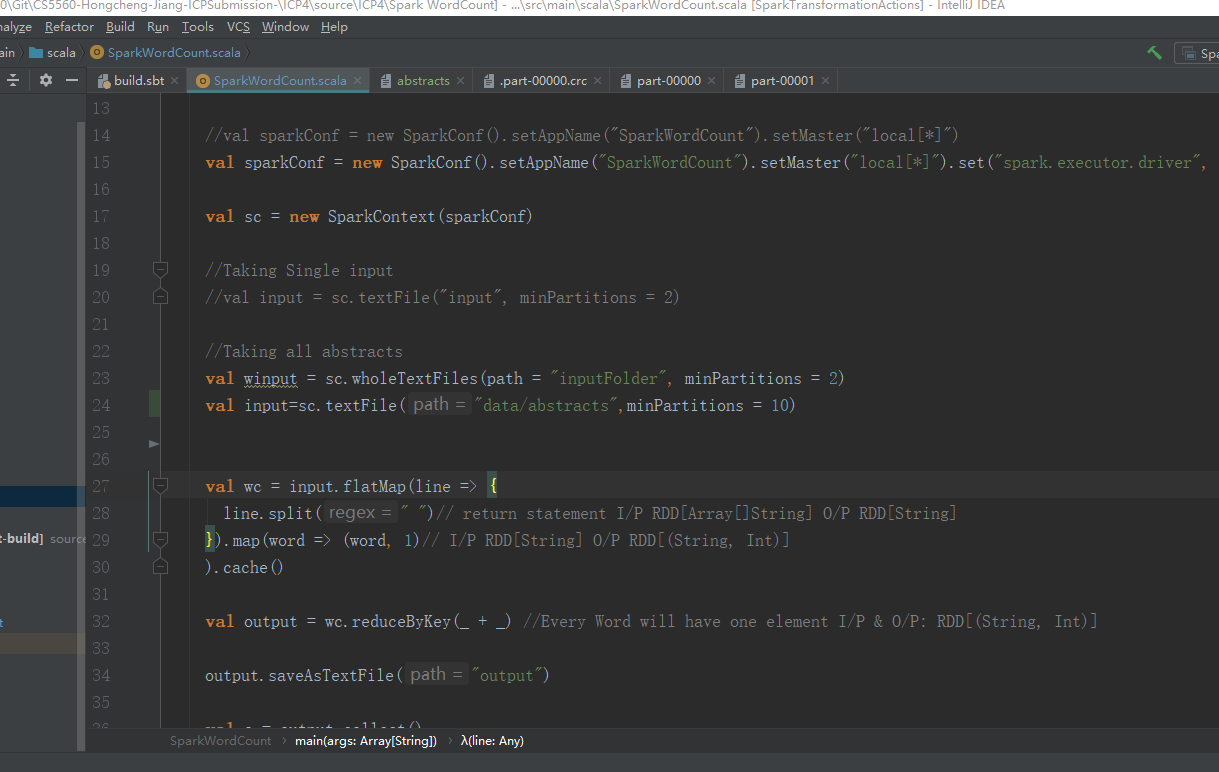
Use at least two transformations and actions.

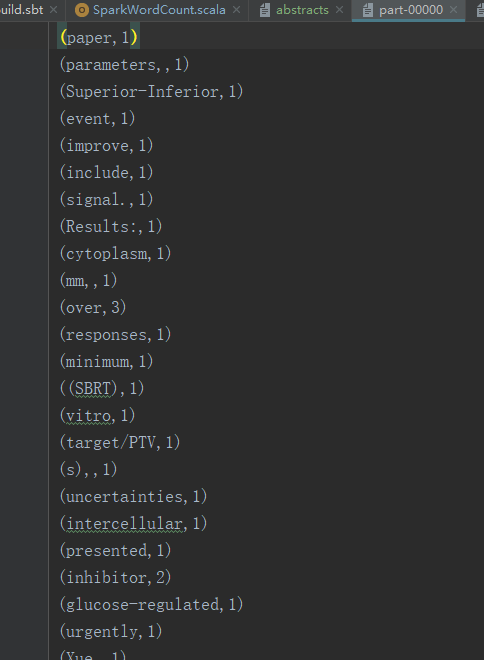
***Input***: Choose 10 of abstracts from your dataset

