Big Data and Large Scale Computing

Lab Report -04

July 21, 2025

Name : Rohan Baghel
Student ID: 202116011

About

Apache Spark
Apache Spark™ is a multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters.

Key Features of Apache Spark:

- Batch/streaming data
- SQL analytics
- Data science at scale
- Machine learning

Question 1

The first task is to install Spark and gain basic working knowledge. For the same, you can refer to chapter 2 in [3] and go through sections "Downloading Spark" and "Introduction to Spark's Python and Scala Shells" (you may also refer to the some initial chapters in [4] as well). Subsequently, do the following.

Answer :

To install Apache Spark first install necessary dependencies.

- JDK
- Scala
- Git

Use the command to download all the packages at once.

```
$ sudo apt install default-jdk scala git -y
```

Verify the installed dependencies

```
$ java -version; javac -version; scala -version; git --version
```

Download Apache spark using the link :-
`https://spark.apache.org/downloads.html`

After Downloading manually extract the file to a specific location where you want to install the spark

Now give the location of the spark to the System, you need ot add path in .bashrc file.

- To open .bashrc file go to home and see the hiden files.
- open the .bashrc file in any text editor and add.
-
```
export SPARK_HOME=/home/rohan/spark
export PATH=$PATH:$SPARK_HOME/bin
```
- Run the command in command line to save the changes in .bashrc file

  ```
  $ source ~/.bashrc
  ```

  Now to start all the clusters use

  ```
  $ start-all.sh
  ```

(a) Create a file "sample.txt" with a few sentences written in English. First, put the file in the local file system and do a word count. Then, with Hadoop installed in pseudo-distributed mode, put the file in HDFS and again do the word count. You can refer to a and b for a demo on solving this problem.

samplefile.txt

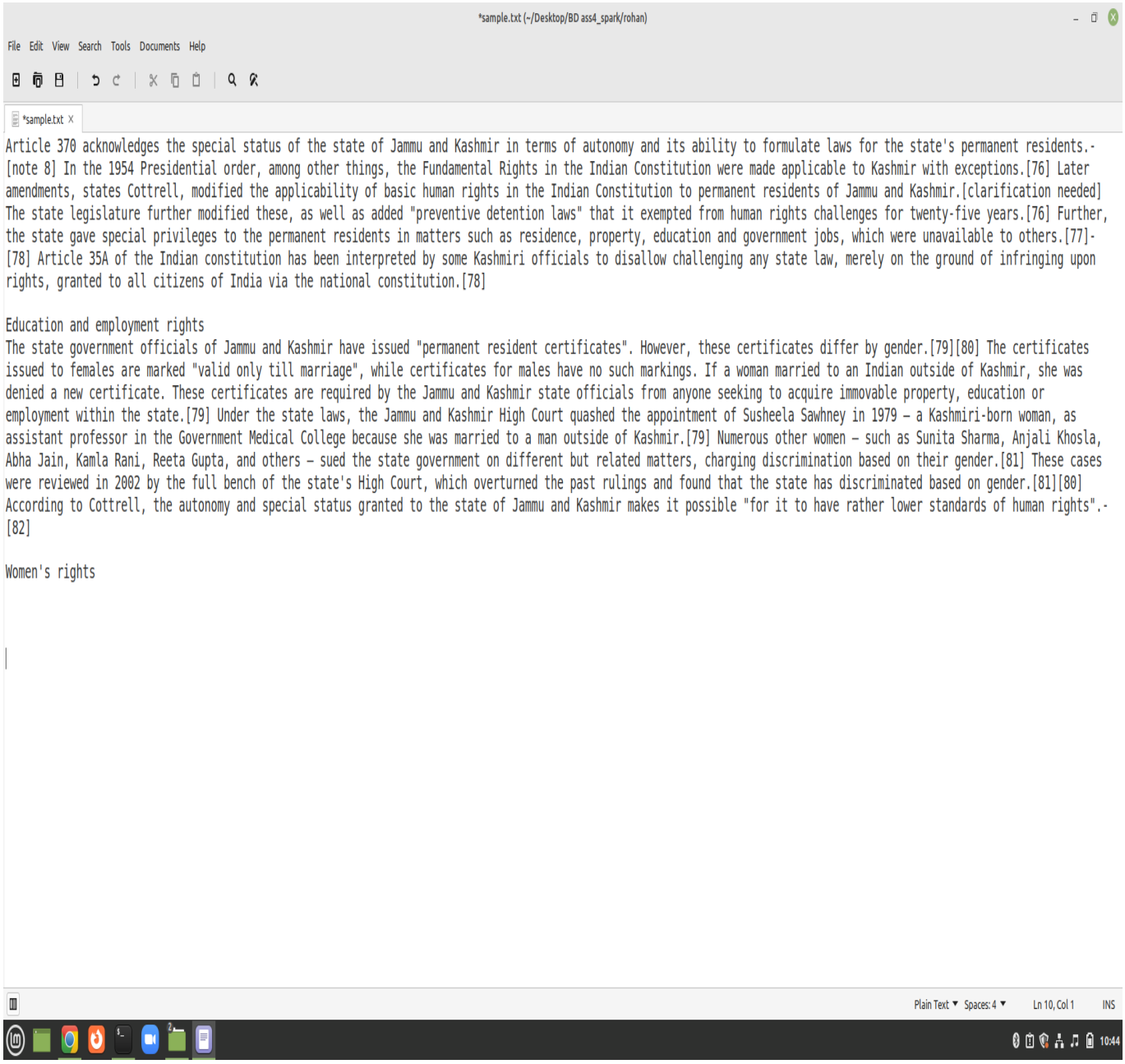


Figure 1: Sample File

Open the spark shell using

```
$ spark-shell
```

Command Prompt will look like

```
rohan@rohan-HP-Laptop-15-bs1xx:~$ spark-shell
22/04/15 17:13:36 WARN Utils: Your hostname, rohan-HP-Laptop-15-bs1xx
resolves to a loopback address: 127.0.1.1; using 192.168.252.38 instead
(on interface wlo1)
22/04/15 17:13:36 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to
another address
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.
properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use
setLogLevel(newLevel).
22/04/15 17:13:41 WARN NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
22/04/15 17:13:42 WARN Utils: Service 'SparkUI' could not bind on port
4040. Attempting port 4041.
Spark context Web UI available at http://192.168.252.38:4041
Spark context available as 'sc' (master = local[*],
app id = local-1650023022362).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.2.1
      /_/

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.
```

To read file from the local machine use

```
val data = sc.textFile("sample.txt")
```

