Course: DS745 -- Visualization and Unstructured Data Analysis

Project III: Text Mining

Topic Modeling to Discover Main Topics in Movie Summaries

Student: *Connie Sosa*

Date: April 28, 2017

Professor: Dr. G. Bansal

Index Page

Table of Contents

[Dataset description 3](#_Toc481263693)

[Problem description / Objective of the project 3](#_Toc481263694)

[Variables: identify relevant variables that are being used in the analysis 3](#_Toc481263695)

[Methodology: steps detailing how the analysis was carried out 5](#_Toc481263696)

[Visually describe the textual data and findings 8](#_Toc481263697)

[Findings along with discussion 11](#_Toc481263698)

[References 13](#_Toc481263699)

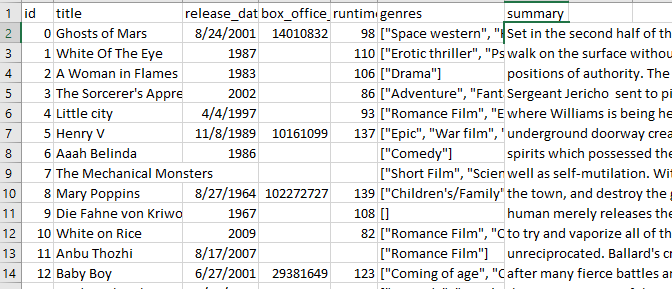
## Dataset description

This text mining project uses the dataset movies10000.csv, it contains summaries of about 10,000 movies scraped from Wikipedia site, along with some metadata about each movie. The original dataset contains over 42,000 movies and is made available by Automated Insights, Inc. (ai)

There are 10,000 movies in this csv file dataset, each with seven variables, the dataset is not in any particular order. Seven variables are

id, title, release date, box office revenue, runtime, genres, summary

Please see Figure 1 below.



**Figure 1:** First thirteen rows of the data in Excel.

## Problem description / Objective of the project

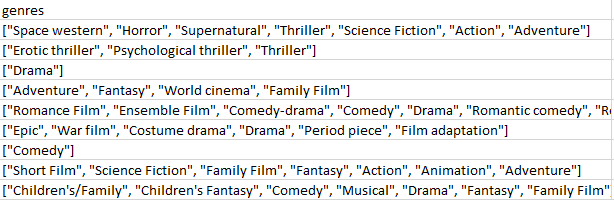
What are the top popular genres? What are the three main topics found in the most popular genre?

We are interested in finding out the popular movie genres in the dataset. After obtaining the popular movie genres, we are to find out the main topics found in the summary field. Essentially, examining what words are the characteristic of the summaries in the movies from the most popular genre.

## Variables: identify relevant variables that are being used in the analysis

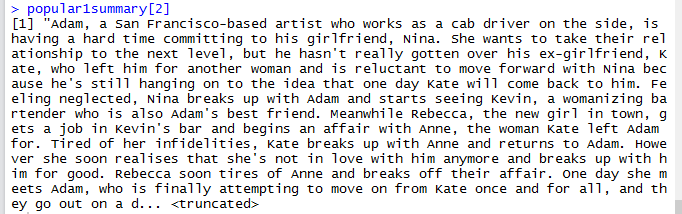
Two relevant variables out of seven total variables are being used in this analysis, they are genres and summary.

1. Genres: The genres that have been categorized for the movie. It is possible for a movie to belong to multiple genres, while some movies have no genres associated with them. Listed in Figure 2A are some examples of the genres for the first nine movies listed in the dataset.



**Figure 2A:** Genres for the first few movies.

1. Summary: The summary field contains the movie plot. Figure 2B shows the plot text in the summary field in R for the film ‘*Little City’*.

