Building a Search Engine using PySpark

Ankit Dhall

PART I: Data Exploration

Reading sample file "shakespeare_small.json" directly from the url:
http://elmokhtari.com/downloads/ds8003/shakespeare_small.json to a dataframe **df1**.
Showing the dataframe content using .show()

Code:

```
import json
import requests
data = requests.get("http://elmokhtari.com/downloads/ds8003/shakespeare_small.json")
df1 = sqlContext.createDataFrame([json.loads(line) for line in data.iter_lines()])
df1.show()
```

Line 1 & 2 - # Import json library to handle json data and import requests library to fetch data directly from the url.
Line 3 - # Gets data from the givel url and stores in 'data'
Line 4 - # Takes each line from the json data and creates a dataframe
Line 5 - # Shows top 20 rows of the dataframe 'df1'

maria_dev@sandbox-hdp:~/ankit

```
[maria_dev@sandbox-hdp ankit]$ pyspark
SPARK_MAJOR_VERSION is set to 2, using Spark2
Python 2.7.5 (default, Aug  7 2019, 00:51:29)
[GCC 4.8.5 20150623 (Red Hat 4.8.5-39)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.3.0.2.6.5.0-292
      /_/

Using Python version 2.7.5 (default, Aug  7 2019 00:51:29)
SparkSession available as 'spark'.
>>> import json
>>> import requests
>>> data = requests.get("http://elmokhtari.com/downloads/ds8003/shakespeare_small.json")
>>> df1 = sqlContext.createDataFrame([json.loads(line) for line in data.iter_lines()])
/usr/hdp/current/spark2-client/python/pyspark/sql/session.py:346: UserWarning: inferring schema from dict is deprecated,please use pyspark.sql.Row instead
  warnings.warn("inferring schema from dict is deprecated,"
>>> df1.show()
+---+-------+-----------+---------+-------------+--------------+--------------------+----+
|_id|line_id|line_number|play_name|      speaker|speech_number|          text_entry|type|
+---+-------+-----------+---------+-------------+--------------+--------------------+----+
|  3|      4|      1.1.1| Henry IV|KING HENRY IV|             1|So shaken as we a...|line|
|  4|      5|      1.1.2| Henry IV|KING HENRY IV|             1|Find we a time fo...|line|
|  5|      6|      1.1.3| Henry IV|KING HENRY IV|             1|And breathe short...|line|
|  6|      7|      1.1.4| Henry IV|KING HENRY IV|             1|To be commenced i...|line|
|  7|      8|      1.1.5| Henry IV|KING HENRY IV|             1|No more the thirs...|line|
|  8|      9|      1.1.6| Henry IV|KING HENRY IV|             1|Shall daub her li...|line|
|  9|     10|      1.1.7| Henry IV|KING HENRY IV|             1|Nor more shall tr...|line|
| 10|     11|      1.1.8| Henry IV|KING HENRY IV|             1|Nor bruise her fl...|line|
| 11|     12|      1.1.9| Henry IV|KING HENRY IV|             1|Of hostile paces:...|line|
| 12|     13|     1.1.10| Henry IV|KING HENRY IV|             1|Which, like the m...|line|
| 13|     14|     1.1.11| Henry IV|KING HENRY IV|             1|All of one nature...|line|
| 14|     15|     1.1.12| Henry IV|KING HENRY IV|             1|Did lately meet i...|line|
| 15|     16|     1.1.13| Henry IV|KING HENRY IV|             1|And furious close...|line|
| 16|     17|     1.1.14| Henry IV|KING HENRY IV|             1|Shall now, in mut...|line|
| 17|     18|     1.1.15| Henry IV|KING HENRY IV|             1|March all one way...|line|
| 18|     19|     1.1.16| Henry IV|KING HENRY IV|             1|Against acquainta...|line|
| 19|     20|     1.1.17| Henry IV|KING HENRY IV|             1|The edge of war, ...|line|
| 20|     21|     1.1.18| Henry IV|KING HENRY IV|             1|No more shall cut...|line|
| 21|     22|     1.1.19| Henry IV|KING HENRY IV|             1|As far as to the ...|line|
| 22|     23|     1.1.20| Henry IV|KING HENRY IV|             1|Whose soldier now...|line|
+---+-------+-----------+---------+-------------+--------------+--------------------+----+
only showing top 20 rows
```

Show the rows count.

Code:

df1.count() #Gives the row count of the dataframe

Output:

Rows Count = 68

maria_dev@sandbox-hdp:~/ankit

```
[maria_dev@sandbox-hdp ankit]$ pyspark
SPARK_MAJOR_VERSION is set to 2, using Spark2
Python 2.7.5 (default, Aug  7 2019, 00:51:29)
[GCC 4.8.5 20150623 (Red Hat 4.8.5-39)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.3.0.2.6.5.0-292
      /_/

Using Python version 2.7.5 (default, Aug  7 2019 00:51:29)
SparkSession available as 'spark'.
>>> import json
>>> import requests
>>> data = requests.get("http://elmokhtari.com/downloads/ds8003/shakespeare_small.json")
>>> df1 = sqlContext.createDataFrame([json.loads(line) for line in data.iter_lines()])
/usr/hdp/current/spark2-client/python/pyspark/sql/session.py:346: UserWarning: inferring schema from dict is deprecated,please use pyspark.sql.Row instead
  warnings.warn("inferring schema from dict is deprecated,"
>>> df1.show()
+---+-------+-----------+---------+-------------+-------------+--------------------+----+
|_id|line_id|line_number|play_name|      speaker|speech_number|          text_entry|type|
+---+-------+-----------+---------+-------------+-------------+--------------------+----+
|  3|      4|      1.1.1| Henry IV|KING HENRY IV|            1|So shaken as we a...|line|
|  4|      5|      1.1.2| Henry IV|KING HENRY IV|            1|Find we a time fo...|line|
|  5|      6|      1.1.3| Henry IV|KING HENRY IV|            1|And breathe short...|line|
|  6|      7|      1.1.4| Henry IV|KING HENRY IV|            1|To be commenced i...|line|
|  7|      8|      1.1.5| Henry IV|KING HENRY IV|            1|No more the thirs...|line|
|  8|      9|      1.1.6| Henry IV|KING HENRY IV|            1|Shall daub her li...|line|
|  9|     10|      1.1.7| Henry IV|KING HENRY IV|            1|Nor more shall tr...|line|
| 10|     11|      1.1.8| Henry IV|KING HENRY IV|            1|Nor bruise her fl...|line|
| 11|     12|      1.1.9| Henry IV|KING HENRY IV|            1|Of hostile paces:...|line|
| 12|     13|     1.1.10| Henry IV|KING HENRY IV|            1|Which, like the m...|line|
| 13|     14|     1.1.11| Henry IV|KING HENRY IV|            1|All of one nature...|line|
| 14|     15|     1.1.12| Henry IV|KING HENRY IV|            1|Did lately meet i...|line|
| 15|     16|     1.1.13| Henry IV|KING HENRY IV|            1|And furious close...|line|
| 16|     17|     1.1.14| Henry IV|KING HENRY IV|            1|Shall now, in mut...|line|
| 17|     18|     1.1.15| Henry IV|KING HENRY IV|            1|March all one way...|line|
| 18|     19|     1.1.16| Henry IV|KING HENRY IV|            1|Against acquainta...|line|
| 19|     20|     1.1.17| Henry IV|KING HENRY IV|            1|The edge of war, ...|line|
| 20|     21|     1.1.18| Henry IV|KING HENRY IV|            1|No more shall cut...|line|
| 21|     22|     1.1.19| Henry IV|KING HENRY IV|            1|As far as to the ...|line|
| 22|     23|     1.1.20| Henry IV|KING HENRY IV|            1|Whose soldier now...|line|
+---+-------+-----------+---------+-------------+-------------+--------------------+----+
only showing top 20 rows

>>> df1.count()
68
>>>
```

Uploading the file: shakespeare_full.json to the linux machine and loading its content to a dataframe df2. Showing the first 20 rows of the dataframe using .show()

Code:

#Put file in Hadoop FS

hadoop fs -mkdir /user/maria_dev/ankit

hadoop fs -put shakespeare_full.json /user/maria_dev/ankit

hadoop fs -ls /user/maria_dev/ankit

maria_dev@sandbox-hdp:~/ankit

```
[maria_dev@sandbox-hdp ankit]$ ls
[maria_dev@sandbox-hdp ankit]$ ls
shakespeare_full.json
[maria_dev@sandbox-hdp ankit]$ hadoop fs -mkdir /user/maria_dev/ankit
[maria_dev@sandbox-hdp ankit]$ hadoop fs -put shakespeare_full.json /user/maria_dev/ankit
[maria_dev@sandbox-hdp ankit]$ hadoop fs -ls /user/maria_dev/ankit
Found 1 items
-rw-r--r--   1 maria_dev hdfs   21317209 2019-12-07 21:00 /user/maria_dev/ankit/shakespeare_full.json
[maria_dev@sandbox-hdp ankit]$
```

from pyspark.sql import SQLContext

sqlContext = SQLContext(sc) #Defining the SQLContext

df2 = sqlContext.read.json("/user/maria_dev/ankit/shakespeare_full.json")

