Big_Data06

May 17, 2022

1 Assignment 6 Report

2 Name - Ambuj Mishra

3 Student ID - 202116003

This is a colab PDF containing both question codes and the description required in the assignment.

4 Spark Initialization

[ ]:
```
# innstall java
!apt-get install openjdk-8-jdk-headless -qq > /dev/null

# install spark (change the version number if needed)
!wget -q https://archive.apache.org/dist/spark/spark-3.0.0/spark-3.0.
 ↪0-bin-hadoop3.2.tgz

# unzip the spark file to the current folder
!tar xf spark-3.0.0-bin-hadoop3.2.tgz

# set your spark folder to your system path environment.
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
os.environ["SPARK_HOME"] = "/content/spark-3.0.0-bin-hadoop3.2"


# install findspark using pip
!pip install -q findspark
```

[ ]:
```
import findspark

findspark.init()
```

[ ]:
```
from pyspark.sql import SparkSession
spark = SparkSession.builder\
        .master("local")\
        .appName("Colab")\
        .config('spark.ui.port', '4050')\
        .getOrCreate()
```

[ ]:
```
type(spark)
```

[ ]:
```
pyspark.sql.session.SparkSession
```

[ ]:
```
sc=spark.sparkContext
```

[1]:
```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

5 Question-1(a)

5.1 From Table 3-2

[ ]:
```
#####  creating a RDD to process further
nums = sc.parallelize([1, 2, 3, 4])
```

5.1.1 Using map()

[ ]:
```
#map()
#Apply a function to each element in the RDD and return an RDD of the result.
squared = nums.map(lambda  x : x**2)
squared.collect()
```

[ ]:
```
[1, 4, 9, 16]
```

5.1.2 Using flatMap()

[ ]:
```
nums.flatMap(lambda x : x.to(3))
```

[ ]:
```
PythonRDD[2] at RDD at PythonRDD.scala:53
```

5.1.3 Using filter()

[ ]:
```
nums.filter(lambda x : x == 1).collect()
```

[ ]:
```
[1]
```

5.1.4 Using Distinct()

[ ]:
```
nums.distinct().collect()
```

[ ]:
```
[1, 2, 3, 4]
```

5.1.5 Using sample()

[ ]:
```
nums.sample(False, 0.5,1).collect()
```

[ ]:
```
[1, 3]
```

5.2 From Table 3-3

[ ]:
```
nums1 = sc.parallelize([1, 2, 3])
nums2 = sc.parallelize([3, 4, 5])
```

5.2.1 Using union()

[ ]:
```
nums1.union(nums2).collect()
```

[ ]:
```
[1, 2, 3, 3, 4, 5]
```

5.2.2 Using intersection()

[ ]:
```
nums1.intersection(nums2).collect()
```

[ ]:
```
[3]
```

5.2.3 Using subtract()

[ ]:
```
nums1.subtract(nums2).collect()
```

[ ]:
```
[2, 1]
```

5.2.4 Using cartesian()

[ ]:
```
nums1.cartesian(nums2).collect()
```

[ ]:
```
[(1, 3), (1, 4), (1, 5), (2, 3), (2, 4), (2, 5), (3, 3), (3, 4), (3, 5)]
```

5.3 From Table 3-4

5.3.1 Using collect()

[ ]:
```
basicRDD = sc.parallelize([1, 2, 3, 3])
```

[ ]:
```
basicRDD.collect()
```

[ ]:
```
[1, 2, 3, 3]
```

5.3.2 Using count()

[ ]:
```
basicRDD.count()
```

[ ]:
```
4
```

5.3.3 Using countByValue()

[ ]:
```
basicRDD.countByValue()
```

[ ]:
```
defaultdict(int, {1: 1, 2: 1, 3: 2})
```

5.3.4 Using take()

[ ]:
```
basicRDD.take(3)
```

[ ]:
```
[1, 2, 3]
```

5.3.5 Using top()

[ ]:
```
basicRDD.top(2)
```

[ ]:
```
[3, 3]
```

5.3.6 Using takeOrdered()

[ ]:
```
basicRDD.takeOrdered(3)
```

[ ]:
```
[1, 2, 3]
```

5.3.7 Using takeSample()

[ ]:
```
basicRDD.takeSample(False, 2)
```

[ ]:
```
[3, 1]
```

5.3.8 Using reduce()

[ ]:
```
basicRDD.reduce(lambda x, y: x+y)
```

[ ]:
```
9
```

5.3.9 Using fold()

[ ]:
```
basicRDD.fold(0, lambda x, y :  x+y)
```

[ ]:
```
9
```

5.3.10 Using aggregate()

[ ]:
```
print(basicRDD.aggregate((0, 0) , (lambda  x, y :  (x[0] + y, x[1] + 1)), ␣
↪(lambda  x, y : (x[0] + y[0], x[1] + y[1]))))
```

