

# CS 6140: Project Proposal

Brian Kimmig, Jesus Zarate

We propose to collect data on a large number of movies in order to predict the genre or combination of genres of a movie. We aim to gather data of movies via the [OMDB API](#), where simple get requests can be made to get the synopses of each movie. Though we are unsure at the moment how many movies titles we need/want we plan to start by getting the synopses from titles in this dataset on Kaggle ([Kaggle dataset](#)).

Once we have all of the movie synopses we will create our features from the text. This will be done by finding words or phrases that occur frequently, but excluding very common words (articles like a, an, the and others). We would like to try using single words to create features as well as k-grams. This will obviously still have a very large number of features so we would like to employ min-hashing to further reduce the number of features we create.

After we have features for each film we plan to classify the genre or genres of the film. Since our feature set is most likely going to be large we would also like to experiment and see the impact on classification of feature reduction techniques like PCA and matrix sketching.