Aaron Squier
Merlin Carson
Parker Moore

Midterm Report
CS 410: Explorations in Data Science

Objective

Inquire and explore steps in the data scientist pipeline including: data wrangling, data cleaning, and predictive modeling with AI. We will use the Speech Recognition data set, found on Kaggle.com, in order to build an algorithm that recognizes simple speech commands.

Approach

Our approach is broken up into three phases: research, modeling, and evaluation. In the research phase we will gather and clean the data. We must also determine which type of machine learning models are best practice for this data set (regression, clustering, …). Search for Kaggle Kernels relating to our topic to avoid mistakes others have made. We will apply an incremental approach to adding more words to classify beginning with two. Then, implement a handful of machine learning models to obtain an accuracy score. This is the point where we may decide to add additional words to try and classify if the earlier phase was a success, Finally, examine our results to determine the precision/accuracy of our models.

Team Structure

Team D is divided into two sub groups: data visualization and machine learning. We represent the machine learning subgroup and the research is divided as follows.

- Parker Moore: Apply machine learning algorithms to dataset to recognize simple speech commands
- Aaron Squier: Apply Deep Learning & Convolutional Networks to dataset to recognize simple speech commands

- Merlin Carson: Implement Data Cleaning & Pre-processing techniques to dataset to remove corrupt or inaccurate records from table in order to improve data quality and overall productivity

Project Milestones Update

1. Research and understand given topics. Present Research Presentation
    a. Each member of our team researched and presented their designated topic.
2. Gather and clean data
    a. Data has been gathered and cleaned. Currently waiting for Aaron and Parker to model the data
3. Apply machine learning model to dataset and record results
    a. Research has been done into the library that we are going to use to develop our machine learning models. We have settled on the Keras library for both modelling techniques
    b. We have agreed to start off with a small proof of concept and train our models to classify just the words “yes” and “no”. When it has been shown this can be done with a high degree of accuracy then we will add additional words up to the full set if successful
4. Apply deep learning models to dataset and record results
    a. See 3a and 3b
5. Investigate results
6. Present final project

Date / Time of Scheduled midpoint meeting with Prof.

- Meeting has been scheduled for Thursday, August 1st, at 11am in our usual classroom.