Project 4 - NLP and Unsupervised Learning

Gregory Lull

At this point in the project lifecycle I don't quite understand what unsupervised learning looks like, but here are a couple of ideas floating in my mind:

Carbon Dating

I have a hypothesis that different grammatical structures and vocabulary is used more commonly in different time periods. Assuming I have books / text documents ranging a time span, say from present to a couple centuries ago, I'm wondering if it could classify which period of time a piece of text comes from.

Problems:

1. My issue with this is that I feel like a target variable such as "year printed" is an important piece of information. A consequence of not having that information might be that the model classifies based on genre or complexity or length or any other number of factors.

Data:

- Plan on finding free/open source books such as Project Gutenberg.

Character Profile Reading Recommendation

When I read fiction what really throws me off is if there's "bad" or unexpected dialogue. If I'm reading a more adult fiction like Game of Thrones, but the dialogue sounds like young-teen, then I do not want to continue to read the book. I'm curious if I could extract the dialogue from the books and create some kind of character profile, and then recommend books that have similar character profiles. For example if you really like a late-teen science-fiction novel, then the recommendation system could produce a list of books that have characters you may enjoy more in the late-teen science-fiction range.

Problems:

1. Not sure how to extract dialogue, or if that's too minor of a detail, maybe just start with similarity first?
2. What is a character profile? Is it sentiment analysis, advanced vocabulary / grammar?

Data:

- Free books like Project Gutenberg
- Maybe need to use something like Goodbooks to measure my "recommendation" system

News Recommendation System

This sounds like a technically challenging task, I would like to create a model that recommends news articles that are very similar or very not similar, and maybe this can be tuned (TF-IDF?). As an MVP it can just use old news articles and spit out something similar, but expanding the project I would like it to consume (scrape) and spit out currently relevant news articles, and deploy a webapp for this.

Problems:

1. My concern is that this doesn't have enough data analysis elements to it, and is more of a webapp project.

Data:

- For a MVP maybe just recent data https://www.kaggle.com/snapcrack/all-the-news