CS 4830: Final Project

Team **Lockdown**: Kavish Shah, Akshara B, Vishal H, Vishnu Harshith

Indian Institute of Technology Madras

Introduction

The purpose of this project is to predict ratings based on massive data set containing Yelp reviews

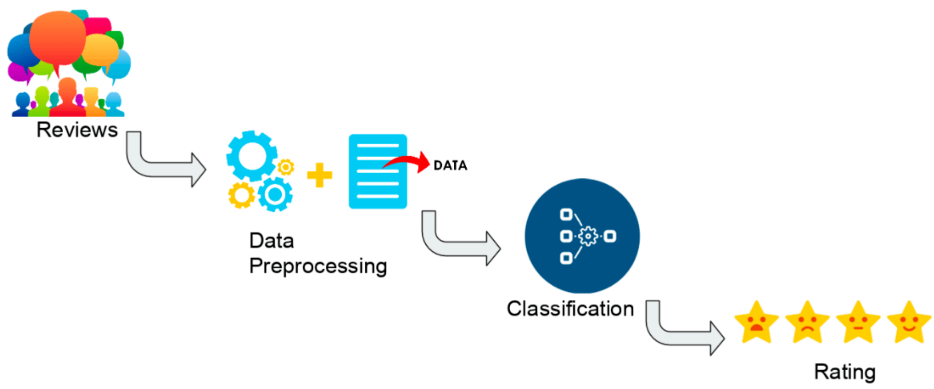


Fig 1: Workflow for finding correlation b/w correlation between reviews and ratings

Steps Involved

1. Batch Computation:
   Used a DataProc Cluster and submit a Spark job for data pre-processing and model training. Store a model in GCS bucket, and submit a spark job for evaluation on validation dataset stored in the bucket

2. Real Time Computation:
   Streamed data stored on the GCS bucket into Kafka and used spark streaming to read the data and make real-time predictions.

Data-Set description

The data set is a Jason file, roughly of 9GB size and has the following columns:


```
{"review_id":"xQY8N_XvtGbearJ5X4QryQ","user_id":"OwjRMXRC0KyPrIlcjaXeFQ","business_id":"-MhfebM0QIsKt87iDN-FNw","stars":
gallery on my most recent trip to Las Vegas. When I saw they would be showing infamous eggs of the House of Faberge from
the Gallery is pretty much hidden from view. It's what real estate agents would call \"cozy\" or \"charming\" - basicall
the two *s you ask? Let me tell you:\n\n* pricing for this, while relatively inexpensive for a Las Vegas attraction, is
kid friendly at all. Seriously, don't bring them.\n* the security is not trained properly for the show. When the curatin
art in a certain sequence, whether it be by historical period or cultural significance (this is how audio guides are usu
helpful. I was told to \"just look around\" and \"do whatever.\" \n\nAt such a *fine* institution, I find the lack of kn
{"review_id":"UmFMZ8PyXZTY2QcwzsfQYA","user_id":"nIJD_7ZXHq-FX8byPMOkMQ","business_id":"lbrU8StCq3yDfr-QMnGrmQ","stars":
son needed a haircut this past summer and the lure of the $7 kids cut signs got me in the door. We had to wait a few min
they had at the time was a pleather sofa with giant holes in it. And my son noticed ants crawling all over the floor and
walked out upon entering but I didn't. So the older black male stylist finishes the haircut he was doing and it's our tu
Really freaking simple, right? WRONG! Rather than use the clippers and go up to actually cut the hair, he went down. Usi
cosmetology license not know how to use a set of freaking clippers??? I realized almost immediately that he had no idea
know what? That's fine.\", paid and left. All I wanted to do was get out of that scummy joint and take my son to a real
{"review_id":"LG2ZaYiOgpr2DK_90pYjNw","user_id":"V34qejxNsCbcgD8C0HVk-Q","business_id":"HQl28KMwrEKHqhFrrDqVNQ","stars":
festive. The shrimp tacos and house fries are my standbys. The fries are sometimes good and sometimes great, and the spi
are mid-level, so it's not a cheap dive you can go to every week, but rather a treat when you do. Try it out. You won't
{"review_id":"i6g_oA9Yf9Y31qt0wibXpw","user_id":"ofKDkJKXSKZXu5xJNGiiBQ","business_id":"5JxlZaqCnk1MnbgRirs40Q","stars":
unengaged waiter;\n\nClueless manager, who seeing us with barely nibbled entrees\non plates shoved forward for pickup, t
\nwho serve up grub this bad,\nfor their own protection.\n\nNever, never, NEVER again\n(Back to Yard House for real food
{"review_id":"6TdNDKywdbjoTkizeMce8A","user_id":"UgMW8bLE0QMJDCkQ1Ax5Mg","business_id":"IS4cv902ykd8wj1TR0N3-A","stars":
are griping about the Drive thru is packed just like most of the other canes in the area but I like to go sit down to en
clean up crew zips around the dining area constantly so it's usually well kept. My only gripe is the one fella with Red
just put on the table to help go with my meal. After I was done he just reached for my tray no \"excuse me or are you do
being said still like to grub here and glad that there's finally a Cane's close to me.","date":"2017-01-14 21:56:57"}
{"review_id":"L2O_INwlrRuoX05KSjc4eg","user_id":"5vD2kmE25YBrbayKhykNxQ","business_id":"nlxHRv1zXGT0c0K51q3jDg","stars":
everything, and everything I see and taste here tells me that they're using top-grade fresh ingredients. The brisket san
menu to try different things. Definitely a step up from Subway, Quiznos, Jimmy Johns, etc in my opinion. As with all of
them.","date":"2013-05-07 07:25:25"}
```