```
(9, 4)
```

5.3.11 Using foreach()

[ ]:
```
def my_print(x):
  print(x+1)

basicRDD.foreach(my_print)
```

6 Question-1(b)

[ ]:
```
from pyspark.storagelevel import StorageLevel

result = basicRDD.map(lambda x : x * x)
result.persist(StorageLevel.DISK_ONLY)

print(result.collect())
print(result.count())
```

```
[1, 4, 9, 9]
4
```

7 Question-2

7.1 Implementing Page Rank algorithm on Spark-shell

[5]:
```
##### The screenshot regarding the page rank problem is added in the next page.
```

[ ]:

[ ]:

[ ]:

[ ]:

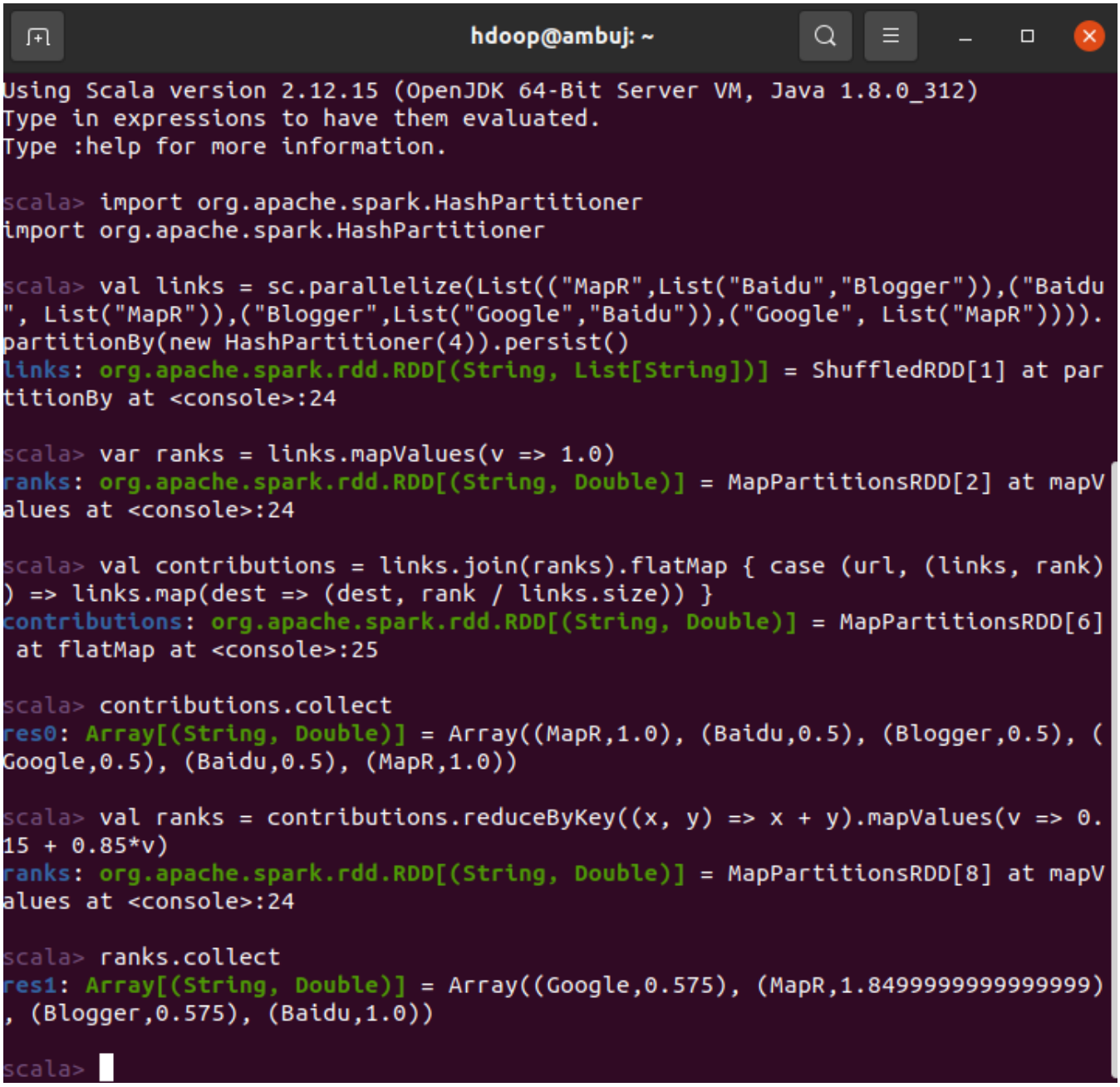
[ ]:

[ ]:

[ ]:

[ ]:

[ ]:


hdoop@ambuj: ~
Using Scala version 2.12.15 (OpenJDK 64-Bit Server VM, Java 1.8.0_312)
Type in expressions to have them evaluated.
Type :help for more information.
scala> import org.apache.spark.HashPartitioner
import org.apache.spark.HashPartitioner
scala> val links = sc.parallelize(List(("MapR",List("Baidu","Blogger")),("Baidu", List("MapR")),("Blogger",List("Google","Baidu")),("Google", List("MapR")))).partitionBy(new HashPartitioner(4)).persist()
links: org.apache.spark.rdd.RDD[(String, List[String])] = ShuffledRDD[1] at partitionBy at <console>:24
scala> var ranks = links.mapValues(v => 1.0)
ranks: org.apache.spark.rdd.RDD[(String, Double)] = MapPartitionsRDD[2] at mapValues at <console>:24
