Report of MAT 596 Directed Research

Fall 2018

Lu Liu

supervised by Prof. George Legrady

Project Description

Naming a movie requires deliberation because the title of a film plays a significant role in attracting audience at first glance. My master project aims to develop an interactive data visualization project to explore the hidden mechanism forming the titles of movies, especially in the context of culture and aesthetics.

This report provides details of the preparation works for my master project **PUZZLES OF NAMING** (A Visualization based on the Analysis of the Titles of Films) from two aspects: 1) data collection and 2) data analysis methods and part of results. Further studies are required to convert extracted information of the titles from data analysis results into visual representations and to evaluate various technical approaches by user studies.

Data Collection

The dataset of the project will be acquired from the Full MovieLens Dataset (https://www.kaggle.com/rounakbanik/the-movies-dataset), a product of member activity in the MovieLens movie recommendation system. The raw data contains information on 45,000 movies featured in the Full MovieLens Dataset. Features include release titles, dates, genres, and overviews.

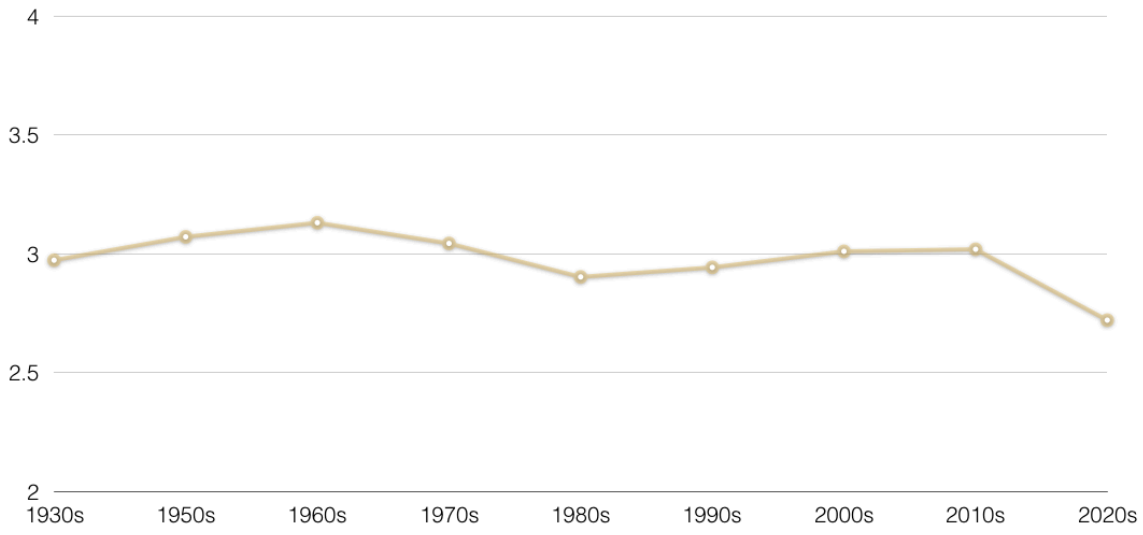
Data analysis Methods & Approaches

Basic Data Analysis for raw data

After pre-processing the raw data, analyze the data by investigating every word in the titles via NLP (Natural Language Processing). NLTK (Natural language Toolkit) is a leading platform for building Python programs to work with human language data. I used these text processing libraries provided by NLTK to finish three kinds of analyses: tokenization to count word frequency, pos (part-of-speech) tagging, and sentiment analysis

1) The trend of the length of titles

This line chart shows the trend of the length of titles over times. The vertical axis is the average number of words in titles every 10 years. It is obvious that the titles of films are getting shorter.