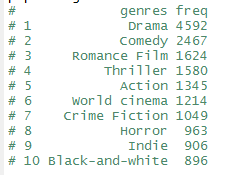


**Figure 2B:** Summary field for “*Little City*” in R

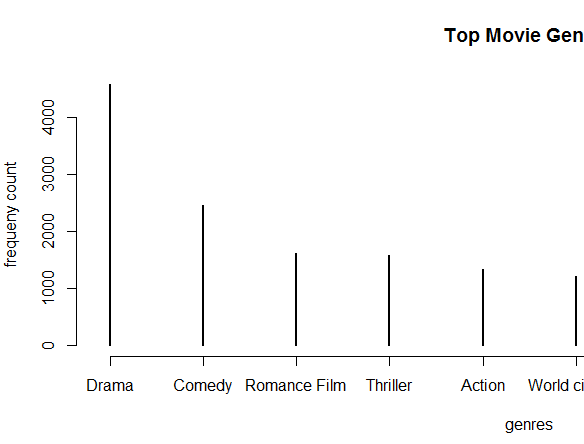
## Methodology: steps detailing how the analysis was carried out

This text mining project applies LDA (Latent Dirichlet Allocation), a topic modeling algorithm, to automatically discover topics that these summaries contain and identify the underlying themes for the most popular movie genres. It also utilizes LDAvis R package to visually examine the topics identified. The details are outlined below.

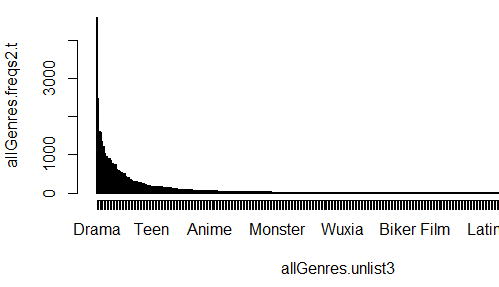
* Read dataset -- Read the movie dataset stored in the comma delimited .csv file.
* Process genres information -- Get all the genres, parse it by commas, then remove the unnecessary characters such as quotes, slashes, spaces, etc. Once genres data is cleaned, a contingency table is created to show the number of movies belong to each of the genres. Last, obtain the top genres. Figure 3 lists the R outputs for the top 10 genres and its associated counts. Figure 4A plots the top 5 genres. It’s interesting to note that ‘all movie genres plot’, as seen in Figure 4B, appears to exhibit Zipf’s Law.



**Figure 3:** R output of the top ten movie genres

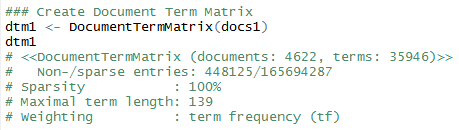


**Figure 4A:** Top five movie genre

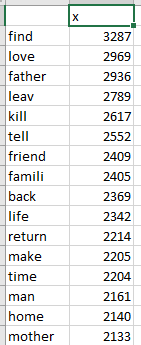


**Figure 4B:** All movie genres plot exhibits Zipf’s Law

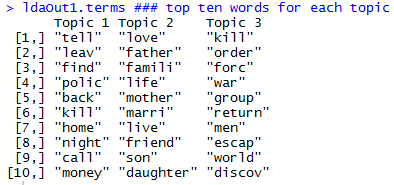
* Process summary information -- Identify the index positions of the movies for the top genre. Get all the text information from the summary variable for the top genres category. In this case, the most popular genre identified is ‘Drama’.
* Create corpus from the summary information. Summary text for ‘Drama’ genre is the focus of this analysis.
* Perform pre-processing of the corpus using tm (text mining) functions. This includes removing of punctuation marks, number, English stop words, blank spaces, converting text to lowercase, and stemming of the document.
* Construct document term matrix.



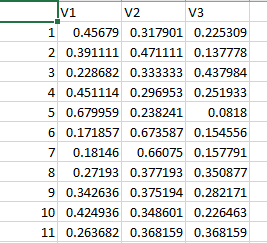
* Compute the word frequencies. See below the snapshot of the word frequency in Excel file. Terms ‘find’ and ‘love’ appear most often.



