CS167 Project D-1 Twitter Analysis

Group D1
Minsoo Kim - mkim410
Rajeev Thundiyil - rthun001
Christian Boroff - cboro002

Task 1 - Data Preparation

- Prepare the data for analysis later in the project.
- To give a brief summary, I first took the given data set and kept only the some of the attributes and created a new output file with it. Then I extracted the top 20 keywords from the newly created dataset.

Loading the file

```
// Load the given input file using the json format.
      val tweetsDF = sparkSession.read.format("json")
        .option("sep", "\t")
        .option("inferSchema", "true")
        .option("header", "true")
        .load(inputFile)
```

Relevant Attributes

```
// Keep only the following attributes {id, text, entities.hashtags.txt, user.description,
retweet_count, reply_count, and quoted_status_id}
     val clean_tweets_df: DataFrame = tweetsDF.selectExpr("id", "text",
"transform(entities.hashtags, x -> x.text) AS hashtags", "user.description AS user_description",
"retweet_count", "reply_count", "quoted_status_id")
```

The relevant attributes mentioned above included the following,

- {id, text, entities.hashtags.txt, user.description, retweet_count, reply_count, and quoted_status_id}

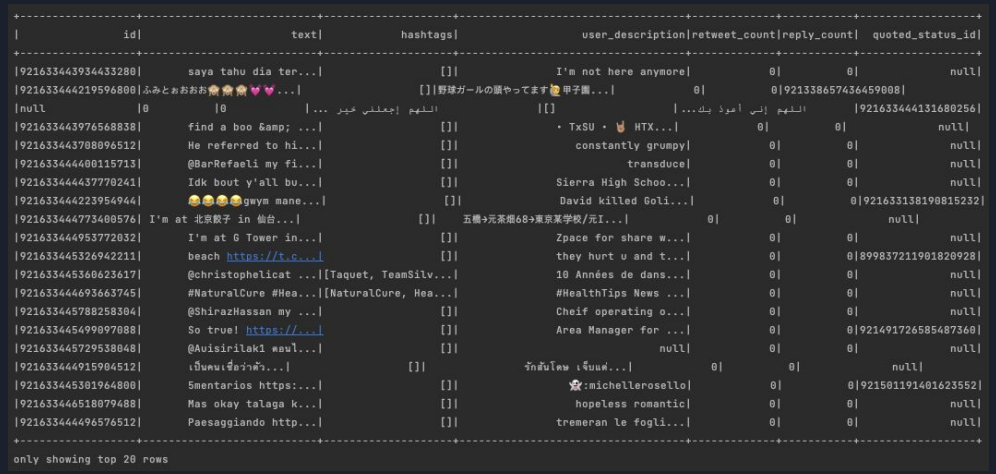
Storing output to a new JSON file to tweets_clean

```
// Store the output in a new JSON file named tweets_clean
     clean_tweets_df.write.mode(SaveMode.Overwrite).json("tweets_clean")
```

Results

Tweets_1K

```
root
 |-- id: long (nullable = true)
 |-- text: string (nullable = true)
 |-- hashtags: array (nullable = true)
 |    |-- element: string (containsNull = true)
 |-- user_description: string (nullable = true)
 |-- retweet_count: long (nullable = true)
 |-- reply_count: long (nullable = true)
 |-- quoted_status_id: long (nullable = true)
```



```
+------------------+--------------------+------------------+--------------------+-------------+-----------+------------------+
|                id|                text|          hashtags|    user_description|retweet_count|reply_count|  quoted_status_id|
+------------------+--------------------+------------------+--------------------+-------------+-----------+------------------+
|9216334439344332 80|       saya tahu dia ter...|                []|I'm not here anymore|            0|          0|              null|
|9216334442195968 00|ふみとぉおおお👧👧👧💗💗...|                []|野球ガールの膜やってます🙋甲子園...|            0|          0|921338657436459008|
|null              |0          |0          |... اللهم إجعلني خير |[]                |...اللهم إني أعوذ بك |9216334441316802 56|
|9216334439765688 38|       find a boo &amp;amp; ...|                []|• TxSU • 🤘 HTX...|            0|          0|              null|
|9216334437080965 12|       He referred to hi...|                []|constantly grumpy|            0|          0|              null|
|9216334444001157 13|       @BarRefaeli my fi...|                []|transduce|            0|          0|              null|
|9216334444377702 41|       Idk bout y'all bu...|                []|Sierra High Schoo...|            0|          0|              null|
|9216334442239549 44|       😂😂😂😂gwym mane...|                []|David killed Goli...|            0|          0|9216331381908152 32|
|9216334447734005 76| I'm at 北京餃子 in 仙台...|                []|五橋→元茶畑68→東京某学校/元I...|            0|          0|              null|
|9216334449537720 32|       I'm at G Tower in...|                []|Zpace for share w...|            0|          0|              null|
|9216334453269422 11|       beach https://t.c...|                []|they hurt u and t...|            0|          0|8998372119018209 28|
|9216334453606236 17|       @christophelicat ...|[Taquet, TeamSilv...|10 Années de dans...|            0|          0|              null|
|9216334446936637 45|       #NaturalCure #Hea...|[NaturalCure, Hea...|#HealthTips News ...|            0|          0|              null|
|9216334457882583 04|       @ShirazHassan my ...|                []|Cheif operating o...|            0|          0|              null|
|9216334454990970 88|       So true! https://...|                []|Area Manager for ...|            0|          0|9216917265854873 60|
|9216334457295380 48|       @Auisirilak1 คอนโ...|                []|null|            0|          0|              null|
|9216334449159045 12|       เป็นคนเชื่อว่าพ้ว...|                []|รักต้นโตพ เจ็บแค่...|            0|          0|              null|
|9216334453019648 00|       5mentarios https:...|                []|👧:michellerosello|            0|          0|9215011914016235 52|
|9216334465180794 88|       Mas okay talaga k...|                []|hopeless romantic|            0|          0|              null|
|9216334449695765 12|       Paesaggiando http...|                []|tremeran le fogli...|            0|          0|              null|
+------------------+--------------------+------------------+--------------------+-------------+-----------+------------------+
only showing top 20 rows
```