To check the content of values in file use

```
data.collect
```

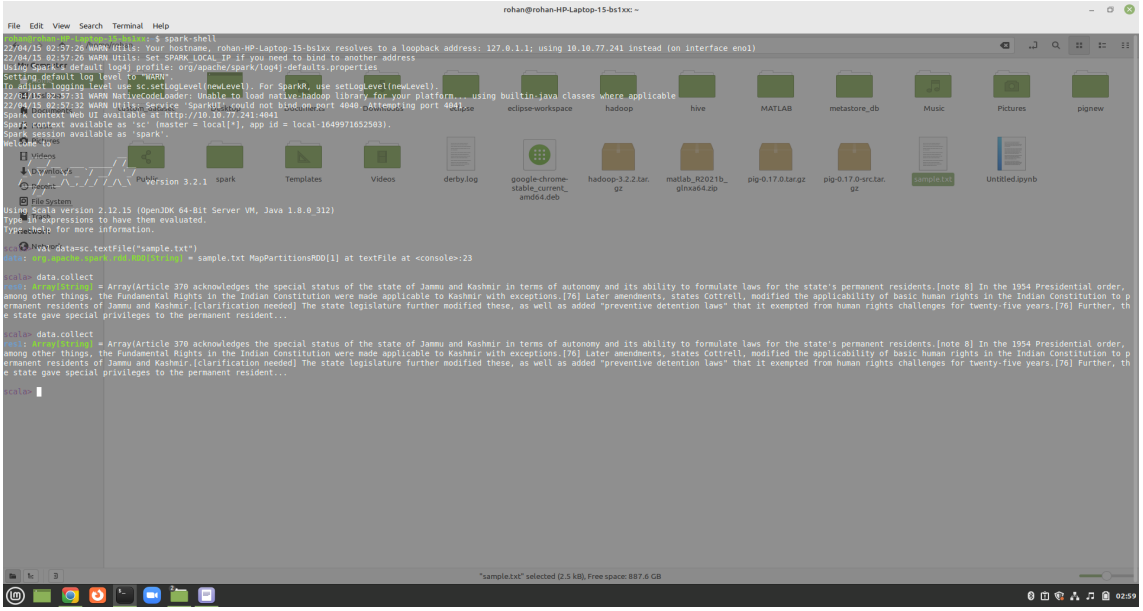


Figure 2: data.collect

To split words use:

```
val splitdata = data.flatMap(line => line.split(' '))
```

