Project 4 Proposal: Unsupervised Learning (NLP)
Disney Remakes: Original Animated vs. Live Action

Problem
Disney has released 14 original animated films into live action remakes so far. I want to use text from IMDB reviews to measure audience sentiment (positive vs. negative) about the most recent remakes that were actually retellings of the same Disney story: Aladdin, Dumbo, The Lion King, and Beauty and the Beast.

Data
- Scrape IMDB pages for reviews of the above 4 films
- Possibly pair with box office information for these films (as well as costs of making these films), if time permits & available

Known Unknowns/Barriers
- Scraping
- Labeling positive vs. negative reviews

Potential Resources
- NumPy, Pandas
- SK Learn
- NLTK, Gensim, spaCy
- Matplotlib, Seaborn
- Possibly other data sources or scraping resources, if time permits/available

MVP
- Dimensionality reduction, and maybe topic modeling
- K-means clustering

Kristen Tokunaga