# Course Code: AIML427

# Course Title: Big Data

# Assignment 3 – Individual Part 2

George Wiles

Student ID: 300586465

Due Date: 8 June 2021, 17:00 pm

Date of Submission: 8 June 2021

# Group Machine Learning Task

1. Task Description and Overview

# Individual Machine Learning Task – AG News Classification Dataset

1. Task Description and Overview  
   According to Anand [1], AG is a collection of more than 1 million new articles gathered by the academic news search engine ComeToMyHead. The dataset contains 3 columns; the class index or response variable, and 2 text columns representing the article title and the article description.

This is an example of a multiclass classification problem, where the response variable (‘Class Index) is a value that represents 4 new classifiers (1-World, 2-Sports, 3-Business, 4-Science and Technology), that represents the title and sentence description of the each news article instance.

Original Dataset Size  
The AG News training set contains 120,00 training dataset instances (29 MB), and the test set contains 7600 testing samples (1.8 MB). It is assumed a typical text article can contain 500-800 words, which when applied to a TF-IDF (term frequency-inverse document frequency) numerical statistic a vectorized feature model containing both title and description fields could contain up to 1500 unique words or features. The Kaggle dataset [1] retrieved at <https://www.kaggle.com/amananandrai/ag-news-classification-dataset> has been pre-processed to ensure that there are no missing columns, the text and description columns contain quotes, commas and new line characters that have been escaped.

Expected Output  
This assignment uses two algorithms, logical regression and decision tree, executed on a spark Hadoop cluster to predict the multiclass response variable of the test set against the trained model using the training data set.

1. Processing Steps

The original dataset contains 2 features being the Title and Description text fields As both the title and description text fields can be represented as 500-800 features respectively, this assignment utilizes reg-ex tokenizers and stop words to reduce this feature account combining the cleaned/transformed features using a Spark vector assembler.

For both Title and Description columns the following pre-fitting processing steps occur:

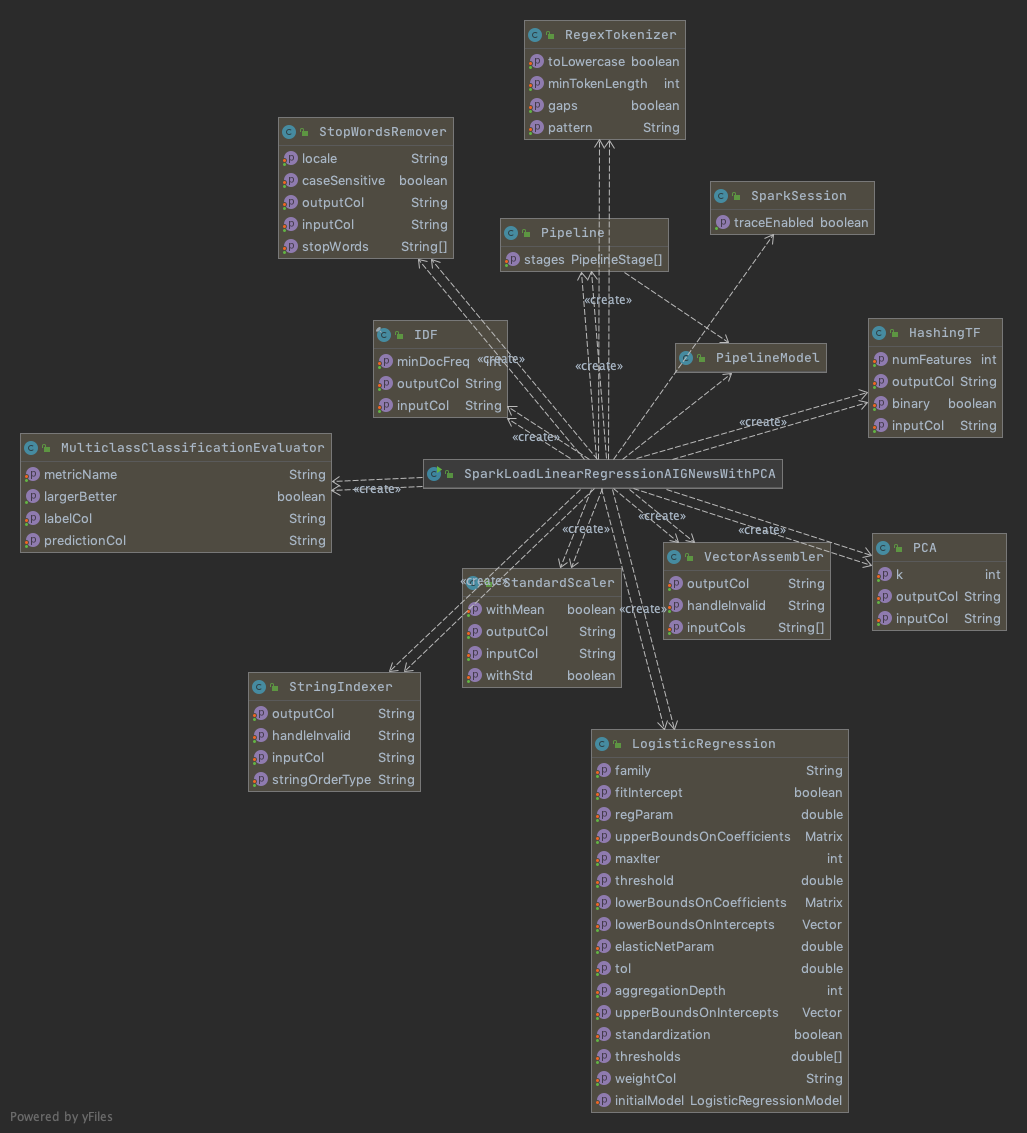
* 1. 1. Regex Tokenizer to
     1. Remove white space characters
     2. Set all word/features as lower case to remove duplication
     3. Set a minimum length token of 1 character to reduce punctuation.  
        *|Reuters - Stocks ...| - to - |[reuters, stocks,...|*
  2. Stop words remover, by default the Spark ML StopWordsRemover contains a list of common stop words (and, of, to et al).

*|stocks, assets, of| - to -.|[stocks, assets, nati...|*

* 1. Hashing term frequency (HashingTF vs CountVectorizer, both were investigated with as a pre-fitting processing data transform step.

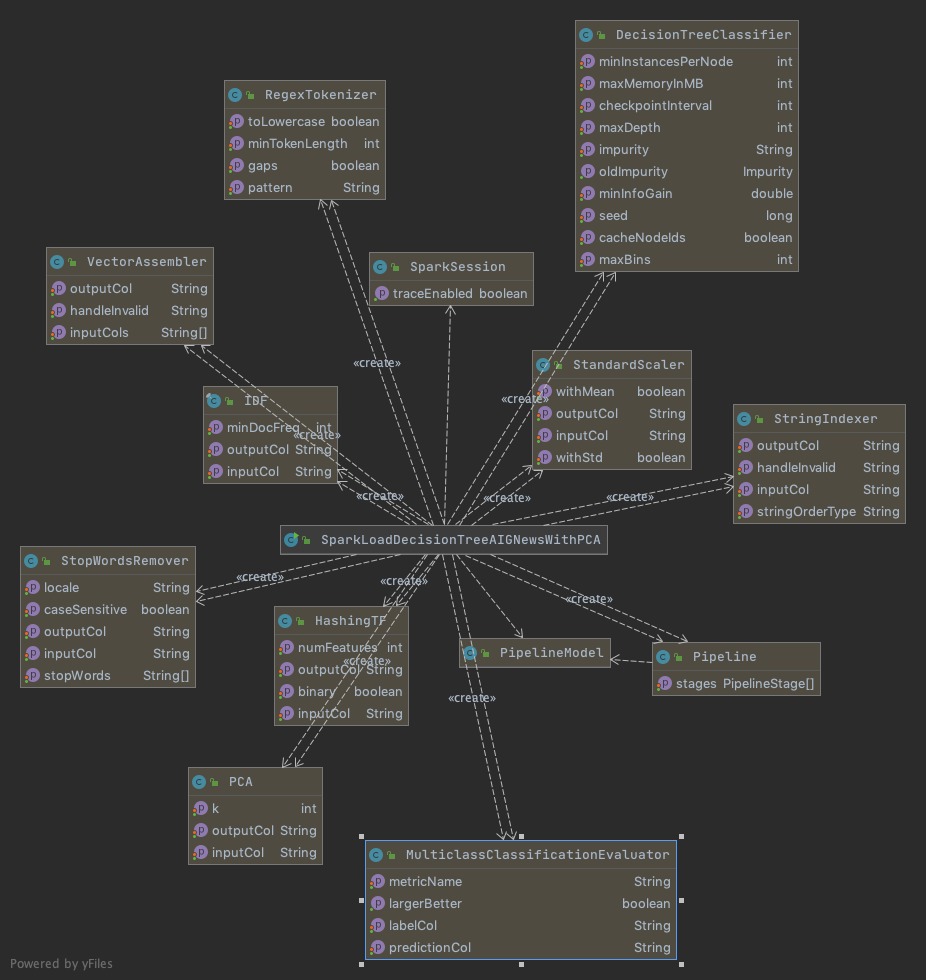
1. Program Description

Linear Regression   
The Linear Regression classification program for the AIG dataset can be seen in Diagram 1, Linear Regression AIG classification program. The main program executed on the spark clusters is called *SparkLoadDecisionTreeAIGNewsWithPCA,* This class contains a number of pre-fitting data processing functions that are executed by the pipleline in a deterministic order. Both original features Title (text) and Description (text) are processed as described in section 2 before fitting the linear regression module



*Diagram 1 Linear Regression AIG classification program.*

Decision Tree



1. Installing and Running the Programs
2. Compare and Discuss Results
3. Compare and Discuss Results with PCA

# Appendix

[1] Aman Anand: AG’s News Topic Classification Set. Retrieved 7 June 2021: https://www.kaggle.com/amananandrai/ag-news-classification-dataset