*My final project group will only consist of myself.*

*I will handle both the coding and the writing of the project.*

### Project 1 – Sentimental Analysis to film reviews on Letterboxd

Most likely to choose

**Questions:**

I am a film fanatic. As a result, I would love to utilize this opportunity to conduct a sentimental analysis about online film reviews to one of my favorite yet controversial director.

* What are the sentimental toward films by that specific director? Are there any splits in opinions? What can we learn from it?
* Is it possible to characterize and generalize the director’s cinematic language and styles by analyzing viewers’ comments?
* Can we build an AI model trained with parts of director’s film reviews to predict review polarity of his/her other films?

**Dataset:**

The website that I am going to getting the film and reviews is called Letterboxd – a social platform focused on sharing opinions about, and love of, film. The platform allows users to rate and record their opinions about films which makes it ideal to conduct my desired sentimental analysis.

**Methodology:**

To answer my first two questions proposed, I will primarily work with the sentimental analysis toolkit (VADER) that we learned in class. I will also use topic modeling in some way that I can extract topics to generalize a director’s cinematic style.

If possible, I wish to use SKlearn to create an AI model to predict review polarity’s other films.

### Project 2 – Classifying Toxic Comment on twitter

**Questions:**

Being anonymous over the internet can sometimes make people say nasty things that they normally would not in real life. I would love to utilize this chance to conduct a text classification to the toxic comments on twitter.

* How can we define toxic comments?
* Should the classification be binary?

**Dataset:**

For this project, I will be using *Sentiment-140* from [Kaggle](https://www.kaggle.com/kazanova/sentiment140). It contains a labels data of 1.6 Million Tweets and I find it a good amount of data to train our model.

**Methodology:**

I am going to divide the dataset into two even proportion for both training and testing. I will use SKlearn to conduct tokenization, and at the same time leveraging the Word2Vec that we learned in the class to perform word embedding, capturing context of a word in a tweet.

Finally, I would love to use deep learning models like Sequence Model, SVD, or Naïve Bayes, to build an AI model that eventually help us classify negative comments on Twitter.

### Project 3 – ANAlysis on MOVIE SCRIPTs

**Questions:**

Instead of conducting an analysis on movie reviews, I am also interested in what a movie scripts – lines speaking by each character – can entail about the general theme.

* How different lines manifests certain character’s personality?
* How does the analysis tie back to theme of the movie?
* Can we predict the genre of the film based on our finding about the script?

**Data:**

I will interest in using the Star Wars Movie Scripts on [Kaggle](https://www.kaggle.com/xvivancos/star-wars-movie-scripts). However, if I can find the script of some of my favorite films, I might as well use those as my primary data to work with.

**Methodology:**

What I can think of right now is to first conduct a sentimental analysis using NLP toolkits like word counts, n-grams, lexicons. I can also find the most relevant words for each character using TF-IDF and PMI/PPMI. Topic-modeling might be helpful in abstracting themes from scenes which consist of multiple lines.