

Aisha Muhammad Nawaz L200921

PySpark QUIZ # 2 8A BSCS MMD

19th March 2024

Instructions: Sentimental Analysis on Tweets 3 files given: tweets,negative,positive Output count and categorized tweets in two files, ignore words whose length is less than 4

```
In [2]: # #Running on Colab
!pip install pyspark
!pip install -U -q PyDrive
!apt install openjdk-8-jdk-headless -qq
import os
os.environ['JAVA_HOME'] = '/usr/lib/jvm/java-8-openjdk-amd64'
```

Collecting pyspark  
 Downloading pyspark-3.5.1.tar.gz (317.0 MB)  
 317.0/317.0 MB 3.5 MB/s eta 0:00:00  
 Preparing metadata (setup.py) ... done  
 Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.10/dist-packages (from pyspark) (0.10.9.7)  
 Building wheels for collected packages: pyspark  
 Building wheel for pyspark (setup.py) ... done  
 Created wheel for pyspark: filename=pyspark-3.5.1-py2.py3-none-any.whl size=317488491 sha256=4f3387353f2c2e7e095aa5204de3fb8425dc23ea4ed1a7269d4659adb01995ff  
 Stored in directory: /root/.cache/pip/wheels/80/1d/60/2c256ed38dddc2fdd93be545214a63e02fbd8d74fb0b7f3a6  
 Successfully built pyspark  
 Installing collected packages: pyspark  
 Successfully installed pyspark-3.5.1  
 The following additional packages will be installed:  
 libxtst6 openjdk-8-jre-headless  
 Suggested packages:  
 openjdk-8-demo openjdk-8-source libnss-mdns fonts-dejavu-extra fonts-nanum fonts-ipafont-gothic  
 fonts-ipafont-mincho fonts-wqy-microhei fonts-wqy-zenhei fonts-indic  
 The following NEW packages will be installed:  
 libxtst6 openjdk-8-jdk-headless openjdk-8-jre-headless  
 0 upgraded, 3 newly installed, 0 to remove and 45 not upgraded.  
 Need to get 39.7 MB of archives.  
 After this operation, 144 MB of additional disk space will be used.  
 Selecting previously unselected package libxtst6:amd64.  
 (Reading database ... 121918 files and directories currently installed.)  
 Preparing to unpack .../libxtst6\_2%3a1.2.3-1build4\_amd64.deb ...  
 Unpacking libxtst6:amd64 (2:1.2.3-1build4) ...  
 Selecting previously unselected package openjdk-8-jre-headless:amd64.  
 Preparing to unpack .../openjdk-8-jre-headless\_8u402-ga-2ubuntu1~22.04\_amd64.deb ...  
 Unpacking openjdk-8-jre-headless:amd64 (8u402-ga-2ubuntu1~22.04) ...  
 Selecting previously unselected package openjdk-8-jdk-headless:amd64.  
 Preparing to unpack .../openjdk-8-jdk-headless\_8u402-ga-2ubuntu1~22.04\_amd64.deb ...  
 Unpacking openjdk-8-jdk-headless:amd64 (8u402-ga-2ubuntu1~22.04) ...  
 Setting up libxtst6:amd64 (2:1.2.3-1build4) ...  
 Setting up openjdk-8-jre-headless:amd64 (8u402-ga-2ubuntu1~22.04) ...  
 update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/orbd to provide /usr/bin/orbd (orbd) in auto mode  
 update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/servertool to provide /usr/bin/servertool (servertool) in auto mode  
 update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/tnameserv to provide /usr/bin/tnameserv (tnameserv) in auto mode  
 Setting up openjdk-8-jdk-headless:amd64 (8u402-ga-2ubuntu1~22.04) ...  
 update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/clhsdb to provide /usr/bin/clhsdb (clhsdb) in auto mode  
 update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/extcheck to provide /usr/bin/extcheck (extcheck) in auto mode  
 update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/hsdb to provide /usr/bin/hsdb (hsdb) in auto mode  
 update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/idlj to provide /usr/bin/idlj (idlj) in auto mode  
 update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/javah to provide /usr/bin/javah (javah) in auto mode  
 update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/jhat to provide /usr/bin/jhat (jhat) in auto mode  
 update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/jsadebugd to provide /usr/bin/jsadebugd (jsadebugd) in auto mode  
 update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/native2ascii to provide /usr/bin/native2ascii (native2ascii) in auto mode  
 update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/schemagen to provide /usr/bin/schemagen (schemagen) in auto mode  
 update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/wsgen to provide /usr/bin/wsgen (wsgen) in auto mode  
 update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/wsimport to provide /usr/bin/wsimport (wsimport) in auto mode  
 update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/xjc to provide /usr/bin/xjc (xjc) in auto mode  
 Processing triggers for libc-bin (2.35-0ubuntu3.4) ...  
 /sbin/ldconfig.real: /usr/local/lib/libtbbbind\_2\_5.so.3 is not a symbolic link  
  
 /sbin/ldconfig.real: /usr/local/lib/libtbb.so.12 is not a symbolic link  
  
 /sbin/ldconfig.real: /usr/local/lib/libtbbmalloc.so.2 is not a symbolic link  
  
 /sbin/ldconfig.real: /usr/local/lib/libtbbbind.so.3 is not a symbolic link  
  
 /sbin/ldconfig.real: /usr/local/lib/libtbbmalloc\_proxy.so.2 is not a symbolic link  
  