***Data Statistics Tasks:***

1. *Word Count Statistics*

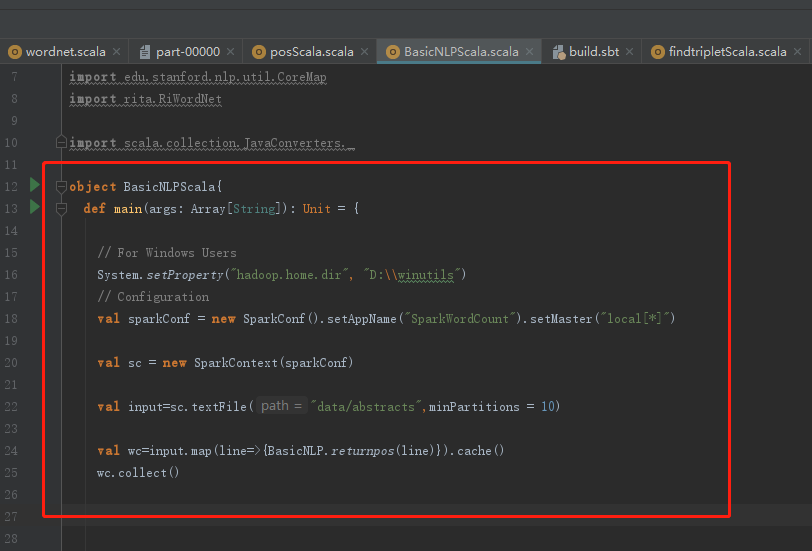
first, using sc.wholeTextFiles to read the data, abstracts. second, using flatMap to split the inputs into word third, using .map(word=>(word,1) to calculate the word counts forth, using reduceByKey(+) to sum the number of occurrences for the same word.