#Loads the json file into a dataframe ‘df2’

df2.show() #Shows the top 20 rows of the dataframe ‘df2’

maria_dev@sandbox-hdp:~/ankit

```
[maria_dev@sandbox-hdp ankit]$ pyspark
SPARK_MAJOR_VERSION is set to 2, using Spark2
Python 2.7.5 (default, Aug  7 2019, 00:51:29)
[GCC 4.8.5 20150623 (Red Hat 4.8.5-39)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.3.0.2.6.5.0-292
      /_/

Using Python version 2.7.5 (default, Aug  7 2019 00:51:29)
SparkSession available as 'spark'.
>>> from pyspark.sql import SQLContext
>>> sqlContext = SQLContext(sc)
>>> df2 = sqlContext.read.json("/user/maria_dev/ankit/shakespeare_full.json")
>>> df2.show()
+---+-------+-----------+---------+-------------+-------------+--------------------+-----+
|_id|line_id|line_number|play_name|      speaker|speech_number|          text_entry| type|
+---+-------+-----------+---------+-------------+-------------+--------------------+-----+
|  0|      1|           | Henry IV|             |         null|               ACT I|  act|
|  1|      2|           | Henry IV|             |         null|SCENE I. London. ...|scene|
|  2|      3|           | Henry IV|             |         null|Enter KING HENRY,...| line|
|  3|      4|      1.1.1| Henry IV|KING HENRY IV|            1|So shaken as we a...| line|
|  4|      5|      1.1.2| Henry IV|KING HENRY IV|            1|Find we a time fo...| line|
|  5|      6|      1.1.3| Henry IV|KING HENRY IV|            1|And breathe short...| line|
|  6|      7|      1.1.4| Henry IV|KING HENRY IV|            1|To be commenced i...| line|
|  7|      8|      1.1.5| Henry IV|KING HENRY IV|            1|No more the thirs...| line|
|  8|      9|      1.1.6| Henry IV|KING HENRY IV|            1|Shall daub her li...| line|
|  9|     10|      1.1.7| Henry IV|KING HENRY IV|            1|Nor more shall tr...| line|
| 10|     11|      1.1.8| Henry IV|KING HENRY IV|            1|Nor bruise her fl...| line|
| 11|     12|      1.1.9| Henry IV|KING HENRY IV|            1|Of hostile paces:...| line|
| 12|     13|     1.1.10| Henry IV|KING HENRY IV|            1|Which, like the m...| line|
| 13|     14|     1.1.11| Henry IV|KING HENRY IV|            1|All of one nature...| line|
| 14|     15|     1.1.12| Henry IV|KING HENRY IV|            1|Did lately meet i...| line|
| 15|     16|     1.1.13| Henry IV|KING HENRY IV|            1|And furious close...| line|
| 16|     17|     1.1.14| Henry IV|KING HENRY IV|            1|Shall now, in mut...| line|
| 17|     18|     1.1.15| Henry IV|KING HENRY IV|            1|March all one way...| line|
| 18|     19|     1.1.16| Henry IV|KING HENRY IV|            1|Against acquainta...| line|
| 19|     20|     1.1.17| Henry IV|KING HENRY IV|            1|The edge of war, ...| line|
+---+-------+-----------+---------+-------------+-------------+--------------------+-----+
only showing top 20 rows

>>> 
```

Showing the count of entries grouped by “speaker” on df2.

Code:

df2.groupBy("speaker").count().show()

Groups ‘df2’ on the field ‘speaker’ & aggregates using count() and shows the top 20 rows

maria_dev@sandbox-hdp:~/ankit

```
   /__ / .__/\_,_/_/ /_/\_\   version 2.3.0.2.6.5.0-292
      /_/

Using Python version 2.7.5 (default, Aug  7 2019 00:51:29)
SparkSession available as 'spark'.
>>> from pyspark.sql import SQLContext
>>> sqlContext = SQLContext(sc)
>>> df2 = sqlContext.read.json("/user/maria_dev/ankit/shakespeare_full.json")
>>> df2.show()
+---+-------+-----------+---------+-------------+-------------+--------------------+-----+
|_id|line_id|line_number|play_name|      speaker|speech_number|          text_entry| type|
+---+-------+-----------+---------+-------------+-------------+--------------------+-----+
|  0|      1|           | Henry IV|             |         null|               ACT I|  act|
|  1|      2|           | Henry IV|             |         null|SCENE I. London. ...|scene|
|  2|      3|           | Henry IV|             |         null|Enter KING HENRY,...| line|
|  3|      4|      1.1.1| Henry IV|KING HENRY IV|            1|So shaken as we a...| line|
|  4|      5|      1.1.2| Henry IV|KING HENRY IV|            1|Find we a time fo...| line|
|  5|      6|      1.1.3| Henry IV|KING HENRY IV|            1|And breathe short...| line|
|  6|      7|      1.1.4| Henry IV|KING HENRY IV|            1|To be commenced i...| line|
|  7|      8|      1.1.5| Henry IV|KING HENRY IV|            1|No more the thirs...| line|
|  8|      9|      1.1.6| Henry IV|KING HENRY IV|            1|Shall daub her li...| line|
|  9|     10|      1.1.7| Henry IV|KING HENRY IV|            1|Nor more shall tr...| line|
| 10|     11|      1.1.8| Henry IV|KING HENRY IV|            1|Nor bruise her fl...| line|
| 11|     12|      1.1.9| Henry IV|KING HENRY IV|            1|Of hostile paces:...| line|
| 12|     13|     1.1.10| Henry IV|KING HENRY IV|            1|Which, like the m...| line|
| 13|     14|     1.1.11| Henry IV|KING HENRY IV|            1|All of one nature...| line|
| 14|     15|     1.1.12| Henry IV|KING HENRY IV|            1|Did lately meet i...| line|
| 15|     16|     1.1.13| Henry IV|KING HENRY IV|            1|And furious close...| line|
| 16|     17|     1.1.14| Henry IV|KING HENRY IV|            1|Shall now, in mut...| line|
| 17|     18|     1.1.15| Henry IV|KING HENRY IV|            1|March all one way...| line|
| 18|     19|     1.1.16| Henry IV|KING HENRY IV|            1|Against acquainta...| line|
| 19|     20|     1.1.17| Henry IV|KING HENRY IV|            1|The edge of war, ...| line|
+---+-------+-----------+---------+-------------+-------------+--------------------+-----+
only showing top 20 rows

>>> df2.groupBy("speaker").count().show()
+-----------------+-----+
|          speaker|count|
+-----------------+-----+
|       EUPHRONIUS|   16|
|Third Conspirator|   12|
|            PETER|   63|
|  First Gentleman|  284|
|           AEGEON|  150|
|        DONALBAIN|   10|
|        LYCHORIDA|   11|
|          QUINTUS|   30|
|           AENEAS|  153|
|           Porter|   97|
|          RUTLAND|   26|
|              NYM|   78|
|   LORD FITZWATER|   27|
|         CARDINAL|  120|
|       Attendants|    2|
|       ANTIPHOLUS|    6|
|    Third Servant|   31|
|        ANNE PAGE|   31|
|        Moonshine|    6|
|       SIR ANDREW|  155|
+-----------------+-----+
only showing top 20 rows

>>> 
```

Usinging spark.sql to show all entries where line_number starts with “1.1.” and text_entry contains the word “sometimes”.

Code:

df2.createOrReplaceTempView("text_data")

sqlDF = spark.sql("SELECT _id, speaker, line_number, text_entry FROM text_data WHERE line_number LIKE '1.1.%' and text_entry LIKE '%sometimes%'")
sqlDF.show()

Line 1 - # Creates a temporary view for ‘df2’ by the name of ‘text_data’
Line 2 - # Using Spark SQL, we select the data based on the required conditions
Line 3 - # Shows the top 20 rows of the filtered data

maria_dev@sandbox-hdp:~/ankit

```
>>> df2.createOrReplaceTempView("text_data")
>>> sqlDF = spark.sql("SELECT _id, speaker, line_number, text_entry FROM text_data WHERE line_number LIKE '1.1.%' and text_entry LIKE '%sometimes%'")
>>> sqlDF.show()
+-----+---------+-----------+--------------------+
|  _id|  speaker|line_number|          text_entry|
+-----+---------+-----------+--------------------+
|18634|    PHILO|     1.1.63|Sir, sometimes, w...|
|32496|  HORATIO|     1.1.59|Did sometimes mar...|
|61534| BASSANIO|    1.1.166|Of wondrous virtu...|
|64418|  SLENDER|    1.1.240|A justice of peac...|
|75845|ANTIOCHUS|     1.1.34|Yon sometimes fam...|
+-----+---------+-----------+--------------------+
```

Generating a list with the number of characters in every text entry where the speaker is “DONALBAIN”

Code:

```
import pyspark.sql.functions as F
temp = df2.where(df2['speaker'] == "DONALBAIN")
res = temp.select(temp['_id'], temp['text_entry'])
res2 = res.withColumn('length', F.length('text_entry'))
final_list = [int(row['length']) for row in res2.collect()]
print final_list
```

Line 1 - # Imports the SQL Functions Library
Line 2 - # Filters dataframe where speaker is ‘DONALBAIN’
Line 3 - # Choosing only the relevant columns from the data
Line 4 - # Creating a new column with the number of characters in ‘text_entry’ field
Line 5 & 6 - # Creating a list of all values of the new ‘length’ columns and printing the list

Output:
[14, 47, 15, 45, 11, 28, 36, 43, 49, 18]

maria_dev@sandbox-hdp:~/ankit

```
>>> import pyspark.sql.functions as F
>>> temp = df2.where(df2['speaker'] == "DONALBAIN")
>>> res = temp.select(temp['_id'], temp['text_entry'])
>>> res2 = res.withColumn('length', F.length('text_entry'))
>>> final_list = [int(row['length']) for row in res2.collect()]
>>> final_list
[14, 47, 15, 45, 11, 28, 36, 43, 49, 18]
```

Considering all text entries of the speaker “DONALBAIN”.
Generating a list of pairs (key, value) where **key** is the _id of the text entry and **value** is the number of words in this text entry.

Code:

```
import pyspark.sql.functions as F
temp = df2.where(df2['speaker'] == "DONALBAIN")
res = temp.select(temp['_id'], temp['text_entry'])
res = res.withColumn("text_entry", F.lower(F.col("text_entry")))
res = res.withColumn("text_entry", F.regexp_replace(F.col("text_entry"), '[^\sa-zA-Z0-9]', ''))
res2 = res.withColumn('wordCount', F.size(F.split(F.col('text_entry'), ' ')))
final_list = [(int(row['_id']), int(row['wordCount'])) for row in res2.collect()]
print final_list
```

Line 1 - # Imports the SQL Functions Library
Line 2 - # Filters dataframe where speaker is ‘DONALBAIN’
Line 3 - # Choosing only the relevant columns from the data
Line 4 & 5 - # Converting the ‘text_entry’ column to lowercase and removing any punctuations
Line 5 - # Splitting the ‘text_entry’ column data on spaces to count number of words and storing it in a new column called ‘wordCount’
Line 6 & 7- # Converting the data into list of key & value pairs and printing the list.