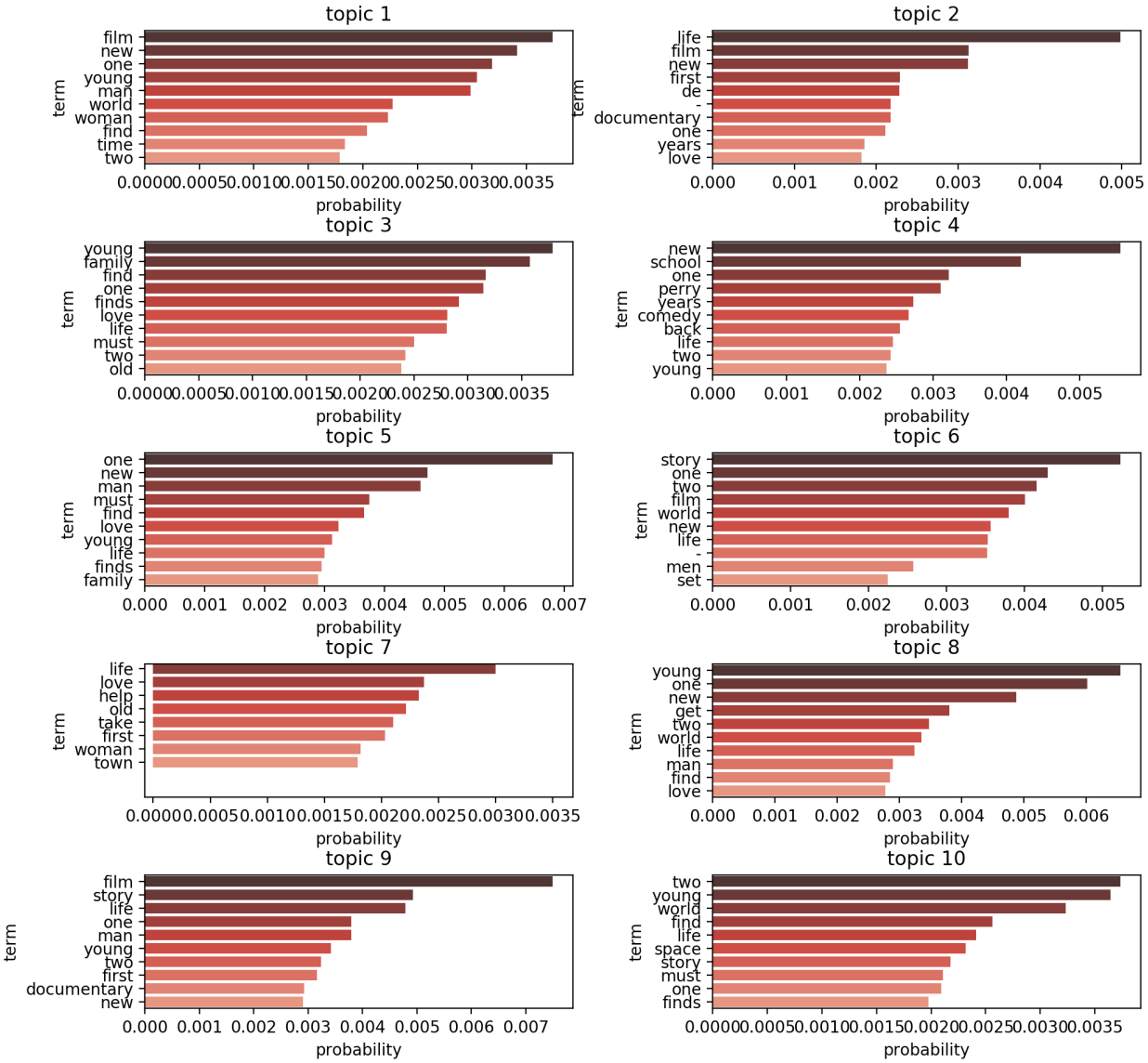


2) The trend of the most frequently appeared word (mp4 file attached in the attachments)

