# **Assignment 3: Apache Spark**

Emily Strong

CSYE 7245 Spring 2018

## Abstract

## Introduction

## Methodology

### Data Set

In Part B I used the Labeled Faces in the Wild (LFW) [1] data set, the data set from my term project. The LFW contains over 13,000 labeled pictures of famous people, with 1680 of the subjects having multiple photos in the set. In Parts C & D I combined the LFW with the CIFAR-10 [2], a labeled object data set with 60,000 images of 10 types of objects. Each image is 32px x 32px and is represented as a numpy array of its RBG values. I chose this data set because 6 of the types are animals. Using this data set should ensure that the CNN is specific to detecting human faces rather than overgeneralizing or detecting eyes.

### Part A: Set Up Apache Spark

### Part B: Structured Streaming

### Part C & D Alternate: Deep Learning with Apache Spark and Tensor Flow

## Results

## References

[1] Huang, Gary B., Manu Ramesh, Tamara Berg, and Erik Learned-Miller. "Labeled faces in the wild: A database for studying face recognition in unconstrained environments." *Vol. 1, no. 2. Technical Report 07-49*, University of Massachusetts, Amherst, 2007.

[2] Learning Multiple Layers of Features from Tiny Images, Alex Krizhevsky, 2009. Retrieved from: https://www.cs.toronto.edu/~kriz/cifar.html