scala> val contributions = links.join(ranks).flatMap { case (url, (links, rank)) => links.map(dest => (dest, rank / links.size)) }
contributions: org.apache.spark.rdd.RDD[(String, Double)] = MapPartitionsRDD[6] at flatMap at <console>:25
scala> contributions.collect
res0: Array[(String, Double)] = Array((MapR,1.0), (Baidu,0.5), (Blogger,0.5), (Google,0.5), (Baidu,0.5), (MapR,1.0))
scala> val ranks = contributions.reduceByKey((x, y) => x + y).mapValues(v => 0.15 + 0.85*v)
ranks: org.apache.spark.rdd.RDD[(String, Double)] = MapPartitionsRDD[8] at mapValues at <console>:24
scala> ranks.collect
res1: Array[(String, Double)] = Array((Google,0.575), (MapR,1.8499999999999999), (Blogger,0.575), (Baidu,1.0))
scala>

8 Question-3(a)

8.1 Importing Dataset

[ ]:
```
# in Python
sales = spark.read.format("csv").option("header", "true").option("inferSchema",␣
 ↪"true").load("/content/drive/MyDrive/Spark-The-Definitive-Guide-master/data/
 ↪retail-data/by-day/*.csv").coalesce(5).where("Description IS NOT NULL")
fakeIntDF = spark.read.parquet("/content/drive/MyDrive/
 ↪Spark-The-Definitive-Guide-master/data/simple-ml-integers")
simpleDF = spark.read.json("/content/drive/MyDrive/
 ↪Spark-The-Definitive-Guide-master/data/simple-ml")
scaleDF = spark.read.parquet("/content/drive/MyDrive/
 ↪Spark-The-Definitive-Guide-master/data/simple-ml-scaling")
```

[ ]:
```
sales.cache()
sales.show()
```