Field name	Type	Mode	Policy tags	Description
text	STRING	NULLABLE		
cool	INTEGER	NULLABLE		
funny	INTEGER	NULLABLE		
stars	FLOAT	NULLABLE		
date	TIMESTAMP	NULLABLE		bq-datetime
business_id	STRING	NULLABLE		
user_id	STRING	NULLABLE		
useful	INTEGER	NULLABLE		
review_id	STRING	NULLABLE		

Fig: 2 Different columns in the yelp data set obtained using BigQuery

Context

This dataset is a subset of Yelp's businesses, reviews, and user data. Various platforms are present for people express their views on a matter. To get a general public opinion one may have to read all the views on the matter. This leads to a subclass of data mining problems known as sentiment analysis. We present below below the various pre-processing techniques adopted in this problem which may typically be used to work on any other massive data set as well.

Pre-processing

Loading the data

Loaded the JSON formated data into the google cloud bucket using gs util command. Imported the data into BigQuery table and defined the schema.

Cleaning the Data

We perform the pre-processing of the data in a DataProc Cluster using Spark.The major hurdle dealing with text based models is that it consumes a lot of time pre-processing it. We have used many pre-processing techniques for this project.

text	cool	funny	stars	date	business_id	user_id	useful	review_id
I just had a pedi...	0	0	2.0	2014-08-03 21:38:45	MJ5xkXJ2UwJL60QHu...	MXhsNk1DQw5qvwieZ...	3	9l2_cfGWNvf0nZudP...
We scheduled an a...	0	0	1.0	2015-05-18 19:22:54	r4cdabWEmTr211OFu...	NXAX_SSLw2hxiiAms...	3	V74eUOstrSH5wpMgF...
This old favorite...	1	1	4.0	2009-01-11 03:35:34	9DTypMuJS6GQZYdUW...	25N2f7Al7Tgu9zQE-...	2	6PyhrNqHrolvxZ6YD...
This place was hu...	0	0	3.0	2017-07-03 03:15:45	7JWWeXqil03Ua_wV9...	s72nhH1Hl_lN_Ju4h...	0	Z_ShCWTEdEEvca9Fs...
Went there to wat...	0	0	3.0	2013-09-20 04:24:55	Y7T9osP-hoqp2nLnm...	xSt--5JG6QqTEOwXJ...	0	jsf1I9Sy1w3VVnkg1...
Played cornhole f...	1	0	4.0	2017-10-07 23:35:14	3Jq5LfJ5fmJ5KmuA6...	dXP2z-Mqrlk2p2uKU...	1	yp5g42ulO-bBEnUhH...
Invited a friend ...	0	0	1.0	2015-05-05 21:08:44	dQWE63uOiSfaf3fup...	_IR48ok0ZkPMWJ2Pl...	0	4kn_D8-XXseuNuGsv...
What an incredibl...	0	0	5.0	2013-03-28 23:11:14	-4TMQnQJW1yd6NqGR...	cf-n5eNXJE7DG0ijb...	0	ytTNQOOfooqMWofal...
Excited to give t...	0	0	2.0	2012-10-19 05:36:10	-4TMQnQJW1yd6NqGR...	7A87zOycTC5D108bh...	0	Q2fHu4Zx0Ts1l5RaZ...
This place is qui...	0	0	4.0	2017-10-08 23:37:14	z5KWNSDvgv-4I62P8...	AisJLgSqMlfkeae9y...	0	84jtzzhtl7Vk8WBmi...

