

Final_Project_1

March 10, 2023

Import and Overview

```
[1]: import sys
      print(sys.version)
```

3.8.15 | packaged by conda-forge | (default, Nov 22 2022, 08:46:39)
[GCC 10.4.0]

```
[2]: spark.version
```

```
[2]: '3.1.3'
```

```
[3]: import pandas as pd
      import numpy as np
      pd.set_option('display.max_colwidth', None)
      pd.reset_option('display.max_rows')
      from itertools import compress
      from pyspark.sql.functions import *
      from pyspark.sql.types import *
      import seaborn as sns
      import matplotlib.pyplot as plt
      warnings.filterwarnings(action='ignore')
```

```
[4]: from pyspark.sql import SparkSession
      from pyspark import SparkContext
      from pyspark.sql import SQLContext
      from pyspark.sql import Row
      from pyspark.sql.functions import col
```

```
[5]: spark.conf.set("spark.sql.repl.eagerEval.enabled", True)
```

```
[5]: !pip install simhash
```

Collecting simhash

Downloading simhash-2.1.2-py3-none-any.whl (4.7 kB)

Requirement already satisfied: numpy in

/opt/conda/miniconda3/lib/python3.8/site-packages (from simhash) (1.19.5)

Installing collected packages: simhash

Successfully installed simhash-2.1.2

WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: <https://pip.pypa.io/warnings/venv>

```
[6]: from simhash import Simhash, SimhashIndex
```

Data Import

```
[28]: %%time
twitter = spark.read.json('gs://chen26-bdp/final_project')
```

23/02/25 18:22:13 WARN

org.apache.spark.sql.execution.datasources.SharedInMemoryCache: Evicting cached table partition metadata from memory due to size constraints (spark.sql.hive.filesourcePartitionFileCacheSize = 262144000 bytes). This may impact query planning performance.

CPU times: user 2.88 s, sys: 688 ms, total: 3.57 s

Wall time: 17min 18s

23/02/25 18:38:13 WARN org.apache.spark.sql.catalyst.util.package: Truncated the string representation of a plan since it was too large. This behavior can be adjusted by setting 'spark.sql.debug.maxToStringFields'.

```
[6]: twitter.count()
```

```
[6]: 99994342
```

```
[7]: twitter.printSchema()
```

```
root
|-- coordinates: struct (nullable = true)
|   |-- coordinates: array (nullable = true)
|   |   |-- element: double (containsNull = true)
|   |-- type: string (nullable = true)
|-- created_at: string (nullable = true)
|-- display_text_range: array (nullable = true)
|   |-- element: long (containsNull = true)
|-- entities: struct (nullable = true)
|   |-- hashtags: array (nullable = true)
|   |   |-- element: struct (containsNull = true)
|   |   |   |-- indices: array (nullable = true)
|   |   |   |   |-- element: long (containsNull = true)
|   |   |-- text: string (nullable = true)
```

```

|  |-- media: array (nullable = true)
|  |  |-- element: struct (containsNull = true)
|  |  |  |-- additional_media_info: struct (nullable = true)
|  |  |  |  |-- description: string (nullable = true)
|  |  |  |  |-- embeddable: boolean (nullable = true)
|  |  |  |  |-- monetizable: boolean (nullable = true)
|  |  |  |  |-- title: string (nullable = true)
|  |  |  |-- description: string (nullable = true)
|  |  |  |-- display_url: string (nullable = true)
|  |  |  |-- expanded_url: string (nullable = true)
|  |  |  |-- id: long (nullable = true)
|  |  |  |-- id_str: string (nullable = true)
|  |  |  |-- indices: array (nullable = true)
|  |  |  |  |-- element: long (containsNull = true)
|  |  |  |-- media_url: string (nullable = true)
|  |  |  |-- media_url_https: string (nullable = true)
|  |  |  |-- sizes: struct (nullable = true)
|  |  |  |  |-- large: struct (nullable = true)
|  |  |  |  |  |-- h: long (nullable = true)
|  |  |  |  |  |-- resize: string (nullable = true)
|  |  |  |  |  |-- w: long (nullable = true)
|  |  |  |  |-- medium: struct (nullable = true)
|  |  |  |  |  |-- h: long (nullable = true)
|  |  |  |  |  |-- resize: string (nullable = true)
|  |  |  |  |  |-- w: long (nullable = true)
|  |  |  |  |-- small: struct (nullable = true)
|  |  |  |  |  |-- h: long (nullable = true)
|  |  |  |  |  |-- resize: string (nullable = true)
|  |  |  |  |  |-- w: long (nullable = true)
|  |  |  |  |-- thumb: struct (nullable = true)
|  |  |  |  |  |-- h: long (nullable = true)
|  |  |  |  |  |-- resize: string (nullable = true)
|  |  |  |  |  |-- w: long (nullable = true)
|  |  |  |-- source_status_id: long (nullable = true)
|  |  |  |-- source_status_id_str: string (nullable = true)
|  |  |  |-- source_user_id: long (nullable = true)
|  |  |  |-- source_user_id_str: string (nullable = true)
|  |  |  |-- type: string (nullable = true)
|  |  |  |-- url: string (nullable = true)
|  |-- symbols: array (nullable = true)
|  |  |-- element: struct (containsNull = true)
|  |  |  |-- indices: array (nullable = true)
|  |  |  |  |-- element: long (containsNull = true)
|  |  |  |-- text: string (nullable = true)
|  |-- urls: array (nullable = true)
|  |  |-- element: struct (containsNull = true)
|  |  |  |-- display_url: string (nullable = true)
|  |  |  |-- expanded_url: string (nullable = true)

