COSC 6339: Big Data Analytics

Homework 3: Pre-processing Text Data to Build a Predictive Model with a Deep Neural Network

Group 10

Team Members	PSID
Okoronkwo, Kelechi P.	2056733
Edara, Pravineeth	2158584
Liu, Xiaoqing	1991841
El Aassal, Ayman	1585185

Introduction

We create a program to analyze text data in the Python language working on top of Hadoop in this project. The program combines stand-alone Python with PySpark to build a predictive (supervised) model, assuming we have a target variable already created.

We use the Amazon Baby reviews dataset to generate two models. The first model is created with just the reviews and the second model with the review and other features from the dataset.

Tools Used

- **Dataset:** amazon_reviews_baby_100k.tsv
- **Libraries:** PySpark, Tensorflow, Elephas, Nltk, Numpy, Pandas, eel

Approach

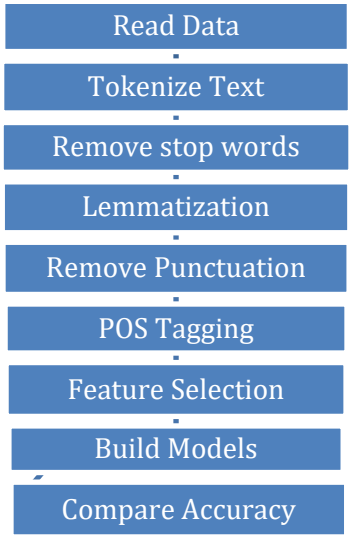
Below are the steps we took for this project:

- Tokenize the text
- Remove non-English words
- Remove stop words
- Lemmatization
- Remove punctuations
- POS Tagging
- Feature Selection
- Build a Predictive Model with Deep Neural Network
- Compare accuracy with other models using F1-Score

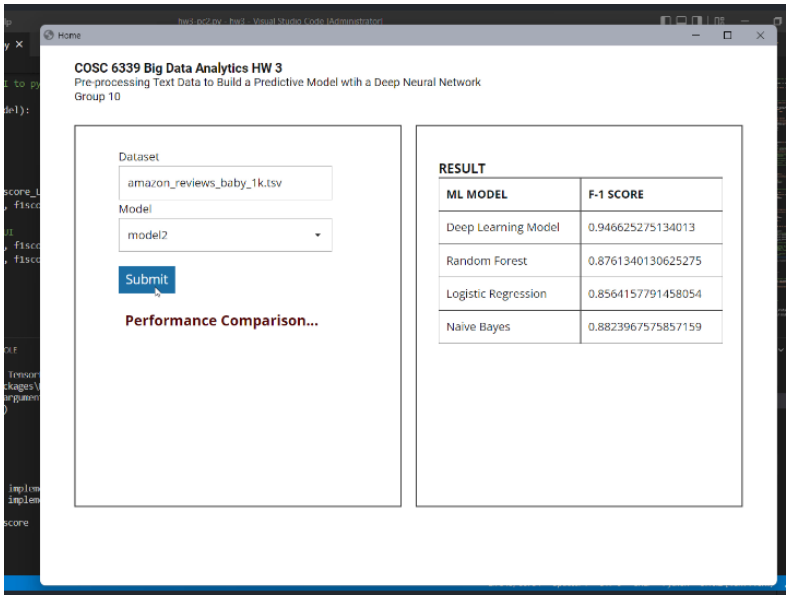
Responsibility Matrix

Team Members	Task
Okoronkwo, Kelechi P.	GUI design, text processing, code review, and modification, Presentation, Report
Edara, Pravineeth	Code review, presentation, report
Liu, Xiaoqing	Deep Neural Network and Comparison with other libraries
El Aassal, Ayman	Text Processing, code review, and modification, Presentation

Process Flow Diagram



Program GUI



Results and deliverables

Dataset: amazon_reviews_baby_100k.tsv

Model 1

Machine Learning Model	F1 Score
Neural Networks	0.93
Random Forest	0.88
Linear Regression	0.90
Naïve Bayes	0.91

Model 2

Machine Learning Model	F1 Score
Neural Networks	0.95
Random Forest	0.89
Linear Regression	0.89
Naïve Bayes	0.91

Conclusion and Recommendation

The dataset used is unbalanced so F1 Score was used to compare the models.
Univariate feature selection was used to reduce the dimension to 100 which was used to build the model and gave over 94% accuracy.
Looking at the results we see that the second model has a better accuracy because there are more columns from the dataset used to generate the model.