

ICP 2

Submitted By: Dhairya Chandra

What you learned in the ICP

I have learned how to setup spark environment in Google Colab, Working with Spark commands, Basic intro to spark.

ICP description what was the task you were performing

Task was to upload a txt file on google colab and then printing the words with same first letter in each row.

first_character	grouped_words
m	more,motorcycle,m...
f	football,fans,fal...
n	november,new
v	virus,vaccine,virus
o	of,officials,over...
h	have,health,have,...
p	politicized,peopl...
d	distribute,death,...
w	wont
c	carolina,currentl...
u	university,univer...
i	in,iowa,in,in,ine...
l	1000
j	judge
b	be,by,berlusconi,...
r	reports,ready,rai...
a	a,a,and,a,after,a...
t	the,that,than,the...
s	south,students,st...

Challenges that you faced

Spark was new to me I haven't worked in Spark before, To understand the working of spark it took me some time.

Screen shots that shows the successful execution of each required step of your code

I have commented each step please refer to below screenshots:

[2] # Install Spark

```
!apt-get install openjdk-8-jdk-headless -qq > /dev/null
!wget -q https://downloads.apache.org/spark/spark-3.0.0/spark-3.0.0-bin-hadoop3.2.tgz
!tar xf spark-3.0.0-bin-hadoop3.2.tgz
!pip install -q findspark
```

[3] # Setting 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"
```

[4] # Installing PySpark

```
!pip install pyspark
```

[5] `from pyspark.sql import SparkSession`
`from pyspark.sql.functions import collect_list, udf, lit, explode, split, col, lower, trim, regexp_replace, substring, concat_ws, concat`

[6] # Creating Spark Application to run on local

```
spark = SparkSession.builder.master("local[*]").appName("ICP2_Dhairya_Chandra").enableHiveSupport().getOrCreate()
```

[7] # Add upload button to upload file from computer on Google Colab

```
from google.colab import files
files.upload()
```

Choose Files icp2.txt
• icp2.txt(text/plain) - 693 bytes, last modified: n/a - 100% done
Saving icp2.txt to icp2.txt
{'icp2.txt': b'The University of South Carolina reports that more than 1,000 students currently have the virus.\r\nThe C.D.C. tells

[8] # Importing uploaded file as a data frame

```
data = spark.read.text('icp2.txt')
```

[8] # Importing uploaded file as a data frame

```
data = spark.read.text('icp2.txt')
```

▶ data.show()

```
┌-----+
|          value|
├-----+
|The University of...|
|The C.D.C. tells ...|
|In Iowa, college ...|
|Virus fallout fro...|
|New studies show ...|
|Silvio Berlusconi...|
|A judge orders th...|
└-----+
```

```
[10] # Converting paragraph to words list
```

```
data_words = data.select(explode(split(regexp_replace(trim(lower(col("value"))), "[\$, '\. \t\n-]", ""), "\s+")).alias("value"))
```

```
[11] data_words.show()
```

```
┌-----+
| value |
├-----+
| the   |
| university |
| of    |
| south |
| carolina |
| reports |
| that   |
| more   |
| than   |
| 1000   |
| students |
| currently |
| have   |
| the    |
| virus  |
| the    |
| cdc    |
| tells  |
| health |
| officials |
└-----+
only showing top 20 rows
```

```
[ ] # Displaying the first letter of each word in separate column
```

```
first_char = data_words.select((substring("value",1,1)).alias("first_character"),col("value").alias("word"))
```

```
▶ first_char.show()
```

```
┌-----+
| first_character | word |
├-----+
| t | the |
| u | university |
| o | of |
| s | south |
| c | carolina |
| r | reports |
| t | that |
| m | more |
| t | than |
| l | 1000 |
| s | students |
| c | currently |
| h | have |
| t | the |
| v | virus |
| t | the |
| c | cdc |
| t | tells |
| h | health |
| o | officials |
└-----+
only showing top 20 rows
```

```
[ ] # Displaying all words starting with letter T in txt document
```

```
first_char.filter(first_char.first_character=="t").show()
```

```
┌-----+
| first_character | word |
├-----+
| t | the |
| t | that |
| t | than |
| t | the |
| t | the |
| t | tells |
| t | to |
| t | to |
| t | timing |
| t | the |
| t | tests |
| t | the |
| t | to |
| t | the |
└-----+
```

```
[ ] # Grouping all words in document with letters their first common letter

words_group = first_char.groupBy("first_character").agg(concat_ws(" ", collect_list("word")).alias("grouped_words"))
```

words_group.show()

```
+-----+-----+
|first_character|grouped_words|
+-----+-----+
|m|more,motorcycle,m...|
|f|football,fans,fal...|
|n|    november,new    |
|v|virus,vaccine,virus|
|o|of,officials,over...|
|h|have,health,have,...|
|p|politicized,peopl...|
|d|distribute,death,...|
|w|    wont            |
|c|carolina,currentl...|
|u|university,univer...|
|i|in,iowa,in,in,ine...|
|l|    1000            |
|j|    judge           |
|b|be,by,berlusconi,...|
|r|reports,ready,rai...|
|a|a,a,and,a,after,a...|
|t|the,that,than,the...|
|s|south,students,st...|
+-----+-----+
```