To check

```
splitdata.collect
```

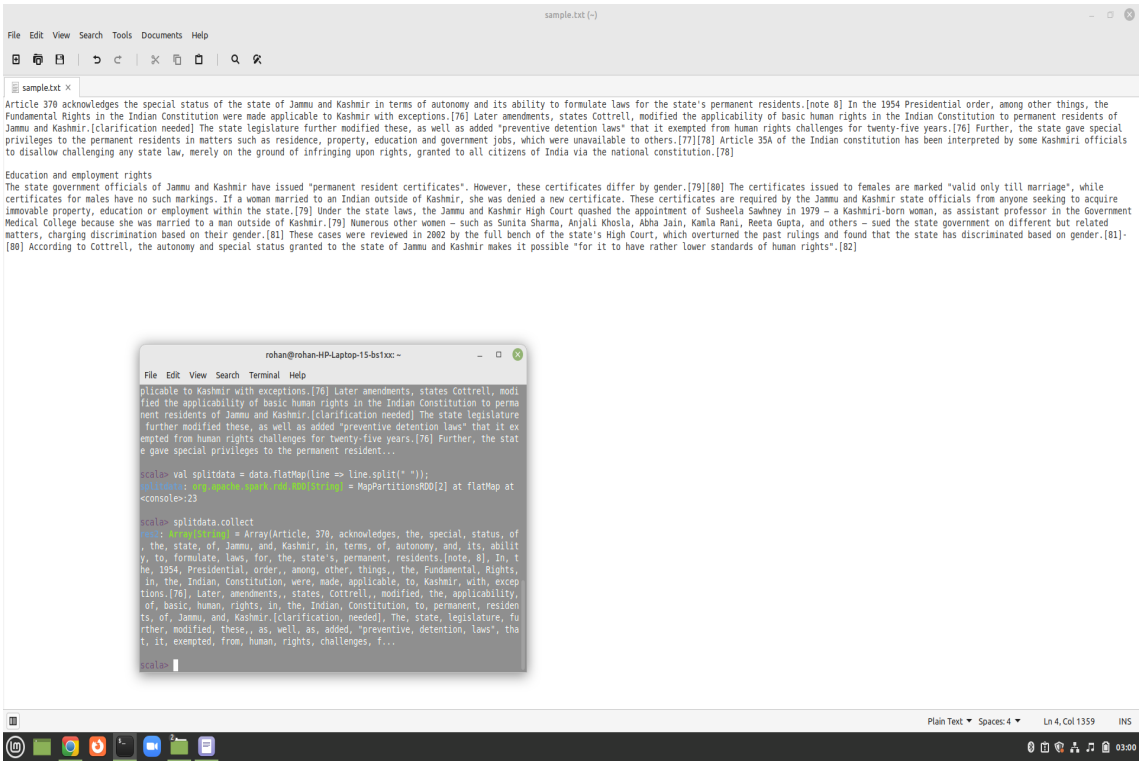


Figure 3: splitdata.collect

Now use map function:

```
val mapdata = splitdata.map(word => (word,1))
```

to chech use

```
mapdata.collect
```

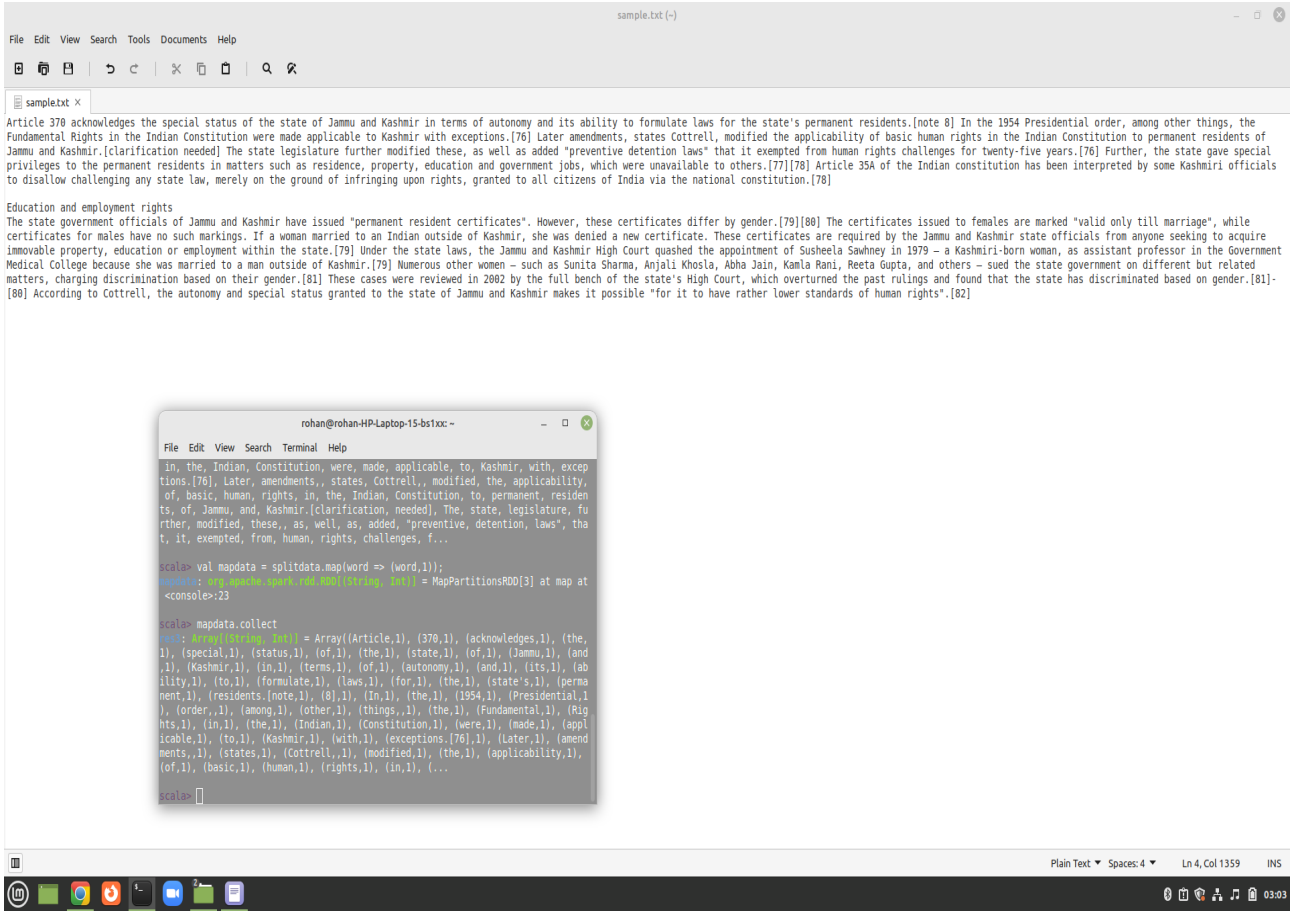


Figure 4: mapdata.collect

Now perform reduce operation:

```
val reducedata = mapdata.reduceByKey(_+_)
```

To check the output use

```
reducedata.collect
```

Output of 1 a) 1st part

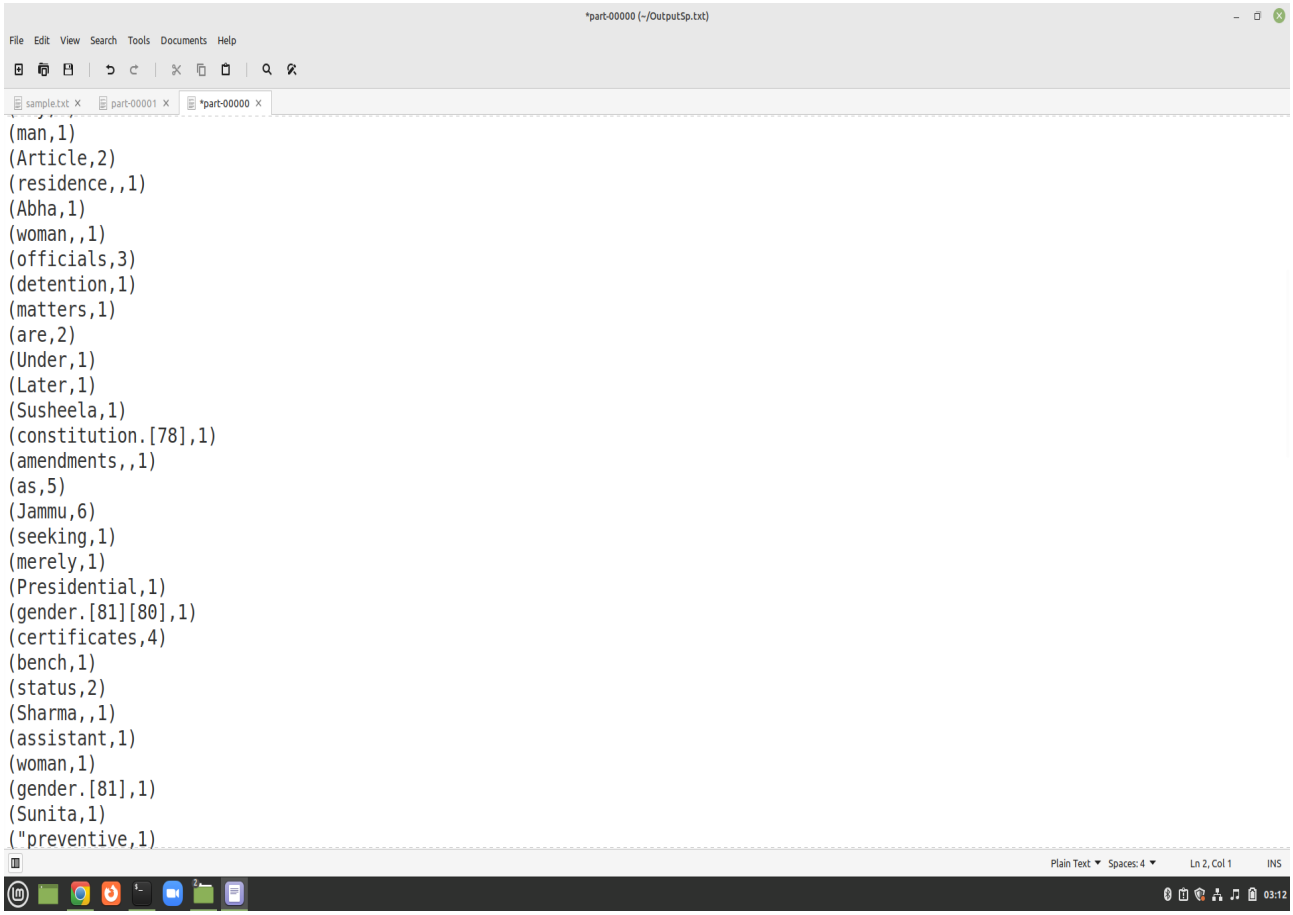


Figure 5: output.collect

0.0.1 For 1 a) 2nd part

To run wordcount in hdfs mode using spark
Need to firstly put the sample.txt file into hdfs using

```
$ hdfs dfs -put sample.txt /
```