```
+---------+---------+--------------------+--------+------------------+---------
+----------+--------------+
|InvoiceNo|StockCode|         Description|Quantity|
InvoiceDate|UnitPrice|CustomerID|       Country|
+---------+---------+--------------------+--------+------------------+---------
+----------+--------------+
|   580538|    23084|  RABBIT NIGHT LIGHT|      48|2011-12-05 08:38:00|
1.79|   14075.0|United Kingdom|
|   580538|    23077| DOUGHNUT LIP GLOSS |      20|2011-12-05 08:38:00|
1.25|   14075.0|United Kingdom|
|   580538|    22906|12 MESSAGE CARDS …|      24|2011-12-05 08:38:00|
1.65|   14075.0|United Kingdom|
|   580538|    21914|BLUE HARMONICA IN…|      24|2011-12-05 08:38:00|
1.25|   14075.0|United Kingdom|
|   580538|    22467|   GUMBALL COAT RACK|       6|2011-12-05 08:38:00|
2.55|   14075.0|United Kingdom|
|   580538|    21544|SKULLS  WATER TRA…|      48|2011-12-05 08:38:00|
0.85|   14075.0|United Kingdom|
|   580538|    23126|FELTCRAFT GIRL AM…|       8|2011-12-05 08:38:00|
4.95|   14075.0|United Kingdom|
|   580538|    21833|CAMOUFLAGE LED TORCH|      24|2011-12-05 08:38:00|
1.69|   14075.0|United Kingdom|
|   580539|    21479|WHITE SKULL HOT W…|       4|2011-12-05 08:39:00|
4.25|   18180.0|United Kingdom|
|   580539|   84030E|ENGLISH ROSE HOT …|       4|2011-12-05 08:39:00|
4.25|   18180.0|United Kingdom|
|   580539|    23355|HOT WATER BOTTLE …|       4|2011-12-05 08:39:00|
4.95|   18180.0|United Kingdom|
|   580539|    22111|SCOTTIE DOG HOT W…|       3|2011-12-05 08:39:00|
```

```
4.95|   18180.0|United Kingdom|
|   580539|    21115|ROSE CARAVAN DOOR…|       8|2011-12-05 08:39:00|
1.95|   18180.0|United Kingdom|
|   580539|    21411|GINGHAM HEART  DO…|       8|2011-12-05 08:39:00|
1.95|   18180.0|United Kingdom|
|   580539|    23235|STORAGE TIN VINTA…|      12|2011-12-05 08:39:00|
1.25|   18180.0|United Kingdom|
|   580539|    23239|SET OF 4 KNICK KN…|       6|2011-12-05 08:39:00|
1.65|   18180.0|United Kingdom|
|   580539|    22197|      POPCORN HOLDER|      36|2011-12-05 08:39:00|
0.85|   18180.0|United Kingdom|
|   580539|    22693|GROW A FLYTRAP OR…|      24|2011-12-05 08:39:00|
1.25|   18180.0|United Kingdom|
|   580539|    22372|AIRLINE BAG VINTA…|       4|2011-12-05 08:39:00|
4.25|   18180.0|United Kingdom|
|   580539|    22375|AIRLINE BAG VINTA…|       4|2011-12-05 08:39:00|
4.25|   18180.0|United Kingdom|
+---------+---------+--------------------+--------+-------------------+---------
+----------+--------------+
only showing top 20 rows
```

8.2 A.) Using StandardScaler

[ ]:
```
from pyspark.ml.feature import StandardScaler

sScaler = StandardScaler().setInputCol("features")
sScaler.fit(scaleDF).transform(scaleDF).show()
```

```
+---+--------------+------------------------------------+
| id|      features|StandardScaler_e63c4ca325e1__output|
+---+--------------+------------------------------------+
|  0|[1.0,0.1,-1.0]|                [1.19522860933439…|
|  1| [2.0,1.1,1.0]|                [2.39045721866878…|
|  0|[1.0,0.1,-1.0]|                [1.19522860933439…|
|  1| [2.0,1.1,1.0]|                [2.39045721866878…|
|  1|[3.0,10.1,3.0]|                [3.58568582800318…|
+---+--------------+------------------------------------+
```

8.3 B.) Using MinMaxScaler

[ ]:
```
from pyspark.ml.feature import MinMaxScaler

minMax = MinMaxScaler().setMin(5).setMax(10).setInputCol("features")
```

```
fittedminMax = minMax.fit(scaleDF)
fittedminMax.transform(scaleDF).show()
```

```
+---+--------------+--------------------------------+
| id|      features|MinMaxScaler_7bc11441f185__output|
+---+--------------+--------------------------------+
|  0|[1.0,0.1,-1.0]|                   [5.0,5.0,5.0]|
|  1| [2.0,1.1,1.0]|                   [7.5,5.5,7.5]|
|  0|[1.0,0.1,-1.0]|                   [5.0,5.0,5.0]|
|  1| [2.0,1.1,1.0]|                   [7.5,5.5,7.5]|
|  1|[3.0,10.1,3.0]|                 [10.0,10.0,10.0]|
+---+--------------+--------------------------------+
```

8.4 C.) MaxAbsScaler

[ ]:
```
from pyspark.ml.feature import MaxAbsScaler

maScaler = MaxAbsScaler().setInputCol("features")
fittedmaScaler = maScaler.fit(scaleDF)
fittedmaScaler.transform(scaleDF).show()
```

```
+---+--------------+--------------------------------+
| id|      features|MaxAbsScaler_756c5bb55422__output|
+---+--------------+--------------------------------+
|  0|[1.0,0.1,-1.0]|             [0.33333333333333…|
|  1| [2.0,1.1,1.0]|             [0.66666666666666…|
|  0|[1.0,0.1,-1.0]|             [0.33333333333333…|
|  1| [2.0,1.1,1.0]|             [0.66666666666666…|
|  1|[3.0,10.1,3.0]|                    [1.0,1.0,1.0]|
+---+--------------+--------------------------------+
```

8.5 D.) Elementwise-Product

[ ]:
```
from pyspark.ml.feature import ElementwiseProduct
from pyspark.ml.linalg import Vectors

scaleUpVec = Vectors.dense(10.0, 15.0, 20.0)
scalingUp = ElementwiseProduct().setScalingVec(scaleUpVec).
 ↪setInputCol("features")
scalingUp.transform(scaleDF).show()
```

```
+---+--------------+-------------------------------------+
| id|      features|ElementwiseProduct_46832125aaa3__output|
```

```
+---+--------------+--------------------------------------+
|  0|[1.0,0.1,-1.0]|                       [10.0,1.5,-20.0]|
|  1| [2.0,1.1,1.0]|                       [20.0,16.5,20.0]|
|  0|[1.0,0.1,-1.0]|                       [10.0,1.5,-20.0]|
|  1| [2.0,1.1,1.0]|                       [20.0,16.5,20.0]|
|  1|[3.0,10.1,3.0]|                      [30.0,151.5,60.0]|
+---+--------------+--------------------------------------+
```

8.6 E.) Normalizer

[ ]:
```
from pyspark.ml.feature import Normalizer

manhattanDistance = Normalizer().setP(1).setInputCol("features")
manhattanDistance.transform(scaleDF).show()
```

```
+---+--------------+------------------------------+
| id|      features|Normalizer_8d0ba3515a2a__output|
+---+--------------+------------------------------+
|  0|[1.0,0.1,-1.0]|           [0.47619047619047…|
|  1| [2.0,1.1,1.0]|           [0.48780487804878…|
|  0|[1.0,0.1,-1.0]|           [0.47619047619047…|
|  1| [2.0,1.1,1.0]|           [0.48780487804878…|
|  1|[3.0,10.1,3.0]|           [0.18633540372670…|
+---+--------------+------------------------------+
```

9 Question-3(b)

9.1 A.) Tokenizing Text Before Processing

[ ]:
```
from pyspark.ml.feature import Tokenizer

tkn = Tokenizer().setInputCol("Description").setOutputCol("DescOut")
tokenized = tkn.transform(sales.select("Description"))
tokenized.show(20, False)
```

```
+---------------------------------+-----------------------------------------+
|Description                      |DescOut                                  |
+---------------------------------+-----------------------------------------+
|RABBIT NIGHT LIGHT               |[rabbit, night, light]                   |
|DOUGHNUT LIP GLOSS               |[doughnut, lip, gloss]                   |
|12 MESSAGE CARDS WITH ENVELOPES  |[12, message, cards, with, envelopes]    |
|BLUE HARMONICA IN BOX            |[blue, harmonica, in, box]               |
|GUMBALL COAT RACK                |[gumball, coat, rack]                    |
```

