Application of von Mises-Fisher mixture model for text clustering IMDB movie synopses

PRESENTER:
Justin Peter

OBJECTIVE

Text mining and clustering are excellent data analytic techniques to manipulate text into interesting insights. Studies also found that mixture modeling is an effective tool for creating meaningful clusters of text. In this project, we utilize the mixture model based on von Mises-Fisher distributions to create movie clusters within the crime genre.

METHODS

1. Extract data from IMDB
   - Focused on crime genre with movie rating of at least 8
2. Transform movie descriptions into document term matrix
   - Tokenization
   - TF-IDF
   - Stopwords
   - Stemming
   - Word lengths
3. Word selection approaches
   - Frequency removal
   - Manual word removal
4. Find the optimal number of clusters using a cross validation approach
5. Apply it to the IMDB data to obtain the results with the table on the center
6. Determine cluster names
   1. Like words
   2. Bigram frequencies

R LIBRARIES

MovMF (mixture model using vMF distribution), tm, dplyr, ggplot, SnowballC, stringr

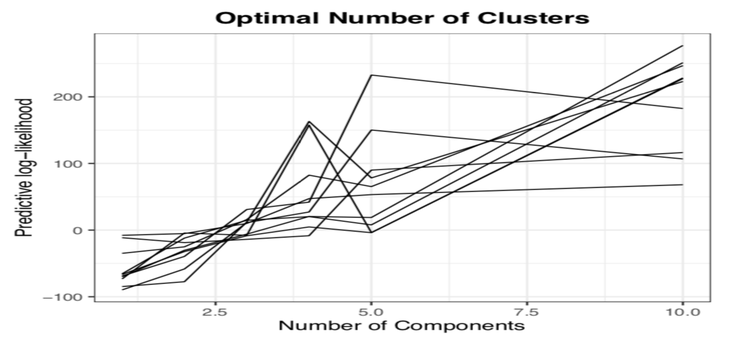
POTENTIAL

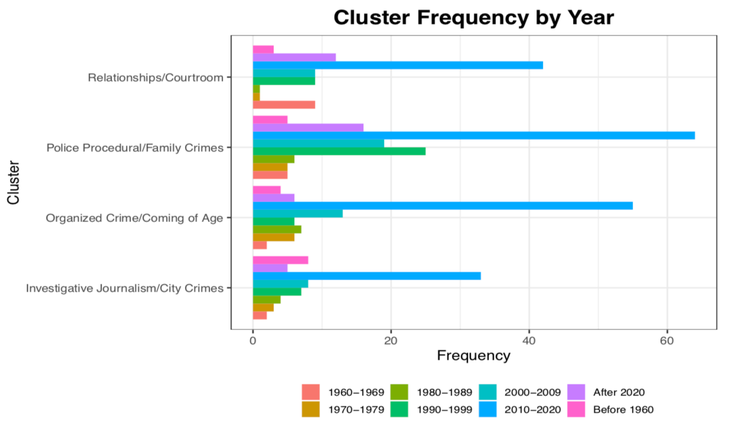
- Should we depend on word selection to improve the algorithm?
  - Time consuming
  - Context

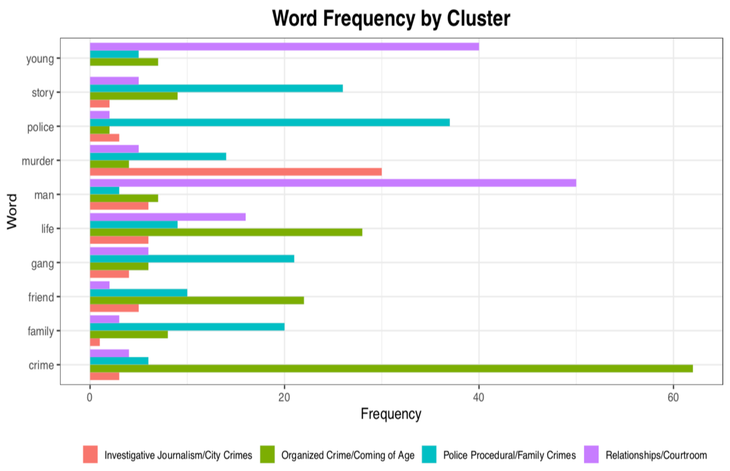
The Godfather and The Godfather Part 2 don't belong in the same category.

Investigative Journalism/City Crimes	Police Procedural/Family Crimes	Relationships/Courtroom	Organized Crimes/Coming of Age
murder	police	man	crime
people	story	young	life
city	family	woman	friend
love	brother	revenge	boy
son	team	life	old
journey	father	girl	mother
journalist	cop	private	school
accused	officer	house	family
blood	gang	lawyer	boss
children	killer	husband	syndicate

Take a picture to download the full paper







Popular Movies
The Dark Knight
The Godfather
To Kill A Mockingbird
The Godfather Part II

Justin Peter, Dr. Shushismita Sarkar

BGSU | DEPARTMENT OF Computer Science
BOWLING GREEN STATE UNIVERSITY