Output:
[(56668, 3), (56698, 9), (56699, 3), (56700, 9), (56701, 3), (56702, 6), (56723, 6), (56724, 9), (56725, 9), (56726, 3)]

maria_dev@sandbox-hdp:~/ankit

```
>>>
>>> import pyspark.sql.functions as F
>>> temp = df2.where(df2['speaker'] == "DONALBAIN")
>>> res = temp.select(temp['_id'], temp['text_entry'])
>>> res = res.withColumn("text_entry", F.lower(F.col("text_entry")))
>>> res = res.withColumn("text_entry", F.regexp_replace(F.col("text_entry"), '[^\sa-zA-Z0-9]', ''))
>>> res2 = res.withColumn('wordCount', F.size(F.split(F.col('text_entry'), ' ')))
>>> final_list = [(int(row['_id']), int(row['wordCount'])) for row in res2.collect()]
>>> print final_list
[(56668, 3), (56698, 9), (56699, 3), (56700, 9), (56701, 3), (56702, 6), (56723, 6), (56724, 9), (56725, 9), (56726, 3)]
>>>
>>>
```

PART II: Building a search engine with PySpark

[Building Index] Compute TFIDF scores for all words in all text entries and build an inverted index. This index will be stored in the dataframe **tokensWithTfIdf** containing the following columns: (token, _id, tf, df, idf, tf_idf).
token is any word in text entries, **_id**: text entry id, **(tf,idf,tf_idf)** scores of the pair (token,_id).
Before creating the index, text entries must be converted to lower case and the punctuation signs removed.
Cache the dataframe **tokensWithTfIdf** in memory for further usage.

Code:

#Code to read the json data into dataframe 'df2'

```
from pyspark.sql import SQLContext
sqlContext = SQLContext(sc)
df2 = sqlContext.read.json("/user/maria_dev/ankit/shakespeare_full.json")
df2.show()
```

#Calculating 'N' (Total Number of Documents in the dataframe 'df2')

```
N = df2.count()
N = float(N)
N
```

maria_dev@sandbox-hdp:~/ankit

```
>>> from pyspark.sql import SQLContext
>>> sqlContext = SQLContext(sc)
>>> df2 = sqlContext.read.json("/user/maria_dev/ankit/shakespeare_full.json")
>>> df2.show()
+---+-------+-----------+---------+-------------+-------------+--------------------+-----+
|_id|line_id|line_number|play_name|      speaker|speech_number|          text_entry| type|
+---+-------+-----------+---------+-------------+-------------+--------------------+-----+
|  0|      1|           | Henry IV|             |         null|               ACT I|  act|
|  1|      2|           | Henry IV|             |         null|SCENE I. London. ...|scene|
|  2|      3|           | Henry IV|             |         null|Enter KING HENRY,...| line|
|  3|      4|      1.1.1| Henry IV|KING HENRY IV|            1|So shaken as we a...| line|
|  4|      5|      1.1.2| Henry IV|KING HENRY IV|            1|Find we a time fo...| line|
|  5|      6|      1.1.3| Henry IV|KING HENRY IV|            1|And breathe short...| line|
|  6|      7|      1.1.4| Henry IV|KING HENRY IV|            1|To be commenced i...| line|
|  7|      8|      1.1.5| Henry IV|KING HENRY IV|            1|No more the thirs...| line|
|  8|      9|      1.1.6| Henry IV|KING HENRY IV|            1|Shall daub her li...| line|
|  9|     10|      1.1.7| Henry IV|KING HENRY IV|            1|Nor more shall tr...| line|
| 10|     11|      1.1.8| Henry IV|KING HENRY IV|            1|Nor bruise her fl...| line|
| 11|     12|      1.1.9| Henry IV|KING HENRY IV|            1|Of hostile paces:...| line|
| 12|     13|     1.1.10| Henry IV|KING HENRY IV|            1|Which, like the m...| line|
| 13|     14|     1.1.11| Henry IV|KING HENRY IV|            1|All of one nature...| line|
| 14|     15|     1.1.12| Henry IV|KING HENRY IV|            1|Did lately meet i...| line|
| 15|     16|     1.1.13| Henry IV|KING HENRY IV|            1|And furious close...| line|
| 16|     17|     1.1.14| Henry IV|KING HENRY IV|            1|Shall now, in mut...| line|
| 17|     18|     1.1.15| Henry IV|KING HENRY IV|            1|March all one way...| line|
| 18|     19|     1.1.16| Henry IV|KING HENRY IV|            1|Against acquainta...| line|
| 19|     20|     1.1.17| Henry IV|KING HENRY IV|            1|The edge of war, ...| line|
+---+-------+-----------+---------+-------------+-------------+--------------------+-----+
only showing top 20 rows

>>> N = df2.count()
>>> N = float(N)
>>> N
111396.0
```

#Convert text_entry to Lower Case

```
from pyspark.sql.functions import lower, col
columnName="text_entry"
df2 = df2.withColumn(columnName, lower(col(columnName)))
df2.show()
```

maria_dev@sandbox-hdp:~/ankit

```
>>> from pyspark.sql.functions import lower, col
>>> columnName="text_entry"
>>> df2 = df2.withColumn(columnName, lower(col(columnName)))
>>> df2.show()
+---+-------+-----------+---------+-------------+-------------+--------------------+-----+
|_id|line_id|line_number|play_name|      speaker|speech_number|          text_entry| type|
+---+-------+-----------+---------+-------------+-------------+--------------------+-----+
|  0|      1|           | Henry IV|             |         null|               act i|  act|
|  1|      2|           | Henry IV|             |         null|scene i. london. ...|scene|
|  2|      3|           | Henry IV|             |         null|enter king henry,...| line|
|  3|      4|      1.1.1| Henry IV|KING HENRY IV|            1|so shaken as we a...| line|
|  4|      5|      1.1.2| Henry IV|KING HENRY IV|            1|find we a time fo...| line|
|  5|      6|      1.1.3| Henry IV|KING HENRY IV|            1|and breathe short...| line|
|  6|      7|      1.1.4| Henry IV|KING HENRY IV|            1|to be commenced i...| line|
|  7|      8|      1.1.5| Henry IV|KING HENRY IV|            1|no more the thirs...| line|
|  8|      9|      1.1.6| Henry IV|KING HENRY IV|            1|shall daub her li...| line|
|  9|     10|      1.1.7| Henry IV|KING HENRY IV|            1|nor more shall tr...| line|
| 10|     11|      1.1.8| Henry IV|KING HENRY IV|            1|nor bruise her fl...| line|
| 11|     12|      1.1.9| Henry IV|KING HENRY IV|            1|of hostile paces:...| line|
| 12|     13|     1.1.10| Henry IV|KING HENRY IV|            1|which, like the m...| line|
| 13|     14|     1.1.11| Henry IV|KING HENRY IV|            1|all of one nature...| line|
| 14|     15|     1.1.12| Henry IV|KING HENRY IV|            1|did lately meet i...| line|
| 15|     16|     1.1.13| Henry IV|KING HENRY IV|            1|and furious close...| line|
| 16|     17|     1.1.14| Henry IV|KING HENRY IV|            1|shall now, in mut...| line|
| 17|     18|     1.1.15| Henry IV|KING HENRY IV|            1|march all one way...| line|
| 18|     19|     1.1.16| Henry IV|KING HENRY IV|            1|against acquainta...| line|
| 19|     20|     1.1.17| Henry IV|KING HENRY IV|            1|the edge of war, ...| line|
+---+-------+-----------+---------+-------------+-------------+--------------------+-----+
only showing top 20 rows
```

#Remove punctuations from text_entry

```
from pyspark.sql.functions import regexp_replace
columnName="text_entry"
df2 = df2.withColumn(columnName, regexp_replace(col(columnName), '[^\sa-zA-Z0-9]', ''))
df2.show()
```

maria_dev@sandbox-hdp:~/ankit

```
>>> from pyspark.sql.functions import regexp_replace
>>> columnName="text_entry"
>>> df2 = df2.withColumn(columnName, regexp_replace(col(columnName), '[^\sa-zA-Z0-9]', ''))
>>> df2.show()
+---+-------+-----------+---------+-------------+-------------+--------------------+-----+
|_id|line_id|line_number|play_name|      speaker|speech_number|          text_entry| type|
+---+-------+-----------+---------+-------------+-------------+--------------------+-----+
|  0|      1|           | Henry IV|             |         null|               act i|  act|
|  1|      2|           | Henry IV|             |         null|scene i london th...|scene|
|  2|      3|           | Henry IV|             |         null|enter king henry ...| line|
|  3|      4|      1.1.1| Henry IV|KING HENRY IV|            1|so shaken as we a...| line|
|  4|      5|      1.1.2| Henry IV|KING HENRY IV|            1|find we a time fo...| line|
|  5|      6|      1.1.3| Henry IV|KING HENRY IV|            1|and breathe short...| line|
|  6|      7|      1.1.4| Henry IV|KING HENRY IV|            1|to be commenced i...| line|
|  7|      8|      1.1.5| Henry IV|KING HENRY IV|            1|no more the thirs...| line|
|  8|      9|      1.1.6| Henry IV|KING HENRY IV|            1|shall daub her li...| line|
|  9|     10|      1.1.7| Henry IV|KING HENRY IV|            1|nor more shall tr...| line|
| 10|     11|      1.1.8| Henry IV|KING HENRY IV|            1|nor bruise her fl...| line|
| 11|     12|      1.1.9| Henry IV|KING HENRY IV|            1|of hostile paces ...| line|
| 12|     13|     1.1.10| Henry IV|KING HENRY IV|            1|which like the me...| line|
| 13|     14|     1.1.11| Henry IV|KING HENRY IV|            1|all of one nature...| line|
| 14|     15|     1.1.12| Henry IV|KING HENRY IV|            1|did lately meet i...| line|
| 15|     16|     1.1.13| Henry IV|KING HENRY IV|            1|and furious close...| line|
| 16|     17|     1.1.14| Henry IV|KING HENRY IV|            1|shall now in mutu...| line|
| 17|     18|     1.1.15| Henry IV|KING HENRY IV|            1|march all one way...| line|
| 18|     19|     1.1.16| Henry IV|KING HENRY IV|            1|against acquainta...| line|
| 19|     20|     1.1.17| Henry IV|KING HENRY IV|            1|the edge of war l...| line|
+---+-------+-----------+---------+-------------+-------------+--------------------+-----+
only showing top 20 rows
```