only showing top 10 rows

Fig 3: Original Dataframe

Most of them are imported from PySpark ML features.Tokenizing the words is where we split the entire sentences into words and the order of the words will not hold any importance any more. We performed following functions over the text reviews:

1. **Convert to Lower case:**
   Converted all the reviews to lower case to remove any discrepancies in the corpus

text	cool	funny	stars	date	business_id	user_id	useful	review_id	words
I just had a pedi...	0	0	2.0	2014-08-03 21:38:45	MJ5xkXJ2UwJL60QHu...	MXhsNk1DQw5qvwieZ...	3	9l2_cfGWNvf0nZudP...	[i, just, had, a,...
We scheduled an a...	0	0	1.0	2015-05-18 19:22:54	r4cdabWEmTr211OFu...	NXAX_SSLw2hxiiAms...	3	V74eUOstrSH5wpMgF...	[we, scheduled, a...
This old favorite...	1	1	4.0	2009-01-11 03:35:34	9DTypMuJS6GQZYdUW...	25N2f7Al7Tgu9zQE-...	2	6PyhrNqHrolvxZ6YD...	[this, old, favor...
This place was hu...	0	0	3.0	2017-07-03 03:15:45	7JWWeXqil03Ua_wV9...	s72nhH1Hl_lN_Ju4h...	0	Z_ShCWTEdEEvca9Fs...	[this, place, was...
Went there to wat...	0	0	3.0	2013-09-20 04:24:55	Y7T9osP-hoqp2nLnm...	xSt--5JG6QqTEOwXJ...	0	jsf1I9Sy1w3VVnkg1...	[went, there, to,...
Played cornhole f...	1	0	4.0	2017-10-07 23:35:14	3Jq5LfJ5fmJ5KmuA6...	dXP2z-Mqrlk2p2uKU...	1	yp5g42ulO-bBEnUhH...	[played, cornhole...
Invited a friend ...	0	0	1.0	2015-05-05 21:08:44	dQWE63uOiSfaf3fup...	_IR48ok0ZkPMWJ2Pl...	0	4kn_D8-XXseuNuGsv...	[invited, a, frie...
What an incredibl...	0	0	5.0	2013-03-28 23:11:14	-4TMQnQJW1yd6NqGR...	cf-n5eNXJE7DG0ijb...	0	ytTNQOOfooqMWofal...	[what, an, incred...
Excited to give t...	0	0	2.0	2012-10-19 05:36:10	-4TMQnQJW1yd6NqGR...	7A87zOycTC5D108bh...	0	Q2fHu4Zx0Ts1l5RaZ...	[excited, to, giv...
This place is qui...	0	0	4.0	2017-10-08 23:37:14	z5KWNSDvgv-4I62P8...	AisJLgSqMlfkeae9y...	0	84jtzzhtl7Vk8WBmi...	[this, place, is,...

only showing top 10 rows

2. **Remove Stop Words**
   Stop words such as prepositions or articles (”a”, ”the”) are also removed so that we only keep the more meaningful words in the corpus.

text	cool	funny	stars	date	business_id	user_id	useful	review_id	words	words_nsw
I just had a pedi...	0	0	2.0	2014-08-03 21:38:45	MJ5xkXJ2UwJL60QHu...	MXhsNk1DQw5qvwieZ...	3	9l2_cfGWNvf0nZudP...	[i, just, had, a,...	[pedi, &, acrylic...
We scheduled an a...	0	0	1.0	2015-05-18 19:22:54	r4cdabWEmTr211OFu...	NXAX_SSLw2hxiiAms...	3	V74eUOstrSH5wpMgF...	[we, scheduled, a...	[scheduled, appoi...
This old favorite...	1	1	4.0	2009-01-11 03:35:34	9DTypMuJS6GQZYdUW...	25N2f7Al7Tgu9zQE-...	2	6PyhrNqHrolvxZ6YD...	[this, old, favor...	[old, favorite, m...
This place was hu...	0	0	3.0	2017-07-03 03:15:45	7JWWeXqil03Ua_wV9...	s72nhH1Hl_lN_Ju4h...	0	Z_ShCWTEdEEvca9Fs...	[this, place, was...	[place, hugely, u...
Went there to wat...	0	0	3.0	2013-09-20 04:24:55	Y7T9osP-hoqp2nLnm...	xSt--5JG6QqTEOwXJ...	0	jsf1I9Sy1w3VVnkg1...	[went, there, to,...	[went, watch, eag...
Played cornhole f...	1	0	4.0	2017-10-07 23:35:14	3Jq5LfJ5fmJ5KmuA6...	dXP2z-Mqrlk2p2uKU...	1	yp5g42ulO-bBEnUhH...	[played, cornhole...	[played, cornhole...
Invited a friend ...	0	0	1.0	2015-05-05 21:08:44	dQWE63uOiSfaf3fup...	_IR48ok0ZkPMWJ2Pl...	0	4kn_D8-XXseuNuGsv...	[invited, a, frie...	[invited, friend,...
What an incredibl...	0	0	5.0	2013-03-28 23:11:14	-4TMQnQJW1yd6NqGR...	cf-n5eNXJE7DG0ijb...	0	ytTNQOOfooqMWofal...	[what, an, incred...	[incredibly, gorg...
Excited to give t...	0	0	2.0	2012-10-19 05:36:10	-4TMQnQJW1yd6NqGR...	7A87zOycTC5D108bh...	0	Q2fHu4Zx0Ts1l5RaZ...	[excited, to, giv...	[excited, give, h...
This place is qui...	0	0	4.0	2017-10-08 23:37:14	z5KWNSDvgv-4I62P8...	AisJLgSqMlfkeae9y...	0	84jtzzhtl7Vk8WBmi...	[this, place, is,...	[place, quite, ha...

only showing top 10 rows

3. **Unigrams & trigrams - Frequency > 20**
   We are going attempt two modeling approaches. The first one is to split the model into trigrams (e.g. every three words in a single review will be split together. We will then pick out the trigrams that appear more than 20 times in the corpus, so that we can eliminate any random phrases that only appear once or twice.
   These phrases, will be then joined together, using an underscore ””, and be replaced in the original text. The pipeline of Tokenize → CountVectorizer (BagOfWords) → TF-IDF . Model will then be applied using the new texts. This step ensures that we have a well-mixed combinations of unigrams and trigrams in our training data.

text	cool	funny	stars	date	business_id	user_id	useful	review_id	words	words_nsw	ngram
I just had a pedi...	0	0	2.0	2014-08-03 21:38:45	MJ5xkXJ2UwJL60QHu...	MXhsNk1DQw5qvwieZ...	3	9l2_cfGWNvf0nZudP...	[i, just, had, a,...	[pedi, &, acrylic...	[i just had, just...
We scheduled an a...	0	0	1.0	2015-05-18 19:22:54	r4cdabWEmTr211OFu...	NXAX_SSLw2hxiiAms...	3	V74eUOstrSH5wpMgF...	[we, scheduled, a...	[scheduled, appoi...	[we scheduled an,...
This old favorite...	1	1	4.0	2009-01-11 03:35:34	9DTypMuJS6GQZYdUW...	25N2f7Al7Tgu9zQE-...	2	6PyhrNqHrolvxZ6YD...	[this, old, favor...	[old, favorite, m...	[this old favorit...
This place was hu...	0	0	3.0	2017-07-03 03:15:45	7JWWeXqil03Ua_wV9...	s72nhH1Hl_lN_Ju4h...	0	Z_ShCWTEdEEvca9Fs...	[this, place, was...	[place, hugely, u...	[this place was, ...
Went there to wat...	0	0	3.0	2013-09-20 04:24:55	Y7T9osP-hoqp2nLnm...	xSt--5JG6QqTEOwXJ...	0	jsf1I9Sy1w3VVnkg1...	[went, there, to,...	[went, watch, eag...	[went there to, t...
Played cornhole f...	1	0	4.0	2017-10-07 23:35:14	3Jq5LfJ5fmJ5KmuA6...	dXP2z-Mqrlk2p2uKU...	1	yp5g42ulO-bBEnUhH...	[played, cornhole...	[played, cornhole...	[played cornhole ...
Invited a friend ...	0	0	1.0	2015-05-05 21:08:44	dQWE63uOiSfaf3fup...	_IR48ok0ZkPMWJ2Pl...	0	4kn_D8-XXseuNuGsv...	[invited, a, frie...	[invited, friend,...	[invited a friend...
What an incredibl...	0	0	5.0	2013-03-28 23:11:14	-4TMQnQJW1yd6NqGR...	cf-n5eNXJE7DG0ijb...	0	ytTNQOOfooqMWofal...	[what, an, incred...	[incredibly, gorg...	[what an incredib...
Excited to give t...	0	0	2.0	2012-10-19 05:36:10	-4TMQnQJW1yd6NqGR...	7A87zOycTC5D108bh...	0	Q2fHu4Zx0Ts1l5RaZ...	[excited, to, giv...	[excited, give, h...	[excited to give,...
This place is qui...	0	0	4.0	2017-10-08 23:37:14	z5KWNSDvgv-4I62P8...	AisJLgSqMlfkeae9y...	0	84jtzzhtl7Vk8WBmi...	[this, place, is,...	[place, quite, ha...	[this place is, p...

only showing top 10 rows

Model Architecture

We have used two basic methods to determine the rating of the reviews:

Support Vector Machine

With a train-test split ratio of 80:20 we then run 50 iterations of Support Vector Machine (SVM) model on the training data. The parallel computing of the Spark attributes to speeding up the entire process. Given an imbalanced data-set, it is important know which classification metrics we are going to optimize. Accuracy in such a scenario will not be representitaive of how well the model is performing, while F1 score – as a weighted average of precision and recall could reveal how well the model performs in identifying both the prediction relevancy and what % of truly relevant results are correctly predicted. The parameters used here are:

1. Number of Iterations = 50
2. Regularization Parameter = 0.3
3. Learning Algorithm: Stochastic Gradient Descent

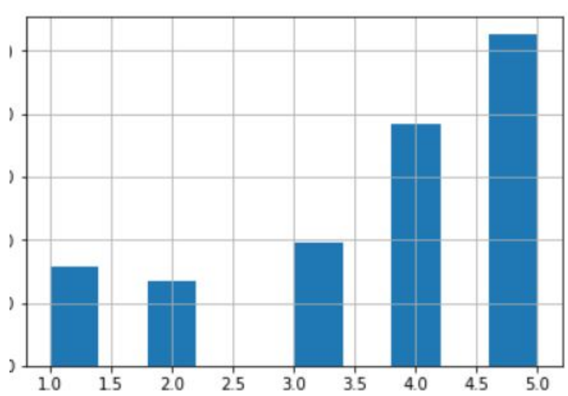
Elastic Net Logistic Regression

Using the same training data, we instead applied a regularized logistic regression. SVM focuses on finding the separating plane that maximizes the distance of the closest points to the margin, whereas Logistic Regression maximizing the probability of the data (i.e. the further it lies away from the hyperplane the better). In addition to a normal logistic regression, we used a linear combination of L1 and L2 regularization, i.e. Elastic Net, to prevent the model from overfitting. Since LASSO (L1) tend to select only one significant coefficients the other, EN adds in the penalty from Ridge Regression (L2) that helps to overcome the disadvantages.

1. Reg Parameter = 0.3
2. ElasticNet Parameter = 0.8
3. Max Iterations = 100

Model performance

The dataset is quite imbalanced and hence we used the F1 score to evaluate the model performance.



textbfFig 4: F1 score using logistic regression

Logistic Regression

Below is the code used -

```
from __future__ import print_function
from pyspark.ml.classification import LogisticRegression
from pyspark.ml import Pipeline, PipelineModel
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
import os
from pyspark import SparkContext
from pyspark.sql.session import SparkSession
from pyspark.streaming import StreamingContext
import pyspark.sql.types as tp
from pyspark.ml.feature import *
from pyspark.ml import Pipeline
from pyspark.ml.feature import StringIndexer, OneHotEncoderEstimator, VectorAssembler
from pyspark.ml.feature import StopWordsRemover, Word2Vec, RegexTokenizer
from pyspark.ml.classification import LogisticRegression
#from pyspark.mllib.feature import HashingTF, IDF

sc = SparkContext(appName="Project")
```

```python
spark = SparkSession(sc)

spark_df = spark.read.format("bigquery").option("table", "bdl_project20.yelp").load().
    toDF("text", "cool", "funny", "stars", "date", "business_id","user_id","useful","
    review_id")

spark_df.show(10)

tok =  RegexTokenizer(inputCol= "text" , outputCol= "words", pattern= '\\W')
stopword_rm = StopWordsRemover(inputCol="words", outputCol="words_nsw")
hashingTF = HashingTF(inputCol="words_nsw", outputCol="rawFeatures")
idf = IDF(inputCol="rawFeatures", outputCol="features")
lr = LogisticRegression(featuresCol="features", labelCol="stars")
pipe = Pipeline(stages=[tok, stopword_rm,hashingTF,idf,lr])
model = pipe.fit(spark_df)
model.save("gs://bdl__project20/model/")
predictions = model.transform(spark_df)

predictions.show(10)

evaluator = MulticlassClassificationEvaluator(labelCol="stars", predictionCol="
    prediction", metricName="f1")

f1 = evaluator.evaluate(predictions)

print("F1 score =", f1)
```

```
20/09/20 06:47:35 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Querying table codelearn-281105.bdl_project20.y
20/09/20 06:47:35 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Going to read from codelearn-281105.bdl_project
20/09/20 06:47:35 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Created read session for table 'codelearn-28110
20/09/20 06:47:35 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Requested 14 max partitions, but only received
F1 score = 0.7076878667
20/09/20 07:00:05 INFO org.spark_project.jetty.server.AbstractConnector: Stopped Spark@1bf2ab2b{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
```

textbfFig 4: F1 score using logistic regression

Real-time computation

We used the trained model to perform real-time predictions of test data streaming to Kafka. We have used a part of the data for streaming.
Below is the code used -

```python
from __future__ import print_function
from pyspark import SparkContext
import pyspark.sql.types as tp
from pyspark.sql import Row
from pyspark.streaming import StreamingContext
from pyspark.ml.classification import LogisticRegression
from pyspark.streaming.kafka import KafkaUtils
import os
from pyspark.sql import *
from pyspark.sql.types import StructType, StructField, FloatType
from pyspark.ml.feature import VectorAssembler,StringIndexer, VectorIndexer,
    StandardScaler,IndexToString
from pyspark.ml.classification import LogisticRegression
from pyspark.sql.session import SparkSession
from pyspark.ml import Pipeline, PipelineModel
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
import json
from pyspark.sql import SQLContext

kafka_topic = 'from-pubsub'
zk = '10.182.0.2:2181' # Apache ZooKeeper quorum
app_name = 'from-pubsub' # Can be some other name
sc = SparkContext(appName="KafkaPubsub")
ssc = StreamingContext(sc, 20)
model = PipelineModel.load("gs://bdl__project20/model/")
schema = StructType([StructField("sl", FloatType(), True), StructField("sw", FloatType
    (), True),StructField("pl", FloatType(), True),StructField("pw", FloatType(), True
    )])
#spark = SparkSession.builder.appName(app_name).config("spark.master", "local").
    getOrCreate()
spark = SparkSession(sc)
kafkaStream = KafkaUtils.createStream(ssc, zk, app_name, {kafka_topic: 1})
#kafkaStream.pprint()
print("Starting")
def get_prediction(temp):
  print("Inside IF")
  temp.toDF().show()
  dstream = temp.map(lambda x: json.loads(x[1]))
  dstream.toDF().show()
  dat = dstream.map(lambda x:x[0])
  dat.toDF().show()
```

```
    #dat1 = dat.map(lambda x: x.split(','))
    #dat2 = dat1.map(lambda x: [float(i) for i in x])
    #dat3 = dat2.map(lambda x: x[1:])
    dat4 = dat.map(lambda x: Row(text=str(x)))
    #df = dat.toDF().toPandas()
    #df.columns = ['text']
    #print(df)
    df1 = spark.createDataFrame(dat4)
    df1.show()
    pred = model.transform(df1)
    pred.show()
kafkaStream.foreachRDD(lambda x: get_prediction(x))
print("Ending")
ssc.start()
ssc.awaitTermination()
```

Publishing rate is 10 seconds.Kafka was ingesting at a rate of 20 seconds.

Conclusion

The Logistic Regression model performs better that SVM. Nowadays, real-time computing generally needs to process large amounts of data, in addition to meeting some of the requirements of non-real-time computing (e.g., accurate results). The most important requirement of real-time computing is the response to computing results in real time. This project enabled us to understand the importance of real time computing.