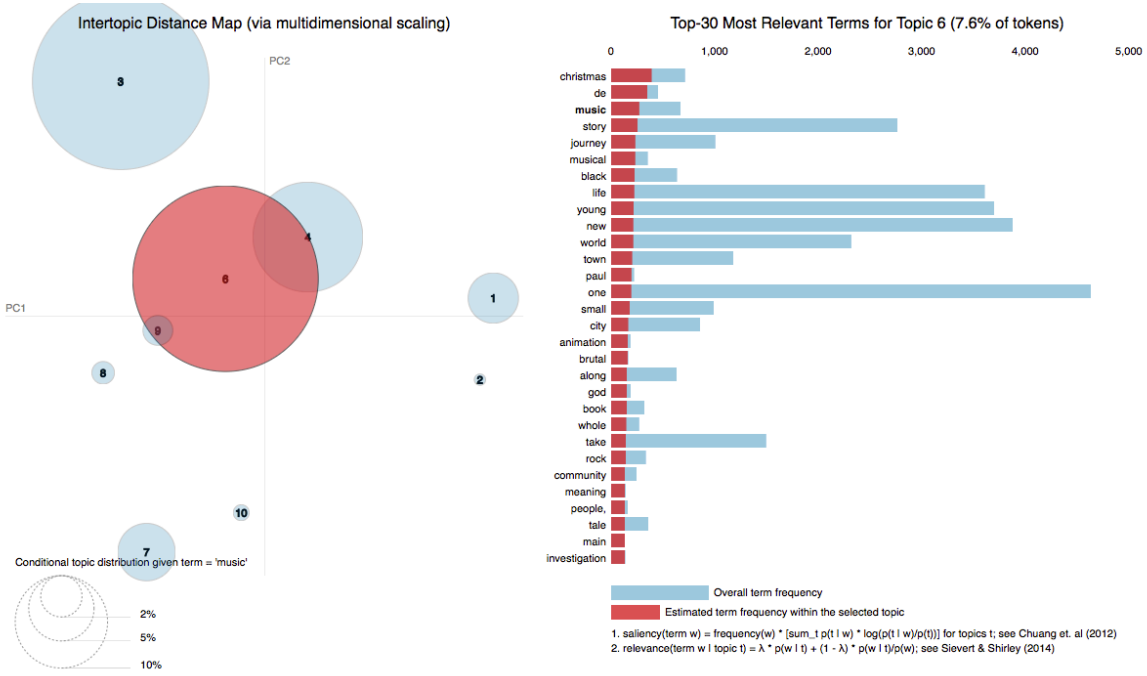
Advanced Statistical Topic Modeling: LDA Modeling (Latent Dirichlet allocation)

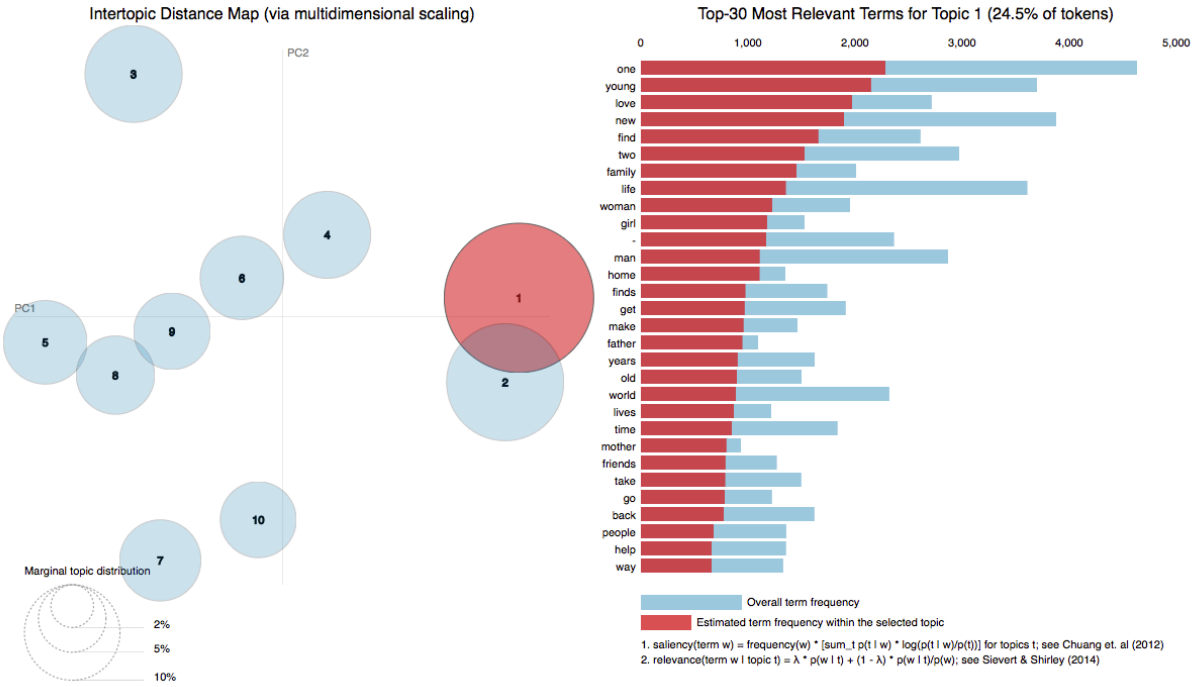
LDA is a popular topic model to investigate the latent topics of each documents in a given corpus.LDA assumes that each document is composed of a number of topics, and each word in the document is attributable to one of those topics [Blei et al. 2003]

