|  |  |
| --- | --- |
| **Practicum Case** |  |
| COMP6579  Big Data Processing |
| **Computer Science** | **E201-COMP6579-DD01-06** |
| ***Valid on*** *Even Semester Year 2019/2020* | **Revision 00** |

## Learning Outcomes

* Understand Big Data Analytics and Visualizations

## Topic

* Session 06 – Introduction to Spark

## Subtopics

* RDD Introduction
* RDD Operation
* Persistent Data
* Shared Variable in Spark

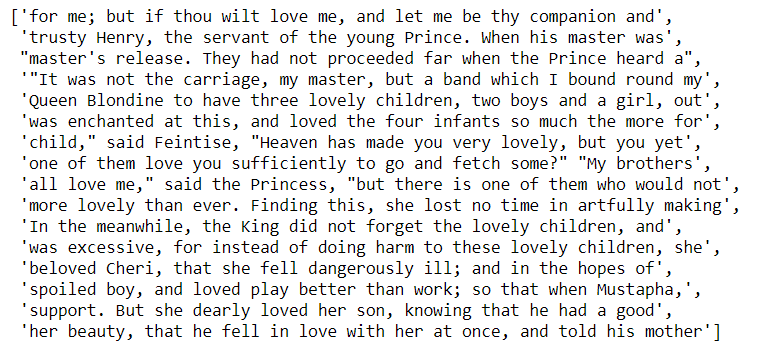
## Soal

*Case*

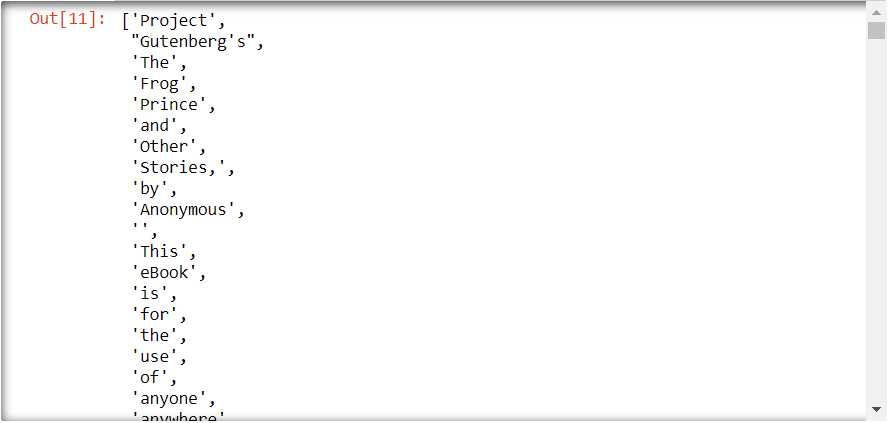
**Learning Spark Programming**

Budi is a fresh graduate at Blujek University and he want to try his luck applying a job as data scientist at Blujek Data Technologies as a Data Scientist. To prepare for the upcoming coding interview, he needs to study about related topics that essential for a Data Scientist. Of course, one of them is **Spark**. After researching for a week, he then asked his friend, Meong, whose an expert in Spark Programming to help him test hist knowledge about Spark. Below is the list of question Budi got from Meong:

1. **Read** a file named “The Frog Prince.txt” and **count** how manylines the file has
2. **Count** how many times the “Prince” word occurs in the file
3. **Filter** any lines contains “master” or “love” in the file and **print** each of them



1. Extract all **unique** words in the file and display it in **single** array



1. Using **accumulators**, **count** how many **blank lines** available in the file

Reference:

- <http://www.gutenberg.org/>

**Please ask your teaching assistant if there are any related questions.**