**CS5560 Knowledge Discovery Management**

**LAB ASSIGNMENT #2**

1. ***In Class Question***

**Generate the output (changes or transformations in the data) manually when the following Spark tasks are applied on the input text. Show your output in details.**

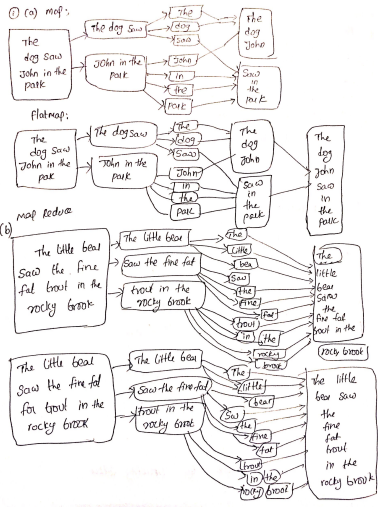
***Input***:

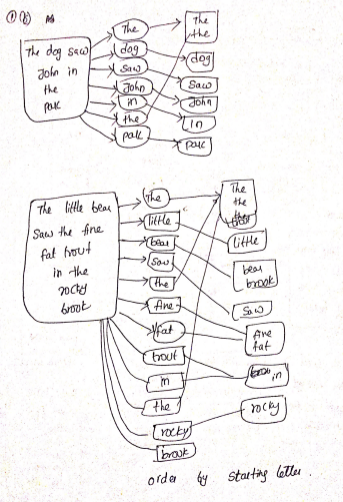
The dog saw John in the park

The little bear saw the fine fat trout in the rocky brook.

**Spark Tasks:**

* 1. Map vs FlatMap
  2. Map Reduce
  3. Group by Starting Letter (Draw diagram how the spark methods used changes the data similar wordcount diagram as shown below)





1. ***In Class Question***

**Write a simple spark program to read a dataset and group each word by the starting letter of its lemmatized word (in this exercise, we assume case-not-sensitive).**

* 1. Write a function ***F*** in Java using CoreNLP to extract Lemmatized Words
  2. Call the function ***F*** from a ***Spark*** ***Transformation*** function
  3. Use 1 (c) diagram to guide the rest of the task.

An example of this task is shown below:

***Input:*** This is a question answering system. The question is from Quora.

***Output:***

T=> This, The

A => a, answer

I => is

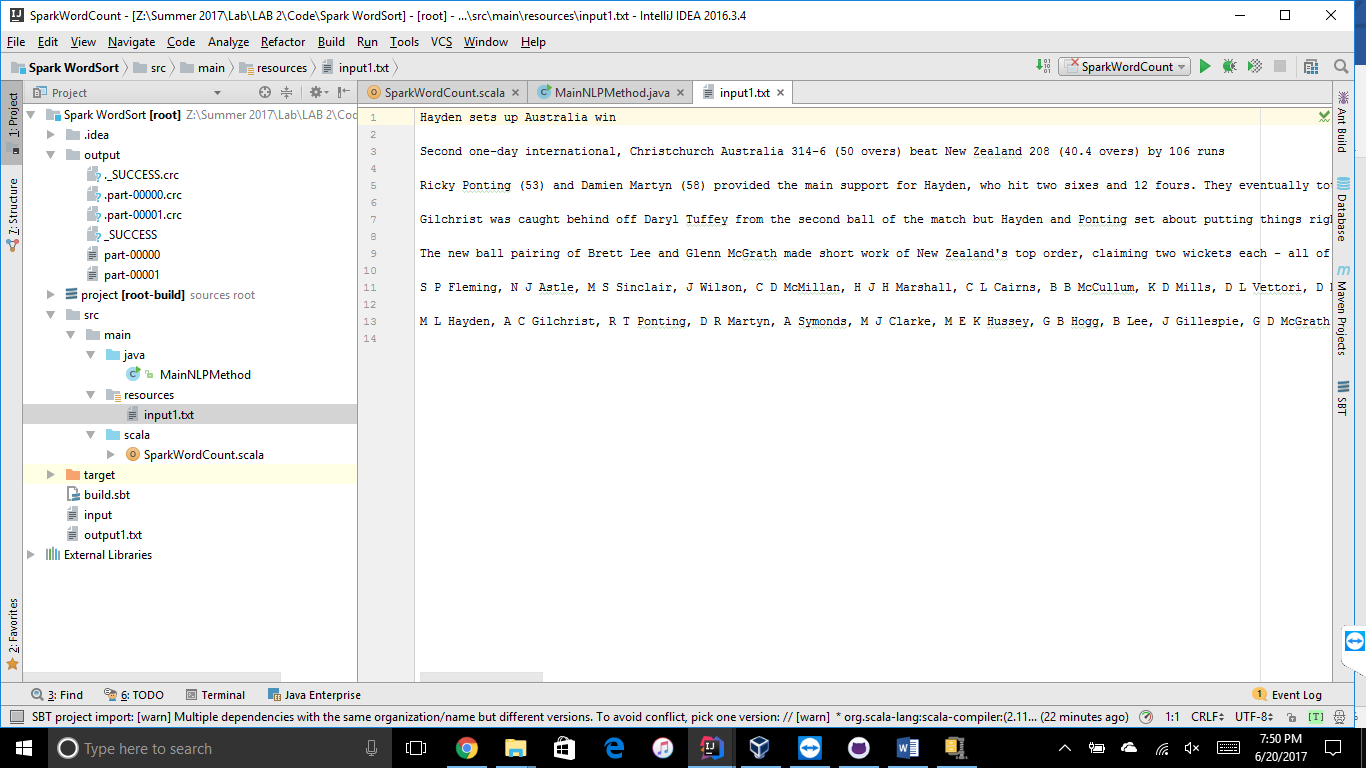
Q => question, Quora

S => system

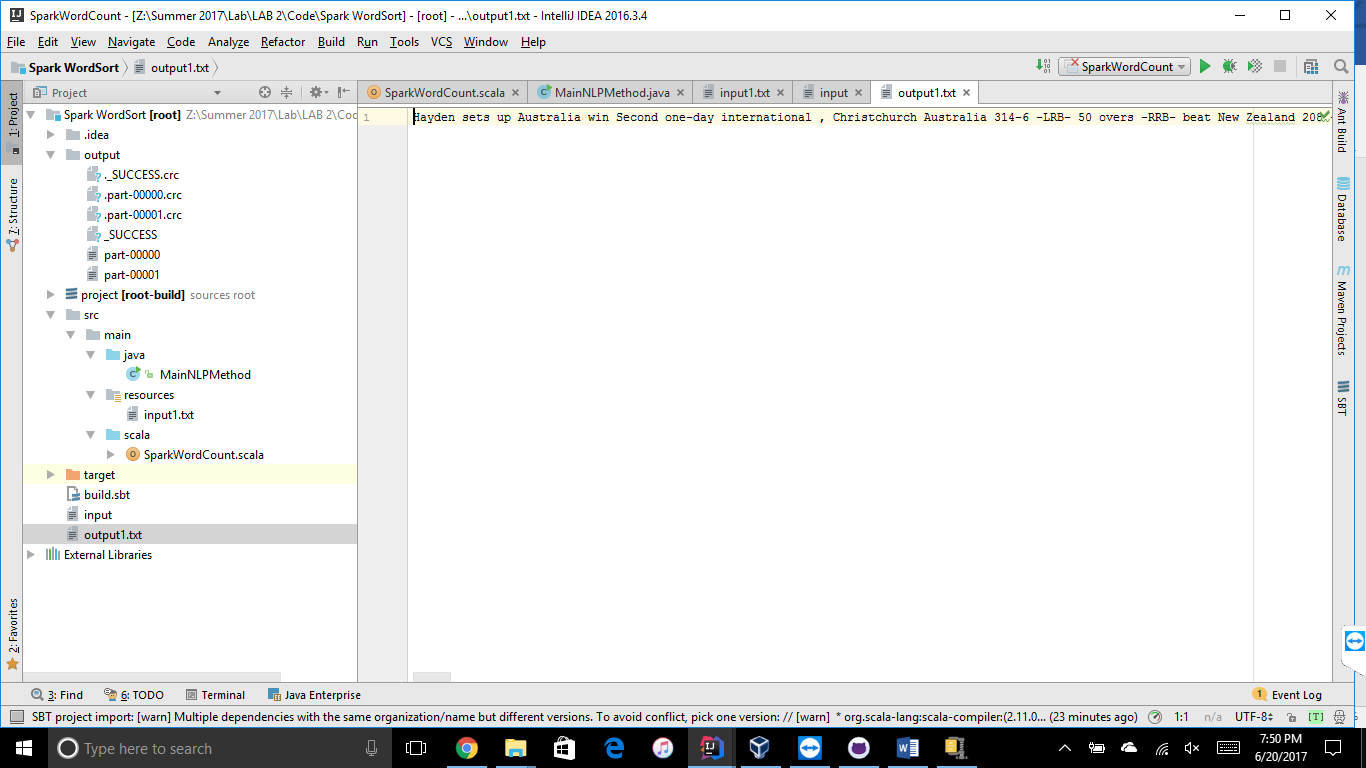
F => from

