Databricks Example Notebooks

The course resources including number of databricks sample notebooks. Following summarizes the contents of those notebooks. These notebooks may be of assistance to you when you are creating your group analytics project.

# Introduction

The introductory notebooks are mainly introduction to the spark language. However the regression and data notebooks provide guidance as to how you can access resources and do a simple regression.

1. GetData shows how to copy data into your notebook
2. Data Sources provides a list of online resources for projects.
3. Other notebook provide demonstration of data frames and regression

# Data operations

These notebooks mainly show how to use spark dataframe and associated operations on data frames.

# ML 1

Please notebooks demonstrate introductory uses of the spark machine learning API.

1. Pipline demonstrates how to use the spark pipeline in comparison to the Python pipeline
2. Transformation demonstrates the use of transformations in a pipeline , illustrated with a random forest model.
3. Logistic Regression demonstrates a relatively simple pipeline that processes natural language text. **Start here if you want to use natural language text for prediction**.
4. LDA demonstrates the use of LDA for clustering.
5. K-means Demonstrates the use of k-means for clustering.
6. These models demonstrates more complex pipelines and their interaction with parameter grid and cross validate.
7. Recommendation demonstrates a recommendation modeling using ALS.

# ML2 (advanced ML)

1. Advanced use of spark NLP.
2. Using a neural network to cluster documents found on Pub Med.
3. Creates a custom sentiment analysis for text.
4. Demonstrates the graph analysis API from Spark.
5. Recommendation demonstrates a recommendation modeling using ALS.
6. Demonstrates Pub Med document classification using tensors.
7. Demonstrates various binary classification models.