Susan Kuretski
skure2@illinois.edu
CS410 - Fall 2020
Course Project Progress Report

Overview

This course project uses a dataset from Kaggle [https://www.kaggle.com/sayangoswami/reddit-memes-dataset] to perform optical character recognition on Reddit memes and do cluster analysis. The original proposal stated doing sentiment analysis, but as the reviewer suggested I have switched to cluster analysis via K-means since upvotes and downvotes are affected by multiple factors. I hope cluster analysis on these memes will uncover unlikely themes.

Tasks Completed

Tasks completed to this day are:

- Download and clean data
- Uploaded images to Google Cloud Platform storage
- Performed optical character recognition with Google Cloud Vision and Translate to determine which language
- Stored results in GCP

Pending Tasks

- Evaluation of OCR data
- Processing data to do cluster analysis - make each meme into a term vector
- Perform cluster analysis
    - Plot out clusters
- Evaluation
    - Determining a “good” number of clusters via elbow method
    - Silhouette analysis to see how “good” the clusters are

Challenges

My biggest challenge is my unfamiliarity with K-means and unsupervised learning. I chose K-means since I don’t have a ground truth or labelled dataset. I pondered halving the dataset of ~3,000 images to label it, but I don’t see it as the best use of my time for this project. While K-means may seem straightforward for those experienced with it, I would imagine it will take some time for me to fine tune the number of clusters, perform evaluation, and plot the results out. As it stands now, I have spent about four hours doing the completed tasks.

Another challenge is working with OCR text data which has originated from memes, which often is sarcastic, misspelled, and infers cultural or societal knowledge. There are multiple layers of ambiguity and room for error.