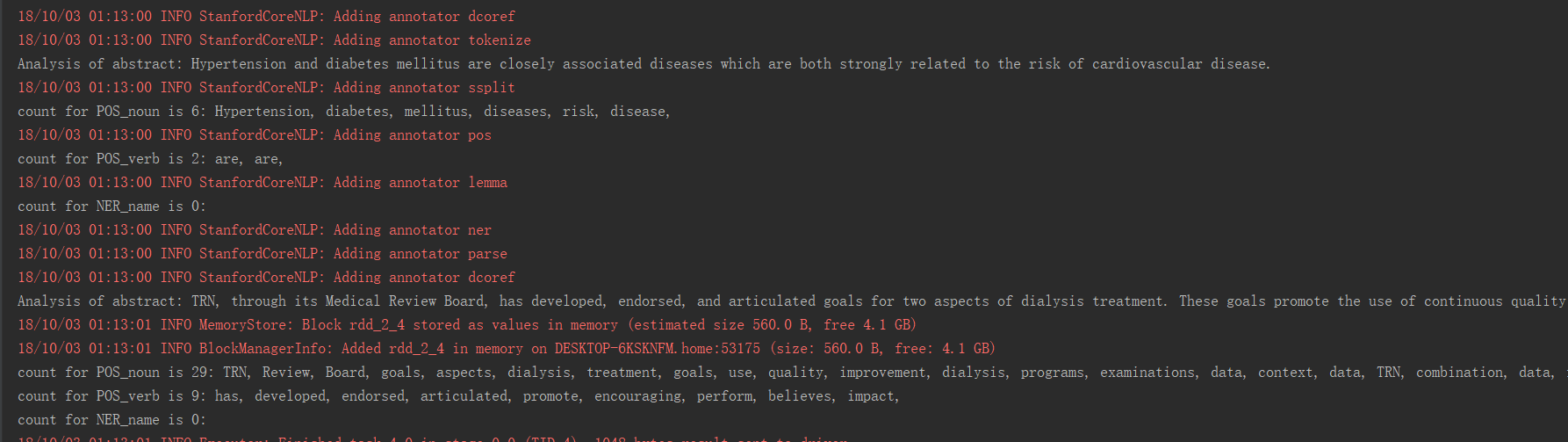




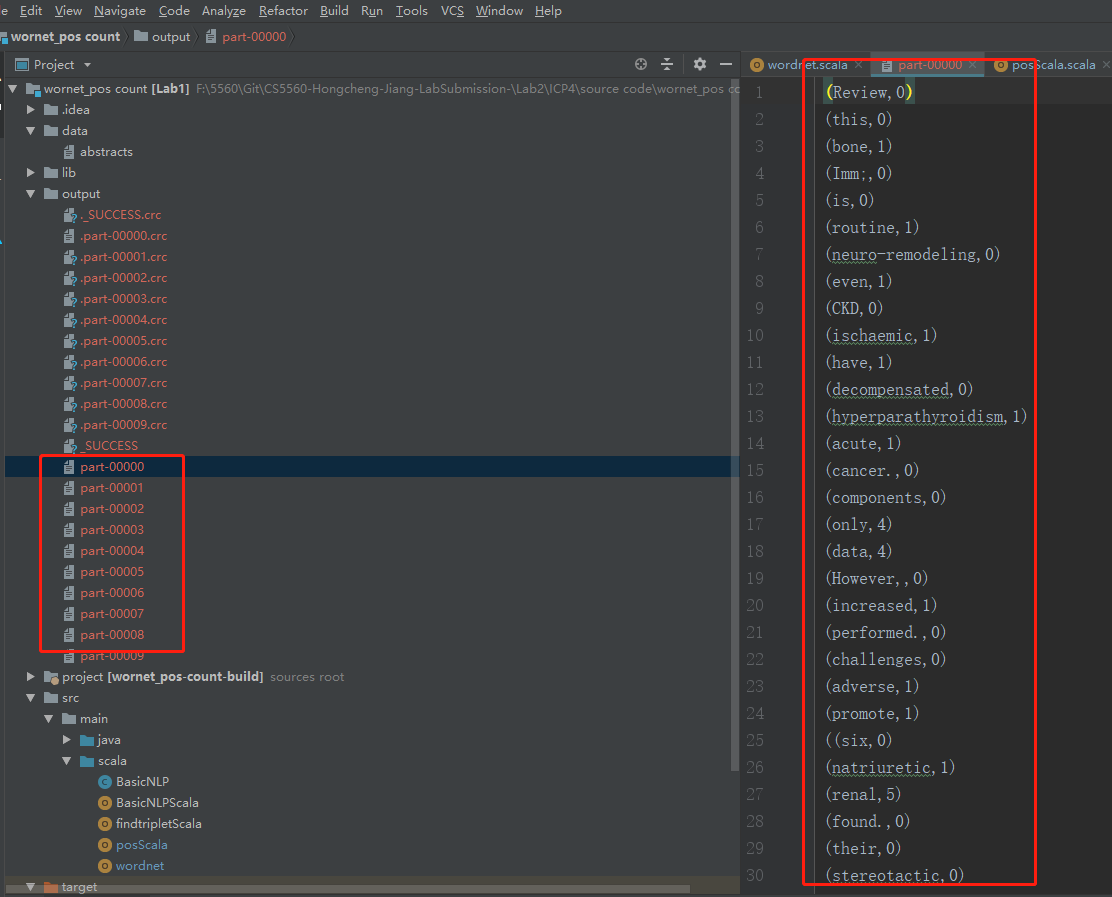
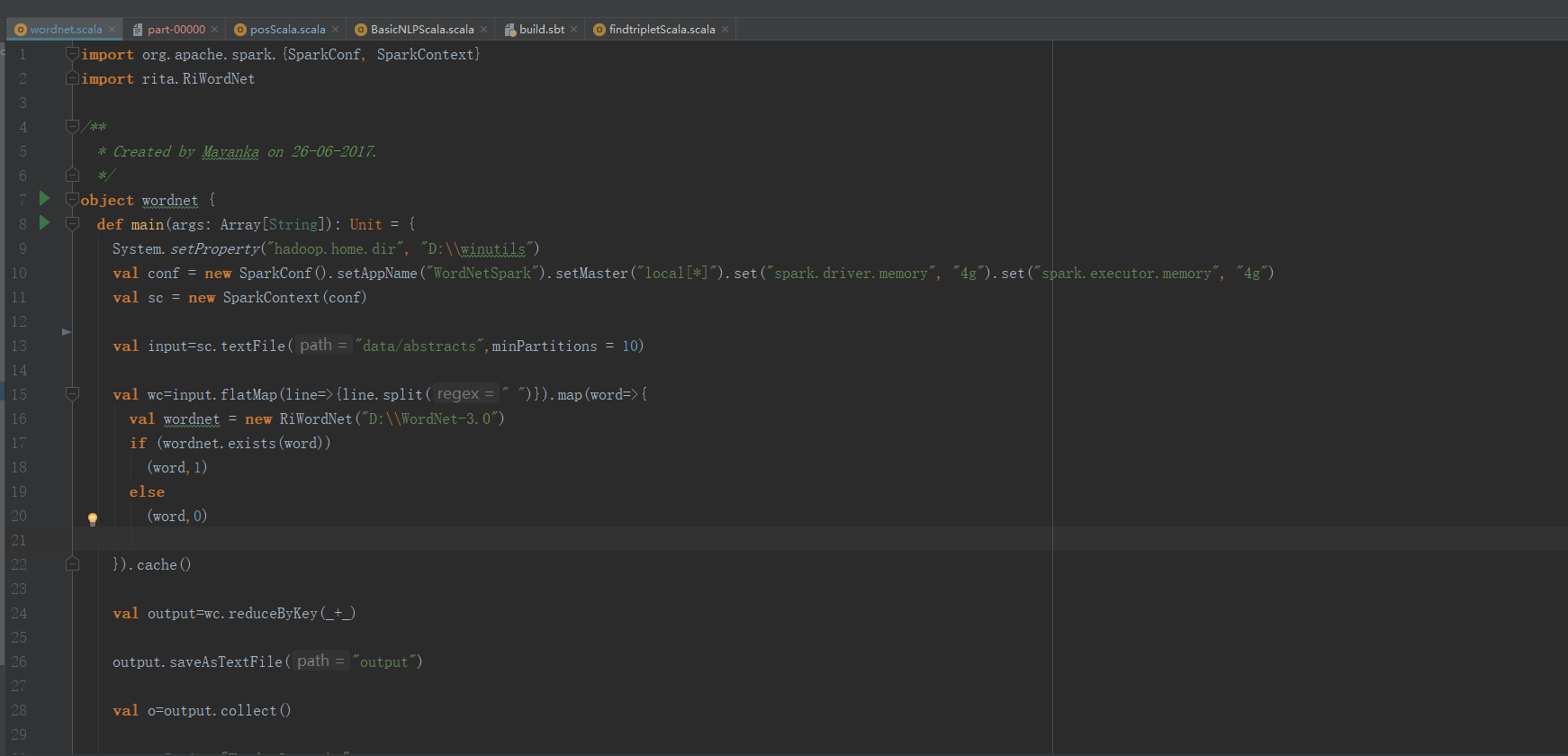
1. *Parts of Speech Statistics*

*We use NLPAPI to get POS.*

**

**

1. *WordNet Statistics*



1. *Medical Word Statistics*

*specifying the input arguments, i.e., ID, bioconcept, inputfile and result format.*