Note that the *lemmatized* word of***answering*** *is* **answer.**

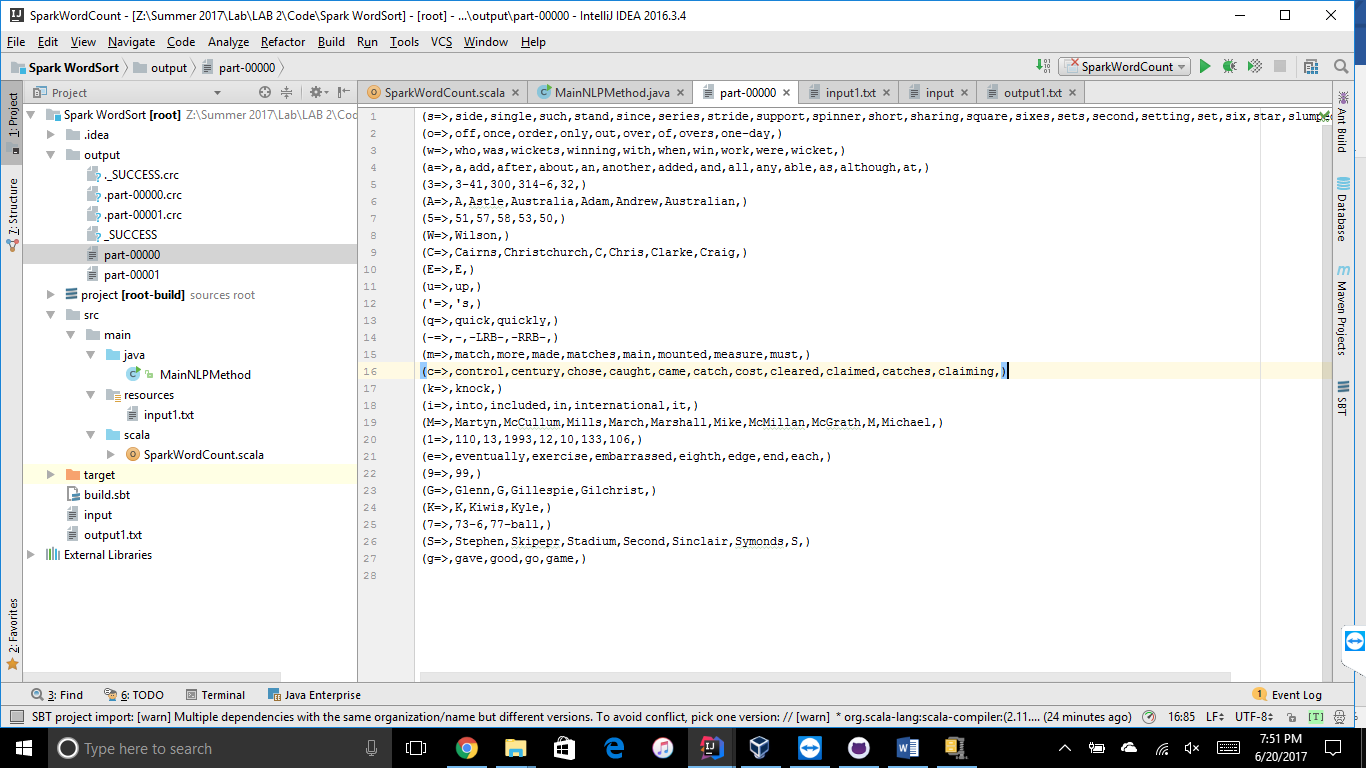
**Input dataset file:**



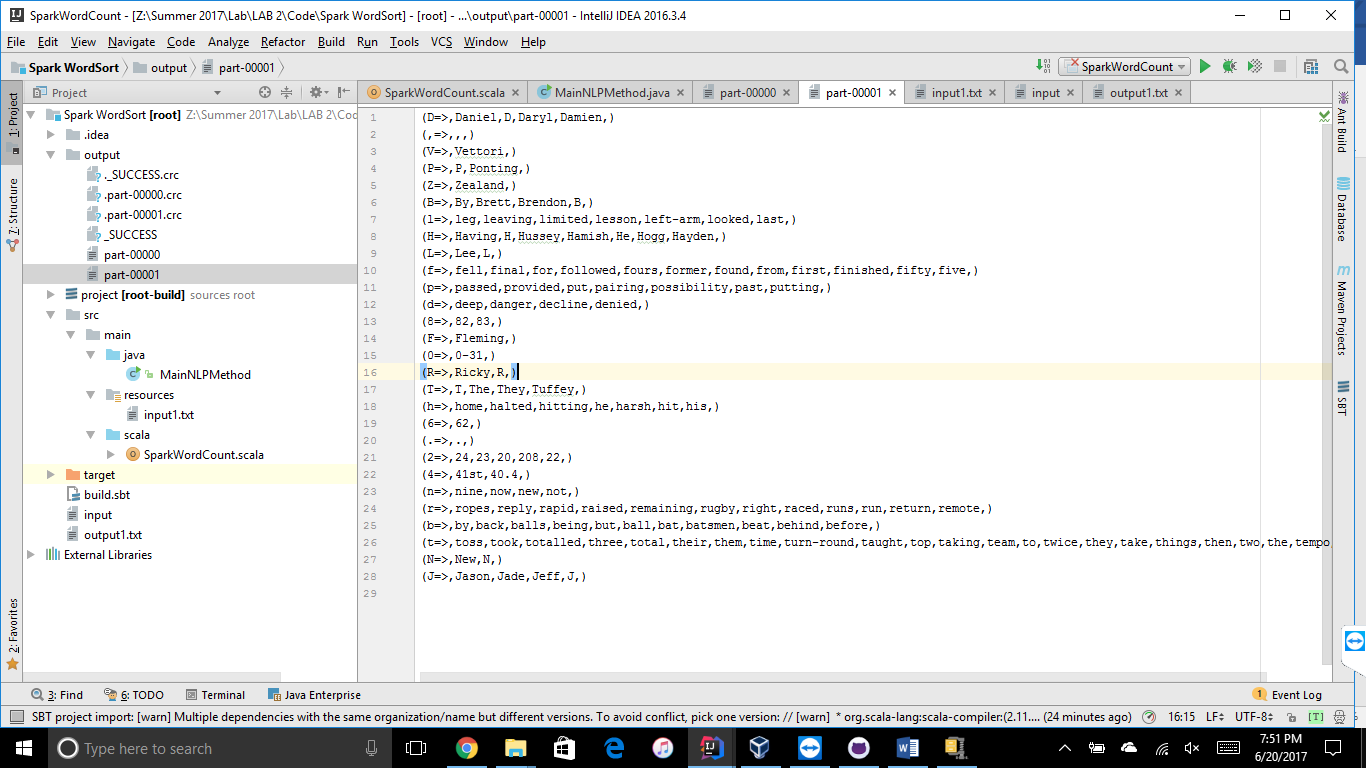
**Processed file:**



**Output file1:**



**Output file2:**



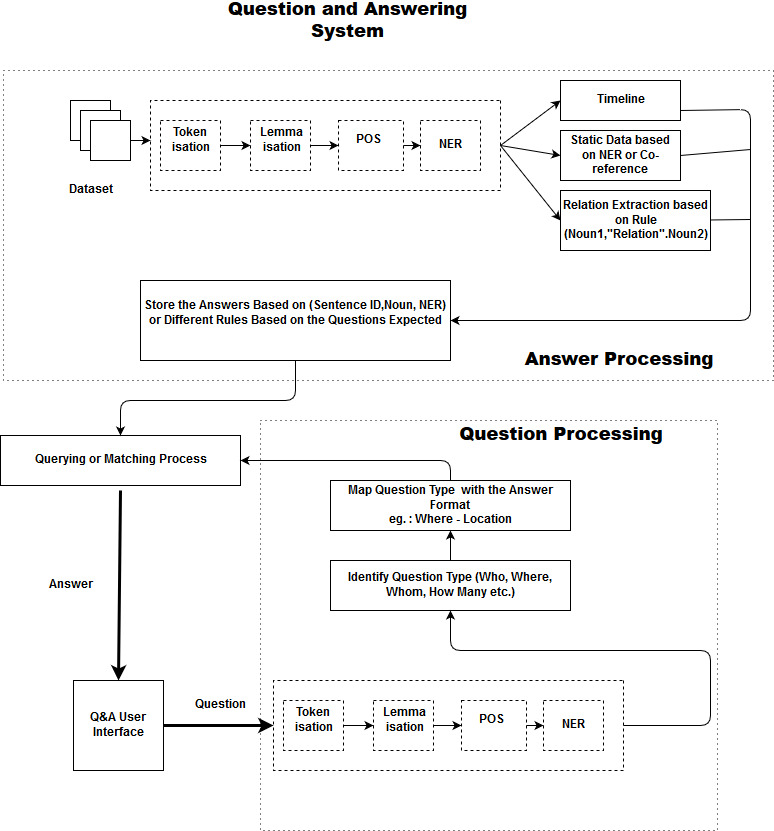
1. ***Take home Question***

**Create a simple question answering system as an extension of the dataset and tasks done in (2). Continuation from Tutorial 1B. Make sure to use at least two *Spark Transformations and two Spark Actions.***