This will copy the sample file from your home directory to hdfs

- Now proceed the same process as above need to specify the location of the file in hdfs us the command

```
val data = sc.textFile("hdfs://localhost:9000/sample.txt")
```

now proceed the same above process

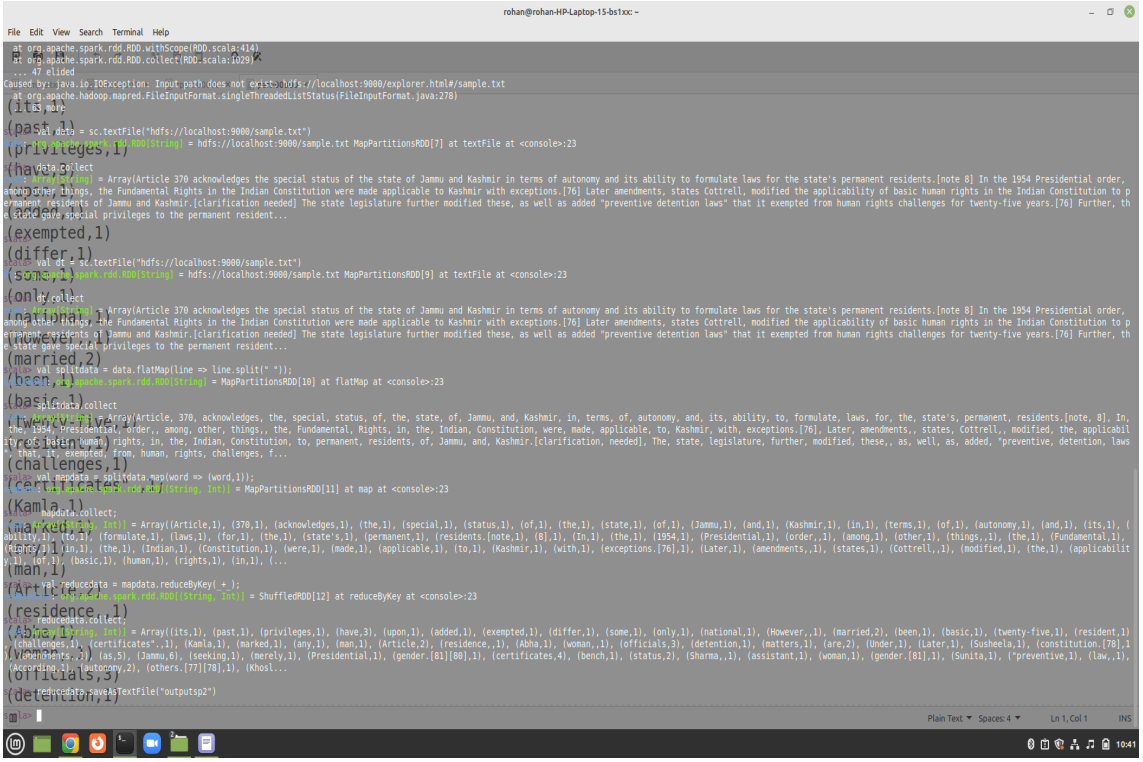


Figure 6: in hdfs mode

output

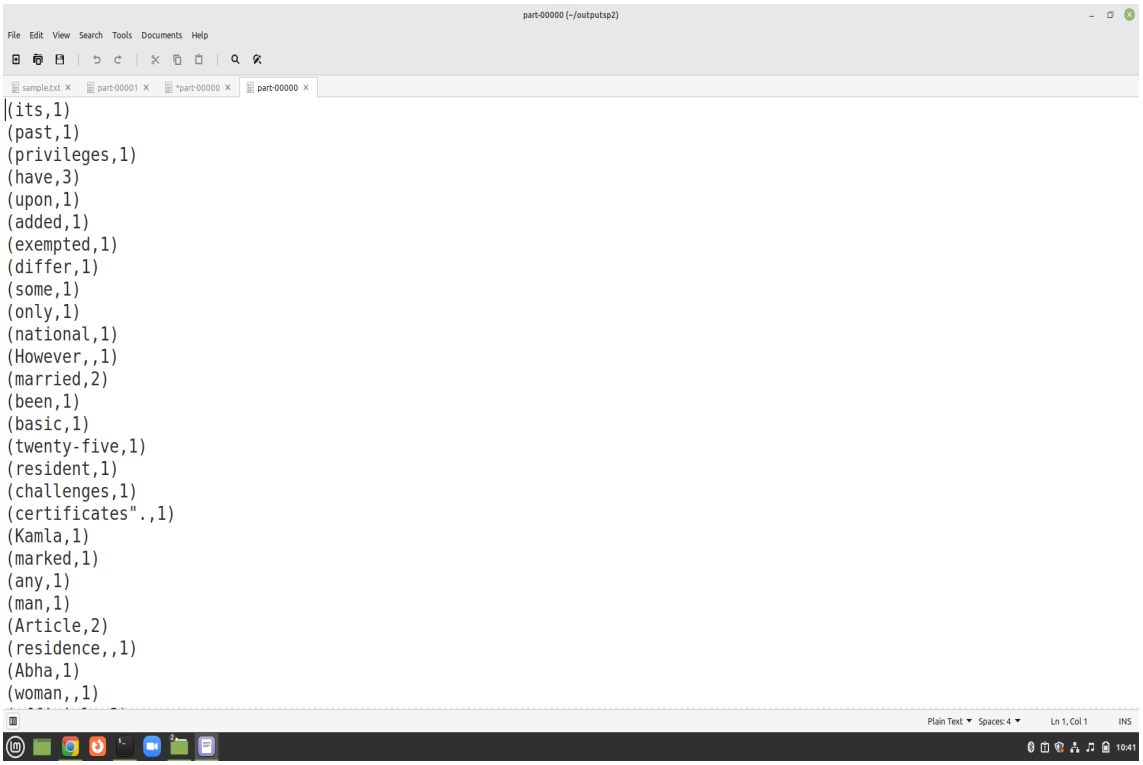


Figure 7: Output

(b) Install PySpark and repeat 1-(a) by doing the word count through it.

