Author : Nur Laili Solichah (Astra Data Scientist Bootcamp Batch 2)

Social Media Analysis with SPARK

Pada practice case SPARK, saya menggunakan data "clean_tweet.csv", sesuai dengan instruksi yang diminta. Saya melakukan percobaan cloudera menggunakan dua tools yang berbeda, yaitu JSLinux dan PuTTY.

Python sangat bagus untuk pemodelan sains data, berkat berbagai modul dan paketnya yang membantu mencapai tujuan sains data. Tetapi bagaimana jika data yang Anda tangani tidak bisa masuk ke dalam satu mesin? Mungkin Anda dapat menerapkan pengambilan sampel secara hati-hati untuk melakukan analisis pada satu mesin, tetapi dengan kerangka kerja komputasi terdistribusi seperti PySpark, PySpark dapat secara efisien mengimplementasikan tugas untuk kumpulan data besar. Contohnya pada implementasi social media analysis menggunakan SPARK pada kali ini.

Namun kendala yang dihadapi adalah beberapa library, salah satunya numpy tidak dapat diinstall dan belum disediakan di server cloudera. Oleh karena itu saya hanya mencoba untuk mencoba PySpark dalam eksplorasi data saja.

Berikut merupakan hasil percobaan social media analysis menggunakan PySpark:

1. **JSlinux atau PuTTY**
    a. JSLinux : Masuk ke link https://bellard.org/jslinux/

Author : Nur Laili Solichah (Astra Data Scientist Bootcamp Batch 2)

JSLinux

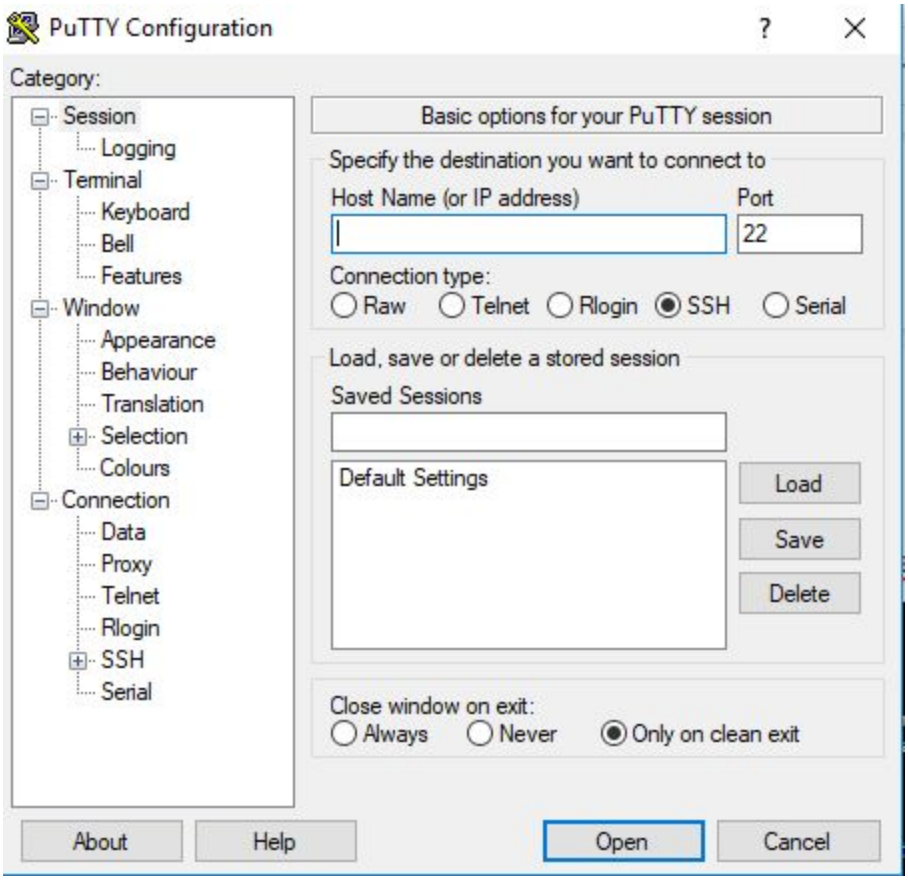
Run Linux or other Operating Systems in your browser!

The following emulated systems are available:

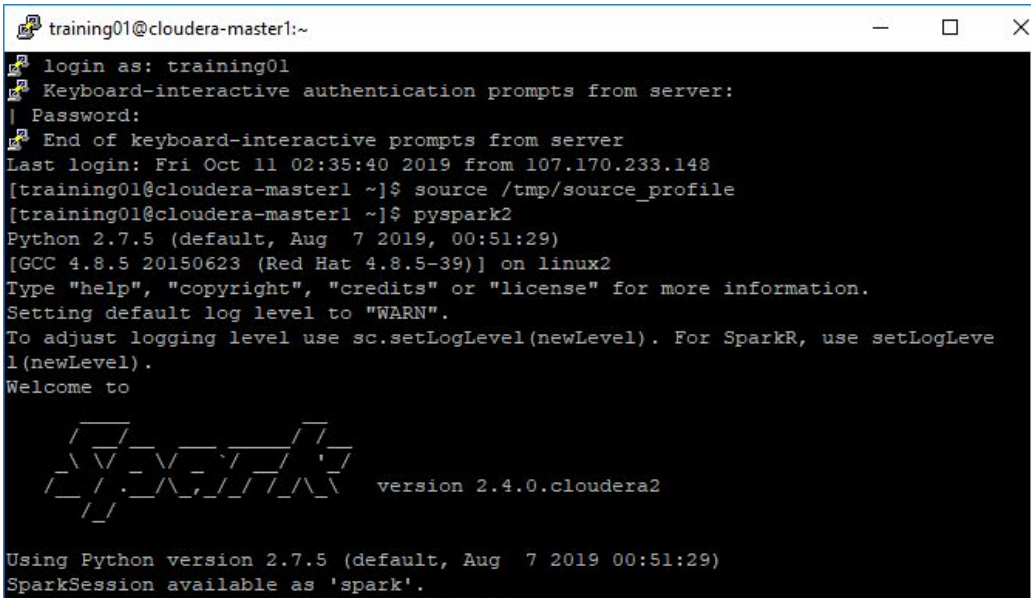
CPU	OS (Distribution)	User Interface	VFsync access	Startup Link	TEMU Config	Comment
x86	Linux 4.12.0 (Buildroot)	Console	Yes	click here	url	
x86	Linux 4.12.0 (Buildroot)	X Window	Yes	click here	url	Right mouse button for the menu.
x86	Windows 2000	Graphical	No	click here	url	Disclaimer.
x86	FreeDOS	VGA Text	No	click here	url	
riscv64	Linux 4.15.0 (Buildroot)	Console	Yes	click here	url	
riscv64	Linux 4.15.0 (Buildroot)	X Window	Yes	click here	url	Right mouse button for the menu.
riscv64	Linux 4.15.0 (Fedora 29)	Console	Yes	click here	url	Warning: longer boot time.
riscv64	Linux 4.15.0 (Fedora 29)	X Window	Yes	click here	url	Warning: longer boot time. Right mouse button for the menu.

Menggunakan Console riscv64 Linux 4.15.0 (Buildroot)

b. PuTTY portable dengan menggunakan Port 22 dan memasukkan Host Name **35.239.158.241**



c. Kemudian masuk ke training01 dan masuk ke pyspark2



d. Setelah itu membaca import SparkSession dan import * agar dapat membaca dataset

```
>>> from pyspark.sql import SparkSession
>>> from pyspark.sql.types import *
```

e. Kemudian membaca dataset dari **/user/cloudera/clean_tweet.csv** dan tampilkan 10 data teratas. Dataset ini memiliki 2 kolom, yaitu text dan target.

```
>>> df = spark.read.csv("/user/cloudera/clean_tweet.csv", header=True)
>>> df.show(10)
+--------------------+------+
|                text|target|
+--------------------+------+
|awww that s a bum...|     0|
|is upset that he ...|     0|
|i dived many time...|     0|
|my whole body fee...|     0|
|no it s not behav...|     0|
|  not the whole crew|     0|
|          need a hug|     0|
|hey long time no ...|     0|
|k nope they didn ...|     0|
|        que me muera|     0|
+--------------------+------+
only showing top 10 rows
```

f. Kemudian **dropna()** untuk menghapus data yang memiliki *null values*, Data awal terdapat 1600000 menjadi 1509626. Data yang berkurang sejumlah 90374.

```
>>> df =  df.dropna()
>>> df.count()
1596754
```

g. Cek tipe dari data atau feature yang terdapat dalam dataframe, menggunakan **dtypes**.

```
>>> df.dtypes
[('text', 'string'), ('target', 'string')]
```

h. Kemudian cek data yang paling teratas dari data Social Media (Twitter) Sentiment Analysis, menggunakan **first()**..

```
>>> df.first()
Row(text=u'awww that s a bummer you shoulda got david carr of third day to do it d
', target=u'0')
```

i. Kemudian implementasi RDD **take(),** dengan memasukkan row kedua dari dataframe.

```
>>> df.take(2)
[Row(text=u'awww that s a bummer you shoulda got david carr of third day to do it
d', target=u'0'), Row(text=u'is upset that he can t update his facebook by texting
 it and might cry as a result school today also blah', target=u'0')]
```

j. Kemudian cek list dari nama kolom dan masing-masing tipe dari kolom, menggunakan **.schema**

```
>>> df.schema
StructType(List(StructField(text,StringType,true),StructField(target,StringType,tr
ue)))
```

k. Kemudian hapus duplikasi data menggunakan **dropDuplicates()**

```
>>> df = df.dropDuplicates()
```

l. Kemudian untuk melihat ringkasan/summary dari data untuk mengidentifikasi *count, mean, stddev, min,* dan *max* menggunakan **describe().**

```
>>> df.describe().show()
+-------+--------------------+--------------------+
|summary|                text|              target|
+-------+--------------------+--------------------+
|  count|             1528140|             1528140|
|   mean|                null|0.49720117266742575|
| stddev|                null| 0.4999923300992921|
|    min|                   a|                   0|
|    max|zzzzzzzzzzzzzzzz...|                   1|
+-------+--------------------+--------------------+
```

m. Kemudian untuk mengetahui jumlah text yang memiliki kelas sentiment positive atau negative dapat menggunakan **groupBy()** berdasarkan target, kemudian dihitung untuk masing-masing target.

```
>>> df.groupBy("target").count().show()
+------+------+
|target| count|
+------+------+
|     0|768347|
|     1|759793|
+------+------+
```

KESIMPULAN

Dari percobaan PySpark yang dilakukan, diketahui bahwa kelas sentiment dapat dijelaskan sebagai berikut:

- Kelas sentimen **0 (negative)** sejumlah 768347
- Kelas sentimen **1 (positive)** sejumlah 759793

Author : Nur Laili Solichah (Astra Data Scientist Bootcamp Batch 2)

Diketahui bahwa jumlah text yang masuk ke target negatif lebih banyak dibandingkan dengan kelas sentimen positif.