Frequent Hashtags

Then, I ran a top-k SQL query to select the top 20 most frequent hashtags. The below shows the SQL and how I collected the result in an array of keywords.

```
// run a top-k SQL query to select the top 20 most frequent hashtags as follows.
    clean_tweets_df.createOrReplaceTempView("clean_tweets")
    val frequent_hashtags = sparkSession.sql(
      """
      SELECT hashtag, count(*) as count
      FROM (
        SELECT explode(hashtags) as hashtag
        FROM clean_tweets
      ) t
      GROUP BY hashtag
      ORDER BY count DESC
      LIMIT 20
    """)

    val keywords: Array[String] = frequent_hashtags.select("hashtag").rdd.map(row => row.getString(0)).collect()
```

```
+--------------------+-----+
|             hashtag|count|
+--------------------+-----+
|      ALDUBxEBLoveis|    7|
|        FurkanPalalı|    5|
|               no309|    5|
|               LalOn|    5|
|             sbhawks|    3|
|DoktorlarDenkliki...|    3|
|        Benimisteğim|    3|
|3    |احتاج_بالوقت_هذا|     |
|                art|    2|
| CNIextravaganza2017|    2|
|                love|    2|
|               happy|    2|
|2    |السعودية        |     |
|          nowplaying|    2|
|           beautiful|    2|
|             türkiye|    1|
|             vegalta|    1|
|           KittyLive|    1|
|                  鯛|    1|
|          tossademar|    1|
+--------------------+-----+
```

Tweets_10K

hashtag	count
ALDUBxEBLoveis	84
FurkanPalalı	51
no309	51
LalOn	51
chien	30
job	28
Hiring	22
sbhawks	16
Top3Apps	16
perdu	15
trouvé	15
CareerArc	14
Job	12
trumprussia	12
trndnl	12
Jobs	11
ShowtimeLetsCelebr8	9
hiring	9
impeachtrumppence	9
music	8

id	text	hashtags	user_description	retweet_count	reply_count	quoted_status_id
921633443934433280	saya tahu dia ter...	[]	I'm not here anymore	0	0	null
921633444219596800	ふみとぉおおお👧👧👧💕💕...	[]	野球ガールの顔やってます🙋甲子園...	0	0	921338657436459008
null	0	0	... اللهم إجعلني خير	[]	اللهم إني أعوذ بك...	921633444131680256
921633443976568838	find a boo & ...	[]	• TxSU • ✌ HTX...	0	0	null
921633443708096512	He referred to hi...	[]	constantly grumpy	0	0	null
921633444400115713	@BarRefaeli my fi...	[]	transduce	0	0	null
921633444437770241	Idk bout y'all bu...	[]	Sierra High Schoo...	0	0	null
921633444223954944	😂😂😂😂gwym mane...	[]	David killed Goli...	0	0	921633138190815232
921633444773400576	I'm at 北京餃子 in 仙台...	[]	五橋→元茶畑68→東京某学校/元I...	0	0	null
921633444953772032	I'm at G Tower in...	[]	Zpace for share w...	0	0	null
921633445326942211	beach https://t.c...	[]	they hurt u and t...	0	0	899837211901820928
921633445360623617	@christophelicat ...	[Taquet, TeamSilv...	10 Années de dans...	0	0	null
921633444693663745	#NaturalCure #Hea...	[NaturalCure, Hea...	#HealthTips News ...	0	0	null
921633445788258304	@ShirazHassan my ...	[]	Cheif operating o...	0	0	null
921633445499097088	So true! https://...	[]	Area Manager for ...	0	0	921491726585487360
921633445729538048	@Auisirilak1 ตอนไ...	[]	null	0	0	null
921633444915904512	เป็นคนเชื่อว่าตัว...	[]	รักสันโดษ เจ็บแค่...	0	0	null
921633445301964800	5mentarios https:...	[]	🌟:michellerosello	0	0	921501191401623552
921633446518079488	Mas okay talaga k...	[]	hopeless romantic	0	0	null
921633444496576512	Paesaggiando http...	[]	tremeran le fogli...	0	0	null