 /sbin/ldconfig.real: /usr/local/lib/libtbbbind\_2\_0.so.3 is not a symbolic link

In [ ]: `!sudo apt update`

```

Get:1 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64 InRelease [1,581 B]
Hit:2 http://archive.ubuntu.com/ubuntu jammy InRelease
Get:3 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ InRelease [3,626 B]
Get:4 http://security.ubuntu.com/ubuntu jammy-security InRelease [110 kB]
Get:5 http://archive.ubuntu.com/ubuntu jammy-updates InRelease [119 kB]
Get:6 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64 Packages [810 kB]
Hit:7 https://ppa.launchpadcontent.net/c2d4u.team/c2d4u4.0+/ubuntu jammy InRelease
Hit:8 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy InRelease
Hit:9 http://archive.ubuntu.com/ubuntu jammy-backports InRelease
Hit:10 https://ppa.launchpadcontent.net/graphics-drivers/ppa/ubuntu jammy InRelease
Hit:11 https://ppa.launchpadcontent.net/ubuntugis/ppa/ubuntu jammy InRelease
Get:12 http://security.ubuntu.com/ubuntu jammy-security/main amd64 Packages [1,695 kB]
Get:13 http://security.ubuntu.com/ubuntu jammy-security/universe amd64 Packages [1,082 kB]
Get:14 http://archive.ubuntu.com/ubuntu jammy-updates/universe amd64 Packages [1,375 kB]
Get:15 http://archive.ubuntu.com/ubuntu jammy-updates/restricted amd64 Packages [2,242 kB]
Get:16 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 Packages [1,975 kB]
Get:17 http://archive.ubuntu.com/ubuntu jammy-updates/multiverse amd64 Packages [51.1 kB]
Fetched 9,465 kB in 4s (2,236 kB/s)
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
46 packages can be upgraded. Run 'apt list --upgradable' to see them.

```

```

In [3]: # Import the Libraries we will need
import pyspark
from pyspark.sql import *
from pyspark.sql.functions import *
from pyspark import SparkContext, SparkConf

# Create Spark session and ContextRun PySpark.
# create the session
conf = SparkConf().set("spark.ui.port", "4050")
# create the context
sc = pyspark.SparkContext(conf=conf)
spark = SparkSession.builder.appName("DataFrame").config('spark.ui.port', '4050').getOrCreate()
spark

```

Out[3]: **SparkSession - in-memory****SparkContext**[Spark UI \(http://f7f289dd7b96:4050/\)](http://f7f289dd7b96:4050/)**Version**

v3.5.1

**Master**

local[\*]

**AppName**

pyspark-shell

```

In [ ]: tweets = sc.parallelize(['Wow! The weather is great today','This instructor hates students','I was not this bad at maths before','She sucks at lying'])
negative=list(sc.parallelize(['bad','hates','sucks']).collect())
positive=list(sc.parallelize(['great']).collect())

def findCat(line,pos,neg):
    pC=0
    pN=0
    for word in line:
        if(len(word)>4):
            if(word in pos):
                pC=pC+1
            elif(word in neg):
                pN=pN+1
    if(pC>pN):
        return 'Positive'
    else:
        return 'Negative'
tweetsCat=tweets.map(lambda x: (x,findCat(x.split(' '),negative,positive)))
tweetsCount=tweetsCat.map(lambda x: (x[1],x[0])).countByKey()

tweetsCat.saveAsTextFile('Category3.txt')
sc.parallelize([(s,c) for s,c in tweetsCount.items()]).saveAsTextFile('Count3.txt')

```

```
In [ ]: tweets = sc.parallelize(['Wow! The weather is great today', 'This instructor hates students', 'I was not this bad at maths before', 'She sucks at lying'])
negative = sc.parallelize(['bad', 'hates', 'sucks']).collect()
positive = sc.parallelize(['great']).collect()

def findCat(line, negative, positive):
    nC = 0
    nP = 0
    for w in line:
        w = w.lower().strip('.,?!') # Convert to lowercase and remove punctuation
        if len(w) > 4:
            if w in negative:
                nC += 1
            elif w in positive:
                nP += 1
    if nP > nC:
        return 'Positive'
    else:
        return 'Negative'

tweetsCat = tweets.map(lambda x: (x, findCat(x.lower().split(), negative, positive)))
tweetsCount = tweetsCat.map(lambda x: (x[1], 1)).reduceByKey(lambda x,y:x+y)

tweetsCat.saveAsTextFile('Category5.txt')
tweetsCount.saveAsTextFile('Count5.txt')
```

Try Before Final Exam

```
In [5]: tweets=sc.textFile('TWEETS.txt')
PositiveWords=sc.broadcast(set(sc.textFile('Positive_words.txt').collect()))
NegativeWords=sc.broadcast(set(sc.textFile('Negative_words.txt').collect()))

def classifyTweet(line,PositiveWords,NegativeWords):
    words=[w for w in line.split(" ") if len(w)>=4]
    positiveCount=0
    negativeCount=0
    for word in words:
        word=word.lower().replace('!', '').replace('.', '').replace(',', '')
        if (word) in list(PositiveWords):
            positiveCount=positiveCount+1
        elif(word in list(NegativeWords)):
            negativeCount=negativeCount+1

    if(positiveCount>=negativeCount):
        return (line,'Positive')

    else:
        return (line,'Negative')

classifiedTweets=tweets.map(lambda line: classifyTweet(line,PositiveWords.value,NegativeWords.value))
counts=classifiedTweets.map(lambda line: (line[1]+' tweets count: ',1)).reduceByKey(lambda x,y: x+y)

classifiedTweets.saveAsTextFile('TweetSentimentAnalysis4.txt')
counts.saveAsTextFile('TweetsCount2.txt')
```

In [ ]: