**Spark Questions**

**Data Download:** <https://github.com/databricks/Spark-The-Definitive-Guide/tree/master/data/flight-data>

**Questions:**

1. Read data with all suitable options like header and others? Also, try to read by different command.

|  |
| --- |
|  |
| flightData2015 = spark\ |
|  | .read\ |
|  | .option("inferSchema", "true")\ |
|  | .option("header", "true")\ |
|  | .csv("/data/flight-data/csv/2015-summary.csv") |
|  |  |

|  |
| --- |
| val staticDataFrame = spark.read.format("csv") |
|  | .option("header", "true") |
|  | .option("inferSchema", "true") |
|  | .load("/data/retail-data/by-day/\*.csv") |

1. Sort the data in descending order of count and store top 20 rows in a new dataframe?

|  |
| --- |
|  |
| val maxSql = spark.sql(""" |
|  | SELECT DEST\_COUNTRY\_NAME, sum(count) as destination\_total |
|  | FROM flight\_data\_2015 |
|  | GROUP BY DEST\_COUNTRY\_NAME |
|  | ORDER BY sum(count) DESC |
|  | LIMIT 5 |
|  | """) |
|  |  |

1. How can we create a temporary view of the dataframe?

|  |
| --- |
|  |
| flightData2015 = spark\ |
|  | .read\ |
|  | .option("inferSchema", "true")\ |
|  | .option("header", "true")\ |
|  | .csv("/data/flight-data/csv/2015-summary.csv") |
|  |  |
|  | # COMMAND ---------- |
|  |  |
|  | flightData2015.createOrReplaceTempView("flight\_data\_2015") |
|  |  |

1. Register dataframe as a table?
2. Count of maximum destination country and sort them. Compare spark SQL with Dataframe.

<https://github.com/databricks/Spark-The-Definitive-Guide/blob/master/code/A_Gentle_Introduction_to_Spark-Chapter_2_A_Gentle_Introduction_to_Spark.scala>

1. Find maximum number of flights to and from any given location? Perform the task using Python and SQL both.

|  |
| --- |
|  |
| val maxSql = spark.sql(""" |
|  | SELECT DEST\_COUNTRY\_NAME, sum(count) as destination\_total |
|  | FROM flight\_data\_2015 |
|  | GROUP BY DEST\_COUNTRY\_NAME |
|  | ORDER BY sum(count) DESC |
|  | LIMIT 5 |
|  | """) |
|  |  |
|  | maxSql.show() |
|  |  |

<https://github.com/databricks/Spark-The-Definitive-Guide/blob/master/code/A_Gentle_Introduction_to_Spark-Chapter_2_A_Gentle_Introduction_to_Spark.scala>

1. Find out, top 5 destination countries in the data using both Python and SQL.

**Text Data Analysis:**

Download public data available:

**Question:**

1. Convert paragraph into the lines of a document using space.

<https://spark.apache.org/docs/2.2.0/ml-features.html>

1. Remove the stop words.
2. Group the words in above rdd based on the first 3 characters of each words.
3. Calculate the word frequency in the corpus. // Task1 complete

Perform some other transformations:

1. Create sample from complete data.
2. Union (create 2 sample and try doing union transformation)
3. Join (Using key and without key)
4. Distinct in rdd

**\*Please try to use as many transformation as you can.**