```
[ ] # Combining all in 1 column

results = words_group.orderBy("first_character").select(concat(col("first_character"),lit(" "),col("grouped_words")).alias("Result"))
```

results.show()

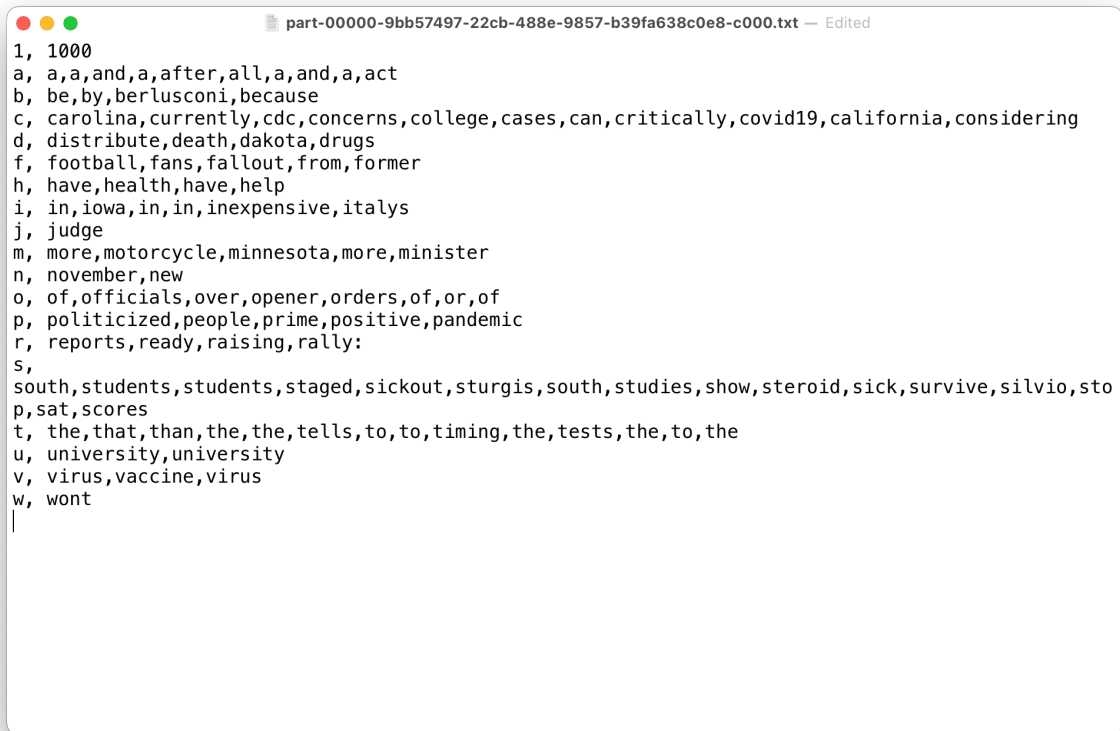
```
+-----+-----+
|Result|
+-----+-----+
|1, 1000|
|a, a,a,and,a,afte...|
|b, be,by,berlusco...|
|c, carolina,curre...|
|d, distribute,dea...|
|f, football,fans,...|
|h, have,health,ha...|
|i, in,iowa,in,in,...|
|j, judge           |
|m, more,motorcycl...|
|n, november,new    |
|o, of,officials,o...|
|p, politicized,pe...|
|r, reports,ready,...|
|s, south,students...|
|t, the,that,than,...|
|u, university,uni...|
|v, virus,vaccine,...|
|w, wont            |
+-----+-----+
```

Saving the results in output.txt file

```
results.coalesce(1).write.format("text").option("header", "false").mode("append").save("output.txt")
```

Output file link if applicable

Here is the screenshot of my output file which is attached in folder



```
part-00000-9bb57497-22cb-488e-9857-b39fa638c0e8-c000.txt — Edited
1, 1000
a, a,a,and,a,after,all,a,and,a,act
b, be,by,berlusconi,because
c, carolina,currently,cdc,concerns,college,cases,can,critically,covid19,california,considering
d, distribute,death,dakota,drugs
f, football,fans,fallout,from,former
h, have,health,have,help
i, in,iowa,in,in,inexpensive,italys
j, judge
m, more,motorcycle,minnesota,more,minister
n, november,new
o, of,officials,over,opener,orders,of,or,of
p, politicized,people,prime,positive,pandemic
r, reports,ready,raising,rally:
s,
south,students,students,staged,sickout,sturgis,south,studies,show,steroid,sick,survive,silvio,sto
p,sat,scores
t, the,that,than,the,the,tells,to,to,timing,the,tests,the,to,the
u, university,university
v, virus,vaccine,virus
w, wont
|
```

Video link (YouTube or any other publicly available video platform)

Video Link: <https://youtu.be/maxqIPPgRdk>