* 1. Use the Question Type

(https://en.wikipedia.org/wiki/Interrogative\_word). Suggestions are When, Who, What. Also try your questions with Google Search engine (<https://google.com>) or the Yahoo AnswerDemo (https://answers.yahoo.com/)

* 1. The three steps are the below diagram is going to be basic idea of your Q&A System

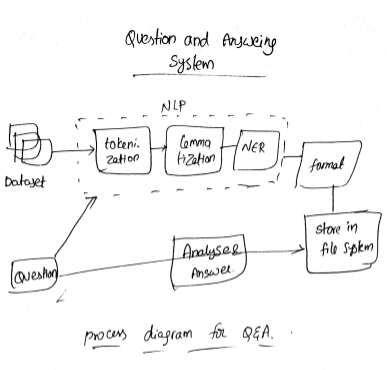


* 1. Start with Step I for the System, it will be easier to process data based on the question.

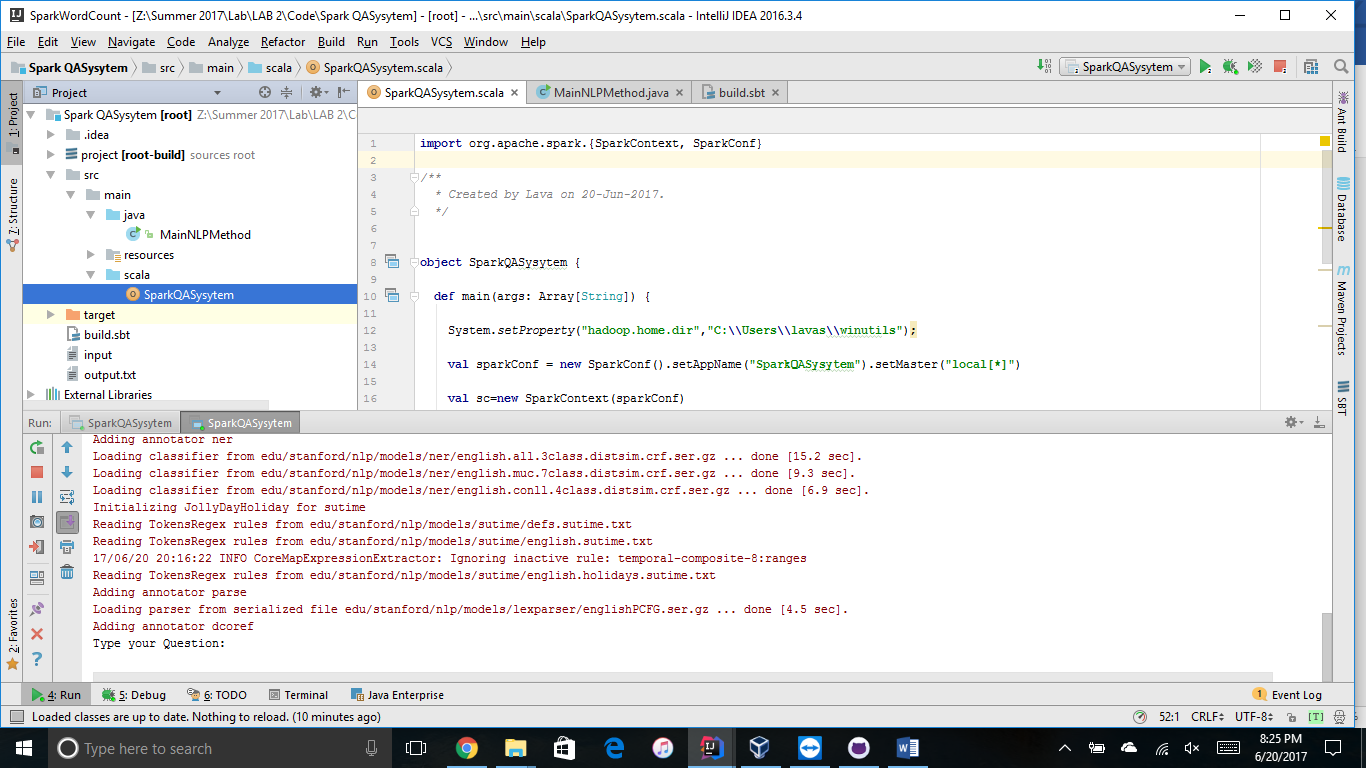
**[note that the first three (a), (b), (c) questions are same as the Tutorial 1(B)]**

* 1. Spark Implementation for Step III in the above diagram (How to process the Data)
     1. ***For Step III, Part 1 Reading Dataset:*** Use spark methods to read the dataset. (Hint: Check Tutorial 2B ppt)
     2. ***For Step III, Part 2 Processing Dataset:*** Use the spark’s transformation and actions to call the coreNLP function.

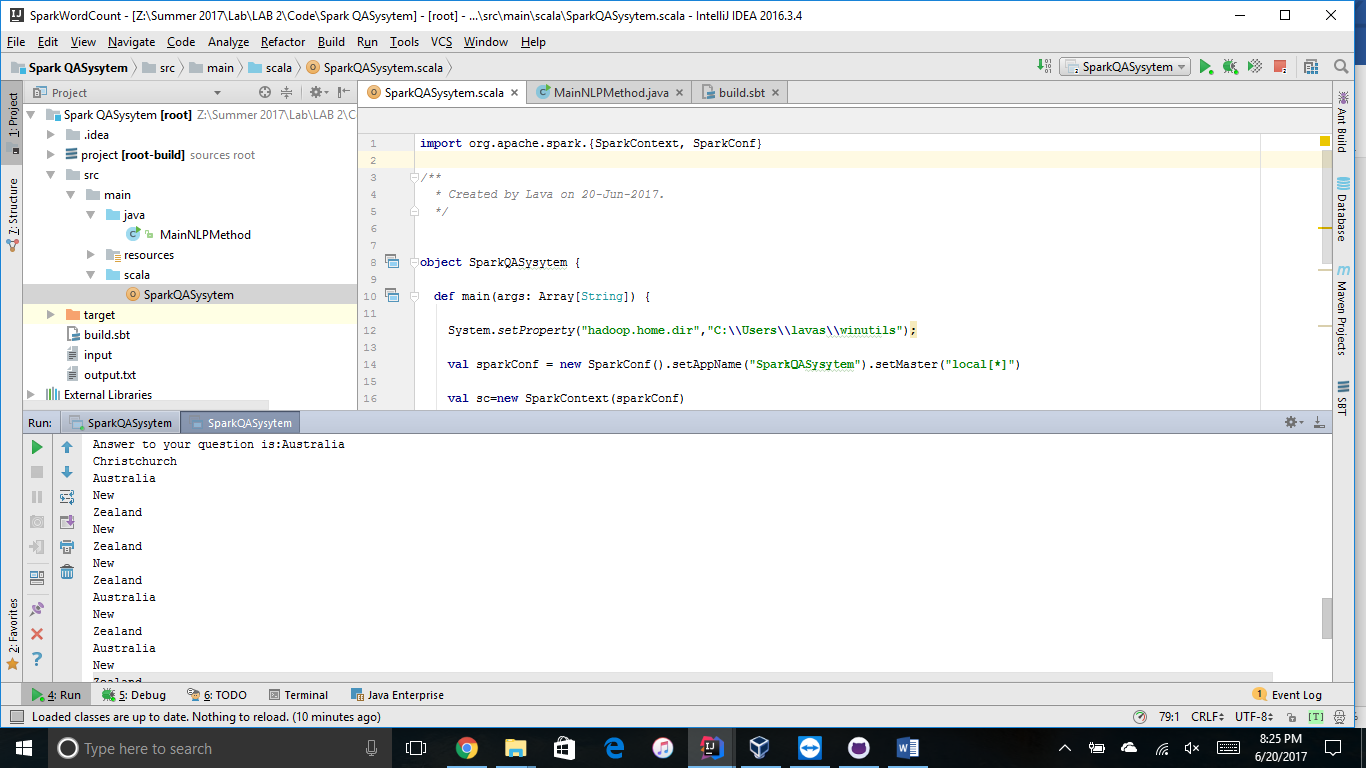
Draw a process diagram for spark methods used (similar to the word count diagram) to guide you through the coding process. Include your diagram in your report.