* Topic modeling -- Set control parameters for Gibbs Sampling method. Discover the three major topics by running Latent Dirichlet Allocation (LDA) model using the Gibbs Sampling algorithm.
* Result Set -- Extract the most likely topics and the most likely terms for each topic. Figure 5A shows the top ten terms for the three main topics discovered. Figure 5B contains the probability for each topic associated with each document.



**Figure 5A:** Top ten terms for each topic.



**Figure 5B:** Probability for each topic associated with each document

## Visually describe the textual data and findings

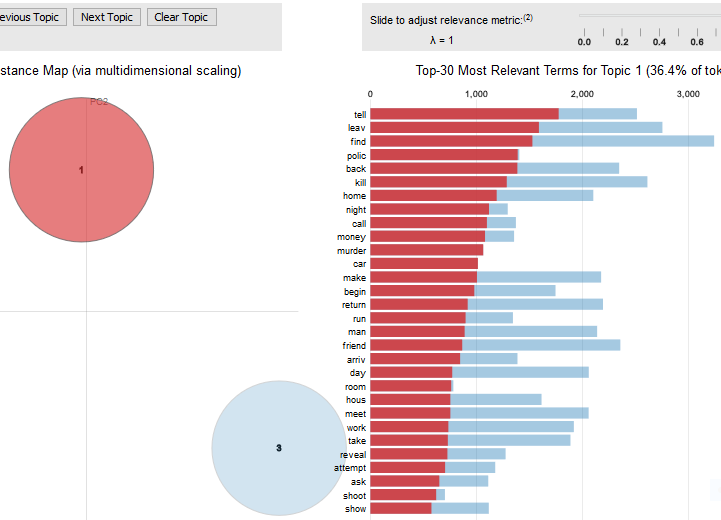
LDAvis, interactive visualization for Topic Models, is used to visualize the findings from LDA topic modeling. Exploring the most popular genre ‘Drama’.

LDAvis reveals that topic 1, with largest circle diameter, has the highest percentage of tokens (36.4%) associated with it. The other two are distributed fairly evenly with 32%, and 31.6% for topics 2 and 3, respectively.

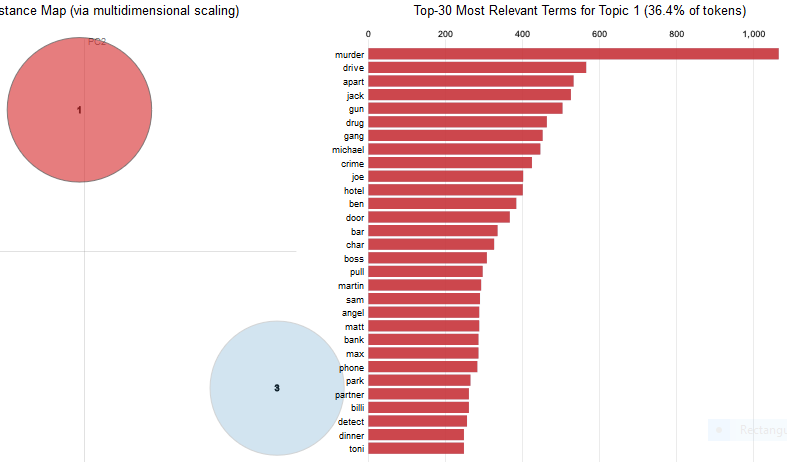
Examining topic 1, Figure 6A shows the top keywords are ‘tell’, ‘leav’, ‘find’, ‘polic’, ‘back’, ‘kill’, ‘home’ when lambda is set to 1. With lambda set to 0, Figure 6B shows the top keywords (excluding proper nouns) that are fully associated with topic 1 are ‘murder’, ‘drive’, ‘apart’, ‘gun’, ‘drug’, ‘gang’, and ‘crime’.

Moving on to topic 2. The graph (not displayed) reveals the terms that are associated with this topic. Some keywords are ‘kill’, ‘forc’, ‘escape’, ‘war’, ‘attack’, ‘return’, ‘soldier’, ‘capture’, ‘surviv’, and ‘battl’.