only showing top 20 rows

```
root
 |-- id: long (nullable = true)
 |-- text: string (nullable = true)
 |-- hashtags: array (nullable = true)
 |    |-- element: string (containsNull = true)
 |-- user_description: string (nullable = true)
 |-- retweet_count: long (nullable = true)
 |-- reply_count: long (nullable = true)
 |-- quoted_status_id: long (nullable = true)
```

Task 2

My task was to show the topic of the tweet by doing an array intersection of the most frequent hashtags from task 1 with the hashtags within the tweet itself. If it's a match, a new column is created with the topic shown, found through the intersection.

```
clean_tweets_df.createOrReplaceTempView( viewName = "tweets_clean")
//convert keywords to an array separated with , so it can be used for array intersect in a query
val topics: String = "'"+ keywords.mkString("','") + "'"

//dataframe
```

Task 2 (cont)

```
val topics_df: DataFrame = sparkSession.sql(
  sqlText = s"""
  SELECT id, text,element_at(t1.tweet_topic,1), user_description, retweet_count, reply_count, quoted_status_id
  FROM ( SELECT *, array_intersect(hashtags, array($topics)) AS tweet_topic FROM tweets_clean) AS t1 WHERE size(tweet_topic) > 0;
   """)


//write to json
topics_df.write.json( path = "tweets_topic.json")
topics_df.show()
val t4 = System.nanoTime()
```

The query picks up all of the data types in order, as given in the project specifications.

root
 |-- id: long (nullable = true)
 |-- text: string (nullable = true)
 |-- topic: string (nullable = true)
 |-- user_description: string (nullable = true)
 |-- retweet_count: long (nullable = true)
 |-- reply_count: long (nullable = true)
 |-- quoted_status_id: long (nullable = true)

Task 2 Results

```
Operations on file 'Tweets_1k.json' took 7.5779234 seconds
+------------------+--------------------------------+------------------------+------------------------------+-------------+-----------+------------------+
|                id|                            text|element_at(tweet_topic, 1)|            user_description|retweet_count|reply_count|  quoted_status_id|
+------------------+--------------------------------+------------------------+------------------------------+-------------+-----------+------------------+
|921633446644080641|              #negramaroofficia...|                   love|          Negramanteinside_...|            0|          0|              null|
|921633445045866497|              #CNIextravaganza2...|       CNIextravaganza2017|          Hebat Produknya H...|            0|          0|              null|
|921633449882128384|              #DoktorlarDenklik...|      DoktorlarDenkliki...|                     emin ben|            0|          0|              null|
|921633451773648896|              Na miss ko mag tw...|            ALDUBxEBLoveis|          Resilient. Object...|            0|          0|              null|
|921633452289642497|              #FurkanPalalı Değ...|              FurkanPalalı|                          null|            0|          0|              null|
|921633451714920448|              #KittyLive penuh ...|                 KittyLive|                      Semangat|            0|          0|              null|
|null              |0         |0           |...انا مقدر علي قلبي            |احتاج_بالوقت_هذا          |...احتاج_بالوقت_هذا#              |921633453715738624|
|921633455766728704|              Don`t run from yo...|            ALDUBxEBLoveis|          BEHIND THE THICK ...|            0|          0|              null|
|921633464276996097|              Künefe Ocağı Rezi...|                  türkiye|          Bym Isı Rezistans...|            0|          0|              null|
|921633465921028096|ブルーマウンテンさんやん！\n追加...|                sbhawks|野球垢です！ホークスファンです！|            0|          0|              null|
|921633468634632194|              Pastinya dong, ka...|       CNIextravaganza2017|          BLOGGER | BUZZER ...|            0|          0|921627555261702144|
|921633470413262848|              #FurkanPalalı Değ...|              FurkanPalalı|                          null|            0|          0|              null|
|921633476964646912|              Start by doing wh...|            ALDUBxEBLoveis|          BEHIND THE THICK ...|            0|          0|              null|
|921633490482946048|              Stay by Rihanna F...|               nowplaying|          A live stream of ...|            0|          0|              null|
|921633493569998848|              #FurkanPalalı Değ...|              FurkanPalalı|                          null|            0|          0|              null|
|921633494530392064|              Saturday let's go...|               nowplaying|              Student of Life!|            0|          0|              null|
|null              |0         |0           |...سياسي - يومية            |السعودية                  |...أ التحقيق مع رجال              |921633495511896065|
|921633502453518336|              There are many th...|            ALDUBxEBLoveis|          BEHIND THE THICK ...|            0|          0|              null|
|921633507558002688|              #happy #Saturday ...|                   happy|          Pastry chef, pers...|            0|          0|              null|
|921633511752257536|              #FurkanPalalı Değ...|              FurkanPalalı|                          null|            0|          0|              null|
+------------------+--------------------------------+------------------------+------------------------------+-------------+-----------+------------------+
```

Task 3 - Machine Learning Model

- For this task we built a machine learning model pipeline using the code below

```
val tokenizer = new Tokenizer().setInputCol("text").setOutputCol("words")

val hashingTF = new HashingTF().setInputCol("words").setOutputCol("features")

val stringIndexer = new StringIndexer().setInputCol("element_at(tweet_topic, 1)").setOutputCol("label").setHandleInvalid("skip")

val logisticRegression = new LogisticRegression()

val pipeline = new Pipeline().setStages(Array(tokenizer, hashingTF, stringIndexer, logisticRegression))

val Array(trainingData, testData) = topics_df.randomSplit(Array(0.7, 0.3))

val logisticModel = pipeline.fit(trainingData)

val predictions = logisticModel.transform(testData)

predictions.select( col= "id",  cols= "text", "element_at(tweet_topic, 1)", "user_description", "label", "prediction").show( numRows= 10)
```

- This produced the following output

id	text	element_at(tweet_topic, 1)	user_description	label	prediction
921633452289642497	#FurkanPalalı Değ...	FurkanPalalı	null	1.0	1.0
921633470413262848	#FurkanPalalı Değ...	FurkanPalalı	null	1.0	1.0
921633493569998848	#FurkanPalalı Değ...	FurkanPalalı	null	1.0	1.0
921633502453518336	There are many th...	ALDUBxEBLoveis	BEHIND THE THICK ...	0.0	0.0
921633525505310722	あー落ちない #sbhawks	sbhawks	野球垢です！ホークスファンです！！	6.0	6.0
921633536628551680	Everyone's deserv...	ShowtimeLetsCelebr8	love + trust	8.0	0.0
921633595747213312	If you're looking...	job	Staffing and Recr...	3.0	3.0
921633595801792512	Can you recommend...	job	Quintiles and IMS...	3.0	3.0
921633599660658688	Sometimes people ...	ALDUBxEBLoveis	BEHIND THE THICK ...	0.0	0.0
921633653129486336	@kuriring01 @Bern...	ALDUBxEBLoveis	Businesswoman, lo...	0.0	0.0
921633666656284672	#FurkanPalalı Değ...	FurkanPalalı	null	1.0	1.0
921633666656284672	#FurkanPalalı Değ...	FurkanPalalı	null	1.0	1.0
921633683789856768	Can you recommend...	job	Follow us for job...	3.0	3.0
921633684708454400	See our latest #R...	job	Join a team that ...	3.0	3.0
921633691469787136	#trumprussia #imp...	trumprussia	#NYC Registered N...	11.0	11.0
921633698423955456	😂🤣\nla ako masa...	ALDUBxEBLoveis	CMSRN	0.0	0.0
921633705260548096	Ganda nila 😍 #Sh...	ShowtimeLetsCelebr8	Hotelier and a 👶...	8.0	8.0

Task 3 - Continued

- We then computed the precision and recall of our machine learning model on the 10k tweets dataset using the following code.

```
// Compute the number of true positives, false positives, and false negatives for each class
val tp = (0 ≤ to ≤ 10).map(c => predictions.filter(col( colName = "label") === c && col( colName = "prediction") === c).count()).sum
val fp = (0 ≤ to ≤ 10).map(c => predictions.filter(col( colName = "label") =!= c && col( colName = "prediction") === c).count()).sum
val fn = (0 ≤ to ≤ 10).map(c => predictions.filter(col( colName = "label") === c && col( colName = "prediction") =!= c).count()).sum

// Compute overall precision and recall
val overallPrecision = tp.toDouble / (tp + fp)
val overallRecall = tp.toDouble / (tp + fn)
```

- Our machine learning model using logistic regression to predict the topic of a tweet had a precision and recall of 0.94382. This means 94.38% of the time, our model correctly identified true positive cases.

Question

Which classifier did we use for our machine learning model?

a. logistic regression classifier

b. multi-layer perceptron classifier

c. decision tree classifier