Use pip to install pyspark

```
pip install pyspark
```

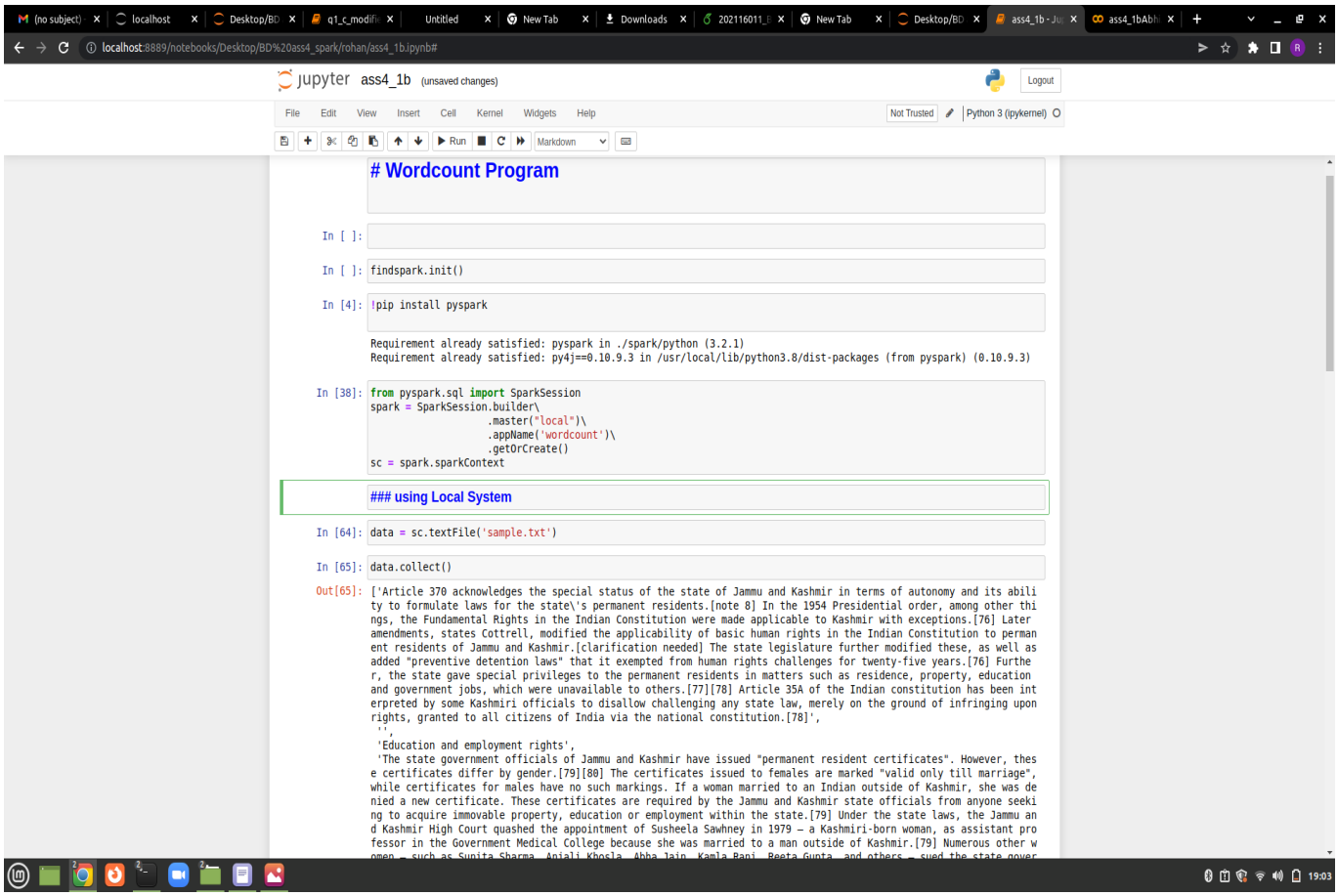
Now add location of pyspark in .bashrc file

```
#
export PYSPARK_PYTHON=/usr/bin/python3.8
#pyspark
export PYSPARK_DRIVER_PYTHON="jupyter"
export PYSPARK_DRIVER_PYTHON_OPTS='notebook'
```

To save the changes done in .bashrc file use the command in command prompt

```
$ sourec ~/.bashrc
```

Jupyter notebook file



jupyter ass4_1b (unsaved changes)

In [46]:
```
splitdata = data.flatMap(lambda line: line.split(' '))
```

In [47]:
```
splitdata.collect()
```

Out[47]:
```
['Article',
 '370',
 'acknowledges',
 'the',
 'special',
 'status',
 'of',
 'the',
 'state',
 'of',
 'Jammu',
 'and',
 'Kashmir',
 'in',
 'terms',
 'of',
 'autonomy',
 'and',
 'its',
```

In [49]:
```
mapdata = splitdata.map(lambda word: (word,1))
```

In [50]:
```
mapdata.collect()
```

Out[50]:
```
[('Article', 1),
 ('370', 1),
 ('acknowledges', 1),
 ('the', 1),
 ('special', 1),
 ('status', 1),
 ('of', 1),
 ('the', 1),
 ('state', 1),
 ('of', 1),
 ('Jammu', 1),
 ('and', 1),
 ('Kashmir', 1),
 ('in', 1),
 ('terms', 1),
 ('of', 1),
 ('autonomy', 1),
 ('and', 1),
 ('its', 1),
```

jupyter ass4_1b (autosaved)

```
 ('8]', 1),
 ('1954', 1),
 ('Presidential', 1),
 ('order,', 1),
 ('other', 2),
 ('things,', 1),
 ('Rights', 1),
 ('Indian', 4),
 ('Cottrell,', 2),
 ('modified', 2),
 ('applicability', 1),
 ('basic', 1),
 ('human', 3),
```

In [ ]:

Using HDFS mode

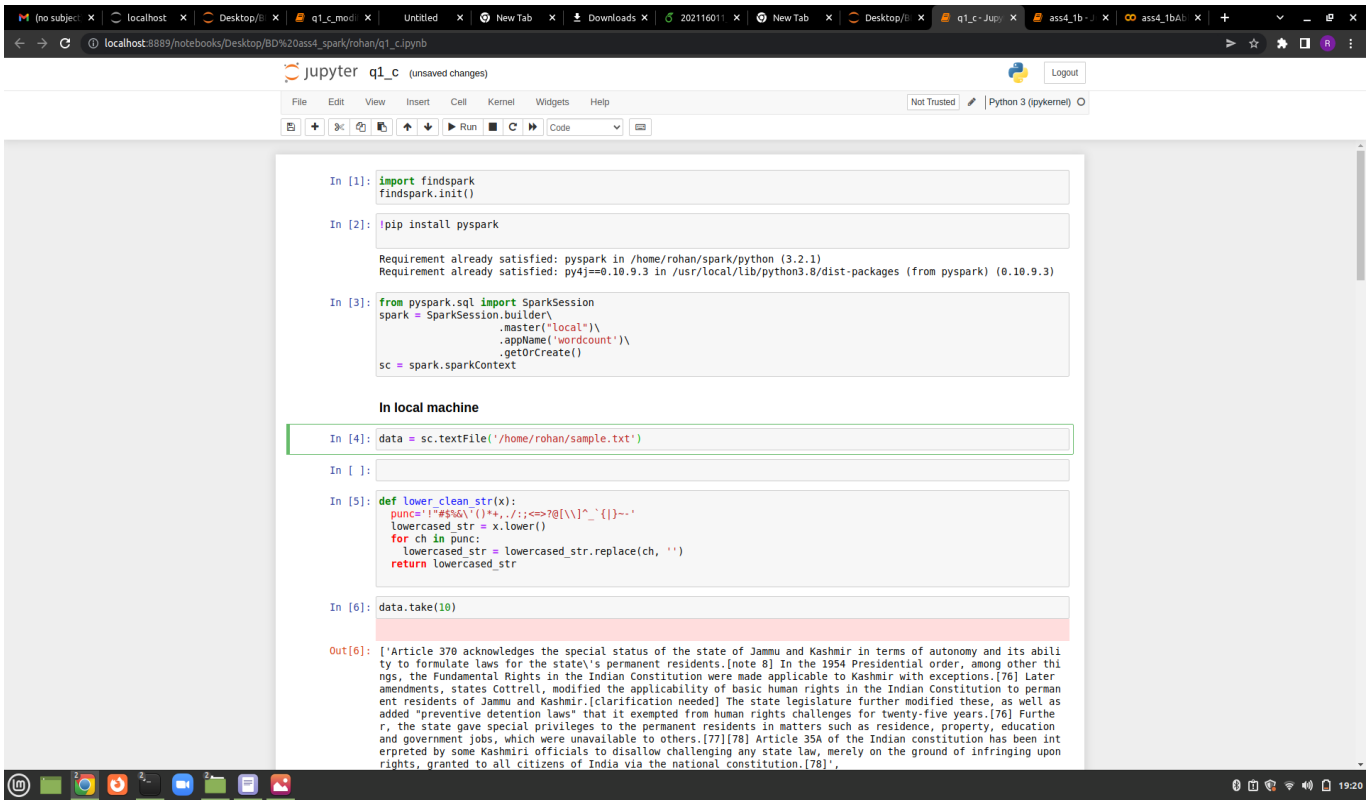
In [59]:
```
# b 2nd part
data2 = sc.textFile('hdfs://localhost:9000/sample.txt')
```

In [62]:
```
out = data2.flatMap(lambda line: line.split(' '))\
           .map(lambda word: (word,1))\
           .reduceByKey(lambda x,y : x+y)
```

In [63]:
```
out.collect()
```

Out[63]:
```
[('370', 1),
 ('of', 15),
 ('state', 10),
 ('Jammu', 6),
 ('in', 7),
 ('permanent', 3),
 ('8]', 1),
 ('1954', 1),
 ('Presidential', 1),
 ('order,', 1),
 ('other', 2),
 ('things,', 1),
 ('Rights', 1),
 ('Indian', 4),
 ('Cottrell,', 2),
 ('modified', 2),
 ('applicability', 1),
 ('basic', 1),
 ('human', 3),
```

(c) Modify your code lines for 1-(a) and 1-(b) such that case-insensitive counting of words is done.



```
In [1]: import findspark
        findspark.init()

In [2]: !pip install pyspark

        Requirement already satisfied: pyspark in /home/rohan/spark/python (3.2.1)
        Requirement already satisfied: py4j==0.10.9.3 in /usr/local/lib/python3.8/dist-packages (from pyspark) (0.10.9.3)

In [3]: from pyspark.sql import SparkSession
        spark = SparkSession.builder\
                            .master("local")\
                            .appName('wordcount')\
                            .getOrCreate()
        sc = spark.sparkContext
```

In local machine

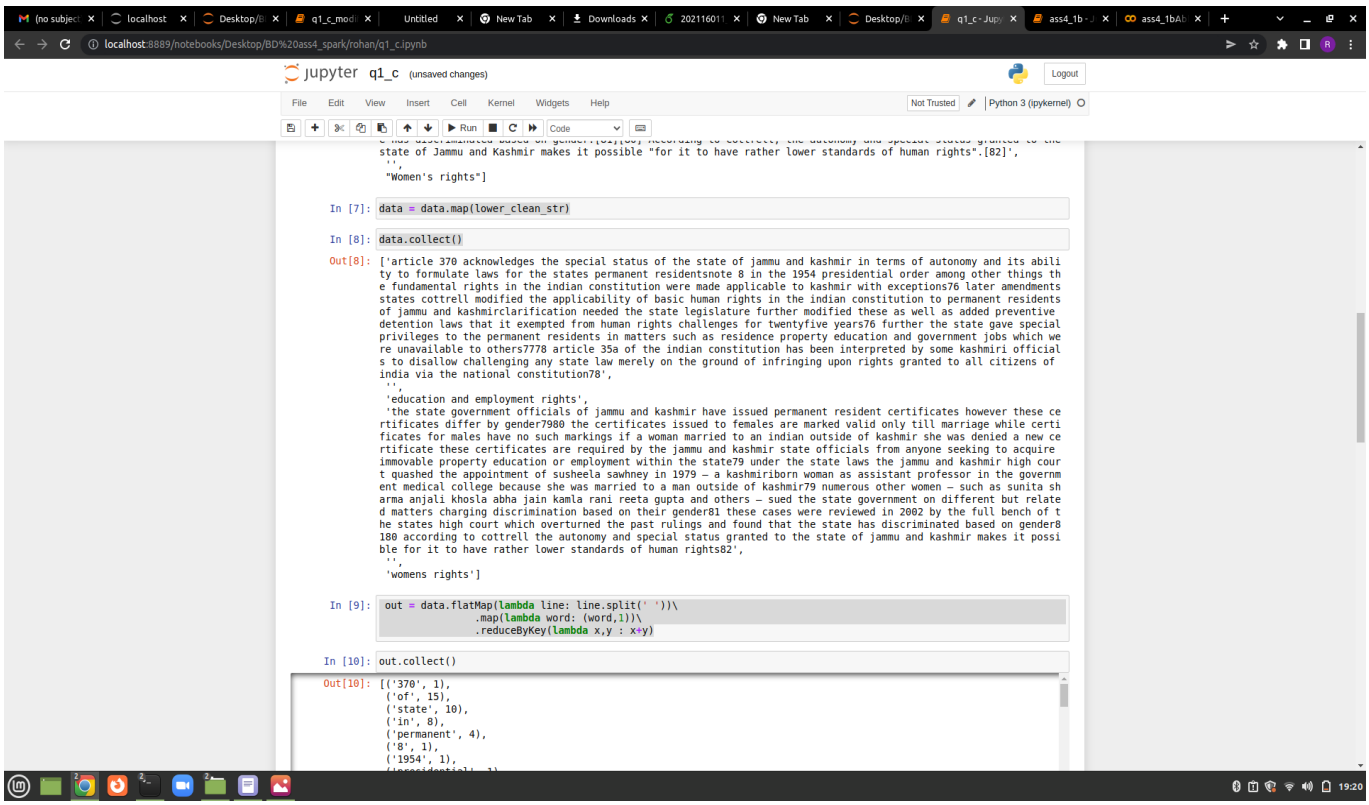
```
In [4]: data = sc.textFile('/home/rohan/sample.txt')

In [ ]:

In [5]: def lower_clean_str(x):
          punc='!"#$%&\'()*+,./:;<=>?@[\\]^_`{|}~-'
          lowercased_str = x.lower()
          for ch in punc:
            lowercased_str = lowercased_str.replace(ch, '')
          return lowercased_str

In [6]: data.take(10)

Out[6]: ['Article 370 acknowledges the special status of the state of Jammu and Kashmir in terms of autonomy and its abili
        ty to formulate laws for the state\'s permanent residents.[note 8] In the 1954 Presidential order, among other thi
        ngs, the Fundamental Rights in the Indian Constitution were made applicable to Kashmir with exceptions.[76] Later
        amendments, states Cottrell, modified the applicability of basic human rights in the Indian Constitution to perman
        ent residents of Jammu and Kashmir.[clarification needed] The state legislature further modified these, as well as
        added "preventive detention laws" that it exempted from human rights challenges for twenty-five years.[76] Furthe
        r, the state gave special privileges to the permanent residents in matters such as residence, property, education
        and government jobs, which were unavailable to others.[77][78] Article 35A of the Indian constitution has been int
        erpreted by some Kashmiri officials to disallow challenging any state law, merely on the ground of infringing upon
        rights, granted to all citizens of India via the national constitution.[78]',
```



```
        state of Jammu and Kashmir makes it possible "for it to have rather lower standards of human rights".[82]',
         '',
         "Women's rights"]

In [7]: data = data.map(lower_clean_str)

In [8]: data.collect()

Out[8]: ['article 370 acknowledges the special status of the state of jammu and kashmir in terms of autonomy and its abili
        ty to formulate laws for the states permanent residentsnote 8 in the 1954 presidential order among other things th
        e fundamental rights in the indian constitution were made applicable to kashmir with exceptions76 later amendments
        states cottrell modified the applicability of basic human rights in the indian constitution to permanent residents
        of jammu and kashmirclarification needed the state legislature further modified these as well as added preventive
        detention laws that it exempted from human rights challenges for twentyfive years76 further the state gave special
        privileges to the permanent residents in matters such as residence property education and government jobs which we
        re unavailable to others7778 article 35a of the indian constitution has been interpreted by some kashmiri official
        s to disallow challenging any state law merely on the ground of infringing upon rights granted to all citizens of
        india via the national constitution78',
         '',
         'education and employment rights',
         'the state government officials of jammu and kashmir have issued permanent resident certificates however these ce
        rtificates differ by gender7980 the certificates issued to females are marked valid only till marriage while certi
        ficates for males have no such markings if a woman married to an indian outside of kashmir she was denied a new ce
        rtificate these certificates are required by the jammu and kashmir state officials from anyone seeking to acquire
        immovable property education or employment within the state79 under the state laws the jammu and kashmir high cour
        t quashed the appointment of susheela sawhney in 1979 – a kashmiriborn woman as assistant professor in the governm
        ent medical college because she was married to a man outside of kashmir79 numerous other women – such as sunita sh
        arma anjali khosla abha jain kamla rani reeta gupta and others – sued the state government on different but relate
        d matters charging discrimination based on their gender81 these cases were reviewed in 2002 by the full bench of t
        he states high court which overturned the past rulings and found that the state has discriminated based on gender8
        180 according to cottrell the autonomy and special status granted to the state of jammu and kashmir makes it possi
        ble for it to have rather lower standards of human rights82',
         '',
         'womens rights']

In [9]: out = data.flatMap(lambda line: line.split(' '))\
                  .map(lambda word: (word,1))\
                  .reduceByKey(lambda x,y : x+y)

In [10]: out.collect()

Out[10]: [('370', 1),
         ('of', 15),
         ('state', 10),
         ('in', 8),
         ('permanent', 4),
         ('8', 1),
         ('1954', 1),
```

localhost:8889/notebooks/Desktop/BD%20ass4_spark/rohan/q1_c.ipynb

jupyter q1_c (autosaved)

File Edit View Insert Cell Kernel Widgets Help

```
('needed', 1),
('further', 2),
```

In [ ]:

In [11]:
```
### In hdfs mode
```

In [12]:
```
data2 = sc.textFile('/home/rohan/sample.txt')
```

In [13]:
```
data2.collect()
```

Out[13]:
```
['Article 370 acknowledges the special status of the state of Jammu and Kashmir in terms of autonomy and its abili
ty to formulate laws for the state\'s permanent residents.[note 8] In the 1954 Presidential order, among other thi
ngs, the Fundamental Rights in the Indian Constitution were made applicable to Kashmir with exceptions.[76] Later
amendments, states Cottrell, modified the applicability of basic human rights in the Indian Constitution to perman
ent residents of Jammu and Kashmir.[clarification needed] The state legislature further modified these, as well as
added "preventive detention laws" that it exempted from human rights challenges for twenty-five years.[76] Furthe
r, the state gave special privileges to the permanent residents in matters such as residence, property, education
and government jobs, which were unavailable to others.[77][78] Article 35A of the Indian constitution has been int
erpreted by some Kashmiri officials to disallow challenging any state law, merely on the ground of infringing upon
rights, granted to all citizens of India via the national constitution.[78]',
 '',
 'Education and employment rights',
 'The state government officials of Jammu and Kashmir have issued "permanent resident certificates". However, thes
e certificates differ by gender.[79][80] The certificates issued to females are marked "valid only till marriage",
while certificates for males have no such markings. If a woman married to an Indian outside of Kashmir, she was de
nied a new certificate. These certificates are required by the Jammu and Kashmir state officials from anyone seeki
ng to acquire immovable property, education or employment within the state.[79] Under the state laws, the Jammu an
d Kashmir High Court quashed the appointment of Susheela Sawhney in 1979 — a Kashmiri-born woman, as assistant pro
fessor in the Government Medical College because she was married to a man outside of Kashmir.[79] Numerous other w
omen — such as Sunita Sharma, Anjali Khosla, Abha Jain, Kamla Rani, Reeta Gupta, and others — sued the state gover
nment on different but related matters, charging discrimination based on their gender.[81] These cases were review
ed in 2002 by the full bench of the state\'s High Court, which overturned the past rulings and found that the stat
e has discriminated based on gender.[81][80] According to Cottrell, the autonomy and special status granted to the
state of Jammu and Kashmir makes it possible "for it to have rather lower standards of human rights".[82]',
 '',
 "Women's rights"]
```

In [14]:
```
def lower_clean_str(x):
  punc='!"#$%&\'()*+,./:;<=>?@[\\]^_`{|}~-'
  lowercased_str = x.lower()
  for ch in punc:
    lowercased_str = lowercased_str.replace(ch, '')
  return lowercased_str
```

