FIT5202: Data Processing for Big Data - Assignment 1 Marking Rubric

Part A	Excellent	Very Good	Good	Satisfactory
Step 01: Import pyspark and initialize Spark **Step 02: Create Resilient Distributed Datasets (RDDs)**	Importing and initializing has been done successfully according to the specification. RDDs have been created and the question has been answered correctly.	Importing and initializing has been done with some minor mistakes. RDDs have been created and the question has been answered correctly.	Importing and initializing has been done with some major mistakes. RDDs have been created and the question has been answered correctly.	Importing and initializing has been done with some major mistakes. RDDs have been created and but the question has not been answered.
Step 03: Cleaning/Manipulating text.	A function has been written that performs all the given operations in the specification. Displays the content of the RDDs after the function has been applied.	A function has been written that performs most of the given operations in the specification. Displays the content of the RDDs after the function has been applied.	A function has been written that performs some of the given operations in the specification. Displays the content of the RDDs after the function has been applied.	A function has been written that performs some of the given operations in the specification. Does not display the content of the RDDs after the function has been applied.
Step 04: Transforming the Data	All the correct transformations has been applied and the top 20 records have been displayed.	Some of the correct transformations has been applied and the top 20 records have been displayed.	Some of the correct transformations has been applied but the top 20 records have not been displayed.	No correct transformations has been applied and the top 20 records have not been displayed.
Step 05: Removing Stop Words	Used NLTK package to remove the stop words. Have answered the question successfully.	Used some other package to remove the stop words. Have answered the question successfully.	Used some other package to remove the stop words. Have answered the question incorrectly.	No stop words removed. Have answered the question incorrectly.
Step 06: Find the average occurrences of a word	Performed the calculation correctly and found the average occurances of a word.	Performed the calculation correctly and did not find the average occurances of a word.	Performed the calculation incorrectly and found the average occurances of a word.	Performed the calculation incorrectly and did not find the average occurances of a word.
Step 07: Exploratory data analysis	Uses matplotlib to analyse the distribution. The visualizations shows the clear understanding of factors such as datatype, scale, clarity, accuracy and efficiency.	Uses matplotlib to analyse the distribution. The visualizations does not show the clear understanding of factors such as datatype, scale, clarity, accuracy and efficiency.	Uses other graphical libraries to analyse the distribution. The visualizations shows the clear understanding of factors such as datatype, scale, clarity, accuracy and efficiency.	Uses other graphical libraries to analyse the distribution. The visualizations does not show the clear understanding of factors such as datatype, scale, clarity, accuracy and efficiency.
Part B				
Step 01: Import pyspark and initialize Spark **Step 02: Create Dataframe**	Importing and initializing has been done successfully according to the specification. Dataframe has been created and the question has been answered correctly.	Importing and initializing has been done with some minor mistakes. Dataframe has been created and the question has been answered correctly.	Importing and initializing has been done with some major mistakes. Dataframe has been created and the question has been answered correctly.	Importing and initializing has been done with some major mistakes. Dataframe has been created and but the question has not been answered.
Step 03: Write to Database **Step 04: Read from Database**	The data has been inserted into the MongoDB database with overwrite mode.	The data has been inserted into the MongoDB database without overwrite mode.	The inserting data into the MongoDB database has some minor issues and work after minor change.	The inserting data into the MongoDB database has some major issues but can work after the change.
Step 05: Calculate the statistics of numeric and string columns	All the statistical measures have been correctly calculated and the correct reasoning has been provided.	Most of the statistical measures have been correctly calculated and the correct reasoning has been provided.	Few of the statistical measures have been correctly calculated and the correct reasoning has been provided.	Few of the statistical measures have been calculated, some of them are wrong and no reasoning has been provided.
Step 06: Change the data type of a column	A user defined function (UDF) is used to convert string datetime into date.	String datetime is converted into date using inline functions.	String datetime is converted into date only using inline functions.	String datetime is not properly converted into date.
Step 07: Preliminary data analysis	All the analysis are correct.	Most of the analysis are correct.	Some of the analysis are correct.	Analysis is done but none of them are correct.
Step 08: Exploratory data analysis	Uses matplotlib to analyse the distribution. The visualizations shows the clear understanding of factors such as datatype, scale, clarity, accuracy and efficiency.	Uses matplotlib to analyse the distribution. The visualizations does not show the clear understanding of factors such as datatype, scale, clarity, accuracy and efficiency.	Uses other graphical libraries to analyse the distribution. The visualizations shows the clear understanding of factors such as datatype, scale, clarity, accuracy and efficiency.	Uses other graphical libraries to analyse the distribution. The visualizations does not show the clear understanding of factors such as datatype, scale, clarity, accuracy and efficiency.