Ben Noriega

CS777

2.22.23

Term Project

Project Repo: https://github.com/bennordi/cs-777-term-project-bennordi

Dataset

The dataset is a CSV file found on Kaggle under [Spam Text Message Classification](https://www.kaggle.com/datasets/team-ai/spam-text-message-classification). It contains 5572 rows of two columned data: Category and Message. Category can simply be described as the ‘class’ or ‘label’ of this dataset as it has a simple binary value: ham or spam. Spam represents the text in the ‘message’ column is from a text deemed to be spam. The opposite ‘ham’ represents the text was deemed to be, not spam.

Research Question

This is a simple classification dataset, which we will leverage PySpark and various libraries to analyze text data to build various machine learning models that will predict `ham or spam` based on the given text data. We will then review each model and its results. There are 5572 rows of data to analyze, and luckily the data was preformatted so that `ham/spam` are the only values in the category column, and the message column will need to be formatted into features columns for us to build models off of. The main need for preprocessing was removing non-letters from the text, lemmatizing the text, and then vectorizing the data (with MLlib’s pipeline using Tokenizer(), StopWordsRemover(), CountVectorizer(), IDF(). I was going to use ChiSqSelector() but we learned in Assignment 5 that with larger datasets this can be very computationally expensive, and knowing that the goal of this assignment is to be scalable, I opted to not perform this selection.

This data will be split into training and test sets (70/30 split) and will be seeded for consistency of the split. The classification models will be trained on the same exact training data and will be tested with the same test dataset. Then during the results, I can do some anecdotal tests for messages I create to test out our models.

Machine Learning Model(s)

Results

**Local run output can be found on <output.md>**