**Our Project**

**Git repository link:** <https://github.com/alexjxela/pstat234-final-project.git>

First, we want to apply natural language processing tools to analyze the movie plot summaries. From this we want to get...

Second, we want to combine with other movie attributes to test for the significance of their effect on movie outcomes (e.g., box office).

Third, ...

Part 1-Data

* [**Data A**] Wikipedia Summary Plots (in git repo): <https://www.kaggle.com/jrobischon/wikipedia-movie-plots>
* I suggest that we use the **movies produced in the US** (Origin/Ethnicity=”American”), so that there will be a lower chance of dealing with movie plots containing other languages than English.
* I also suggest that we focus on the movies that were released between 1987 and 2017, whose box office is recorded more completely than the earlier ones. This will be about 6,700 movies in the sample.
* [**Data B**] Box office and other attributes
* We will conduct the NLP analysis on **Data A.** Then we will merge the result with **Data B** that contains the box office and other attributes for our regression analysis.

Part 2-Exploratory Analysis

* Summary statistics for box office by genre, release year, MPAA rating, etc.
* Within each cluster, difference in attributes

Part 3-NLP tools

1. TF-IDF (Preliminary code in git repo)

* Description of the Method
  + <https://medium.com/analytics-vidhya/how-good-or-bad-is-traditional-tf-idf-text-mining-technique-304aec920009>
  + <https://janav.wordpress.com/2013/10/27/tf-idf-and-cosine-similarity/>
* Reference Code
  + <https://towardsdatascience.com/using-nlp-to-find-similar-movies-based-on-plot-summaries-b1481a2ba49b>
  + <https://towardsdatascience.com/how-to-cluster-similar-sentences-using-tf-idf-and-graph-partitioning-in-python-5fb70627a4e0>
  + Datacamp exercise: <https://www.datacamp.com/projects/648>

1. LDA

* Description of the Method
  + <https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation>
  + <https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-latent-dirichlet-allocation-437c81220158>
* Reference Code
  + <https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24>

Part 5--Regression Analysis

Side notes:

* Final Project Guidelines by the professor:
  + <https://github.com/UCSB-PSTAT-234/Spring2021/blob/public/group-project.md>
* NLP, AI, and Machine Learning: What’s The Difference?
  + <https://monkeylearn.com/blog/nlp-ai/#:~:text=Natural%20Language%20Processing%20(NLP)%20is%20the%20part%20of%20AI%20that,machines%20interact%20with%20human%20language.&text=Combined%20with%20machine%20learni%20ng%20algorithms,and%20get%20better%20through%20experience>
* NLTK (Natural Language Toolkit)
  + <https://www.nltk.org/>
* More generally on Topic Analysis
  + <https://monkeylearn.com/topic-analysis/>
* *Recommendations Without User Preferences: A Natural Language Processing Approach*. <https://dl.acm.org/doi/10.1145/604045.604087>
  + Note: It is not a well cited paper, but we can do some similar comparative analysis for TF-IDF and LDA with a small group of movies.

In the next step:

1, finding more attributes for the movies, cast and crew members, e.g. complete the director’s race information, find big star names, etc.

* IMDb: <https://www.imdb.com/list/ls058011111/?sort=list_order,asc&mode=detail&page=1>
* The Numbers: <https://www.the-numbers.com/bankability>

2, LDA

---for Friday’s meeting

3, Regression and results by Friday/Saturday

4, Start writing report next Monday, finish by June 11

5, Make slides, record presentation (schedule a time), finish by June 11