Final Project Outline

**Project Scope:** The goal of this project is to create a model to measure the box office success of movies that have been revealed but not yet released to the public. I hope to design and deploy an analytic solution to consume unstructured data such as comments from a social media post to create a measurement of box-office success based on previous data.

Task List:

***Task A (Data Gathering):***

*Part 1 -*

To begin this project, create or use an online tool to scrape the web for data regarding movie announcement posts. Some sources could include Instagram and Twitter, but I recommend that you begin with Instagram and design a scraping method that could be extended to alternative platforms if necessary. You may also find scraping multiple profiles on a social media platform to be difficult, so first design a method that consumes data from one profile and could be extended to multiple profiles if necessary. Suggestions for good profiles to scrape are: IMDB, Rotten Tomatoes, HBO, etc.

Things to keep in mind:

1. The data should include comments regarding a film that has not yet been released to the public. It may be that comments could be posted prior to the movie's release which I would need to filter. One idea would be to use the metadata gathered from each post to filter the ones posted prior to the box office release date.

2. Multiple profiles would be good, however this could prove to be very difficult. Begin with one profile and one social media platform to scrape for data

3. Store all data in either a CSV file or JSON file. You may find it easier to use JSON but use whatever seems best for the shape of the data

Part 2 (Optional? LMK what you think) -

Later on in the project, I will calculate the sentiment of a movie from the average sentiment coefficient calculated for each comment relative to that movie. To have something to compare it to, it would be interesting to scrape the post-launch reviews of the movies selected in Part 1 and calculate the sentiment to then be compared to the pre-launch sentiment analysis. For example, if I observe that the sentiment found from the data in Part 1 is positive, and the sentiment of the post-launch reviews are positive, then I can consider our *model* accurate at predicting the success of a movie before launch. You can choose to scrape either IMDB’s website under the “[user reviews](https://www.imdb.com/title/tt7975244/reviews/?ref_=tt_ql_urv)” section; or the [reviews provided on Google](https://www.google.com/search?q=jumanji+the+next+level&sxsrf=AOaemvIKp51FGn64JvGPfoRnWGij6Kc0xA%3A1638063166732&ei=PtyiYYyILOykqtsPuKiI4AM&ved=0ahUKEwiMn7Dp9Ln0AhVskmoFHTgUAjwQ4dUDCA4&uact=5&oq=jumanji+the+next+level&gs_lcp=Cgdnd3Mtd2l6EAMyBAguEEMyBAgAEEMyBAgAEEMyBQgAEIAEMgYIABAHEB4yBQgAEIAEMgUIABCABDIFCAAQgAQyBQgAEIAEMgUIABCABEoECEEYAFAAWABgvAFoAHACeACAAT6IAT6SAQExmAEAoAEBwAEB&sclient=gws-wiz#wptab=s:H4sIAAAAAAAAAONgVuLVT9c3NEwzMzQ1SaosfsRowS3w8sc9YSn9SWtOXmPU5OIKzsgvd80rySypFJLmYoOyBKX4uVB18uxi4nLLzMkNSi3LTC1fxGqcnJmXmpuokFWakp4KJHMT87IyrRTKU3OS83NTFUryFUoyQOJ56TmpCkVgTcUAE3UXZI8AAAA) when searching a title.

***Task B (Data Cleaning):***

Now that the data has been gathered, go through and clean the comments for stop words, punctuation, special characters, etc. Make sure that all of the data is formatted synonymously such that everything is lowercase, unnecessary whitespace is deleted, @’s are deleted, and non-english languages are filtered out. This is an important step as the analytics will depend on the quality of the data being filtered in.

Things to keep in mind:

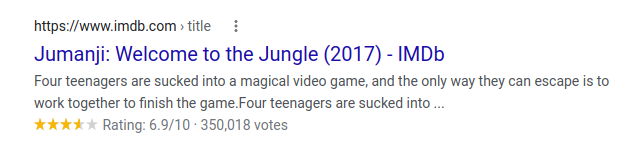
1. Store the clean comments in another file similar to the file created from task A

2. Look into techniques I have used in the past such as Lemmatization, Tokenization, Stemming, etc.

There is code for text cleaning in this [tutorial](https://towardsdatascience.com/nlp-preprocessing-and-latent-dirichlet-allocation-lda-topic-modeling-with-gensim-713d516c6c7d)

***Task C (Sentiment Analysis):***

To get a better understanding of what makes a *successful movie* prior to launch, design an algorithm that will calculate the average sentiment per movie post using the scraped data. Create a visualization with your results and display it in the notebook (please also save this as an image and store it in your working directory). To differentiate between movies which I will consider successful and non-successful, add the rating of each movie to your visualization and display it along with it’s sentiment calculated. To be consistent, find the rating of the movie via IMDB’s website; It should look something like this:

You do not have to do this for every movie in our sample, however it could be interesting to label a few so I can see if there is a correlation between sentiment before launch and the overall rating of the movie post launch.

If Task A, Part 2 was completed, compute the sentiment analysis of the reviews for each movie scraped. Create a visualization to compare the sentiment before the movie's release calculated by the comments, and the sentiment after the release generated by the reviews. Save this image in the notebook and in your local directory.

***Task D (Topic Modeling):***

To support our sentiment analysis above, I can use Topic Modeling to help describe a corpus generated by combining all of the comments relative to a movie post. The movies will serve as the *documents* which I will analyze and the topics generated for each document (movie) could reveal insights in what topics correlate to a successful movie. Here is a good link which describes how to do this [using Gensim.](https://towardsdatascience.com/nlp-preprocessing-and-latent-dirichlet-allocation-lda-topic-modeling-with-gensim-713d516c6c7d) This is hit or miss and might not provide good insights, but could be a good addition to the sentiment analysis performed above. After you have extracted topics for each movie, look into this article on [Towards Data Science that talks about text clustering](https://towardsdatascience.com/a-friendly-introduction-to-text-clustering-fa996bcefd04) with LDA. This may be a good way to visualize the data gathered from this step, but if it turns out to be more of a headache than anything, find an alternative method to visualize the insights gained from this task.

***Task E (Final Thoughts & Insights):***

Now that I have gathered quite a bit of information regarding our project scope, create a document which contains the methods I used in our analysis, any images / graphs I generated and their meaning, and a brief overview of what I could do to continue our exploration of movie success predictions. Feel free to put this either on a PDF document or a slide deck, whatever you feel fits the content the best.