Below are the 10 topics I trained from the dataset. Each topic consists of 10 most frequently appeared key words.



pyLDAvis is designed to help users interpret the topics in a topic model that has been fit to a corpus of text data. The package extracts information from a fitted LDA topic model to inform an interactive web-based visualization. Below is the visualization based on the previous result developed with pyLDAvis Library.





References

Idea

Three trends to watch oscars https://theconversation.com/three-trends-to-watch-at-the-oscars-this-weekend-92639

Data Visualization References:

silent visualization https://vimeo.com/39114742
https://vimeo.com/101531290
https://vimeo.com/2320363
Name Voyager http://www.bewitched.com/namevoyager.html
CINEMETRICS. Film data visualization. Frederic Brodbeck http://cinemetrics.fredericbrodbeck.de
Refik Anadol http://refikanadol.com
A data visualization application for exploring IT infrastructure automation https://vimeo.com/80556363
why data visualization important https://www.youtube.com/watch?v=nLy3OQYsXWA
Topic Modeling and Word Vector Visualization of Open Library Fiction Subjects, Hannah Wolfe http://vislab.mat.ucsb.edu/2017/p3/HannahWolfe/index.html

Projects References:

Learning To Split and Rephrase From Wikipedia Edit History http://aclweb.org/anthology/D18-1080

Resources

WordNet https://wordnet.princeton.edu/
Google word2vec https://code.google.com/archive/p/word2vec/
latent Dirichlet allocation (LDA) https://radimrehurek.com/gensim/models/ldamodel.html
Movie Info API http://www.omdbapi.com
Word2Vec http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/
Chinese Poetry Analysis https://zhuanlan.zhihu.com/p/45415824
NLTK Simply Applications and Examples https://www.nltk.org/book/ch01.html
NLTK with sentiment analysis https://www.kaggle.com/ngyptr/python-nltk-sentiment-analysis
NLP Online Course https://web.stanford.edu/class/cs224n/syllabus.html
NLTK Simply Application https://likegeeks.com/nlp-tutorial-using-python-nltk/
The Movie Dataset (Metadata on over 45,000 movies, 26 million ratings from over 270,000 users) https://www.kaggle.com/rounakbanik/the-movies-dataset/home
TMDB https://www.themoviedb.org/documentation/api
Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

Python Library

Golve https://nlp.stanford.edu/projects/glove/
Textblob https://textblob.readthedocs.io/en/dev/install.html#with-conda
NLTK (Natural Language Toolkit) https://www.nltk.org/
Gensim https://radimrehurek.com/gensim/index.html
LDA model https://radimrehurek.com/gensim/models/ldamodel.html
LDA Visualisation Library https://github.com/bmabey/pyLDAvis
LDAvis Example http://www.shichaoji.com/tag/topic-modeling-python-lda-visualization-gensim-pyldavis-nltk/