```
|SKULLS  WATER TRANSFER TATTOOS     |[skulls, , water, transfer, tattoos]      |
|FELTCRAFT GIRL AMELIE KIT          |[feltcraft, girl, amelie, kit]            |
|CAMOUFLAGE LED TORCH               |[camouflage, led, torch]                  |
|WHITE SKULL HOT WATER BOTTLE       |[white, skull, hot, water, bottle]        |
|ENGLISH ROSE HOT WATER BOTTLE      |[english, rose, hot, water, bottle]       |
|HOT WATER BOTTLE KEEP CALM         |[hot, water, bottle, keep, calm]          |
|SCOTTIE DOG HOT WATER BOTTLE       |[scottie, dog, hot, water, bottle]        |
|ROSE CARAVAN DOORSTOP              |[rose, caravan, doorstop]                 |
|GINGHAM HEART  DOORSTOP RED        |[gingham, heart, , doorstop, red]         |
|STORAGE TIN VINTAGE LEAF           |[storage, tin, vintage, leaf]             |
|SET OF 4 KNICK KNACK TINS POPPIES  |[set, of, 4, knick, knack, tins, poppies] |
|POPCORN HOLDER                     |[popcorn, holder]                         |
|GROW A FLYTRAP OR SUNFLOWER IN TIN |[grow, a, flytrap, or, sunflower, in, tin]|
|AIRLINE BAG VINTAGE WORLD CHAMPION |[airline, bag, vintage, world, champion]  |
|AIRLINE BAG VINTAGE JET SET BROWN  |[airline, bag, vintage, jet, set, brown]  |
+-----------------------------------+------------------------------------------+
only showing top 20 rows
```

[ ]:
```
from pyspark.ml.feature import RegexTokenizer

rt = RegexTokenizer().setInputCol("Description").setOutputCol("DescOut").
 ↪setPattern(" ").setToLowercase(True)
rt.transform(sales.select("Description")).show(20, False)
```

```
+-----------------------------------+------------------------------------------+
|Description                        |DescOut                                   |
+-----------------------------------+------------------------------------------+
|RABBIT NIGHT LIGHT                 |[rabbit, night, light]                    |
|DOUGHNUT LIP GLOSS                 |[doughnut, lip, gloss]                    |
|12 MESSAGE CARDS WITH ENVELOPES    |[12, message, cards, with, envelopes]     |
|BLUE HARMONICA IN BOX              |[blue, harmonica, in, box]                |
|GUMBALL COAT RACK                  |[gumball, coat, rack]                     |
|SKULLS  WATER TRANSFER TATTOOS     |[skulls, water, transfer, tattoos]        |
|FELTCRAFT GIRL AMELIE KIT          |[feltcraft, girl, amelie, kit]            |
|CAMOUFLAGE LED TORCH               |[camouflage, led, torch]                  |
|WHITE SKULL HOT WATER BOTTLE       |[white, skull, hot, water, bottle]        |
|ENGLISH ROSE HOT WATER BOTTLE      |[english, rose, hot, water, bottle]       |
|HOT WATER BOTTLE KEEP CALM         |[hot, water, bottle, keep, calm]          |
|SCOTTIE DOG HOT WATER BOTTLE       |[scottie, dog, hot, water, bottle]        |
|ROSE CARAVAN DOORSTOP              |[rose, caravan, doorstop]                 |
|GINGHAM HEART  DOORSTOP RED        |[gingham, heart, doorstop, red]           |
|STORAGE TIN VINTAGE LEAF           |[storage, tin, vintage, leaf]             |
|SET OF 4 KNICK KNACK TINS POPPIES  |[set, of, 4, knick, knack, tins, poppies] |
|POPCORN HOLDER                     |[popcorn, holder]                         |
|GROW A FLYTRAP OR SUNFLOWER IN TIN |[grow, a, flytrap, or, sunflower, in, tin]|
|AIRLINE BAG VINTAGE WORLD CHAMPION |[airline, bag, vintage, world, champion]  |
|AIRLINE BAG VINTAGE JET SET BROWN  |[airline, bag, vintage, jet, set, brown]  |
```

```
+-----------------------------------+----------------------------------------+
only showing top 20 rows
```

[ ]:
```python
# in Python
from pyspark.ml.feature import RegexTokenizer
rt = RegexTokenizer()\
.setInputCol("Description")\
.setOutputCol("DescOut")\
.setPattern(" ")\
.setGaps(False)\
.setToLowercase(True)
rt.transform(sales.select("Description")).show(20, False)
```