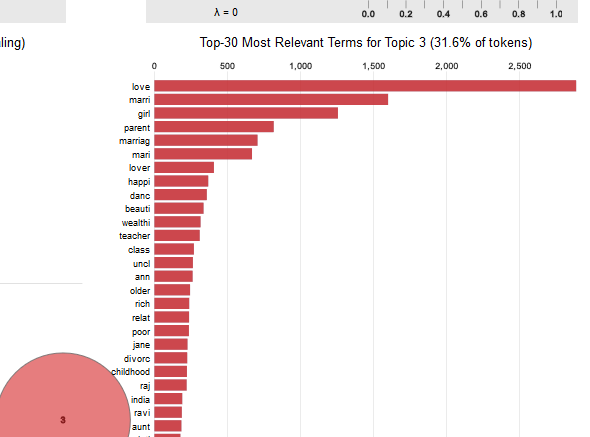
The third main topic discovered from LDA. As it’s shown in Figures 6C and 6D, the top keywords (for both lambda of 0 and 1) are ‘love’, ‘marriag’, ‘girl’, ‘parent’, ‘mother’, ‘famili’, ‘life’, ‘father’, and ‘friend’.



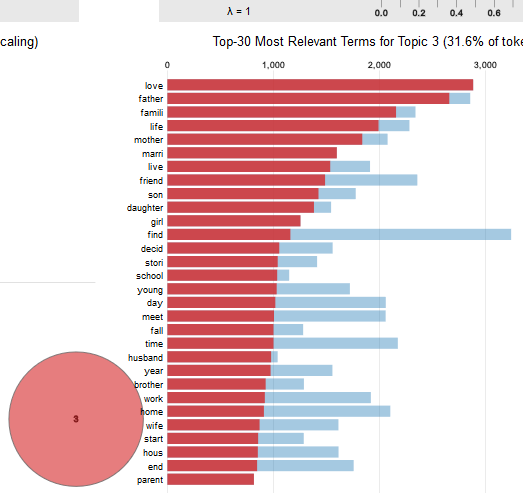
**Figure 6A**: Topic 1 with Lambda set to 1



**Figure 6B**: Topic 1 with Lambda set to 0



**Figure 6C:** Topic 3 with Lambda set to 0



**Figure 6D:** Topic 3 with Lambda set to 1

## Findings along with discussion

In this dataset, the most popular movie genre is ‘Drama’. Out of 10,000 movies, about 45% of them are labeled as such. This is not an exclusive labeling, a movie can belong to multiple genres.

Using the LDA topic modeling method, for movies that belong to the ‘Drama’ genre, I’d describe the characteristics for the three main topics found in the movie plots as follows:

Topic 1: These seem to be crime movies that are related to gangs, drugs, and guns. Many of these involve killing and murder as well.

Topic 2: These appear to be war related movies which involve words such as battle, command, and soldiers. Words such as kill, attack, survival, capture, and escape portray the brutality of wars.

Topic 3**:** About 31% of the movies are about family, love, marriage, and friendship.

Future work and further analysis can be done to validate if many of the movies identified in topic 1 also belong to the ‘crime fiction’ and ‘action’ genres, topic 2 ‘war film’ and ‘action’ genres, and topic 3 ‘romance film’ genre. Perhaps finer topic modeling (increase the number of topics) might even reveal more information.

## References

M. Jockers (2014). *Text Analysis with R for Students of Literature*

Dr. G. Bansal (Spring semester 2017)*. Lecture notes on Text Mining topics*

“*Little CIty*”. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Little\_City

C. Gandrud, R Bloggers (2015). *A Link Between topicmodels LDA and LDAvis.* Retrieved from

https://www.r-bloggers.com/a-link-between-topicmodels-lda-and-ldavis/

Original dataset movie\_dataset.csv, content from Wikipedia, data file made available by Automated Insights, Inc.