```

```

|         |         |         |-- indices: array (nullable = true)
|         |         |         |-- element: long (containsNull = true)
|         |         |         |-- url: string (nullable = true)
|         |-- user_mentions: array (nullable = true)
|         |         |-- element: struct (containsNull = true)
|         |         |         |-- id: long (nullable = true)
|         |         |         |-- id_str: string (nullable = true)
|         |         |         |-- indices: array (nullable = true)
|         |         |         |         |-- element: long (containsNull = true)
|         |         |         |-- name: string (nullable = true)
|         |         |         |-- screen_name: string (nullable = true)
|-- extended_entities: struct (nullable = true)
|   |-- media: array (nullable = true)
|   |   |-- element: struct (containsNull = true)
|   |   |   |-- additional_media_info: struct (nullable = true)
|   |   |   |   |-- description: string (nullable = true)
|   |   |   |   |-- embeddable: boolean (nullable = true)
|   |   |   |   |-- monetizable: boolean (nullable = true)
|   |   |   |   |-- title: string (nullable = true)
|   |   |   |-- description: string (nullable = true)
|   |   |   |-- display_url: string (nullable = true)
|   |   |   |-- expanded_url: string (nullable = true)
|   |   |   |-- id: long (nullable = true)
|   |   |   |-- id_str: string (nullable = true)
|   |   |   |-- indices: array (nullable = true)
|   |   |   |   |-- element: long (containsNull = true)
|   |   |   |-- media_url: string (nullable = true)
|   |   |   |-- media_url_https: string (nullable = true)
|   |   |   |-- sizes: struct (nullable = true)
|   |   |   |   |-- large: struct (nullable = true)
|   |   |   |   |   |-- h: long (nullable = true)
|   |   |   |   |   |-- resize: string (nullable = true)
|   |   |   |   |   |-- w: long (nullable = true)
|   |   |   |   |-- medium: struct (nullable = true)
|   |   |   |   |   |-- h: long (nullable = true)
|   |   |   |   |   |-- resize: string (nullable = true)
|   |   |   |   |   |-- w: long (nullable = true)
|   |   |   |   |-- small: struct (nullable = true)
|   |   |   |   |   |-- h: long (nullable = true)
|   |   |   |   |   |-- resize: string (nullable = true)
|   |   |   |   |   |-- w: long (nullable = true)
|   |   |   |   |-- thumb: struct (nullable = true)
|   |   |   |   |   |-- h: long (nullable = true)
|   |   |   |   |   |-- resize: string (nullable = true)
|   |   |   |   |   |-- w: long (nullable = true)
|   |   |   |-- source_status_id: long (nullable = true)
|   |   |   |-- source_status_id_str: string (nullable = true)
|   |   |   |-- source_user_id: long (nullable = true)

```



```

|   |   |-- element: double (containsNull = true)
|   |-- type: string (nullable = true)
|-- id: long (nullable = true)
|-- id_str: string (nullable = true)
|-- in_reply_to_screen_name: string (nullable = true)
|-- in_reply_to_status_id: long (nullable = true)
|-- in_reply_to_status_id_str: string (nullable = true)
|-- in_reply_to_user_id: long (nullable = true)
|-- in_reply_to_user_id_str: string (nullable = true)
|-- is_quote_status: boolean (nullable = true)
|-- lang: string (nullable = true)
|-- place: struct (nullable = true)
|   |-- bounding_box: struct (nullable = true)
|   |   |-- coordinates: array (nullable = true)
|   |   |   |-- element: array (containsNull = true)
|   |   |   |   |-- element: array (containsNull = true)
|   |   |   |   |   |-- element: double (containsNull = true)
|   |   |   |   |-- type: string (nullable = true)
|   |   |-- country: string (nullable = true)
|   |   |-- country_code: string (nullable = true)
|   |   |-- full_name: string (nullable = true)
|   |   |-- id: string (nullable = true)
|   |   |-- name: string (nullable = true)
|   |   |-- place_type: string (nullable = true)
|   |   |-- url: string (nullable = true)
|-- possibly_sensitive: boolean (nullable = true)
|-- quote_count: long (nullable = true)
|-- quoted_status: struct (nullable = true)
|   |-- coordinates: struct (nullable = true)
|   |   |-- coordinates: array (nullable = true)
|   |   |   |-- element: double (containsNull = true)
|   |   |-- type: string (nullable = true)
|   |-- created_at: string (nullable = true)
|   |-- display_text_range: array (nullable = true)
|   |   |-- element: long (containsNull = true)
|   |-- entities: struct (nullable = true)
|   |   |-- hashtags: array (nullable = true)
|   |   |   |-- element: struct (containsNull = true)
|   |   |   |   |-- indices: array (nullable = true)
|   |   |   |   |   |-- element: long (containsNull = true)
|   |   |   |   |-- text: string (nullable = true)
|   |   |-- media: array (nullable = true)
|   |   |   |-- element: struct (containsNull = true)
|   |   |   |   |-- additional_media_info: struct (nullable = true)
|   |   |   |   |   |-- description: string (nullable = true)
|   |   |   |   |   |-- embeddable: boolean (nullable = true)
|   |   |   |   |   |-- monetizable: boolean (nullable = true)
|   |   |   |   |-- title: string (nullable = true)