```
+-----------------------------------+------------------+
|Description                        |DescOut           |
+-----------------------------------+------------------+
|RABBIT NIGHT LIGHT                 |[ ,  ]            |
|DOUGHNUT LIP GLOSS                 |[ ,  ,  ]         |
|12 MESSAGE CARDS WITH ENVELOPES    |[ ,  ,  ,  ]      |
|BLUE HARMONICA IN BOX              |[ ,  ,  ,  ]      |
|GUMBALL COAT RACK                  |[ ,  ]            |
|SKULLS  WATER TRANSFER TATTOOS     |[ ,  ,  ,  ,  ]   |
|FELTCRAFT GIRL AMELIE KIT          |[ ,  ,  ]         |
|CAMOUFLAGE LED TORCH               |[ ,  ]            |
|WHITE SKULL HOT WATER BOTTLE       |[ ,  ,  ,  ,  ]   |
|ENGLISH ROSE HOT WATER BOTTLE      |[ ,  ,  ,  ]      |
|HOT WATER BOTTLE KEEP CALM         |[ ,  ,  ,  ]      |
|SCOTTIE DOG HOT WATER BOTTLE       |[ ,  ,  ,  ]      |
|ROSE CARAVAN DOORSTOP              |[ ,  ]            |
|GINGHAM HEART  DOORSTOP RED        |[ ,  ,  ,  ]      |
|STORAGE TIN VINTAGE LEAF           |[ ,  ,  ]         |
|SET OF 4 KNICK KNACK TINS POPPIES  |[ ,  ,  ,  ,  ,  ]|
|POPCORN HOLDER                     |[ ]               |
|GROW A FLYTRAP OR SUNFLOWER IN TIN |[ ,  ,  ,  ,  ,  ]|
|AIRLINE BAG VINTAGE WORLD CHAMPION |[ ,  ,  ,  ,  ]   |
|AIRLINE BAG VINTAGE JET SET BROWN  |[ ,  ,  ,  ,  ]   |
+-----------------------------------+------------------+
only showing top 20 rows
```

9.2 B.) Removing Common Words

[ ]:
```python
from pyspark.ml.feature import StopWordsRemover

englishStopWords = StopWordsRemover.loadDefaultStopWords("english")
```

```
stops = StopWordsRemover().setStopWords(englishStopWords).setInputCol("DescOut")
stops.transform(tokenized).show()
```

```
+--------------------+--------------------+-----------------------------------
+
|         Description|
DescOut|StopWordsRemover_2ff782587456__output|
+--------------------+--------------------+-----------------------------------
+
|  RABBIT NIGHT LIGHT|[rabbit, night, l…|                 [rabbit, night,
l…|
| DOUGHNUT LIP GLOSS |[doughnut, lip, g…|                 [doughnut, lip,
g…|
|12 MESSAGE CARDS …|[12, message, car…|               [12, message,
car…|
|BLUE HARMONICA IN…|[blue, harmonica,…|               [blue,
harmonica,…|
|   GUMBALL COAT RACK|[gumball, coat, r…|                 [gumball, coat,
r…|
|SKULLS  WATER TRA…|[skulls, , water,…|               [skulls, ,
water,…|
|FELTCRAFT GIRL AM…|[feltcraft, girl,…|               [feltcraft,
girl,…|
|CAMOUFLAGE LED TORCH|[camouflage, led,…|                 [camouflage,
led,…|
|WHITE SKULL HOT W…|[white, skull, ho…|               [white, skull,
ho…|
|ENGLISH ROSE HOT …|[english, rose, h…|               [english, rose,
h…|
|HOT WATER BOTTLE …|[hot, water, bott…|               [hot, water,
bott…|
|SCOTTIE DOG HOT W…|[scottie, dog, ho…|               [scottie, dog,
ho…|
|ROSE CARAVAN DOOR…|[rose, caravan, d…|               [rose, caravan,
d…|
|GINGHAM HEART  DO…|[gingham, heart, …|               [gingham, heart,
…|
|STORAGE TIN VINTA…|[storage, tin, vi…|               [storage, tin,
vi…|
|SET OF 4 KNICK KN…|[set, of, 4, knic…|               [set, 4, knick,
k…|
|      POPCORN HOLDER|   [popcorn, holder]|                      [popcorn,
holder]|
|GROW A FLYTRAP OR…|[grow, a, flytrap…|               [grow, flytrap,
s…|
|AIRLINE BAG VINTA…|[airline, bag, vi…|               [airline, bag,
vi…|
```

```
|AIRLINE BAG VINTA…|[airline, bag, vi…|                    [airline, bag,
vi…|
+------------------+------------------+-----------------------------------
+
only showing top 20 rows
```

9.3 C.) Creating Word Combinations

[ ]:
```
from pyspark.ml.feature import NGram

unigram = NGram().setInputCol("DescOut").setN(1)
bigram = NGram().setInputCol("DescOut").setN(2)
```

[ ]:
```
unigramRDD = unigram.transform(tokenized.select("DescOut"))
unigramRDD.show()
```

