Big Data Frameworks

M2

Efrei Paris

Data Engineering

2022/2023

Final project Data Frames, Machine Learning And NLP with Spark

Salim NAHLE

Organization:

- You can work on any Spark environment
- Deliverables:
-
    - 15 minutes presentation (7 minutes pitch + 8 minutes Q&A): Tuesday 28/02/2023
    - A **PDF report** is expected. It shall contain the code, explanations and necessary screenshots.
- Please work in **pairs**! Each group (composed of 2 persons at most) shall submit one report. Do not forget to indicate your names in the report.
- The report shall be uploaded on Teams before **Monday 27/02/2023 midnight.**
- Late reports are not accepted

Abstract:

- This miniproject is divided into 3 parts:
    - Part 1:
        - Movielens dataset analysis using (Dataframes)
    - Part 2:
        - Bike Rental prediction using Spark machine learning library
    - Part 3:
        - SMS classification into Spam/Safe using spark NLP featurization tools and spark classification algorithms
        - This part can be replaced by any idea of your choice including natural language processing with Spark (movielens tag analysis, any topic related chatGPT).

    - Part 4: Bonus (2 points)
        - Finish the AWS cloud practitioner track labs (100% → 2 points bonus)

Part 1: Movielens Data Analysis

Data set:

https://grouplens.org/datasets/movielens/

- Create an S3 bucket (s3-de1-<name>) , with several folders inside it:
    - Input
    - Output
    - Scripts
    - Logs
- Upload the movielens data to <your bucket>/input/movielens (alternatively use my bucket) s3://nahle-bucket-datalake/emr/input/movielens/[file-name].csv
- Create a key-pair ((key_de1_<name>)
- Create an EMR cluster (1 master and 2 workers)
- Cluster creation will take few minutes, once the cluster is created, connect to the master node using ssh
- Project :
    - **We want to create an ETL process with spark:**
        - Load the data into your Hadoop cluster and create the necessary dataframes
        - Explore the different DFs
        - The objective is to analyze the movielens data, transform them and create a new dataset, in which we have the following fields:
            - Movie ID
            - Movie name
            - Year of release
            - Number of ratings
            - Genre
            - Rating average
        - Load the new dataset into a parquet file (or CSV) <your bucket>/output/movielens/
    - Load the new dataset (CSV or parquet) into apache Spark and create a dataframe and an SQL view: Then using Spark DataFrames and SQL, write the following queries:
        - Best movie per year
        - Best movie per genre
        - For 'action' movie per year
        - Best romance per year

Part 2: Bike Rental Data Set from UCI Machine Learning Repository

1. Citations

Reconsider the Bike Rental data set and the provided notebook.

Fanaee-T, Hadi, and Gama, Joao, 'Event labeling combining ensemble detectors and background knowledge', Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelber

2. Attributes on original data

- season : season (1:springer, 2:summer, 3:fall, 4:winter)
- yr : year (0: 2011, 1:2012)
- mnth : month (1 to 12)
- hr : hour (0 to 23)
- holiday : weather day is holiday or not (extracted from [Web Link])
- weekday : day of the week
- workingday : if day is neither weekend nor holiday is 1, otherwise is 0.
- weathersit :
  - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
  - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp : Normalized temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-8, t_max=+39 (only in hourly scale)
- hum: Normalized humidity. The values are divided to 100 (max)
- windspeed: Normalized wind speed. The values are divided to 67 (max)

3. URL:

https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

4. Consulting Project

You have been contacted to build a predictive model to help Bike Rental companies in predicting the hourly and daily demand on bikes.
A first linear model to predict the 'demands' was provided. The evaluation of this first model (meanAbsoluteError and r2) was very far from being satisfactory.

- Try to get some insights from the results you obtained:
    - Display, for instance, the average real demand versus the average predicted demand and the standard deviation of both by grouping your data by:
        - hour
        - season
        - other features that you think useful
- Add dummy variables to improve your model. (use MAE and R2 for comparison)
- Improve the model by doing cross validation. You shall tune and cross-validate the model using:
    - pyspark.ml.Pipeline
    - pyspark.ml.tuning.ParamGridBuilder
    - pyspark.ml.tuning.CrossValidator
- Try other machine learning algorithms and compare.

Part 3: SMS Spam Collection Data Set

1. Understanding the data set

a) Direct Link

Direct link on UCI's machine learning repository:

https://archive.ics.uci.edu/ml/machine-learning-databases/00228/

b) Data Set Information :

This corpus has been collected from free or free for research sources at the Internet:

- A collection of 425 SMS spam messages was manually extracted from the Grumbletext Web site. This is a UK forum in which cell phone users make public claims about SMS spam messages, most of them without reporting the very spam message received. The identification of the text of spam messages in the claims is a very hard and time-consuming task, and it involved carefully scanning hundreds of web pages.
- A subset of 3,375 SMS randomly chosen ham messages of the NUS SMS Corpus (NSC), which is a dataset of about 10,000 legitimate messages collected for research at the Department of Computer Science at the National University of Singapore. The messages largely originate from Singaporeans and mostly from students attending the University. These messages were collected from volunteers who were made aware that their contributions were going to be made publicly available.
- A list of 450 SMS ham messages collected from Caroline Tag's PhD Thesis
- Finally, we have incorporated the SMS Spam Corpus v.0.1 Big. It has 1,002 SMS ham messages and 322 spam messages and it is public. This corpus has been used in the following academic researches:

[1] GÃ³mez Hidalgo, J.M., Cajigas Bringas, G., Puertas Sanz, E., Carrero GarcÃa, F. Content Based SMS Spam Filtering. Proceedings of the 2006 ACM Symposium on Document Engineering (ACM DOCENG'06), Amsterdam, The Netherlands, 10-13, 2006.

[2] Cormack, G. V., GÃ³mez Hidalgo, J. M., and Puertas SÃ¡nz, E. Feature engineering for mobile (SMS) spam filtering. Proceedings of the 30th Annual international ACM Conference on Research and Development in information Retrieval (ACM SIGIR'07), New York, NY, 871-872, 2007.

[3] Cormack, G. V., GÃ³mez Hidalgo, J. M., and Puertas SÃ¡nz, E. Spam filtering for short messages. Proceedings of the 16th ACM Conference on Information and Knowledge Management (ACM CIKM'07). Lisbon, Portugal, 313-320, 2007.

c) Attribute Information:

The collection is composed by just one text file, where each line has the correct class followed by the raw message. We offer some examples bellow:

ham What you doing?how are you?

ham Ok lar... Joking wif u oni...

ham dun say so early hor... U c already then say...

ham MY NO. IN LUTON 0125698789 RING ME IF UR AROUND! H*

spam FreeMsg: Txt: CALL to No: 86888 & claim your reward of 3 hours talk time to use from your phone now! ubscribe6GBP/ mnth inc 3hrs 16 stop?txtStop

spam Sunshine Quiz! Win a super Sony DVD recorder if you canname the capital of Australia? Text MQUIZ to 82277. B

spam URGENT! Your Mobile No 07808726822 was awarded a L2,000 Bonus Caller Prize on 02/09/03! This is our 2nd attempt to contact YOU! Call 0871-872-9758 BOX95QU

Note: the messages are not chronologically sorted.

d) Relevant Papers:

We offer a comprehensive study of this corpus in the following paper. This work presents a number of statistics, studies and baseline results for several machine learning methods.

Almeida, T.A., GÃ³mez Hidalgo, J.M., Yamakami, A. Contributions to the Study of SMS Spam Filtering: New Collection and Results. Proceedings of the 2011 ACM Symposium on Document Engineering (DOCENG'11), Mountain View, CA, USA, 2011.

e) Citation

http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/

2. Consulting Project

You have been contacted to build a predictive model to classify an incoming sms as Spam or Safe. You are provided the dataset described above. For each observation in the dataset, the row sms as well as the correct class is given.

a) The first step is to change the text features into numeric using the suitable classes (StringIndexer, Tokenizer, StopWordsRemover, CountVectorizer, IDF, VectorAssembler).

b) Then You shall train 4 classifiers and compare them. These are:
   1. LogisticRegression,
   2. DecisionTreeClassifier
   3. RandomForestClassifier
   4. NaiveBayes

c) For one of these classifiers, you shall tune at least one important hyper parameter using ParamGridBuilder and CrossValidator

d) Conclusions: Compare and comment the obtained results (you may use a comparison table).

Required libraries:

- pyspark.ml.feature.StringIndexer
- pyspark.ml.feature.Tokenizer
- pyspark.ml.feature.StopWordsRemover
- pyspark.ml.feature.CountVectorizer
- pyspark.ml.feature.IDF
- pyspark.ml.feature.VectorAssembler
- pyspark.ml.Pipeline
- pyspark.ml.classification.LogisticRegression
- pyspark.ml.classification.RandomForestClassifier
- pyspark.ml.classification.DecisionTreeClassifier
- pyspark.ml.classification.NaiveBayes
- pyspark.ml.tuning.ParamGridBuilder
- pyspark.ml.tuning.CrossValidator
- pyspark.ml.evaluation.MulticlassClassificationEvaluator