Big Data Analytics without SQL
Pre-processing Text Data to Build a Predictive Model with a Deep Neural Network

1 Introduction

You will create a program to analyze text data in the Python language working on top of Hadoop. Your program will combine stand-alone Python with PySpark. The goal is to build a predictive (supervised) model, assuming we have a target variable already created.

You can use the below dataset:

https://www.kaggle.com/datasets/cynthiarempel/amazon-us-customer-reviews-dataset?select=amazon_reviews_us_Grocery_v1_00.tsv

2 Program input and output specification

The input is a large text file stored on HDFS. This file should be pre-processed in PySpark to build a machine learning data set, compatible with a deep neural network.

storage: Plain text file, which must be pre-processed to get keywords.

processing: raw file with PySpark. Data set with Python math libraries including NumPy, SciPy and TensorFlow.

Syntax for program call

Since this project is open ended you have freedom to show how to call your program. It is acceptable to take parameters from the command line or set parameters in the Python code, as long as it is easy. You can optionally build a GUI so that your program can work on a browser.

Expected output

1. After pre-procesing and cleaning the input file indicate the input data set for a predictive machine learning model. The dimensions (numeric features) should be keyword frequencies plus any other available attributes in the raw file.

2. a predictive model for the product rating into 2 classes. The goal will be to predict which keywords tend to give a rating of 1 (bad) or not. That is, we want to predict bad products.

Verification: partition the input data set into train/test subsets (no need to do 10 fold cross-validation). You should verify your model by comparing your prediction with the actual class (given by the rating). You should also compare the DNN accuracy with some other classifier like NB nr SVM.

3 Requirements

- language: Python. Other languages possible, but discouraged (e.g. Java, Scala, C+..contact instructor).
- Big Data platform: Hadoop HDFS and Apache Spark.
- Class target variable: bad product Y/N. You must create this variable, which is not directly available in the input file. Logic explained in detail by TAs.
- Competing system: none, but you can attempt to run some existing Py library or prototype to compute ML models on documents.
- To get started all you have to is copy the file to your HDFS folder.
- Programming: raw file must be pre-processed with PySpark (not Python alone). ML data set can be processed exclusively in Python in your personal computer, but you can also use MLlib in our server.
- The program should not halt when encountering errors.
- Deliverables: source code and README file.