```
+--------------------+--------------------------+
|             DescOut|NGram_6679e9d34f1a__output|
+--------------------+--------------------------+
|[rabbit, night, l…|      [rabbit, night, l…|
|[doughnut, lip, g…|      [doughnut, lip, g…|
|[12, message, car…|      [12, message, car…|
|[blue, harmonica,…|      [blue, harmonica,…|
|[gumball, coat, r…|      [gumball, coat, r…|
|[skulls, , water,…|      [skulls, , water,…|
|[feltcraft, girl,…|      [feltcraft, girl,…|
|[camouflage, led,…|      [camouflage, led,…|
|[white, skull, ho…|      [white, skull, ho…|
|[english, rose, h…|      [english, rose, h…|
|[hot, water, bott…|      [hot, water, bott…|
|[scottie, dog, ho…|      [scottie, dog, ho…|
|[rose, caravan, d…|      [rose, caravan, d…|
|[gingham, heart, …|      [gingham, heart, …|
|[storage, tin, vi…|      [storage, tin, vi…|
|[set, of, 4, knic…|      [set, of, 4, knic…|
|   [popcorn, holder]|         [popcorn, holder]|
|[grow, a, flytrap…|      [grow, a, flytrap…|
|[airline, bag, vi…|      [airline, bag, vi…|
|[airline, bag, vi…|      [airline, bag, vi…|
+--------------------+--------------------------+
only showing top 20 rows
```

[ ]:
```
bigramRDD = bigram.transform(tokenized.select("DescOut"))
bigramRDD.show()
```

```
+--------------------+--------------------------+
|             DescOut|NGram_96fe20ee8337__output|
+--------------------+--------------------------+
|[rabbit, night, l…|      [rabbit night, ni…|
|[doughnut, lip, g…|      [doughnut lip, li…|
|[12, message, car…|      [12 message, mess…|
|[blue, harmonica,…|      [blue harmonica, …|
|[gumball, coat, r…|      [gumball coat, co…|
|[skulls, , water,…|      [skulls ,  water,…|
|[feltcraft, girl,…|      [feltcraft girl, …|
|[camouflage, led,…|      [camouflage led, …|
|[white, skull, ho…|      [white skull, sku…|
|[english, rose, h…|      [english rose, ro…|
|[hot, water, bott…|      [hot water, water…|
|[scottie, dog, ho…|      [scottie dog, dog…|
|[rose, caravan, d…|      [rose caravan, ca…|
|[gingham, heart, …|      [gingham heart, h…|
|[storage, tin, vi…|      [storage tin, tin…|
|[set, of, 4, knic…|      [set of, of 4, 4 …|
|   [popcorn, holder]|          [popcorn holder]|
|[grow, a, flytrap…|      [grow a, a flytra…|
|[airline, bag, vi…|      [airline bag, bag…|
|[airline, bag, vi…|      [airline bag, bag…|
+--------------------+--------------------------+
only showing top 20 rows
```

10 THANK YOU!

[4]:
```
!wget -nc https://raw.githubusercontent.com/brpy/colab-pdf/master/colab_pdf.py
from colab_pdf import colab_pdf
colab_pdf('Big_Data06.ipynb')
```

```
File ‘colab_pdf.py’ already there; not retrieving.


WARNING: apt does not have a stable CLI interface. Use with caution in scripts.


WARNING: apt does not have a stable CLI interface. Use with caution in scripts.

[NbConvertApp] Converting notebook /content/drive/MyDrive/Colab
Notebooks/Big_Data06.ipynb to pdf
[NbConvertApp] Writing 464609 bytes to ./notebook.tex
[NbConvertApp] Building PDF
[NbConvertApp] Running xelatex 3 times: ['xelatex', './notebook.tex', '-quiet']
```

```
[NbConvertApp] Running bibtex 1 time: ['bibtex', './notebook']
[NbConvertApp] WARNING | bibtex had problems, most likely because there were no
citations
[NbConvertApp] PDF successfully created
[NbConvertApp] Writing 562092 bytes to /content/drive/My Drive/Big_Data06.pdf
```

```
<IPython.core.display.Javascript object>
```

```
<IPython.core.display.Javascript object>
```

[4]: 'File ready to be Downloaded and Saved to Drive'

[ ]: