Jonathon Bowyer
Metis Project 4 Proposal – NLP and Unsupervised Learning

For this project, I'll be looking at a dataset of articles from the New York Times (NYT) as well as a dataset of comments on these articles. The goal is to create a model that will be highly predictive for determining if an article about a given topic will have a high or low number of comments. This will be highly actionable as the news organization will likely want to publish articles that generate a lot of discussion.

From there, I'll utilize unsupervised learning to determine if there are topics that these high-comment and low-comment articles center around. This can help foster an understanding of what sections generate most buzz for the news organization, and which sections aren't as riling.

That will be my MVP. From there I'd like to do some sentiment analysis to see if there are certain articles/areas that incite positive or negative emotions.

Domain:

The domain for this topic is quite straightforward: news articles from the NYT and comments on these articles.

Variables:

There are two datasets for this project: one that contains a row for each article (including headline, keywords, and so forth), and one that contains a row for each comment (including what the comment said, what article it was on, and so forth). Each of these datasets has quite a few different variables - here I'll outline the ones I might be using in my project:

Variable Name	Variable Type	Variable Description
Article ID	String	Unique ID for the article
Article Headline	String	Headline for the article
Article Keywords	String	Keywords in the article
Comment Body	String	What the comment says

Known Unknowns:

The first is whether there is an actual relationship between what the article is about and the number of comments that an article will receive. It might be that what actually causes more comments is the author (i.e. people follow a certain writer), or it might be that people are more likely to comment on articles on the weekend, and so forth.

Another known unknown is whether topics will arise using unsupervised learning. Though I can try to pull out topics, the topics that the algorithms pull out may be too scattered to be helpful.