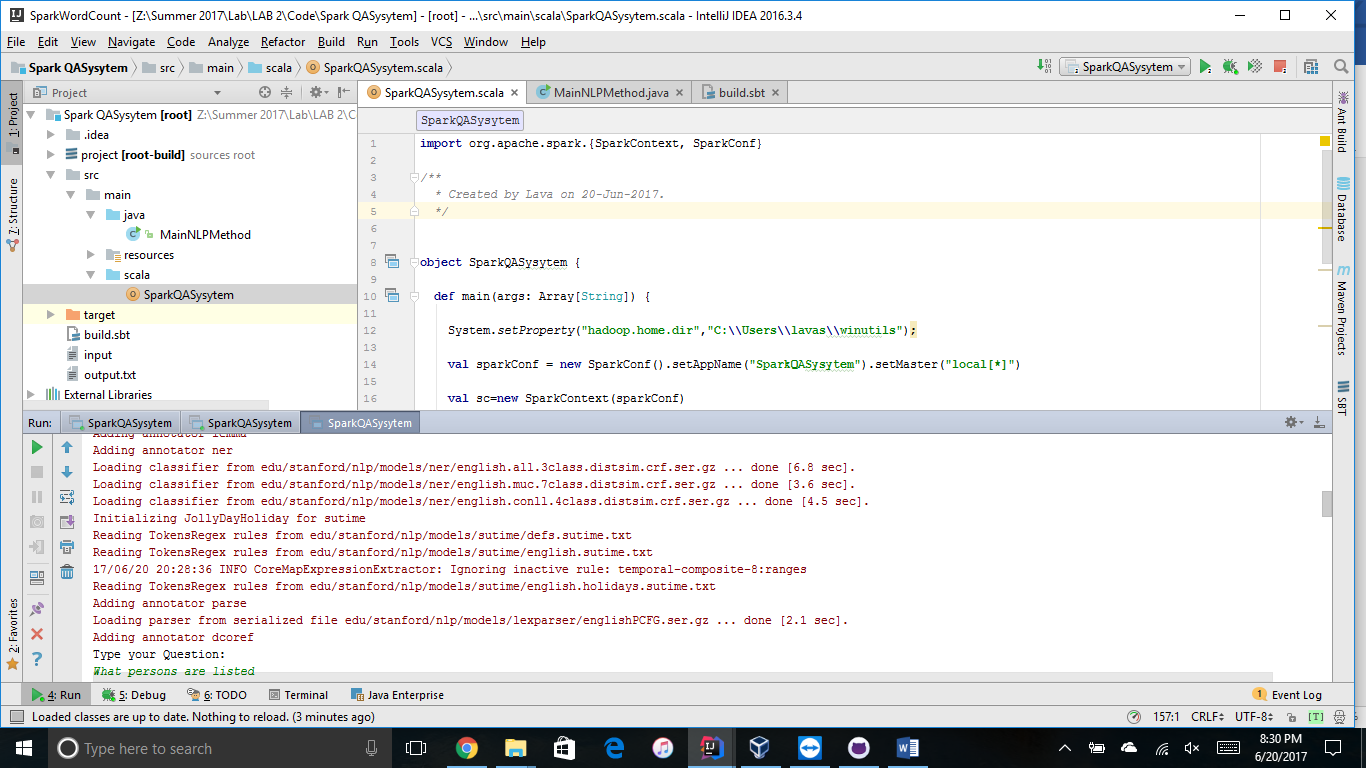
**Input Question1:**



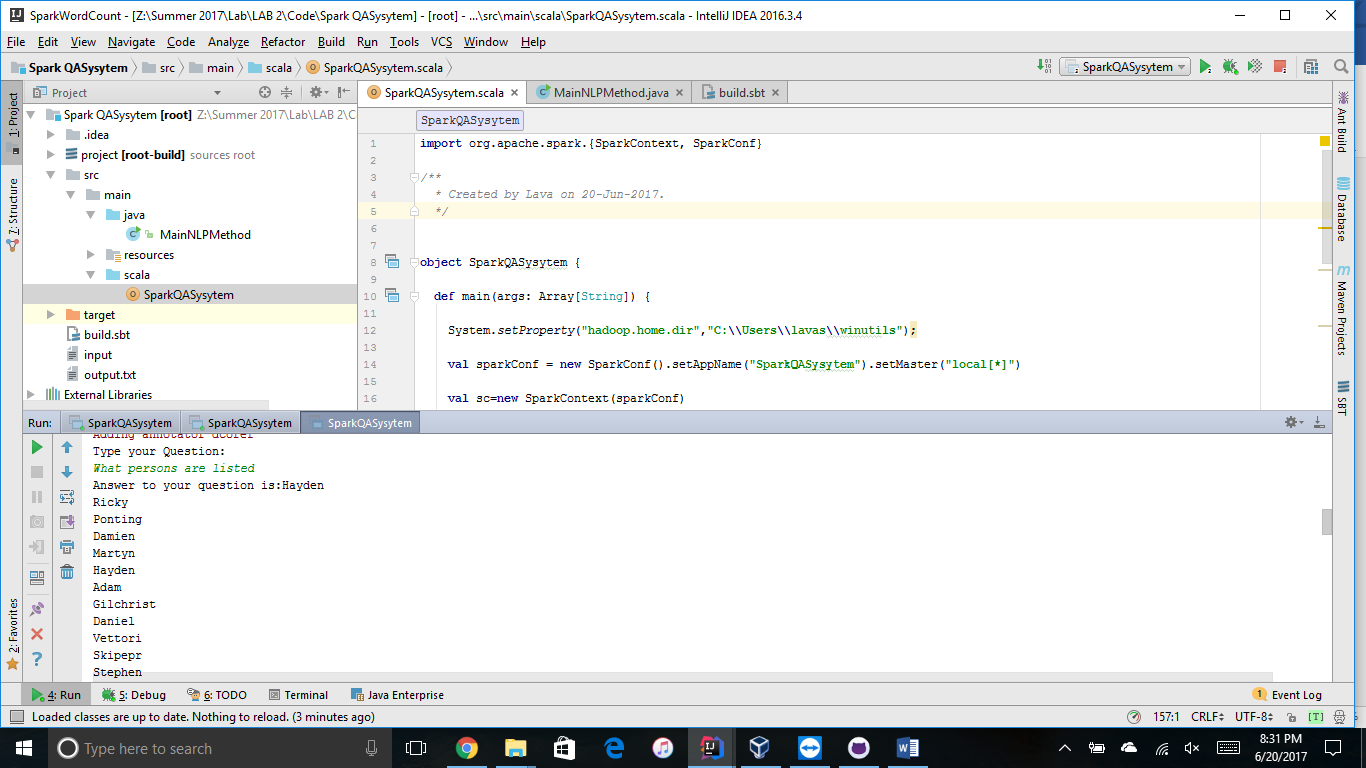
