CISC 5950 — Project 2

In CISC 5950, we have learned the following topics,

1. Set up a 3-node cluster with Hadoop Distributed File System and run examples.
2. On top of HDFS, set up the cluster with MapReduce programming framework.
3. Run examples of MapReduce programs.
4. Scheuling on the Cloud.
5. Spark Programming.
6. Spark MLlib / ML with DataFrame
7. Spark Streaming

In this project, we are going to use a spark cluster for data processing jobs. The project consist of 3+1 parts.

P1: Toxic Comment Classification

Discussing things you care about can be difficult. The threat of abuse and harassment online means that many people stop expressing themselves and give up on seeking different opinions. Platforms struggle to effectively facilitate conversations, leading many communities to limit or completely shut down user comments.

In this part, we are going to practise basic text processing using the Apache Spark with the use of the toxic comment text classification dataset. The machine learning and text processing used here are at a poor standard. The goal was mainly to convert the column comment_text into a column of sparse vectors for use in a classification algorithm in the spark ml library.

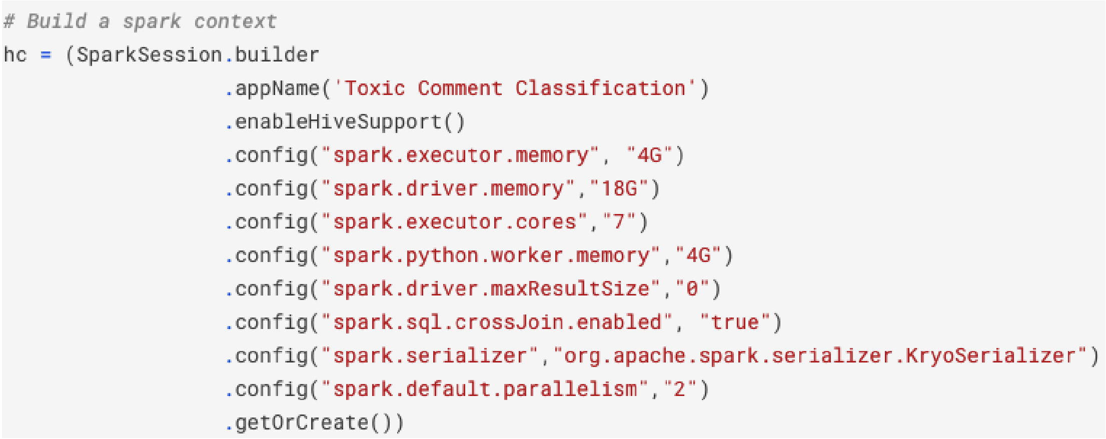
The **pyspark.ml** library is used for machine learning with Spark DataFrames. For machine learning with Spark RDDs use the **pyspark.mllib** library.

You can fine the sample solution from this link, Simple Text Mining. You can download the data from this link, dataset.

To start the project, you have to,

1. Start the 3-node spark cluster based on HDFS
2. Store the data in HDFS
3. Use DataFrame to prepare the data for Logistic Regression APIs.

You have to study the sample code and make it work on your spark cluster.

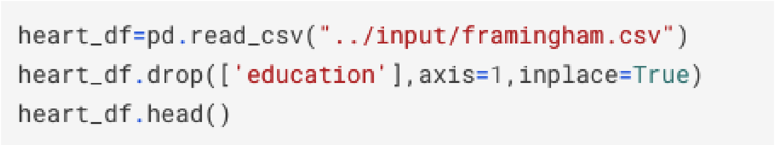


```
# Build a spark context
hc = (SparkSession.builder
                  .appName('Toxic Comment Classification')
                  .enableHiveSupport()
                  .config("spark.executor.memory", "4G")
                  .config("spark.driver.memory","18G")
                  .config("spark.executor.cores","7")
                  .config("spark.python.worker.memory","4G")
                  .config("spark.driver.maxResultSize","0")
                  .config("spark.sql.crossJoin.enabled", "true")
                  .config("spark.serializer","org.apache.spark.serializer.KryoSerializer")
                  .config("spark.default.parallelism","2")
                  .getOrCreate())
```

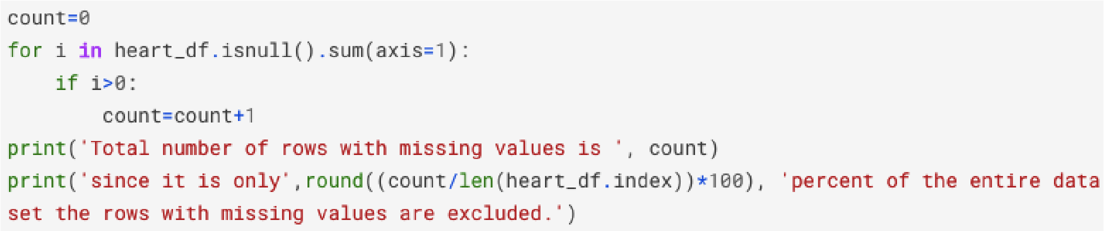
P2: Heart Disease Prediction using Logistic Regression

Logistic regression is a type of regression analysis in statistics used for prediction of outcome of a categorical dependent variable from a set of predictor or independent variables. In logistic regression the dependent variable is always binary. Logistic regression is mainly used to for prediction and also calculating the probability of success.

In this part, we are going to use Logistic Regression to to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk. We use the Framingham Heart dataset (link) for the study. The webpage (link) provides a sample solution with **python only**.



```
heart_df=pd.read_csv("../input/framingham.csv")
heart_df.drop(['education'],axis=1,inplace=True)
heart_df.head()
```



```
count=0
for i in heart_df.isnull().sum(axis=1):
    if i>0:
        count=count+1
print('Total number of rows with missing values is ', count)
print('since it is only',round((count/len(heart_df.index))*100), 'percent of the entire data
set the rows with missing values are excluded.')
```

You are expected to study the sample code and convert it into Spark version (ML or ML-lib).

P3: Logistic Regression classifier on Census Income Data

In this section, we will analyze the Census Dataset from the UCI Machine Learning Repository (link). It provides the following data,

1. Training Set: adult.data (link)
2. Test Set: adult.test (link)
3. Data Description: adult.name (link)

The data contains anonymous information such as age, occupation, education, working class, etc. The goal is to train a binary classifier to predict the income which has two possible values $> 50K$ and $< 50K$. There are 48842 instances and 14 attributes in the dataset. The data contains a good blend of categorical, numerical and missing values.

In this part, we are going use logistic regression on Spark ML / MLlib to train your model and evaluate it. You are expected to an accuracy rate after your evaluation.

P4 (Bonus Question)

Spark ML and MLlib provide a rich set of machine learning libraries, e.g. Random Forest Classifier(link) and Decision Tree Classifier (link).

As the bonus part, you are expected to redo P3 with Random Forest Classifier and Decision Tree Classifier by using the Apache Spark ML/MLlib.

Project Summary

Upon finishing the project, we have completed the following,

- P1: Learn the Logistic Regression in Spark based on the existing code.
- P2: Convert the python code to Apache Spark, which is very common in industry as their data grows.
- P3: Write your own code to utilize logisitic regression based on Spark ML/MLlib.
- P4(Bonus): Explore other machine learning algorithms in Spark.

Grading Rubric

You should complete the lab in groups of 4 students.

(25%) P1: Toxic Comment Classification;
(25%) P2: Heart Disease Prediction using Logistic Regression;
(30%) P3: Logistic Regression classifier on Census Income Data.
(15%) Detailed reports and README files;
(5%) Submission format;
(20%) P4: Random Forest (10%) and Decision Tree (10%).

Submission

You are expected to email me a zip(or tar) file by the deadline (May 5th, 2019). The zip file should include three (or four) folders,

- P1: your codes, report and README
- P2: your codes, report and README
- P3: your codes, report and README
- Bonus: your codes, report and README