```



```

| | | | | -- description: string (nullable = true)
| | | | | -- display_url: string (nullable = true)
| | | | | -- expanded_url: string (nullable = true)
| | | | | -- id: long (nullable = true)
| | | | | -- id_str: string (nullable = true)
| | | | | -- indices: array (nullable = true)
| | | | | | -- element: long (containsNull = true)
| | | | | -- media_url: string (nullable = true)
| | | | | -- media_url_https: string (nullable = true)
| | | | | -- sizes: struct (nullable = true)
| | | | | | -- large: struct (nullable = true)
| | | | | | | -- h: long (nullable = true)
| | | | | | | -- resize: string (nullable = true)
| | | | | | | -- w: long (nullable = true)
| | | | | | -- medium: struct (nullable = true)
| | | | | | | -- h: long (nullable = true)
| | | | | | | -- resize: string (nullable = true)
| | | | | | | -- w: long (nullable = true)
| | | | | | -- small: struct (nullable = true)
| | | | | | | -- h: long (nullable = true)
| | | | | | | -- resize: string (nullable = true)
| | | | | | | -- w: long (nullable = true)
| | | | | | -- thumb: struct (nullable = true)
| | | | | | | -- h: long (nullable = true)
| | | | | | | -- resize: string (nullable = true)
| | | | | | | -- w: long (nullable = true)
| | | | | -- source_status_id: long (nullable = true)
| | | | | -- source_status_id_str: string (nullable = true)
| | | | | -- source_user_id: long (nullable = true)
| | | | | -- source_user_id_str: string (nullable = true)
| | | | | -- type: string (nullable = true)
| | | | | -- url: string (nullable = true)
| | | -- symbols: array (nullable = true)
| | | | -- element: struct (containsNull = true)
| | | | | -- indices: array (nullable = true)
| | | | | | -- element: long (containsNull = true)
| | | | | | -- text: string (nullable = true)
| | | -- urls: array (nullable = true)
| | | | -- element: struct (containsNull = true)
| | | | | -- display_url: string (nullable = true)
| | | | | -- expanded_url: string (nullable = true)
| | | | | -- indices: array (nullable = true)
| | | | | | -- element: long (containsNull = true)
| | | | | | -- url: string (nullable = true)
| | | -- user_mentions: array (nullable = true)
| | | | -- element: struct (containsNull = true)
| | | | | -- id: long (nullable = true)
| | | | | -- id_str: string (nullable = true)

```



```

|   |-- in_reply_to_status_id: long (nullable = true)
|   |-- in_reply_to_status_id_str: string (nullable = true)
|   |-- in_reply_to_user_id: long (nullable = true)
|   |-- in_reply_to_user_id_str: string (nullable = true)
|   |-- is_quote_status: boolean (nullable = true)
|   |-- lang: string (nullable = true)
|   |-- place: struct (nullable = true)
|       |-- bounding_box: struct (nullable = true)
|           |-- coordinates: array (nullable = true)
|               |-- element: array (containsNull = true)
|                   |-- element: double (containsNull = true)
|                       |-- type: string (nullable = true)
|               |-- country: string (nullable = true)
|               |-- country_code: string (nullable = true)
|               |-- full_name: string (nullable = true)
|               |-- id: string (nullable = true)
|               |-- name: string (nullable = true)
|               |-- place_type: string (nullable = true)
|               |-- url: string (nullable = true)
|   |-- possibly_sensitive: boolean (nullable = true)
|   |-- quote_count: long (nullable = true)
|   |-- quoted_status_id: long (nullable = true)
|   |-- quoted_status_id_str: string (nullable = true)
|   |-- reply_count: long (nullable = true)
|   |-- retweet_count: long (nullable = true)
|   |-- retweeted: boolean (nullable = true)
|   |-- scopes: struct (nullable = true)
|       |-- followers: boolean (nullable = true)
|   |-- source: string (nullable = true)
|   |-- text: string (nullable = true)
|   |-- truncated: boolean (nullable = true)
|   |-- user: struct (nullable = true)
|       |-- contributors_enabled: boolean (nullable = true)
|       |-- created_at: string (nullable = true)
|       |-- default_profile: boolean (nullable = true)
|       |-- default_profile_image: boolean (nullable = true)
|       |-- description: string (nullable = true)
|       |-- favourites_count: long (nullable = true)
|       |-- followers_count: long (nullable = true)
|       |-- friends_count: long (nullable = true)
|       |-- geo_enabled: boolean (nullable = true)
|       |-- id: long (nullable = true)
|       |-- id_str: string (nullable = true)
|       |-- is_translator: boolean (nullable = true)
|       |-- listed_count: long (nullable = true)
|       |-- location: string (nullable = true)
|       |-- name: string (nullable = true)

```

```

|     |     |-- profile_background_color: string (nullable = true)
|     |     |-- profile_background_image_url: string (nullable = true)
|     |     |-- profile_background_image_url_https: string (nullable = true)
|     |     |-- profile_background_tile: boolean (nullable = true)
|     |     |-- profile_banner_url: string (nullable = true)
|     |     |-- profile_image_url: string (nullable = true)
|     |     |-- profile_image_url_https: string (nullable = true)
|     |     |-- profile_link_color: string (nullable = true)
|     |     |-- profile_sidebar_border_color: string (nullable = true)
|     |     |-- profile_sidebar_fill_color: string (nullable = true)
|     |     |-- profile_text_color: string (nullable = true)
|     |     |-- profile_use_background_image: boolean (nullable = true)
|     |     |-- protected: boolean (nullable = true)
|     |     |-- screen_name: string (nullable = true)
|     |     |-- statuses_count: long (nullable = true)
|     |     |-- translator_type: string (nullable = true)
|     |     |-- url: string (nullable = true)
|     |     |-- verified: boolean (nullable = true)
|     |     |-- verified_type: string (nullable = true)
|     |     |-- withheld_in_countries: array (nullable = true)
|     |     |     |-- element: string (containsNull = true)
|     |-- withheld_copyright: boolean (nullable = true)
|     |-- withheld_in_countries: array (nullable = true)
|     |     |-- element: string (containsNull = true)
|-- quoted_status_id: long (nullable = true)
|-- quoted_status_id_str: string (nullable = true)
|-- quoted_status_permalink: struct (nullable = true)
|     |-- display: string (nullable = true)
|     |-- expanded: string (nullable = true)
|     |-- url: string (nullable = true)
|-- quoted_text: string (nullable = true)
|-- reply_count: long (nullable = true)
|-- retweet_count: long (nullable = true)
|-- retweeted: string (nullable = true)
|-- retweeted_from: string (nullable = true)
|-- retweeted_status: struct (nullable = true)
|     |-- coordinates: struct (nullable = true)
|     |     |-- coordinates: array (nullable = true)
|     |     |     |-- element: double (containsNull = true)
|     |     |-- type: string (nullable = true)
|     |-- created_at: string (nullable = true)
|     |-- display_text_range: array (nullable = true)
|     |     |-- element: long (containsNull = true)
|     |-- entities: struct (nullable = true)
|     |     |-- hashtags: array (nullable = true)
|     |     |     |-- element: struct (containsNull = true)
|     |     |     |     |-- indices: array (nullable = true)
|     |     |     |     |     |-- element: long (containsNull = true)

```



```

|         |-- embeddable: boolean (nullable = true)
|         |-- monetizable: boolean (nullable = true)
|         |-- title: string (nullable = true)
|         |-- description: string (nullable = true)
|         |-- display_url: string (nullable = true)
|         |-- expanded_url: string (nullable = true)
|         |-- id: long (nullable = true)
|         |-- id_str: string (nullable = true)
|         |-- indices: array (nullable = true)
|         |   |-- element: long (containsNull = true)
|         |-- media_url: string (nullable = true)
|         |-- media_url_https: string (nullable = true)
|         |-- sizes: struct (nullable = true)
|         |   |-- large: struct (nullable = true)
|         |   |   |-- h: long (nullable = true)
|         |   |   |-- resize: string (nullable = true)
|         |   |   |-- w: long (nullable = true)
|         |   |-- medium: struct (nullable = true)
|         |   |   |-- h: long (nullable = true)
|         |   |   |-- resize: string (nullable = true)
|         |   |   |-- w: long (nullable = true)
|         |   |-- small: struct (nullable = true)
|         |   |   |-- h: long (nullable = true)
|         |   |   |-- resize: string (nullable = true)
|         |   |   |-- w: long (nullable = true)
|         |   |-- thumb: struct (nullable = true)
|         |   |   |-- h: long (nullable = true)
|         |   |   |-- resize: string (nullable = true)
|         |   |   |-- w: long (nullable = true)
|         |-- source_status_id: long (nullable = true)
|         |-- source_status_id_str: string (nullable = true)
|         |-- source_user_id: long (nullable = true)
|         |-- source_user_id_str: string (nullable = true)
|         |-- type: string (nullable = true)
|         |-- url: string (nullable = true)
|         |-- video_info: struct (nullable = true)
|         |   |-- aspect_ratio: array (nullable = true)
|         |   |   |-- element: long (containsNull = true)
|         |   |-- duration_millis: long (nullable = true)
|         |   |-- variants: array (nullable = true)
|         |   |   |-- element: struct (containsNull = true)
|         |   |   |   |-- bitrate: long (nullable = true)
|         |   |   |   |-- content_type: string (nullable =
true)
|         |   |   |   |-- url: string (nullable = true)
|         |   |-- full_text: string (nullable = true)
|         |-- favorite_count: long (nullable = true)
|         |-- favorited: boolean (nullable = true)

```

```

|   |-- filter_level: string (nullable = true)
|   |-- geo: struct (nullable = true)
|   |   |-- coordinates: array (nullable = true)
|   |   |   |-- element: double (containsNull = true)
|   |   |-- type: string (nullable = true)
|   |-- id: long (nullable = true)
|   |-- id_str: string (nullable = true)
|   |-- in_reply_to_screen_name: string (nullable = true)
|   |-- in_reply_to_status_id: long (nullable = true)
|   |-- in_reply_to_status_id_str: string (nullable = true)
|   |-- in_reply_to_user_id: long (nullable = true)
|   |-- in_reply_to_user_id_str: string (nullable = true)
|   |-- is_quote_status: boolean (nullable = true)
|   |-- lang: string (nullable = true)
|   |-- place: struct (nullable = true)
|   |   |-- bounding_box: struct (nullable = true)
|   |   |   |-- coordinates: array (nullable = true)
|   |   |   |   |-- element: array (containsNull = true)
|   |   |   |   |   |-- element: array (containsNull = true)
|   |   |   |   |   |   |-- element: double (containsNull = true)
|   |   |   |   |-- type: string (nullable = true)
|   |   |-- country: string (nullable = true)
|   |   |-- country_code: string (nullable = true)
|   |   |-- full_name: string (nullable = true)
|   |   |-- id: string (nullable = true)
|   |   |-- name: string (nullable = true)
|   |   |-- place_type: string (nullable = true)
|   |   |-- url: string (nullable = true)
|   |-- possibly_sensitive: boolean (nullable = true)
|   |-- quote_count: long (nullable = true)
|   |-- quoted_status: struct (nullable = true)
|   |   |-- coordinates: struct (nullable = true)
|   |   |   |-- coordinates: array (nullable = true)
|   |   |   |   |-- element: double (containsNull = true)
|   |   |   |-- type: string (nullable = true)
|   |   |-- created_at: string (nullable = true)
|   |   |-- display_text_range: array (nullable = true)
|   |   |   |-- element: long (containsNull = true)
|   |   |-- entities: struct (nullable = true)
|   |   |   |-- hashtags: array (nullable = true)
|   |   |   |   |-- element: struct (containsNull = true)
|   |   |   |   |   |-- indices: array (nullable = true)
|   |   |   |   |   |   |-- element: long (containsNull = true)
|   |   |   |   |   |   |-- text: string (nullable = true)
|   |   |   |-- media: array (nullable = true)
|   |   |   |   |-- element: struct (containsNull = true)
|   |   |   |   |   |-- additional_media_info: struct (nullable = true)
|   |   |   |   |   |   |-- description: string (nullable = true)

```

```

|         |-- embeddable: boolean (nullable = true)
|         |-- monetizable: boolean (nullable = true)
|         |-- title: string (nullable = true)
|         |-- description: string (nullable = true)
|         |-- display_url: string (nullable = true)
|         |-- expanded_url: string (nullable = true)
|         |-- id: long (nullable = true)
|         |-- id_str: string (nullable = true)
|         |-- indices: array (nullable = true)
|         |   |-- element: long (containsNull = true)
|         |-- media_url: string (nullable = true)
|         |-- media_url_https: string (nullable = true)
|         |-- sizes: struct (nullable = true)
|         |   |-- large: struct (nullable = true)
|         |   |   |-- h: long (nullable = true)
|         |   |   |-- resize: string (nullable = true)
|         |   |   |-- w: long (nullable = true)
|         |   |-- medium: struct (nullable = true)
|         |   |   |-- h: long (nullable = true)
|         |   |   |-- resize: string (nullable = true)
|         |   |   |-- w: long (nullable = true)
|         |   |-- small: struct (nullable = true)
|         |   |   |-- h: long (nullable = true)
|         |   |   |-- resize: string (nullable = true)
|         |   |   |-- w: long (nullable = true)
|         |   |-- thumb: struct (nullable = true)
|         |   |   |-- h: long (nullable = true)
|         |   |   |-- resize: string (nullable = true)
|         |   |   |-- w: long (nullable = true)
|         |-- source_status_id: long (nullable = true)
|         |-- source_status_id_str: string (nullable = true)
|         |-- source_user_id: long (nullable = true)
|         |-- source_user_id_str: string (nullable = true)
|         |-- type: string (nullable = true)
|         |-- url: string (nullable = true)
|         |-- symbols: array (nullable = true)
|         |   |-- element: struct (containsNull = true)
|         |   |   |-- indices: array (nullable = true)
|         |   |   |   |-- element: long (containsNull = true)
|         |   |   |   |-- text: string (nullable = true)
|         |-- urls: array (nullable = true)
|         |   |-- element: struct (containsNull = true)
|         |   |   |-- display_url: string (nullable = true)
|         |   |   |-- expanded_url: string (nullable = true)
|         |   |   |-- indices: array (nullable = true)
|         |   |   |   |-- element: long (containsNull = true)
|         |   |   |   |-- url: string (nullable = true)
|         |-- user_mentions: array (nullable = true)

```



```

|         |         |         |-- default_profile_image: boolean (nullable = true)
|         |         |         |-- description: string (nullable = true)
|         |         |         |-- favourites_count: long (nullable = true)
|         |         |         |-- followers_count: long (nullable = true)
|         |         |         |-- friends_count: long (nullable = true)
|         |         |         |-- geo_enabled: boolean (nullable = true)
|         |         |         |-- id: long (nullable = true)
|         |         |         |-- id_str: string (nullable = true)
|         |         |         |-- is_translator: boolean (nullable = true)
|         |         |         |-- listed_count: long (nullable = true)
|         |         |         |-- location: string (nullable = true)
|         |         |         |-- name: string (nullable = true)
|         |         |         |-- profile_background_color: string (nullable = true)
|         |         |         |-- profile_background_image_url: string (nullable = true)
|         |         |         |-- profile_background_image_url_https: string (nullable = true)
|         |         |         |-- profile_background_tile: boolean (nullable = true)
|         |         |         |-- profile_banner_url: string (nullable = true)
|         |         |         |-- profile_image_url: string (nullable = true)
|         |         |         |-- profile_image_url_https: string (nullable = true)
|         |         |         |-- profile_link_color: string (nullable = true)
|         |         |         |-- profile_sidebar_border_color: string (nullable = true)
|         |         |         |-- profile_sidebar_fill_color: string (nullable = true)
|         |         |         |-- profile_text_color: string (nullable = true)
|         |         |         |-- profile_use_background_image: boolean (nullable = true)
|         |         |         |-- protected: boolean (nullable = true)
|         |         |         |-- screen_name: string (nullable = true)
|         |         |         |-- statuses_count: long (nullable = true)
|         |         |         |-- translator_type: string (nullable = true)
|         |         |         |-- url: string (nullable = true)
|         |         |         |-- verified: boolean (nullable = true)
|         |         |         |-- verified_type: string (nullable = true)
|         |         |         |-- withheld_in_countries: array (nullable = true)
|         |         |         |         |-- element: string (containsNull = true)
|         |         |-- withheld_in_countries: array (nullable = true)
|         |         |         |-- element: string (containsNull = true)
|         |-- quoted_status_id: long (nullable = true)
|         |-- quoted_status_id_str: string (nullable = true)
|         |-- quoted_status_permalink: struct (nullable = true)
|         |         |-- display: string (nullable = true)
|         |         |-- expanded: string (nullable = true)
|         |         |-- url: string (nullable = true)
|         |-- reply_count: long (nullable = true)
|         |-- retweet_count: long (nullable = true)
|         |-- retweeted: boolean (nullable = true)
|         |-- scopes: struct (nullable = true)
|         |         |-- followers: boolean (nullable = true)
|         |         |-- place_ids: array (nullable = true)
|         |         |         |-- element: string (containsNull = true)

```

```

|   |-- source: string (nullable = true)
|   |-- text: string (nullable = true)
|   |-- truncated: boolean (nullable = true)
|   |-- user: struct (nullable = true)
|       |-- contributors_enabled: boolean (nullable = true)
|       |-- created_at: string (nullable = true)
|       |-- default_profile: boolean (nullable = true)
|       |-- default_profile_image: boolean (nullable = true)
|       |-- description: string (nullable = true)
|       |-- favourites_count: long (nullable = true)
|       |-- followers_count: long (nullable = true)
|       |-- friends_count: long (nullable = true)
|       |-- geo_enabled: boolean (nullable = true)
|       |-- id: long (nullable = true)
|       |-- id_str: string (nullable = true)
|       |-- is_translator: boolean (nullable = true)
|       |-- listed_count: long (nullable = true)
|       |-- location: string (nullable = true)
|       |-- name: string (nullable = true)
|       |-- profile_background_color: string (nullable = true)
|       |-- profile_background_image_url: string (nullable = true)
|       |-- profile_background_image_url_https: string (nullable = true)
|       |-- profile_background_tile: boolean (nullable = true)
|       |-- profile_banner_url: string (nullable = true)
|       |-- profile_image_url: string (nullable = true)
|       |-- profile_image_url_https: string (nullable = true)
|       |-- profile_link_color: string (nullable = true)
|       |-- profile_sidebar_border_color: string (nullable = true)
|       |-- profile_sidebar_fill_color: string (nullable = true)
|       |-- profile_text_color: string (nullable = true)
|       |-- profile_use_background_image: boolean (nullable = true)
|       |-- protected: boolean (nullable = true)
|       |-- screen_name: string (nullable = true)
|       |-- statuses_count: long (nullable = true)
|       |-- translator_type: string (nullable = true)
|       |-- url: string (nullable = true)
|       |-- verified: boolean (nullable = true)
|       |-- verified_type: string (nullable = true)
|       |-- withheld_in_countries: array (nullable = true)
|           |-- element: string (containsNull = true)
|   |-- withheld_in_countries: array (nullable = true)
|       |-- element: string (containsNull = true)
|-- source: string (nullable = true)
|-- text: string (nullable = true)
|-- timestamp_ms: string (nullable = true)
|-- truncated: boolean (nullable = true)
|-- tweet_text: string (nullable = true)
|-- user: struct (nullable = true)

```

```

|   |-- contributors_enabled: boolean (nullable = true)
|   |-- created_at: string (nullable = true)
|   |-- default_profile: boolean (nullable = true)
|   |-- default_profile_image: boolean (nullable = true)
|   |-- description: string (nullable = true)
|   |-- favourites_count: long (nullable = true)
|   |-- followers_count: long (nullable = true)
|   |-- friends_count: long (nullable = true)
|   |-- geo_enabled: boolean (nullable = true)
|   |-- id: long (nullable = true)
|   |-- id_str: string (nullable = true)
|   |-- is_translator: boolean (nullable = true)
|   |-- listed_count: long (nullable = true)
|   |-- location: string (nullable = true)
|   |-- name: string (nullable = true)
|   |-- profile_background_color: string (nullable = true)
|   |-- profile_background_image_url: string (nullable = true)
|   |-- profile_background_image_url_https: string (nullable = true)
|   |-- profile_background_tile: boolean (nullable = true)
|   |-- profile_banner_url: string (nullable = true)
|   |-- profile_image_url: string (nullable = true)
|   |-- profile_image_url_https: string (nullable = true)
|   |-- profile_link_color: string (nullable = true)
|   |-- profile_sidebar_border_color: string (nullable = true)
|   |-- profile_sidebar_fill_color: string (nullable = true)
|   |-- profile_text_color: string (nullable = true)
|   |-- profile_use_background_image: boolean (nullable = true)
|   |-- protected: boolean (nullable = true)
|   |-- screen_name: string (nullable = true)
|   |-- statuses_count: long (nullable = true)
|   |-- translator_type: string (nullable = true)
|   |-- url: string (nullable = true)
|   |-- verified: boolean (nullable = true)
|   |-- verified_type: string (nullable = true)
|   |-- withheld_in_countries: array (nullable = true)
|       |-- element: string (containsNull = true)
|-- withheld_in_countries: array (nullable = true)
|   |-- element: string (containsNull = true)

```

```

[9]: data = twitter.select([twitter.created_at,
                           twitter.id_str.alias('tweet_id'),
                           twitter.user['id_str'].alias('user_id'),
                           twitter.user['name'].alias('user_name'),
                           twitter.user['verified'].alias('verified'),
                           twitter.user['followers_count'],
                           ↪alias('followers_count'),

```

```

        twitter.user['location'],
        twitter.user['created_at'],

        twitter.text,
        twitter.lang.alias('tweet_language'),
        twitter.retweet_count,
        twitter.favorite_count,
        twitter.quote_count,
        twitter.entities.hashtags['text'].

↪alias('hashtag_text'),

        twitter.retweeted,
        twitter.retweeted_from,
        #twitter.retweeted_status
        twitter.retweeted_status.retweet_count.

↪alias('rt_count'),

        twitter.retweeted_status.id_str.alias('rt_id'),
        twitter.retweeted_status.created_at.

↪alias('rt_create'),

        twitter.retweeted_status.favorite_count.

↪alias('rt_fav'),

        twitter.retweeted_status.quote_count.

↪alias('rt_quo'),

        twitter.retweeted_status.entities.
↪hashtags['text'].alias('rt_hashtag_text'),
        twitter.retweeted_status.user['id_str'].

↪alias('rt_user_id'),

        twitter.retweeted_status.user['name'].

↪alias('rt_user_name'),

        #twitter.quoted_status.alias('tweet_quote'),
        twitter.place.country.alias('tweet_country'),
        twitter.place.full_name.alias('tweet_location']]

```

```
[10]: data.printSchema()
```

```

root
|-- created_at: string (nullable = true)
|-- tweet_id: string (nullable = true)
|-- user_id: string (nullable = true)
|-- user_name: string (nullable = true)
|-- verified: boolean (nullable = true)
|-- followers_count: long (nullable = true)
|-- user.location: string (nullable = true)
|-- user.created_at: string (nullable = true)
|-- text: string (nullable = true)

```

```

|-- tweet_language: string (nullable = true)
|-- retweet_count: long (nullable = true)
|-- favorite_count: long (nullable = true)
|-- quote_count: long (nullable = true)
|-- hashtag_text: array (nullable = true)
|   |-- element: string (containsNull = true)
|-- retweeted: string (nullable = true)
|-- retweeted_from: string (nullable = true)
|-- rt_count: long (nullable = true)
|-- rt_id: string (nullable = true)
|-- rt_create: string (nullable = true)
|-- rt_fav: long (nullable = true)
|-- rt_quo: long (nullable = true)
|-- rt_hashtag_text: array (nullable = true)
|   |-- element: string (containsNull = true)
|-- rt_user_id: string (nullable = true)
|-- rt_user_name: string (nullable = true)
|-- tweet_country: string (nullable = true)
|-- tweet_location: string (nullable = true)

```

```
[34]: data.select('retweeted', 'retweeted_from').show(5)
```

```

+-----+-----+
|retweeted|retweeted_from|
+-----+-----+
|      RT|   AaronParnas|
|      |           null|
|      RT|   MoneyMiaaaa|
|      RT|   MoneyMiaaaa|
|      |           null|
+-----+-----+
only showing top 5 rows

```

```
[36]: place = data.select('tweet_country', 'tweet_location')
place.filter(col('tweet_country').isNotNull()).show(5)
```

```

+-----+-----+
| tweet_country| tweet_location|
+-----+-----+
| United States|   Smyrna, TN|
| United States| Pompano Beach, FL|
| United States|   Keizer, OR|
| United States| Los Angeles, CA|
|The Netherlands|Rotterdam, The Ne...|
+-----+-----+
only showing top 5 rows

```



```
[35]: retweets = data.select('retweeted','retweeted_from','rt_count','rt_fav')
retweets = retweets.filter(col('retweeted_from').isNotNull())
retweets.orderBy(col('rt_count').desc()).show(5)
```

[Stage 21:=====>(5739 + 2) / 5741]

retweeted	retweeted_from	rt_count	rt_fav
RT	nickjr	516855	2036787
RT	nickjr	516850	2036713
RT	nickjr	516795	2036562
RT	nickjr	516791	2036584
RT	nickjr	516779	2036505

only showing top 5 rows

```
[37]: country_count = place.groupby('tweet_country').agg(count('*').
↳alias('Number_of_tweets'))
country_count.orderBy(col('Number_of_tweets').desc()).show(5)
```

[Stage 25:=====> (15 + 5) / 20]

tweet_country	Number_of_tweets
null	99112826
United States	554287
United Kingdom	73793
India	46194
Kingdom of Saudi ...	38942

only showing top 5 rows

```
[39]: language = data.select('tweet_language').groupBy('tweet_language').
↳agg(count('*').alias('count_by_language'))
language.orderBy(col('count_by_language').desc()).show(5)
```

[Stage 30:> (0 + 1) / 1]

tweet_language	count_by_language
----------------	-------------------

```
+-----+-----+
|          en|          99994342|
+-----+-----+
```

```
[43]: data.select('text').filter(col('text').isNull()).show(5)
```

```
+-----+
|text|
+-----+
+-----+
```

First, clean the text

```
[5]: !pip uninstall -y nltk
!pip install nltk --upgrade --no-cache-dir
```

Found existing installation: nltk 3.6.4

Uninstalling nltk-3.6.4:

Successfully uninstalled nltk-3.6.4

WARNING: Running pip as the 'root' user can result in broken permissions
and conflicting behaviour with the system package manager. It is recommended to
use a virtual environment instead: <https://pip.pypa.io/warnings/venv>

Collecting nltk

Downloading nltk-3.8.1-py3-none-any.whl (1.5 MB)

1.5/1.5 MB

26.6 MB/s eta 0:00:00 0:00:01

Requirement already satisfied: tqdm in

/opt/conda/miniconda3/lib/python3.8/site-packages (from nltk) (4.64.1)

Requirement already satisfied: joblib in

/opt/conda/miniconda3/lib/python3.8/site-packages (from nltk) (1.2.0)

Collecting regex>=2021.8.3

Downloading

regex-2022.10.31-cp38-cp38-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (772
kB)

772.3/772.3 kB

286.2 MB/s eta 0:00:00

Requirement already satisfied: click in

/opt/conda/miniconda3/lib/python3.8/site-packages (from nltk) (7.1.2)

Installing collected packages: regex, nltk

Attempting uninstall: regex

Found existing installation: regex 2021.4.4

Uninstalling regex-2021.4.4:

Successfully uninstalled regex-2021.4.4
Successfully installed nltk-3.8.1 regex-2022.10.31
WARNING: Running pip as the 'root' user can result in broken permissions
and conflicting behaviour with the system package manager. It is recommended to
use a virtual environment instead: <https://pip.pypa.io/warnings/venv>

```
[6]: import re
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

```
[11]: text_data = data.select('text')

text = text_data.rdd.map(lambda x : x['text']).filter(lambda x: x is not None)

StopWords = stopwords.words("english")

tokens = text\
    .map( lambda document: document.strip().lower())\
    .map( lambda document: re.split(" ", document))\
    .map( lambda word: [x for x in word if x.isalnum()])\
    .map( lambda word: [x for x in word if len(x) > 3] )\
    .map( lambda word: [x for x in word if x not in StopWords])\
    #.zipWithIndex()
```

```
[61]: tokens.take(5)
```

```
[61]: [['upset', 'beto', 'interrupting', 'press', 'conference', 'school'],
       ['homeschooling', 'effectively', 'must', 'send', 'child', 'accredited'],
       ['school', 'shootings', 'worried', 'ridiculous'],
       ['school', 'shootings', 'worried', 'ridiculous'],
       ['uncompleted', 'structure', 'fenced', 'gated', 'adebisi']]
```

```
[12]: wordCounts = tokens.flatMap(lambda x: x) \
    .map(lambda x: (x, 1)) \
    .reduceByKey(lambda x, y: x+y) \
    .map(lambda x: (x[1],x[0]))
```

```
[13]: wordCountsSorted = wordCounts.sortByKey(ascending=False)
```

```
[14]: wordCountsSorted.take(20)
```

```
[14]: [(35321364, 'school'),
      (10222592, 'college'),
      (7962524, 'high'),
      (7763989, 'university'),
      (5370633, 'schools'),
      (4097465, 'students'),
      (3793172, 'like'),
      (2894026, 'kids'),
      (2851083, 'professor'),
      (2734429, 'people'),
      (2344286, 'student'),
      (2329987, 'first'),
      (2313212, 'back'),
      (2311217, 'children'),
      (2150610, 'public'),
      (2055366, 'year'),
      (2021828, 'would'),
      (1954768, 'time'),
      (1873296, 'know'),
      (1814856, 'going')]
```

```
[19]: word_counts_desc = wordCountsSorted.map(lambda x : (x[1], x[0]))\
      .toDF(["Words", "Counts"])
```

```
[21]: word_counts_desc.show(30)
```

```
+-----+-----+
|   Words| Counts|
+-----+-----+
|  school|35321364|
| college|10222592|
|   high| 7962524|
|university| 7763989|
|  schools| 5370633|
| students| 4097465|
|   like| 3793172|
|   kids| 2894026|
| professor| 2851083|
|  people| 2734429|
| student| 2344286|
|   first| 2329987|
|   back| 2313212|
| children| 2311217|
|  public| 2150610|
```

```

|      year| 2055366|
|      would| 2021828|
|      time| 1954768|
|      know| 1873296|
|      going| 1814856|
|      want| 1782043|
|  football| 1750798|
|      years| 1711423|
|      state| 1709843|
|      need| 1701804|
|      went| 1667016|
|  teacher| 1538729|
|      live| 1530527|
|     every| 1522680|
|     today| 1506946|
+-----+-----+
only showing top 30 rows

```

```
[40]: keywords = ['college', 'high', 'university', 'students',
                  , 'public', 'private', 'secondary', 'primary', 'education',
                  ↪ 'undergraduate', 'graduate']
```

```
[41]: #filter out rows that do not contain words in keywords
data = data.withColumn('lower', lower(col('text')))
filter_df = data.filter(col('lower').rlike(''.join(keywords)))
```

```
[42]: data_eng = filter_df.filter(col('tweet_language') == 'en')
from pyspark.sql import functions as F
from pyspark.sql import types as t
from pyspark.sql.types import ArrayType, IntegerType, BooleanType

eng_ord=F.udf(lambda x: [ord(a) for a in x],t.ArrayType(IntegerType()))

def english_filter(x):
    for index in range(len(x)):
        if x[index] > 128:
            return False
        else:
            return True

filter_udf = F.udf(english_filter, BooleanType())
tweets_en = data_eng.filter(filter_udf(eng_ord('text')) == True)
```

```
[43]: tweets_en.count()
```

[43]: 39621120

```
[8]: tweets_en.printSchema()
```

```
root
 |-- created_at: string (nullable = true)
 |-- tweet_id: string (nullable = true)
 |-- user_id: string (nullable = true)
 |-- user_name: string (nullable = true)
 |-- verified: boolean (nullable = true)
 |-- followers_count: long (nullable = true)
 |-- user.location: string (nullable = true)
 |-- user.created_at: string (nullable = true)
 |-- text: string (nullable = true)
 |-- tweet_language: string (nullable = true)
 |-- retweet_count: long (nullable = true)
 |-- favorite_count: long (nullable = true)
 |-- quote_count: long (nullable = true)
 |-- hashtag_text: array (nullable = true)
 |   |-- element: string (containsNull = true)
 |-- retweeted: string (nullable = true)
 |-- retweeted_from: string (nullable = true)
 |-- rt_count: long (nullable = true)
 |-- rt_id: string (nullable = true)
 |-- rt_create: string (nullable = true)
 |-- rt_fav: long (nullable = true)
 |-- rt_quo: long (nullable = true)
 |-- rt_hashtag_text: array (nullable = true)
 |   |-- element: string (containsNull = true)
 |-- rt_user_id: string (nullable = true)
 |-- rt_user_name: string (nullable = true)
 |-- tweet_country: string (nullable = true)
 |-- tweet_location: string (nullable = true)
 |-- lower: string (nullable = true)
```

```
[21]: tweets_en.select('retweeted').show(5)
```

```
+-----+
|retweeted|
+-----+
|      RT|
|      |
|      |
|      RT|
|      RT|
+-----+
```

only showing top 5 rows

```
[48]: retweets = tweets_en.select('retweeted', 'retweeted_from', 'rt_count', 'rt_fav')
retweets = retweets.filter(col('retweeted_from').isNotNull())
retweets.orderBy(col('rt_count').desc()).show(5)
```

[Stage 57:=====>(1319 + 4) / 1323]

```
+-----+-----+-----+-----+
|retweeted|retweeted_from|rt_count| rt_fav|
+-----+-----+-----+-----+
|      RT|      nickjr|  516855|2036787|
|      RT|      nickjr|  516850|2036713|
|      RT|      nickjr|  516795|2036562|
|      RT|      nickjr|  516791|2036584|
|      RT|      nickjr|  516779|2036505|
+-----+-----+-----+-----+
```

only showing top 5 rows

Save the data to parquet

```
[44]: %%time
tweets_en.write.format("parquet").\
mode('overwrite').\
save('gs://chen26-bdp/filtered')
```

CPU times: user 1.08 s, sys: 273 ms, total: 1.35 s
Wall time: 10min 15s

```
[ ]: %%time
twitter.write.format("parquet").\
mode('overwrite').\
save('gs://chen26-bdp/original_data')
```

CPU times: user 5.04 s, sys: 1.05 s, total: 6.09 s
Wall time: 57min 31s

```
[22]: %%time
word_counts_desc.write.format("parquet").\
mode('overwrite').\
save('gs://chen26-bdp/word_count')
```

```
CPU times: user 163 ms, sys: 101 ms, total: 265 ms
Wall time: 33.9 s
```

```
[24]: %%time
twitter = spark.read.parquet('gs://chen26-bdp/original_data')
```

```
CPU times: user 3.47 ms, sys: 468 μs, total: 3.94 ms
Wall time: 1.49 s
```

```
[7]: %%time
      tweets_en = spark.read.parquet('gs://chen26-bdp/filtered')
```

```
CPU times: user 7.26 ms, sys: 456 µs, total: 7.71 ms
Wall time: 7.21 s
```


[]:	
[]:	
[]:	
[]:	
[]:	
[]:	