Emily Hillenbrand

Metis Data Science
Final Project
10 December 2019

Cult Films Recommender

Natural Language Processing, Topic Modeling, Recommendation Systems

Overview

I love cult films. I am a third-generation supernerd who was raised on Mel Brooks, Monty Python and heavy irony. The things I like about a movie do not necessarily match up with mainstream opinions, and so I tend to have trouble finding new media to enjoy.

I don't like automated recommendation systems. In this age of machine learning, I still prefer to talk to people who have seen the movie and can answer questions like "was it silly enough?" or "do I need to worry about jumpscares?" and unless I crawl through IMDb reviews and try to find that information myself (also not reliable because people tend to look for negative things to say or are less sensitive than me or etc.), I have to go in blind or refrain from trying new things.

So I created my own recommendation system for cult films. I generated features by scraping reviews from IMDb and doing NLP on them until topics fell out. Then I created an interactive tool for modifying the topic distribution so I could indicate the features I want for the movies I want to see.

Pipeline

1. **Scrape Data:** I scraped the list of cult films from Wikipedia and user reviews for each film from IMDb. I had originally planned to also use a dataset of film plots from Kaggle, but I ultimately found that (1) the dataset didn't match well with my scraped data and (2) all I needed to build a highly functional recommender was the user reviews.

2. **Topic Modeling:** I removed stop words, numbers, and punctuation from the reviews, and then I iteratively removed a long list of names until my topics made sense. I used Tf-Idf to vectorize, and reduced dimensions with NMF so that the topic distribution would be interpretable by humans. (See Appendix for topics created in this process)

3. **Recommend:** I found most similar films using cosine similarity, and then I created a Jupyter interactive widget that allows the user to modify the topic distribution and get the most similar movies relative to a distribution that doesn't necessarily exist.

Other Applications

This technique for creating features from user reviews can be applied to pretty much anything that has reviews, such as comic books, Amazon products, or restaurants.

Big Takeaways

1. New Domain Knowledge!

   I learned a TON about cult films in ways I did not expect. Even when I was coming up with my proposal, I realized what makes something a cult film is what people say about it (more than the film's rating, plot, director, etc.) so I chose to build my recommender from user reviews. It worked even better than I expected!

   I learned a lot more about topic modeling, e.g. a defining feature of zombie movies is that there is almost always gore, so a movie that scores highly in the "Living Dead" topic won't necessarily score very highly in "Gore/Scary" because it wouldn't have many gory elements independent of the zombie parts.

   I now have a list of 10+ movies I want to watch that I wouldn't have heard about if I hadn't done this project, so my unintentional goal of finding new movies to check out has been met.

2. If you see a problem, solve it!

This is partly a side effect of learning so many interesting skills at Metis, but it took me until this final project to realize I can now create automated solutions to some modern problems I've identified. For example, I could scrape Amazon reviews of sewing machines to focus on the features I care about (e.g. durable, fast, simple) and place unimportant features (e.g. lightweight, noisy) at a lower priority. Otherwise I would have to crawl through reviews and try to make a decision manually. I'm excited to see what else I can do with NLP.

Appendix: Topics

Topic: ' Art/Classic '

french, art, cinema, war, image, camera, leave, experience, book, beautiful, dream, woman, mind, documentary, reality

Topic: ' Gory/Scary '

horror, gore, slasher, halloween, house, scary, killer, scare, craven, creepy, murr, blood, kill, effect, ghost

Topic: ' Sci Fi '

alien, sci_fi, planet, space, science, ear, ship, effect, mars, fiction, robot, spaceship, special, scientist, budget

Topic: ' Vampires '

vampire, dracula, horror, hammer, blood, count, carmilla, cushing, castle, martin, nosferatu, helse, kronos, sarandon, susan

Topic: ' Silly '

comedy, funny, laugh, joke, humor, hilarious, gag, fun, sketch, python, funniest, guy, stupid, monty_python, humour

Topic: ' Animated '

animation, animate, anime, disney, cartoon, cat, fritz, voice, japanese, ralph, ult, animated, wizard, fantasy, rabbit

Topic: ' Martial Arts '

martial, kung_fu, fight, action, kong, art, hong, dragon, chinese, japanese, master, chow, fighting, tournament, fighter

Topic: ' Music Documentary '

band, music, concert, rock, song, punk, beatle, documentary, musical, album, stone, fan, musician, footage, zappa

Topic: ' Living Dead '

zombie, horror, gore, fulci, shaun, corpse, dawn, lugosi, live, night, budget, cemetery, island, rob, lionel

Topic: ' Giant Monster '

monster, godzilla, creature, frankenstein, giant, dracula, beast, scientist, japanese, island, effect, horror, gamera, dr, attack

Topic: ' Russia '

russian, soviet, polish, russia, poland, communist, comedy, moscow, katya, communism, shurik, union, regime, nikulin, leningr

Topic: ' Sex '

sex, woman, porn, gay, sexual, rape, erotic, girl, bian, male, explicit, female, relationship, nudity, violence

Topic: ' Gold '

leprechaun, lep, gold, cody, vegas, bridget, hood, horror, space, coin, rap, clover, series, mack, las

Topic: ' Action '

action, cop, gang, western, noir, violence, police, kill, killer, crime, car, gangster, performance, gun, guy

Topic: ' Kids '

school, teen, kid, girl, high, teacher, teenager, boy, young, stunt, parent, friend, child, teenage, heaer

Topic: ' Holiday '

christmas, santa, claus, holiday, martian, halloween, arur, child, ralphie, slasher, mars, toy, kid, burton, night