NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

School of Computer Science and Engineering
College of Engineering

Text Classification and Information Extraction from Reddit

Student: David Ryosuke Harui

Supervisor: Asst Prof Ke Yiping, Kelly

Text Classification

The project aimed to develop a mechanism that would classify Reddit submissions into their most relevant subreddit. It delves into data pre-processing, NLP technique, and machine learning algorithms to determine which algorithm is optimal in terms of accuracy and efficiency.

Information Extraction (Part I)

Reddit dataset does not have a sentiment class label for supervised learning models to train from. The project aimed to utilize K-means for sentiment analysis and compare the accuracy results with pre-built libraries like VADER and TextBlob.

Information Extraction (Part II)

The project aimed to perform sentiment analysis on the politics subreddit to capture the users' collective thinking (political bias) using the methodology from Part I, and uncover which political party or politician the community favors.