Author : Hania Maghfira, Astra Data Scientist Bootcamp 2019

PRACTICE CASE

SOCIAL MEDIA SENTIMENT ANALYSIS WITH SPARK

1. Overview

The purpose of this practice case is analyzing social media especially twitter with Spark. The dataset of this use case is clean_data.csv (can be found on Jupyter Notebook folder). Because of numpy library can not be installed on this Spark, I will tell about data exploration with Spark only.

2. Getting Started

Before we go into the step of analysis, we should login on Spark :

- Open https://bellard.org/jslinux/, then choose Windows 2000.
- Login with the command below :

```
[root@localhost ~]# ssh training02@35.239.158.241
```

- If successful, then input the password Cl0ud3r4*. The successful login status can be seen below.

```
Host '35.239.158.241' is not in the trusted hosts file.
(ecdsa-sha2-nistp256 fingerprint md5 2c:bd:12:14:68:df:e2:27:4e:78:8d:7a:ce:5c:a
2:c9)
Do you want to continue connecting? (y/n) y
training02@35.239.158.241's password:
Last login: Fri Oct 11 03:19:08 2019 from 107.170.233.148
```

- Type the command bellow to use Spark :

```
[training02@cloudera-master1 ~]$ source /tmp/source_profile
[training02@cloudera-master1 ~]$ pyspark2
```

- If successful, the Spark will be shown as below :

```
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.4.0.cloudera2
      /_/
```

3. Steps

The step of analysis social media with Spark are :

1. First of all, import library to load dataframe on Spark.
   /user/cloudera/clean_tweet.csv is the file directory of dataset.

```
>>> from pyspark.sql import SparkSession
>>> from pyspark.sql.types import *
>>> df = spark.read.csv("/user/cloudera/clean_tweet.csv")
```

2. Show dataset by using df.show()

```
>>> df.show()
+--------------------+------+
|                 _c0|   _c1|
+--------------------+------+
|                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|
|spring break in p...|     0|
|i just re pierced...|     0|
|i couldn t bear t...|     0|
|it it counts idk ...|     0|
|i would ve been t...|     0|
|i wish i got to w...|     0|
|hollis death scen...|     0|
| about to file taxes|     0|
|ahh ive always wa...|     0|
+--------------------+------+
only showing top 20 rows
```

3. Count the row of dataset use df.count()

```
>>> df.count()
1600001
```

From the above we can know that dataset has 1600K rows.

4. We need to drop duplicates in dataset, use df.dropna()

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

After we drop duplicates, the dataset now has 1.596.754 rows.