CSC 510 Assignment 1 Report

Justin Oakley

Contents

I. Introduction

This report is about knowing and preparing given data in order to understand its characteristics. It also features a summary of a paper on the problems and challenges of machine learning in big data classification. The reason that these analyses are important to go through is that it will assist in developing skills that are needed in the field of data science and big data. Such attributes that necessary in this field of study are understandings of the importance of data, systems, machine learning techniques, scalability and complexity as well as big data systems and programming languages that are most commonly associated with said systems. In this assignment, four different data files were used for analyzation, *carpet.csv, hardwood.csv, bgg_db_2017_04.csv, and bgg_db_2018_01.csv*, via the Python 3 language (using libraries such as Pandas and Matplotlib) and its respective version of the Apache Spark technology. All of these were a gateway for an in-depth and readable analysis on fundamental processes used in big data science, and a greater knowledge of how technologies such as Apache Spark function.

II. Carpet and Hardwood Datasets

The following datasets used in the first portion of this analysis are called *Carpet Floor* and *Hardwood Floor*, and are computationally derived from the following .jpeg files. Both of these datasets individually contain one thousand twenty-four observations and sixty-four feature columns where each feature observation is a numerical value. Knowing this, all sixty-four features per dataset are processed for statistical information, then utilized for the steps of data visualization. Immediately following these parts, both datasets are labeled based on which file each observation originates from, then is merged and randomized and finally split into two different sets, the training and testing sets.

A. Statistical Information

B. Data Visualization

C. Data Inference

D. Constructing Training and Test Sets

III. Board Game Data From 2018 and 2019

A. Statistical Information

B. Data Visualization

C. Data Inference

IV. Summary of ”Big Data Classification”

V. Results