

CS 167 Project; Twitter Data Analysis

22 March 2023

Task 1: Trisha Agrawal

Task 2: Daniel Boules

Task 3: Shwena Kak

Introduction

The given data sets contain information about tweets, including the text of the tweet, a description of the user, retweet count, like count, etc. Task 1 consisted of cleaning the data set so the input file would only contain the attributes needed to make the prediction model. SparkSQL was used to select the required attributes which were later written to a JSON file to be used for task 2 and task 3. Task 2 included adding a new column to the previously cleaned dataset, where the topic is stored. A topic is defined as a hashtag used most frequently, and if a tweet contains multiple hashtags, then any of the hashtags can be assigned as a topic. Task 3 consisted of designing a machine-learning model to assign a topic to a tweet. The model is trained by teaching it the relationship between all features of a tweet and the given topic based on the additional column that was created in task 2. Then a machine learning model with the following pipeline, Tokenzier, HashingTF transformer, StringIndexer, and LogisticRegression, predicts one topic for each tweet.

Big Data System

SparkSQL was used for task 1 and task 2 because it makes querying large data sets easier.

SparkML was used for task 3 because we needed to extract, transform, and evaluate features based on the tweets input. Furthermore, we needed a supervised learning algorithm, to follow through with the topic classification.

Task 1 Results (Trisha Agrawal):

The results from task one outputted to the terminal the top 20 keywords for the 10k dataset:

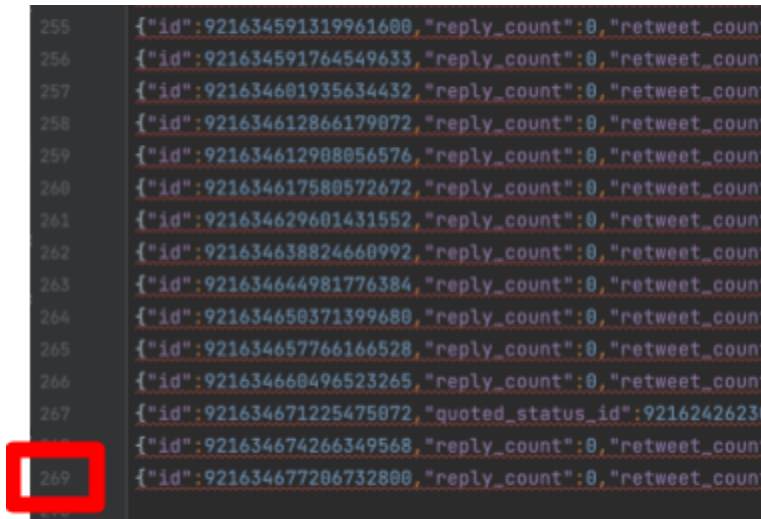
Comma separated list:

ALDUBxEBLoveis, FurkanPalalı, no309, LalOn, chien, job, Hiring, sbhawks, Top3Apps, perdu, trouvé, CareerArc, Job, trumprussia, trndnl, Jobs, hiring, impeachtrumppence, ShowtimeLetsCelebr8, music

Hashtag	Aggregate Sum
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
hiring	9
impeachtrumppence	9
ShowtimeLetsCelebr8	9
music	8

Task 2 Results (Daniel Boules):

Total number of records in the tweets_topic dataset for the 10k dataset: 269



255	{"id":921634591319961600,"reply_count":0,"retweet_count":0,"text":"...","user_description":"..."}
256	{"id":921634591764549633,"reply_count":0,"retweet_count":0,"text":"...","user_description":"..."}
257	{"id":921634601935634432,"reply_count":0,"retweet_count":0,"text":"...","user_description":"..."}
258	{"id":921634612866179072,"reply_count":0,"retweet_count":0,"text":"...","user_description":"..."}
259	{"id":921634612908056576,"reply_count":0,"retweet_count":0,"text":"...","user_description":"..."}
260	{"id":921634617580572672,"reply_count":0,"retweet_count":0,"text":"...","user_description":"..."}
261	{"id":921634629601431552,"reply_count":0,"retweet_count":0,"text":"...","user_description":"..."}
262	{"id":921634638824660992,"reply_count":0,"retweet_count":0,"text":"...","user_description":"..."}
263	{"id":921634644981776384,"reply_count":0,"retweet_count":0,"text":"...","user_description":"..."}
264	{"id":921634650371399680,"reply_count":0,"retweet_count":0,"text":"...","user_description":"..."}
265	{"id":921634657766166528,"reply_count":0,"retweet_count":0,"text":"...","user_description":"..."}
266	{"id":921634660496523265,"reply_count":0,"retweet_count":0,"text":"...","user_description":"..."}
267	{"id":921634671225475072,"quoted_status_id":92162426230,"reply_count":0,"retweet_count":0,"text":"...","user_description":"..."}
268	{"id":921634674266349568,"reply_count":0,"retweet_count":0,"text":"...","user_description":"..."}
269	{"id":921634677206732800,"reply_count":0,"retweet_count":0,"text":"...","user_description":"..."}

Schema after the topic column has been created:

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

Task 3 Results (Shwena Kak):

Overall Precision: 0.9838709677419355

Overall Recall 0.9607843137254901

With the label and prediction columns added:

id	text	topic	user_description	label	prediction
921633465921828896	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!	5.0	5.0
921633502453518336	There are many th...	ALDUBxEBLoveis	BEHIND THE THICK ...	0.0	0.0
921633595747213312	If you're looking...	job	Staffing and Recr...	3.0	3.0
921633637484847105	Want to work in #...	Job	Follow this accou...	11.0	11.0
921633638353195008	When the past cal...	ALDUBxEBLoveis	BEHIND THE THICK ...	0.0	0.0
921633653129486336	@kuring01 @Bern...	ALDUBxEBLoveis	Businesswoman, lo...	0.0	0.0
921633675531509760	Cheer up, tomorro...	ALDUBxEBLoveis	BEHIND THE THICK ...	0.0	0.0
921633685681594368	Want to work in #...	Job	We're a leading #...	11.0	3.0
921633687447498752	#FurkanPalalı Değ...	FurkanPalalı	null	1.0	1.0
921633730678026240	#Top3Apps de ayer...	Top3Apps	Las tendencias de...	4.0	4.0
921633762655453185	One of the best s...	ALDUBxEBLoveis	MaiChard Shipper ...	0.0	0.0
921633778138099712	37 福田秀平 嵐呼べよ福田 風...	sbhawks	いちおう ホークス専用のアカウント...	5.0	5.0
921633803147157504	Interested in a #...	job	HCR ManorCare is ...	3.0	3.0
921633805873459200	It's been a long ...	ALDUBxEBLoveis	Resilient. Object...	0.0	0.0
921633836269649920	#Top3Apps for Fri...	Top3Apps	Information about...	4.0	4.0
921633939026046978	Cried bucket on t...	ALDUBxEBLoveis	Certified Aldub a...	0.0	0.0
921633941039321089	Dont talk if you ...	ALDUBxEBLoveis	BEHIND THE THICK ...	0.0	0.0
921633982789304321	Want to work at 6...	hiring	Follow this accou...	8.0	8.0
921634052674813953	If someone wants ...	ALDUBxEBLoveis	BEHIND THE THICK ...	0.0	0.0
92163406060826114	Questa sera start...	music	#sfamalamusicadel...	9.0	9.0

only showing top 20 rows

Overall Precision: 0.9838709677419355
Overall Recall 0.9607843137254901

Code to show Training-Test split / MulticlassClassificationEvaluator:

```
val cv = new TrainValidationSplit()
    .setEstimator(pipeline)
    .setEvaluator(new MulticlassClassificationEvaluator)
    .setEstimatorParamMaps(paramGrid)
    .setTrainRatio(0.8)
    .setParallelism(2)

val Array(trainingData, testData) = tweetsDF.randomSplit(Array(0.8, 0.2))

val logisticModel: TrainValidationSplitModel = cv.fit(trainingData)
```

Source Used: <https://spark.apache.org/docs/2.2.0/mllib-evaluation-metrics.html>

<https://spark.apache.org/docs/latest/ml-features.html#tf-idf>

Slides: [link to presentation slides](#)