NLP Unsupervised Write Up

Topic Model for Books

Derek Call

Abstract

Judging a book by its cover is only a good idea if you use the back cover. The back cover blurbs are there to draw in readers so that they will feel they need to know more. The genre will be hit on and the main characters are introduced with issue that will be addressed in the book. To this end I have done topic modeling to see which words differentiate most between each topic. Whether it be a murder mystery thriller, a cookbook, or self-improvement literature I aim to show how natural language processing tools can be leveraged to find patterns in otherwise unlabeled data.

Design

Book data was collected along with the back cover blurb. I sorted the words so that the most unique words would carry a higher weight, while more common words would be less impactful. I have topics available to review using the pyLDAvis visualization from the sklearn library in python. This shows the separability between the different topics, as well as their similarity.

Data

This dataset was acquired from Kaggle.com (<https://www.kaggle.com/jdobrow/57000-books-with-metadata-and-blurbs?select=books_with_blurbs.csv>). There are 57,510 books in the dataset with variables for the title, year published, publisher, ISBN, author, and the blurb. This data will lend well to continued work on this project as ISBN numbers are universal so additional data can be found on specific books with ease.

Algorithms

I used python for the data manipulation, text preprocessing, model creation and visualizations.

1. Using sklearn I was able to fit data on a CountVectorizer instance, a TF-IDF instance, and an LDA instance.
2. pyLDAvis was used with sklearn for interactive visualizations.

Models

I did many different tests of different vectorizing methods along with hyperparameter tuning of the number of topics, min\_df and max\_df. The best categorized topic models were when using 12 topics, a min\_df of 65and a max\_df of 0.05. This was done with a TF-IDF vectorizer so that the most frequent words would be weighed less than the rarer words.

Findings

Using and LDA model with the TF-IDF vectorizer I was able to view the top words associated with each of the 12 topics. The topic names that I assigned are:

Topic Names

Mystery/Thriller, Contemporary Fiction, Classic Literature/Fantasy, Self-Improvement, Science/Science Fiction, Cooking, Historical, French, How-To, German, Buzzwords, and Nazi/Horror.

It performed well and made sense on several the books that I manually checked against. Two specific books I included in the analysis are *The Count of Monte Cristo* and *Dune*. Each book had a high probability assigned to one topic, with the rest being equally unlikely. *The Count of Monte Cristo* was assigned 87.8% for the Mystery/Thriller topic, while *Dune* was considered 81.9% a part of the Science/Science Fiction topic.

Tools

* Pandas, Numpy, and SciKitLearn Python packages were used for data cleaning and modeling
* pyLDAvis was used for data visualization

Topics and Top Words

Topic: ' Cooking '

recipes, collection, guide, food, includes, color, books, edition, cooking, photos, information, full, dishes, film, including

Topic: ' German '

und, der, die, den, ein, zu, sie, ist, von, das, mit, sich, er, auf, eine

Topic: ' Nazi/Horror '

vampire, nazi, jessie, german, jewish, holocaust, dracula, poland, hitler, vampires, camp, jews, concentration, killer, anita

Topic: ' Buzzwords '

purchase, released, available, 24, copies, jo, pages, 25, paperback, 2003, million, hardcover, 2002, 2001, 000

Topic: ' Classic Literature/Fantasy '

collection, fiction, published, classic, written, literary, edition, writers, tales, literature, award, volume, characters, poems, writer

Topic: ' Mystery/Thriller '

killer, dead, wife, case, mystery, husband, daughter, crime, dark, town, passion, soon, detective, truth, secrets

Topic: ' Self Improvement '

guide, offers, spiritual, self, practical, shows, god, use, personal, health, questions, learn, provides, today, information

Topic: ' Science/SciFi '

earth, planet, space, alien, ship, battle, captain, mission, crew, evil, land, adventure, race, future, forces

Topic: ' How-To '

color, illustrations, guide, full, easy, step, reference, information, photographs, illustrated, text, includes, instructions, features, projects

Topic: ' Contemporary Fiction '

friends, school, girl, town, house, things, friend, big, baby, something, wants, go, boy, away, always

Topic: ' French '

de, la, un, en, que, et, le, el, les, una, se, il, une, est, del

Topic: ' Historical '

political, president, hollywood, tells, memoir, portrait, account, civil, childhood, personal, age, white, country, nation, unforgettable