In [15]:
```
data2 = data2.map(lower_clean_str)
```

localhost:8889/notebooks/Desktop/BD%20ass4_spark/rohan/q1_c.ipynb

jupyter q1_c (autosaved)

File Edit View Insert Cell Kernel Widgets Help

```
 '',
 "Women's rights"]
```

In [14]:
```
def lower_clean_str(x):
  punc='!"#$%&\'()*+,./:;<=>?@[\\]^_`{|}~-'
  lowercased_str = x.lower()
  for ch in punc:
    lowercased_str = lowercased_str.replace(ch, '')
  return lowercased_str
```

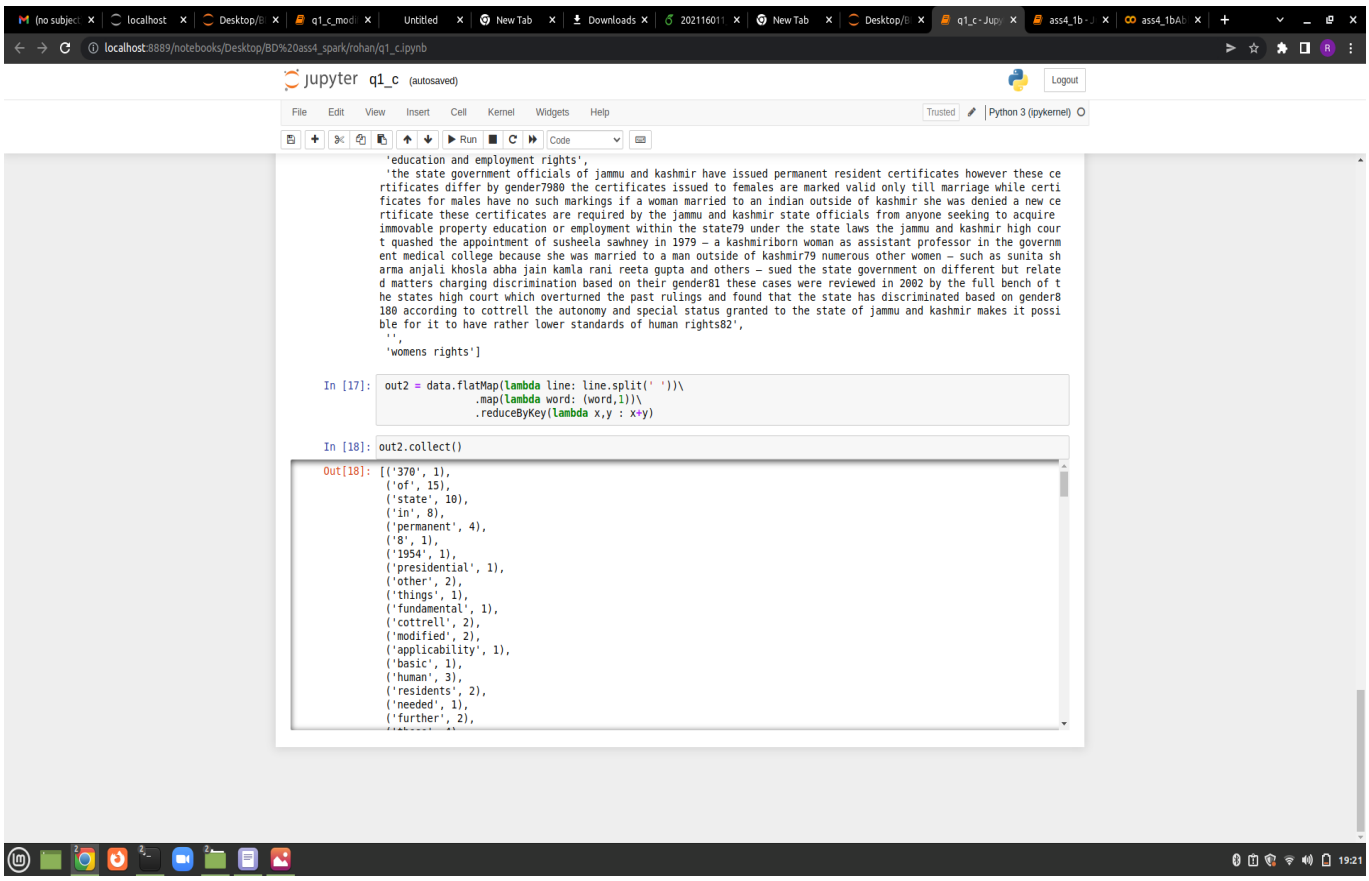
In [15]:
```
data2 = data2.map(lower_clean_str)
```

In [16]:
```
data2.collect()
```

Out[16]:
```
['article 370 acknowledges the special status of the state of jammu and kashmir in terms of autonomy and its abili
ty to formulate laws for the states permanent residentsnote 8 in the 1954 presidential order among other things th
e fundamental rights in the indian constitution were made applicable to kashmir with exceptions76 later amendments
states cottrell modified the applicability of basic human rights in the indian constitution to permanent residents
of jammu and kashmirclarification needed the state legislature further modified these as well as added preventive
detention laws that it exempted from human rights challenges for twentyfive years76 further the state gave special
privileges to the permanent residents in matters such as residence property education and government jobs which we
re unavailable to others7778 article 35a of the indian constitution has been interpreted by some kashmiri official
s to disallow challenging any state law merely on the ground of infringing upon rights granted to all citizens of
india via the national constitution78',
 '',
 'education and employment rights',
 'the state government officials of jammu and kashmir have issued permanent resident certificates however these ce
rtificates differ by gender7980 the certificates issued to females are marked valid only till marriage while certi
ficates for males have no such markings if a woman married to an indian outside of kashmir she was denied a new ce
rtificate these certificates are required by the jammu and kashmir state officials from anyone seeking to acquire
immovable property education or employment within the state79 under the state laws the jammu and kashmir high cour
t quashed the appointment of susheela sawhney in 1979 — a kashmiriborn woman as assistant professor in the governm
ent medical college because she was married to a man outside of kashmir79 numerous other women — such as sunita sh
arma anjali khosla abha jain kamla rani reeta gupta and others — sued the state government on different but relate
d matters charging discrimination based on their gender81 these cases were reviewed in 2002 by the full bench of t
he states high court which overturned the past rulings and found that the state has discriminated based on gender8
180 according to cottrell the autonomy and special status granted to the state of jammu and kashmir makes it possi
ble for it to have rather lower standards of human rights82',
 '',
 'womens rights']
```

In [17]:
```
out2 = data.flatMap(lambda line: line.split(' '))\
            .map(lambda word: (word,1))\
            .reduceByKey(lambda x,y : x+y)
```

In [18]:
```
out2.collect()
```



Note:
Python file is attached with the submission files
Screen shots are from the following python file