#Drop extra columns from df2

df2 = df2.drop('line_id', 'line_number', 'play_name', 'speaker', 'speech_number', 'type')
df2.show()

maria_dev@sandbox-hdp:~/ankit

```
>>> df2 = df2.drop('line_id', 'line_number', 'play_name', 'speaker', 'speech_number', 'type')
>>> df2.show()
+---+--------------------+
|_id|          text_entry|
+---+--------------------+
|  0|               act i|
|  1|scene i london th...|
|  2|enter king henry ...|
|  3|so shaken as we a...|
|  4|find we a time fo...|
|  5|and breathe short...|
|  6|to be commenced i...|
|  7|no more the thirs...|
|  8|shall daub her li...|
|  9|nor more shall tr...|
| 10|nor bruise her fl...|
| 11|of hostile paces ...|
| 12|which like the me...|
| 13|all of one nature...|
| 14|did lately meet i...|
| 15|and furious close...|
| 16|shall now in mutu...|
| 17|march all one way...|
| 18|against acquainta...|
| 19|the edge of war l...|
+---+--------------------+
only showing top 20 rows
```

#Split text_entry column into words by using the split function

```
from pyspark.sql.functions import split
df2 = df2.withColumn("text_entry", split("text_entry", " "))
df2.show()
```

maria_dev@sandbox-hdp:~/ankit

```
>>> from pyspark.sql.functions import split
>>> df2 = df2.withColumn("text_entry", split("text_entry", " "))
>>> df2.show()
+---+--------------------+
|_id|          text_entry|
+---+--------------------+
|  0|            [act, i]|
|  1|[scene, i, london...|
|  2|[enter, king, hen...|
|  3|[so, shaken, as, ...|
|  4|[find, we, a, tim...|
|  5|[and, breathe, sh...|
|  6|[to, be, commence...|
|  7|[no, more, the, t...|
|  8|[shall, daub, her...|
|  9|[nor, more, shall...|
| 10|[nor, bruise, her...|
| 11|[of, hostile, pac...|
| 12|[which, like, the...|
| 13|[all, of, one, na...|
| 14|[did, lately, mee...|
| 15|[and, furious, cl...|
| 16|[shall, now, in, ...|
| 17|[march, all, one,...|
| 18|[against, acquain...|
| 19|[the, edge, of, w...|
+---+--------------------+
only showing top 20 rows
```

#Explode eachtext_entry value into multiple rows to get _id with each word of text_entry

```
from pyspark.sql.functions import explode
df2 = df2.withColumn("token", explode(col("text_entry")))
df2.show()
```

maria_dev@sandbox-hdp:~/ankit

```
>>> from pyspark.sql.functions import explode
>>> df2 = df2.withColumn("token", explode(col("text_entry")))
>>> df2.show()
+---+--------------------+------------+
|_id|          text_entry|       token|
+---+--------------------+------------+
|  0|            [act, i]|         act|
|  0|            [act, i]|           i|
|  1|[scene, i, london...|       scene|
|  1|[scene, i, london...|           i|
|  1|[scene, i, london...|      london|
|  1|[scene, i, london...|         the|
|  1|[scene, i, london...|      palace|
|  2|[enter, king, hen...|       enter|
|  2|[enter, king, hen...|        king|
|  2|[enter, king, hen...|       henry|
|  2|[enter, king, hen...|        lord|
|  2|[enter, king, hen...|        john|
|  2|[enter, king, hen...|          of|
|  2|[enter, king, hen...|   lancaster|
|  2|[enter, king, hen...|         the|
|  2|[enter, king, hen...|        earl|
|  2|[enter, king, hen...|          of|
|  2|[enter, king, hen...|westmoreland|
|  2|[enter, king, hen...|         sir|
|  2|[enter, king, hen...|      walter|
+---+--------------------+------------+
only showing top 20 rows
```

#Calculating Term Frequency by grouping based on '_id' and 'token' and counting how many times each token occurs in each document

```
from pyspark.sql import functions as F
df_tf = df2.groupby("_id", "token").agg(F.count("text_entry").alias("tf"))
df_tf.show()
```

maria_dev@sandbox-hdp:~/ankit

```
>>> from pyspark.sql import functions as F
>>> df_tf = df2.groupby("_id", "token").agg(F.count("text_entry").alias("tf"))
>>> df_tf.show()
+---+-----------+---+
|_id|      token| tf|
+---+-----------+---+
| 55|        and|  1|
| 87|        see|  1|
|116|       upon|  1|
|188|        the|  1|
|190|        god|  1|
|191|       good|  1|
|201|          a|  1|
|239|      tarry|  1|
|245|        not|  1|
|273|     before|  1|
|275|       will|  1|
|282|        our|  1|
|307|    through|  1|
|361|      there|  1|
|401|     retold|  1|
|506|         in|  1|
|509|bolingbroke|  1|
|521|     cousin|  1|
|537|       dive|  1|
|601|    reasons|  1|
+---+-----------+---+
only showing top 20 rows

>>> df_tf.filter(df_tf['_id'] == 3).show()
+---+------+---+
|_id| token| tf|
+---+------+---+
|  3|  care|  1|
|  3|shaken|  1|
|  3|    we|  1|
|  3|    so|  2|
|  3|   wan|  1|
|  3|   are|  1|
|  3|  with|  1|
|  3|    as|  1|
+---+------+---+
```

#Calculating Document Frequency by grouping on each token and counting the number of documents it occurs in

df_idf = df2.groupby("token").agg(F.countDistinct("_id").alias("df"))

df_idf.show()

maria_dev@sandbox-hdp:~/ankit

```
>>> df_idf = df2.groupby("token").agg(F.countDistinct("_id").alias("df"))
>>> df_idf.show()
+-----------+----+
|      token|  df|
+-----------+----+
|      spoil|  23|
|       some|1227|
|        art| 829|
|       hope| 343|
|     ransom|  51|
|      still| 498|
|     doubts|  13|
|    speakst|  33|
|      those| 506|
|      joind|  27|
|        few|  60|
|     voyage|  23|
|ingratitude|  21|
|    governd|  11|
|    blossom|   9|
|    embrace|  68|
|       guts|  12|
|      cramp|   3|
| lieutenant|  47|
|     travel|  32|
+-----------+----+
only showing top 20 rows
```

#Converting 'df' column to Double Type in order for easy calculation later on

```
from pyspark.sql.types import DoubleType
df_idf = df_idf.withColumn("df", df_idf["df"].cast(DoubleType()))
df_idf.show()
```

maria_dev@sandbox-hdp:~/ankit

```
>>> from pyspark.sql.types import DoubleType
>>> df_idf = df_idf.withColumn("df", df_idf["df"].cast(DoubleType()))
>>> df_idf.show()
+-----------+------+
|      token|    df|
+-----------+------+
|      spoil|  23.0|
|       some|1227.0|
|        art| 829.0|
|       hope| 343.0|
|     ransom|  51.0|
|      still| 498.0|
|     doubts|  13.0|
|    speakst|  33.0|
|      those| 506.0|
|      joind|  27.0|
|        few|  60.0|
|     voyage|  23.0|
|ingratitude|  21.0|
|    governd|  11.0|
|    blossom|   9.0|
|    embrace|  68.0|
|       guts|  12.0|
|      cramp|   3.0|
| lieutenant|  47.0|
|     travel|  32.0|
+-----------+------+
only showing top 20 rows
```

#Calculating IDF values

```
df_idf = df_idf.withColumn("idf", F.log10(N/df_idf["df"]))
df_idf.show()
```

maria_dev@sandbox-hdp:~/ankit

```
>>> df_idf = df_idf.withColumn("idf", F.log10(N/df_idf["df"]))
>>> df_idf.show()
+-----------+------+------------------+
|      token|    df|               idf|
+-----------+------+------------------+
|      spoil|  23.0|3.6851417604833445|
|       some|1227.0| 1.958025033773933|
|        art| 829.0|2.1283150659506638|
|       hope| 343.0| 2.511575476458167|
|     ransom|  51.0| 3.339299420403001|
|      still| 498.0|  2.34964025374122|
|     doubts|  13.0|3.9329262441941006|
|    speakst|  33.0|  3.52835565662305|
|      those| 506.0|2.3427190796611383|
|      joind|  27.0|  3.61550583234195|
|        few|  60.0|3.2687183461172937|
|     voyage|  23.0|3.6851417604833445|
|ingratitude|  21.0| 3.724650301767018|
|    governd|  11.0| 4.005476911342712|
|    blossom|   9.0| 4.092627087061612|
|    embrace|  68.0| 3.214360683794701|
|       guts|  12.0|3.9676883504533125|
|      cramp|   3.0| 4.569748341781275|
| lieutenant|  47.0|  3.37477173856522|
|     travel|  32.0|3.5417196181810313|
+-----------+------+------------------+
only showing top 20 rows
```

#Joining df_tf and df_idf based on token columns

tokensWithTfIdf = df_tf.join(df_idf, df_tf["token"] == df_idf["token"], how='left').drop(df_idf["token"])

tokensWithTfIdf.show()

maria_dev@sandbox-hdp:~/ankit

```
>>> tokensWithTfIdf = df_tf.join(df_idf, df_tf["token"] == df_idf["token"], how='left').drop(df_idf["token"])
19/12/08 18:45:42 WARN Column: Constructing trivially true equals predicate, 'token#1389 = token#1389'. Perhaps you need to use aliases.
>>> tokensWithTfIdf.show()
+---+---+-----------+-------+------------------+
|_id| tf|      token|     df|               idf|
+---+---+-----------+-------+------------------+
| 55|  1|        and|23621.0|0.6735713168860664|
| 87|  1|        see| 1329.0|1.9233446155582055|
|116|  1|       upon| 1659.0|1.8270232104765767|
|188|  1|        the|23978.0|0.6670566406399929|
|190|  1|        god|  700.0|2.2017715564866807|
|191|  1|       good| 2620.0| 1.628568305181192|
|201|  1|          a|12793.0| 0.939897196614264|
|239|  1|      tarry|   44.0|  3.40341692001475|
|245|  1|        not| 7966.0|1.1456292944276283|
|273|  1|     before|  789.0| 2.149792593291517|
|275|  1|       will| 4712.0|1.3736643147218923|
|282|  1|        our| 2846.0|1.5926347007526718|
|307|  1|    through|  244.0| 2.659479770162208|
|361|  1|      there| 1708.0|1.8143817301479512|
|401|  1|     retold|    2.0| 4.745839600836956|
|506|  1|         in|10162.0|1.0398904059266607|
|509|  1|bolingbroke|   67.0| 3.220794793800111|
|521|  1|     cousin|  228.0|2.6889347495004836|
|537|  1|       dive|    7.0| 4.201771556486681|
|601|  1|    reasons|   66.0| 3.227325660959069|
+---+---+-----------+-------+------------------+
only showing top 20 rows
```

#Calculating TF-IDF Score

```
tokensWithTfIdf = tokensWithTfIdf.withColumn("tf_idf", col("tf") * col("idf"))
tokensWithTfIdf.show()
```

maria_dev@sandbox-hdp:~/ankit

```
>>> tokensWithTfIdf = tokensWithTfIdf.withColumn("tf_idf", col("tf") * col("idf"))
>>> tokensWithTfIdf.show()
+---+---+----------+-------+------------------+------------------+
|_id| tf|     token|     df|               idf|            tf_idf|
+---+---+----------+-------+------------------+------------------+
| 55|  1|       and|23621.0|0.6735713168860664|0.6735713168860664|
| 87|  1|       see| 1329.0|1.9233446155582055|1.9233446155582055|
|116|  1|      upon| 1659.0|1.8270232104765767|1.8270232104765767|
|188|  1|       the|23978.0|0.6670566406399929|0.6670566406399929|
|190|  1|       god|  700.0|2.2017715564866807|2.2017715564866807|
|191|  1|      good| 2620.0| 1.628568305181192| 1.628568305181192|
|201|  1|         a|12793.0| 0.939897196614264| 0.939897196614264|
|239|  1|     tarry|   44.0|  3.40341692001475|  3.40341692001475|
|245|  1|       not| 7966.0|1.1456292944276283|1.1456292944276283|
|273|  1|    before|  789.0| 2.149792593291517| 2.149792593291517|
|275|  1|      will| 4712.0|1.3736643147218923|1.3736643147218923|
|282|  1|       our| 2846.0|1.5926347007526718|1.5926347007526718|
|307|  1|   through|  244.0| 2.659479770162208| 2.659479770162208|
|361|  1|     there| 1708.0|1.8143817301479512|1.8143817301479512|
|401|  1|    retold|    2.0| 4.745839600836956| 4.745839600836956|
|506|  1|        in|10162.0|1.0398904059266607|1.0398904059266607|
|509|  1|bolingbroke|   67.0| 3.220794793800111| 3.220794793800111|
|521|  1|    cousin|  228.0|2.6889347495004836|2.6889347495004836|
|537|  1|      dive|    7.0| 4.201771556486681| 4.201771556486681|
|601|  1|   reasons|   66.0| 3.227325660959069| 3.227325660959069|
+---+---+----------+-------+------------------+------------------+
only showing top 20 rows
```

#Change ordering of Columns & Caching the Inverted Index

```
tokensWithTfIdf = tokensWithTfIdf.select("token", "_id", "tf", "df", "idf", "tf_idf")
tokensWithTfIdf.show()

tokensWithTfIdf.cache()
```

maria_dev@sandbox-hdp:~/ankit

```
>>> tokensWithTfIdf = tokensWithTfIdf.select("token", "_id", "tf", "df", "idf", "tf_idf")
>>> tokensWithTfIdf.show()
+-----------+---+---+-------+------------------+------------------+
|      token|_id| tf|     df|               idf|            tf_idf|
+-----------+---+---+-------+------------------+------------------+
|        and| 55|  1|23621.0|0.6735713168860664|0.6735713168860664|
|        see| 87|  1| 1329.0|1.9233446155582055|1.9233446155582055|
|       upon|116|  1| 1659.0|1.8270232104765767|1.8270232104765767|
|        the|188|  1|23978.0|0.6670566406399929|0.6670566406399929|
|        god|190|  1|  700.0|2.2017715564866807|2.2017715564866807|
|       good|191|  1| 2620.0| 1.628568305181192| 1.628568305181192|
|          a|201|  1|12793.0| 0.939897196614264| 0.939897196614264|
|      tarry|239|  1|   44.0|  3.40341692001475|  3.40341692001475|
|        not|245|  1| 7966.0|1.1456292944276283|1.1456292944276283|
|     before|273|  1|  789.0| 2.149792593291517| 2.149792593291517|
|       will|275|  1| 4712.0|1.3736643147218923|1.3736643147218923|
|        our|282|  1| 2846.0|1.5926347007526718|1.5926347007526718|
|    through|307|  1|  244.0| 2.659479770162208| 2.659479770162208|
|      there|361|  1| 1708.0|1.8143817301479512|1.8143817301479512|
|     retold|401|  1|    2.0| 4.745839600836956| 4.745839600836956|
|         in|506|  1|10162.0|1.0398904059266607|1.0398904059266607|
|bolingbroke|509|  1|   67.0| 3.220794793800111| 3.220794793800111|
|     cousin|521|  1|  228.0|2.6889347495004836|2.6889347495004836|
|       dive|537|  1|    7.0| 4.201771556486681| 4.201771556486681|
|    reasons|601|  1|   66.0| 3.227325660959069| 3.227325660959069|
+-----------+---+---+-------+------------------+------------------+
only showing top 20 rows

>>> tokensWithTfIdf.cache()
DataFrame[token: string, _id: bigint, tf: bigint, df: double, idf: double, tf_idf: double]
```

[Search] Given a query and a value N, retrieve the top N matching text entries with their score (use TFIDF scores to retrieve the matching text entries)

Constructing a function **search_words (query, N)** where query is a string and N, an integer. The result will display the top N text entries ordered by their score in descending order.

Showing the results of each of the following queries, show three sets of results N=1, 3, 5:
query1 = "to be or not"
query2 = "so far so"
query = "if you said so"

Code:

def search_words(query, N):

1. print(query, N) #Printing the Query and the number of documents to be retrieved
2. import string #Importing the string library to use some string functions
3. query = query.lower() #Making the query to lower case
4. query = query.translate(None, string.punctuation) #Removing any punctuations
5. words = query.split(" ") #Splitting the query to words based on spaces
6. num_of_words = len(words) #Calculating the number of words in the query
7. query_df = sc.parallelize(words).map(lambda x:(x,)).toDF(["query_words"]) #Converting the query to a dataframe containing the query words

8. query_df = query_df.dropDuplicates() #Dropping duplicate words from the query dataframe
9. query_subset = tokensWithTfIdf.join(query_df, query_df["query_words"] == tokensWithTfIdf["token"], how = "inner") #Gets only those words from the Inverted Index that are present in the query
10. scored1 = query_subset.groupBy("_id").agg({"*":"count", "tf_idf":"sum"}) #Counting the number of times a query word occurs in a document as well as summing up the tf-idf scores of the words that are present
11. scored1 = scored1.withColumnRenamed("count(1)", "num_of_matched_words") #Renaming the count column to num_of_matched_words
12. scored1 = scored1.withColumnRenamed("sum(tf_idf)", "temp_score") #Renaming the sum(tf_idf) column to temp_score
13. scored2 = scored1.select(scored1["_id"], (scored1["temp_score"] * scored1["num_of_matched_words"]) / num_of_words) #Finds the actual score using the relevance scoring formula
14. scored = scored2.withColumnRenamed("((temp_score * num_of_matched_words) / " + str(num_of_words) + ")", "score") #Renaming the score column
15. from pyspark.sql import functions as F #Imports SQL functions library
16. result_docs = scored.sort("score", ascending=False) #Sorts the document id's in descending order based on calculated scores
17. result_docs = result_docs.withColumn("score", F.round(result_docs["score"], 3)) #Rounding the scores to 3 decimal places
18. data_df = sqlContext.read.json("/user/maria_dev/ankit/shakespeare_full.json") #Loading the actual data file to view results
19. result_df = result_docs.join(data_df, result_docs["_id"] == data_df["_id"], how="inner").drop(data_df["_id"]) #Joining the retrieved document dataframe with the actual data file in order to view results
20. result_df = result_df.drop("line_id", "line_number", "play_name", "speaker", "speech_number", "type") #Dropping the unnecessary columns
21. result_df = result_df.sort("score", ascending=False).limit(int(N)) #Sorting again after join according to scores and then keeping only N number of required documents
22. final_tuples = tuple((row['_id'], row['score'], row['text_entry']) for row in result_df.collect()) #Storing the retrieved results in tuples
23. from pprint import pprint #Importing Pretty Print Library
24. pprint(final_tuples) # Printing the result tuples

NOTE: A dummy run of the above code has been shown below

maria_dev@sandbox-hdp:~/ankit

```
>>>
>>> query = "To be, or Not"
>>>
>>> import string
>>> query = query.lower()
>>> query = query.translate(None, string.punctuation)
>>>
>>> print query
to be or not
>>>
>>> words = query.split(" ")
>>>
>>> print words
['to', 'be', 'or', 'not']
>>>
>>> num_of_words = len(words)
>>>
>>> print num_of_words
4
>>>
>>> query_df = sc.parallelize(words).map(lambda x:(x,)).toDF(["query_words"])
>>> query_df.show()
+-----------+
|query_words|
+-----------+
|         to|
|         be|
|         or|
|        not|
+-----------+

>>> query_df = query_df.dropDuplicates()
>>> query_df.show()
+-----------+
|query_words|
+-----------+
|        not|
|         be|
|         or|
|         to|
+-----------+
```

maria_dev@sandbox-hdp:~/ankit

```
>>> query_subset = tokensWithTfIdf.join(query_df, query_df["query_words"] == tokensWithTfIdf["token"], how = "inner")
>>> query_subset.show()
+-----+-----+---+------+------------------+------------------+-----------+
|token|  _id| tf|    df|               idf|            tf_idf|query_words|
+-----+-----+---+------+------------------+------------------+-----------+
|  not|  245|  1|7966.0|1.1456292944276283|1.1456292944276283|        not|
|  not|  980|  2|7966.0|1.1456292944276283|2.2912585888552566|        not|
|  not| 1688|  1|7966.0|1.1456292944276283|1.1456292944276283|        not|
|  not| 2083|  1|7966.0|1.1456292944276283|1.1456292944276283|        not|
|  not| 3087|  1|7966.0|1.1456292944276283|1.1456292944276283|        not|
|  not| 9874|  1|7966.0|1.1456292944276283|1.1456292944276283|        not|
|  not|10379|  1|7966.0|1.1456292944276283|1.1456292944276283|        not|
|  not|17525|  1|7966.0|1.1456292944276283|1.1456292944276283|        not|
|  not|18407|  1|7966.0|1.1456292944276283|1.1456292944276283|        not|
|  not|23371|  1|7966.0|1.1456292944276283|1.1456292944276283|        not|
|  not|23840|  1|7966.0|1.1456292944276283|1.1456292944276283|        not|
|  not|23976|  1|7966.0|1.1456292944276283|1.1456292944276283|        not|
|  not|25291|  1|7966.0|1.1456292944276283|1.1456292944276283|        not|
|  not|27982|  1|7966.0|1.1456292944276283|1.1456292944276283|        not|
|  not|32967|  1|7966.0|1.1456292944276283|1.1456292944276283|        not|
|  not|34292|  1|7966.0|1.1456292944276283|1.1456292944276283|        not|
|  not|41063|  1|7966.0|1.1456292944276283|1.1456292944276283|        not|
|  not|42691|  1|7966.0|1.1456292944276283|1.1456292944276283|        not|
|  not|43622|  1|7966.0|1.1456292944276283|1.1456292944276283|        not|
|  not|48242|  1|7966.0|1.1456292944276283|1.1456292944276283|        not|
+-----+-----+---+------+------------------+------------------+-----------+
only showing top 20 rows

>>> scored1 = query_subset.groupBy("_id").agg({"*":"count", "tf_idf":"sum"})
>>> scored1 = scored1.withColumnRenamed("count(1)", "num_of_matched_words")
>>> scored1 = scored1.withColumnRenamed("sum(tf_idf)", "temp_score")
>>> scored1.show()
+------+------------------+--------------------+
|   _id|        temp_score|num_of_matched_words|
+------+------------------+--------------------+
| 95284|1.1456292944276283|                   1|
| 64873|1.1456292944276283|                   1|
| 61135|3.1921374256243946|                   3|
| 60756|1.1456292944276283|                   1|
| 54536|1.1456292944276283|                   1|
|102092|1.1456292944276283|                   1|
| 49967|3.1921374256243946|                   3|
|  1806|1.1456292944276283|                   1|
|107123|1.1456292944276283|                   1|
| 29841|1.1456292944276283|                   1|
| 57651|1.1456292944276283|                   1|
|106002|1.1456292944276283|                   1|
| 45726|1.9564405241900114|                   2|
| 55671|2.3813261958620116|                   2|
| 49983|1.1456292944276283|                   1|
| 39713|1.1456292944276283|                   1|
| 60033|1.1456292944276283|                   1|
| 50287|1.1456292944276283|                   1|
| 81068|1.1456292944276283|                   1|
| 32954|1.1456292944276283|                   1|
+------+------------------+--------------------+
only showing top 20 rows
```

maria_dev@sandbox-hdp:~/ankit

```
>>>
>>> scored2 = scored1.select(scored1["_id"], (scored1["temp_score"] * scored1["num_of_matched_words"]) / num_of_words)
>>> scored2.show()
+------+-----------------------------------------+
|   _id|((temp_score * num_of_matched_words) / 4)|
+------+-----------------------------------------+
| 95284|                      0.28640732360690707|
| 64873|                      0.28640732360690707|
| 61135|                        2.394103069218296|
| 60756|                      0.28640732360690707|
| 54536|                      0.28640732360690707|
|102092|                      0.28640732360690707|
| 49967|                        2.394103069218296|
|  1806|                      0.28640732360690707|
|107123|                      0.28640732360690707|
| 29841|                      0.28640732360690707|
| 57651|                      0.28640732360690707|
|106002|                      0.28640732360690707|
| 45726|                       0.9782202620950057|
| 55671|                       1.1906630979310058|
| 49983|                      0.28640732360690707|
| 39713|                      0.28640732360690707|
| 60033|                      0.28640732360690707|
| 50287|                      0.28640732360690707|
| 81068|                      0.28640732360690707|
| 32954|                      0.28640732360690707|
+------+-----------------------------------------+
only showing top 20 rows

>>>
>>> scored = scored2.withColumnRenamed("((temp_score * num_of_matched_words) / " + str(num_of_words) + ")", "score")
>>> scored.show()
+------+--------------------+
|   _id|               score|
+------+--------------------+
| 95284|0.28640732360690707|
| 64873|0.28640732360690707|
| 61135|  2.394103069218296|
| 60756|0.28640732360690707|
| 54536|0.28640732360690707|
|102092|0.28640732360690707|
| 49967|  2.394103069218296|
|  1806|0.28640732360690707|
|107123|0.28640732360690707|
| 29841|0.28640732360690707|
| 57651|0.28640732360690707|
|106002|0.28640732360690707|
| 45726| 0.9782202620950057|
| 55671| 1.1906630979310058|
| 49983|0.28640732360690707|
| 39713|0.28640732360690707|
| 60033|0.28640732360690707|
| 50287|0.28640732360690707|
| 81068|0.28640732360690707|
| 32954|0.28640732360690707|
+------+--------------------+
only showing top 20 rows
```

maria_dev@sandbox-hdp:~/ankit

```
>>> from pyspark.sql import functions as F
>>> ndocs = 3
>>> result_docs = scored.sort("score", ascending=False)
>>> result_docs = result_docs.withColumn("score", F.round(result_docs["score"], 3))
>>> result_docs.show()
+------+-----+
|   _id|score|
+------+-----+
| 34229|6.946|
|103117|6.135|
|109930|6.045|
|101007|4.899|
| 64679|4.899|
| 36448|4.788|
| 24102|4.096|
|109341|4.096|
| 33365|4.096|
| 99944|4.096|
| 19673|4.028|
| 19954|3.929|
| 93540|3.926|
| 16243|3.926|
|  7789|3.742|
| 93378|3.423|
| 39330|3.321|
|103398|3.321|
| 46440|3.321|
| 86710|3.321|
+------+-----+
only showing top 20 rows

>>> data_df = sqlContext.read.json("/user/maria_dev/ankit/shakespeare_full.json")
>>> result_df = result_docs.join(data_df, result_docs["_id"] == data_df["_id"], how="inner").drop(data_df["_id"])
>>> result_df = result_df.drop("line_id", "line_number", "play_name", "speaker", "speech_number", "type")
>>> result_df = result_df.sort("score", ascending=False).limit(int(ndocs))
>>> result_df.show()
+------+-----+--------------------+
|   _id|score|          text_entry|
+------+-----+--------------------+
| 34229|6.946|To be, or not to ...|
|103117|6.135|will not be seen;...|
|109930|6.045|Not like a corse;...|
+------+-----+--------------------+

>>> final_tuples = tuple((row['_id'], row['score'], row['text_entry']) for row in result_df.collect())
>>> from pprint import pprint
>>> pprint(final_tuples)
((34229, 6.946, u'To be, or not to be: that is the question:'),
 (103117, 6.135, u'will not be seen; or if she be, its four to one'),
 (109930, 6.045, u'Not like a corse; or if, not to be buried,'))
>>>
```

The function was defined and then run on the 3 queries for N = 1, 3, and 5

maria_dev@sandbox-hdp:~/ankit

```
>>>
>>> def search_words(query, N):
...     print(query, N)
...     import string
...     query = query.lower()
...     query = query.translate(None, string.punctuation)
...     words = query.split(" ")
...     num_of_words = len(words)
...     query_df = sc.parallelize(words).map(lambda x:(x,)).toDF(["query_words"])
...     query_df = query_df.dropDuplicates()
...     query_subset = tokensWithTfIdf.join(query_df, query_df["query_words"] == tokensWithTfIdf["token"], how = "inner")
...     scored1 = query_subset.groupBy("_id").agg({"*":"count", "tf_idf":"sum"})
...     scored1 = scored1.withColumnRenamed("count(1)", "num_of_matched_words")
...     scored1 = scored1.withColumnRenamed("sum(tf_idf)", "temp_score")
...     scored2 = scored1.select(scored1["_id"], (scored1["temp_score"] * scored1["num_of_matched_words"]) / num_of_words)
...     scored = scored2.withColumnRenamed("((temp_score * num_of_matched_words) / " + str(num_of_words) + ")", "score")
...     from pyspark.sql import functions as F
...     result_docs = scored.sort("score", ascending=False)
...     result_docs = result_docs.withColumn("score", F.round(result_docs["score"], 3))
...     data_df = sqlContext.read.json("/user/maria_dev/ankit/shakespeare_full.json")
...     result_df = result_docs.join(data_df, result_docs["_id"] == data_df["_id"], how="inner").drop(data_df["_id"])
...     result_df = result_df.drop("line_id", "line_number", "play_name", "speaker", "speech_number", "type")
...     result_df = result_df.sort("score", ascending=False).limit(int(N))
...     final_tuples = tuple((row['_id'], row['score'], row['text_entry']) for row in result_df.collect())
...     from pprint import pprint
...     pprint(final_tuples)
...
>>>
>>>
```

maria_dev@sandbox-hdp:~/ankit

```
>>> query = "to be or not"
>>> search_words(query, 1)
('to be or not', 1)
((34229, 6.946, u'To be, or not to be: that is the question:'),)
>>> search_words(query, 3)
('to be or not', 3)
((34229, 6.946, u'To be, or not to be: that is the question:'),
 (103117, 6.135, u'will not be seen; or if she be, its four to one'),
 (109930, 6.045, u'Not like a corse; or if, not to be buried,'))
>>> search_words(query, 5)
('to be or not', 5)
((34229, 6.946, u'To be, or not to be: that is the question:'),
 (103117, 6.135, u'will not be seen; or if she be, its four to one'),
 (109930, 6.045, u'Not like a corse; or if, not to be buried,'),
 (101007, 4.899, u'Or else you love not, for to be wise and love'),
 (64679, 4.899, u'to meddle or make. You may be gone; it is not good'))
>>>
```

maria_dev@sandbox-hdp:~/ankit

```
>>> query = "so far so"
>>> search_words(query, 1)
('so far so', 1)
((68413, 3.593, u'And so far am I glad it so did sort'),)
>>> search_words(query, 3)
('so far so', 3)
((11154, 3.593, u'So do I wish the crown, being so far off;'),
 (68413, 3.593, u'And so far am I glad it so did sort'),
 (51283, 2.764, u'so, so, so. Well go to supper i he morning. So, so, so.'))
>>> search_words(query, 5)
('so far so', 5)
((68413, 3.593, u'And so far am I glad it so did sort'),
 (11154, 3.593, u'So do I wish the crown, being so far off;'),
 (51283, 2.764, u'so, so, so. Well go to supper i he morning. So, so, so.'),
 (110732, 2.671, u'nothing, let him call me rogue for being so far'),
 (96897, 2.671, u'That brought her for this high good turn so far?'))
>>>
>>>
```

maria_dev@sandbox-hdp:~/ankit

```
>>> query = "if you said so"
>>> search_words(query, 1)
('if you said so', 1)
((18430, 11.773, u'of an If, as, If you said so, then I said so; and'),)
>>> search_words(query, 3)
('if you said so', 3)
((18430, 11.773, u'of an If, as, If you said so, then I said so; and'),
 (29571, 6.37, u'If you but said so, twere as deep with me:'),
 (61123,
  5.089,
  u'O, did you so? And do you remember what you said of the duke?'))
>>> search_words(query, 5)
('if you said so', 5)
((18430, 11.773, u'of an If, as, If you said so, then I said so; and'),
 (29571, 6.37, u'If you but said so, twere as deep with me:'),
 (61123,
  5.089,
  u'O, did you so? And do you remember what you said of the duke?'),
 (106075, 5.073, u'And if it please you, so; if not, why, so.'),
 (10471, 4.364, u'You said so much before, and yet you fled.'))
>>>
```

Outputs Obtained:

query = "to be or not"

search_words(query, 1)
('to be or not', 1)
((34229, 6.946, u'To be, or not to be: that is the question:'),)

search_words(query, 3)
('to be or not', 3)
((34229, 6.946, u'To be, or not to be: that is the question:'),
 (103117, 6.135, u'will not be seen; or if she be, its four to one'),

(109930, 6.045, u'Not like a corse; or if, not to be buried,'))

search_words(query, 5)
('to be or not', 5)
((34229, 6.946, u'To be, or not to be: that is the question:'),
(103117, 6.135, u'will not be seen; or if she be, its four to one'),
(109930, 6.045, u'Not like a corse; or if, not to be buried,'),
(101007, 4.899, u'Or else you love not, for to be wise and love'),
(64679, 4.899, u'to meddle or make. You may be gone; it is not good'))

query = "so far so"

search_words(query, 1)
('so far so', 1)
((68413, 3.593, u'And so far am I glad it so did sort'),)

search_words(query, 3)
('so far so', 3)
((11154, 3.593, u'So do I wish the crown, being so far off;'),
(68413, 3.593, u'And so far am I glad it so did sort'),
(51283, 2.764, u'so, so, so. Well go to supper i he morning. So, so, so.'))

search_words(query, 5)
('so far so', 5)
((68413, 3.593, u'And so far am I glad it so did sort'),
(11154, 3.593, u'So do I wish the crown, being so far off;'),
(51283, 2.764, u'so, so, so. Well go to supper i he morning. So, so, so.'),
(110732, 2.671, u'nothing, let him call me rogue for being so far'),
(96897, 2.671, u'That brought her for this high good turn so far?'))

query = "if you said so"

search_words(query, 1)
('if you said so', 1)
((18430, 11.773, u'of an If, as, If you said so, then I said so; and'),)

search_words(query, 3)
('if you said so', 3)
((18430, 11.773, u'of an If, as, If you said so, then I said so; and'),
(29571, 6.37, u'If you but said so, twere as deep with me:'),
(61123, 5.089, u'O, did you so? And do you remember what you said of the duke?'))

search_words(query, 5)
('if you said so', 5)
((18430, 11.773, u'of an If, as, If you said so, then I said so; and'),
 (29571, 6.37, u'If you but said so, twere as deep with me:'),
 (61123, 5.089, u'O, did you so? And do you remember what you said of the duke?'),
 (106075, 5.073, u'And if it please you, so; if not, why, so.'),
 (10471, 4.364, u'You said so much before, and yet you fled.'))

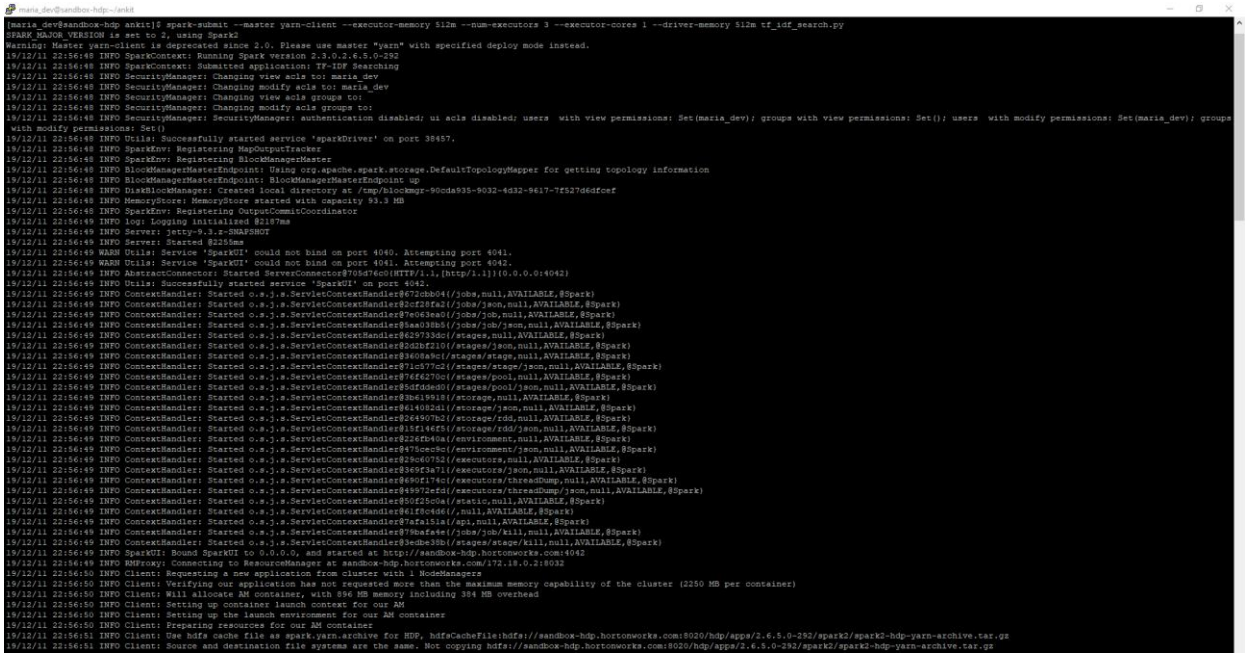
[Job] Write a file **search.py** that will run using **spark-submit**.

Code:

The file 'tf_idf_search.py' was run using spark-submit using the following command

spark-submit --master yarn-client --executor-memory 512m --num-executors 3 --executor-cores 1 --driver-memory 512m tf_idf_search.py

Output:
The output received is displayed below



```
maria_dev@sandbox-hdp:~/ankit
[maria_dev@sandbox-hdp ankit]$ spark-submit --master yarn-client --executor-memory 512m --num-executors 3 --executor-cores 1 --driver-memory 512m tf_idf_search.py
SPARK_MAJOR_VERSION is set to 2, using Spark2
Warning: Master yarn-client is deprecated since 2.0. Please use master "yarn" with specified deploy mode instead.
19/12/11 22:56:48 INFO SparkContext: Running Spark version 2.3.0.2.6.5.0-292
19/12/11 22:56:48 INFO SparkContext: Submitted application: TF-IDF Searching
19/12/11 22:56:48 INFO SecurityManager: Changing view acls to: maria_dev
19/12/11 22:56:48 INFO SecurityManager: Changing modify acls to: maria_dev
19/12/11 22:56:48 INFO SecurityManager: Changing view acls groups to:
19/12/11 22:56:48 INFO SecurityManager: Changing modify acls groups to:
19/12/11 22:56:48 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(maria_dev); groups with view permissions: Set(); users  with modify permissions: Set(maria_dev); groups
 with modify permissions: Set()
19/12/11 22:56:48 INFO Utils: Successfully started service 'sparkDriver' on port 38457.
19/12/11 22:56:48 INFO SparkEnv: Registering MapOutputTracker
19/12/11 22:56:48 INFO SparkEnv: Registering BlockManagerMaster
19/12/11 22:56:48 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
19/12/11 22:56:48 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
19/12/11 22:56:48 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-90cda935-9032-4d32-9617-7f527d6dfcef
19/12/11 22:56:48 INFO MemoryStore: MemoryStore started with capacity 93.3 MB
19/12/11 22:56:48 INFO SparkEnv: Registering OutputCommitCoordinator
19/12/11 22:56:49 INFO log: Logging initialized @2107ms
19/12/11 22:56:49 INFO Server: jetty-9.3.z-SNAPSHOT
19/12/11 22:56:49 INFO Server: Started @2255ms
19/12/11 22:56:49 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
19/12/11 22:56:49 WARN Utils: Service 'SparkUI' could not bind on port 4041. Attempting port 4042.
19/12/11 22:56:49 INFO AbstractConnector: Started ServerConnector@705d76c0{HTTP/1.1,[http/1.1]}{0.0.0.0:4042}
19/12/11 22:56:49 INFO Utils: Successfully started service 'SparkUI' on port 4042.
19/12/11 22:56:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@672dbb04{/jobs,null,AVAILABLE,@Spark}
19/12/11 22:56:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@2cf28fa2{/jobs/json,null,AVAILABLE,@Spark}
19/12/11 22:56:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@7e063ea0{/jobs/job,null,AVAILABLE,@Spark}
19/12/11 22:56:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@5aa038b5{/jobs/job/json,null,AVAILABLE,@Spark}
19/12/11 22:56:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@625933dc{/stages,null,AVAILABLE,@Spark}
19/12/11 22:56:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@2d2bf210{/stages/json,null,AVAILABLE,@Spark}
19/12/11 22:56:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@3608a9c{/stages/stage,null,AVAILABLE,@Spark}
19/12/11 22:56:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@71c577c2{/stages/stage/json,null,AVAILABLE,@Spark}
19/12/11 22:56:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@74f6270c{/stages/pool,null,AVAILABLE,@Spark}
19/12/11 22:56:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@5dfdde00{/stages/pool/json,null,AVAILABLE,@Spark}
19/12/11 22:56:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@3b619918{/storage,null,AVAILABLE,@Spark}
19/12/11 22:56:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@614082d1{/storage/json,null,AVAILABLE,@Spark}
19/12/11 22:56:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@264907b2{/storage/rdd,null,AVAILABLE,@Spark}
19/12/11 22:56:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@15f146f5{/storage/rdd/json,null,AVAILABLE,@Spark}
19/12/11 22:56:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@226fb40a{/environment,null,AVAILABLE,@Spark}
19/12/11 22:56:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@475cec5c{/environment/json,null,AVAILABLE,@Spark}
19/12/11 22:56:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@29c60752{/executors,null,AVAILABLE,@Spark}
19/12/11 22:56:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@369f3a71{/executors/json,null,AVAILABLE,@Spark}
19/12/11 22:56:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@690f174c{/executors/threadDump,null,AVAILABLE,@Spark}
19/12/11 22:56:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@49972efd{/executors/threadDump/json,null,AVAILABLE,@Spark}
19/12/11 22:56:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@50f25c0a{/static,null,AVAILABLE,@Spark}
19/12/11 22:56:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@61f0c4d6{/,null,AVAILABLE,@Spark}
19/12/11 22:56:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@7afa151a{/api,null,AVAILABLE,@Spark}
19/12/11 22:56:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@78baf4e{/jobs/job/kill,null,AVAILABLE,@Spark}
19/12/11 22:56:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@3edbe38b{/stages/stage/kill,null,AVAILABLE,@Spark}
19/12/11 22:56:49 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://sandbox-hdp.hortonworks.com:4042
19/12/11 22:56:49 INFO RMProxy: Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.18.0.2:8032
19/12/11 22:56:50 INFO Client: Requesting a new application from cluster with 1 NodeManagers
19/12/11 22:56:50 INFO Client: Verifying our application has not requested more than the maximum memory capability of the cluster (2250 MB per container)
19/12/11 22:56:50 INFO Client: Will allocate AM container, with 896 MB memory including 384 MB overhead
19/12/11 22:56:50 INFO Client: Setting up container launch context for our AM
19/12/11 22:56:50 INFO Client: Setting up the launch environment for our AM container
19/12/11 22:56:50 INFO Client: Preparing resources for our AM container
19/12/11 22:56:51 INFO Client: Use hdfs cache file as spark.yarn.archive for HDP, hdfsCacheFile:hdfs://sandbox-hdp.hortonworks.com:8020/hdp/apps/2.6.5.0-292/spark2/spark2-hdp-yarn-archive.tar.gz
19/12/11 22:56:51 INFO Client: Source and destination file systems are the same. Not copying hdfs://sandbox-hdp.hortonworks.com:8020/hdp/apps/2.6.5.0-292/spark2/spark2-hdp-yarn-archive.tar.gz
```

maria_dev@sandbox-hdp:~/ankit

```
########################################### Creating Inverted Index ###########################################


+-------+-----+---+-------+------------------+------------------+
|  token|  _id| tf|     df|               idf|            tf_idf|
+-------+-----+---+-------+------------------+------------------+
|   this|66946|  1| 6275.0|1.2492558663478617|1.2492558663478617|
|     my|66947|  1|11022.0|1.0046091898114855|1.0046091898114855|
|     is|66973|  1| 8498.0| 1.117552869747442| 1.117552869747442|
|evitate|66976|  1|    1.0| 5.046869596500938| 5.046869596500938|
|  steep|67005|  1|    7.0| 4.201771556486681| 4.201771556486681|
| nights|67006|  1|   56.0| 3.298681569494737| 3.298681569494737|
|     be|67047|  1| 6474.0| 1.235696901434383| 1.235696901434383|
|    the|67079|  1|23978.0|0.6670566406399929|0.6670566406399929|
|   this|67112|  1| 6275.0|1.2492558663478617|1.2492558663478617|
|   some|67118|  1| 1227.0| 1.958025033773933| 1.958025033773933|
|   love|67183|  1| 1809.0|1.7894310296411238|1.7894310296411238|
|  heart|67198|  1|  900.0|2.0926270870616124|2.0926270870616124|
|   hath|67212|  1| 1821.0|1.7865596507060175|1.7865596507060175|
|      a|67212|  2|12793.0| 0.939897196614264| 1.879794393228528|
|pyramus|67281|  1|   48.0|3.3656283591253504|3.3656283591253504|
| fitted|67321|  1|   13.0|3.9329262441941006|3.9329262441941006|
|   from|67428|  2| 2479.0| 1.652593069733116| 3.305186139466232|
|buskind|67440|  1|    1.0| 5.046869596500938| 5.046869596500938|
|  loves|67539|  1|  252.0|2.6454690557193934|2.6454690557193934|
|   with|67556|  1| 7174.0|1.1911082241609896|1.1911082241609896|
+-------+-----+---+-------+------------------+------------------+
only showing top 20 rows

##################################### Inverted Index Created and Saved #####################################
```

maria_dev@sandbox-hdp:~/ankit

```
##################################### Inverted Index Created and Saved #####################################


############################################## Searching ##############################################

Searching for :
('to be or not', 1)
((34229, 6.946, u'To be, or not to be: that is the question:'),)

Searching for :
('to be or not', 3)
((34229, 6.946, u'To be, or not to be: that is the question:'),
 (103117, 6.135, u'will not be seen; or if she be, its four to one'),
 (109930, 6.045, u'Not like a corse; or if, not to be buried,'))

Searching for :
('to be or not', 5)
((34229, 6.946, u'To be, or not to be: that is the question:'),
 (103117, 6.135, u'will not be seen; or if she be, its four to one'),
 (109930, 6.045, u'Not like a corse; or if, not to be buried,'),
 (101007, 4.899, u'Or else you love not, for to be wise and love'),
 (64679, 4.899, u'to meddle or make. You may be gone; it is not good'))

Searching for :
('so far so', 1)
((68413, 3.593, u'And so far am I glad it so did sort'),)

Searching for :
('so far so', 3)
((68413, 3.593, u'And so far am I glad it so did sort'),
 (11154, 3.593, u'So do I wish the crown, being so far off;'),
 (51283, 2.764, u'so, so, so. Well go to supper i he morning. So, so, so.'))

Searching for :
('so far so', 5)
((68413, 3.593, u'And so far am I glad it so did sort'),
 (11154, 3.593, u'So do I wish the crown, being so far off;'),
 (51283, 2.764, u'so, so, so. Well go to supper i he morning. So, so, so.'),
 (43877, 2.671, u'But thou from loving England art so far,'),
 (110732, 2.671, u'nothing, let him call me rogue for being so far'))

Searching for :
('if you said so', 1)
((18430, 11.773, u'of an If, as, If you said so, then I said so; and'),)

Searching for :
('if you said so', 3)
((18430, 11.773, u'of an If, as, If you said so, then I said so; and'),
 (29571, 6.37, u'If you but said so, twere as deep with me:'),
 (61123,
  5.089,
  u'O, did you so? And do you remember what you said of the duke?'))

Searching for :
('if you said so', 5)
((18430, 11.773, u'of an If, as, If you said so, then I said so; and'),
 (29571, 6.37, u'If you but said so, twere as deep with me:'),
 (61123,
  5.089,
  u'O, did you so? And do you remember what you said of the duke?'),
 (106075, 5.073, u'And if it please you, so; if not, why, so.'),
 (10471, 4.364, u'You said so much before, and yet you fled.'))
[maria_dev@sandbox-hdp ankit]$
```