

Assignment-2

1) Explain the differences between RDD and a traditional Relational Database System.

Consider principles like architecture, data access/retrieval/storage and key differences.

Architecture

RDD(Resilient Distributed Dataset) is an in memory data structure used by Spark. It is an immutable data structure.


Traditional databases are mutable as it needs to insert new records into it.

Data access/Retrieval/Storage

Spark has loaded data in memory in a specific structure and that structure is called RDD. Once your spark job stops, there is no RDD existence.

Database on the other hand are storage systems. You can store your data and query that later.

In short, Spark can load data from a file system or database and create an RDD to process the data whereas traditional databases are used to store the data.

2) Using pyspark create a word count application of all the words of the file [assignment_2_datafile.txt](#)  (located in the files/datasets tab). Avoid counting trivial words such as vowels and pronouns.

Hint: A simple technique to filter out trivial words would be to filter words having length less than 3.

```
((base) Sumeets-Macbook-Air:data sumeetmishra$ python3 word_count.py
19/11/16 18:01:33 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
(929, 'the')
(680, 'and')
(361, 'you')
(358, 'Ham.')
(284, 'his')
(259, 'And')
(259, 'not')
(225, 'that')
(224, 'with')
(217, 'your')
(203, 'this')
(162, 'for')
(160, 'have')
(152, 'The')
(139, 'but')
(139, 'will')
(121, 'him')
(118, 'That')
(118, 'are')
(116, 'King.')
(115, 'But')
(108, 'Hor.')
(107, 'our')
(103, 'shall')
(100, 'what')
(90, 'What')
(86, 'Pol.')
(83, 'from')
```

This is just the one screenshot of my results. The result needs to be captured in multiple screenshots. Please let me know if it is needed.

One key difference I observed is lowercase letters and uppercase letters are counted as 2 different words for example 'what' and 'What'. I could convert all the words to lowercase and take count again if that is what is asked in the exercise.