# **Capstone 2 Project - Proposal Ideas**

1. IMDB Movie Reviews

<http://ai.stanford.edu/~amaas/data/sentiment/>

This is a dataset for binary sentiment classification. It contains movie reviews along with their associated binary sentiment polarity labels. The core dataset contains 50,000 reviews split evenly into 25k train and 25k test sets. The overall distribution of labels is balanced (25k pos and 25k neg). It also includes an additional 50,000 unlabeled

documents for unsupervised learning.

1. Automated Essay Scoring

<https://www.kaggle.com/c/asap-aes/overview>

This is a Kaggle competition to develop an automated scoring algorithm for student-written essays. It is sponsored by The William and Flora Hewlett Foundation (Hewlett).

They provide hand scored essays, so that we can build, train and test model can deliver scores closely to those of human graders.

1. 20 Newsgroups

<http://qwone.com/~jason/20Newsgroups/>

The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned across 20 different newsgroups. It is a popular data set for experiments in text applications of machine learning techniques, such as text classification and text clustering.