**Response:**



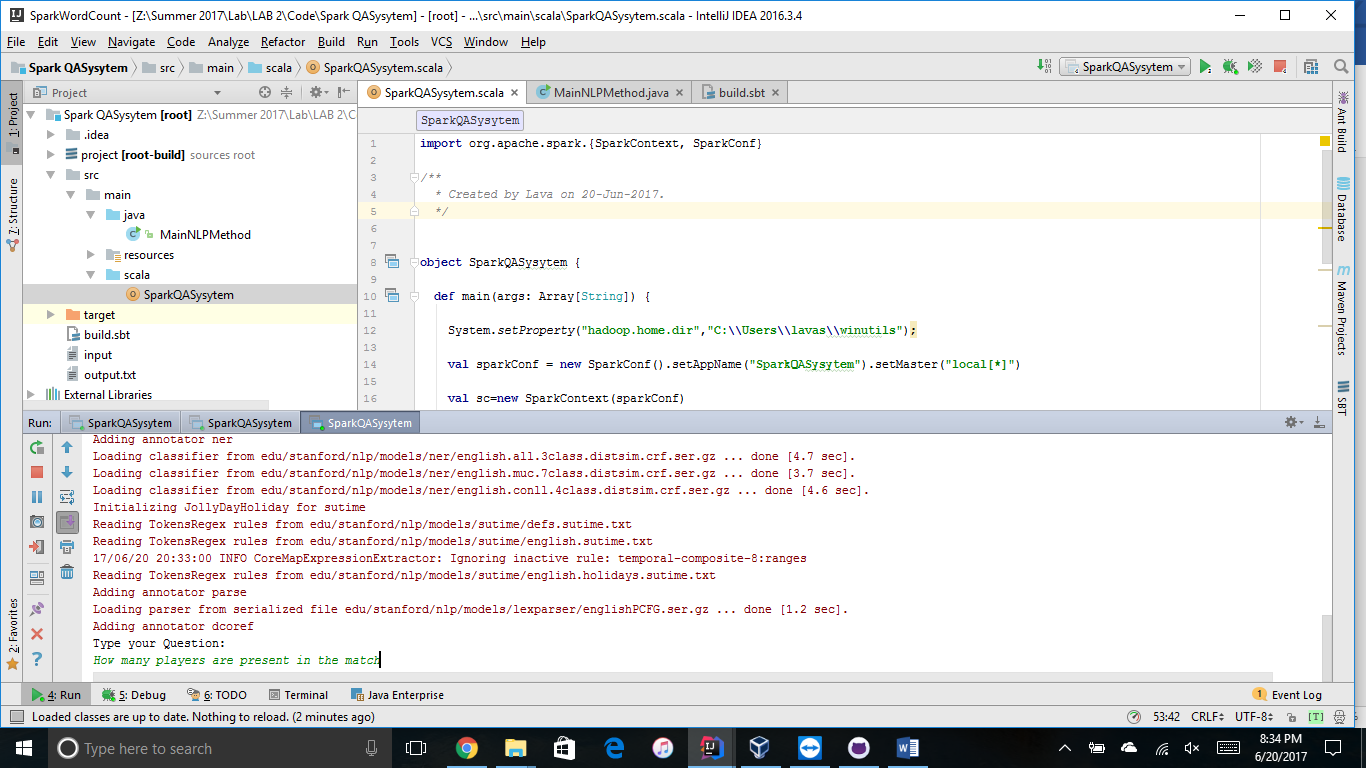
**Input Question2:**



**Response:**



**Input Question3:**



**Response:**

