Are Romantic Comedies Dead?

Naren Prakash, Fathima Shaikh, Caleb Williams

Research Questions

How are we comparing rom-coms across different eras?

- Are recent romantic comedies more correlated with each other than those of the past?
- Have romantic comedy scripts become less complex (in terms of vocabulary)?
- Can the era of a romantic comedy be easily predicted based on aspects of the script?

Agenda

Cleaning, Data Overview, and Word-Based Analysis

Sentiment Analysis and Clustering

Topic Modeling, LLM Integration, Prediction, and Results

Data cleaning, EDA and Word-based Analysis

Cleaning the Data

- Removed common English names using genderdata package
- Removed punctuation, special characters
- Set everything to lowercase
- Removed English Stopwords
- Removed all whitespace
- Specific transformations, like “mmyydd” to a space
- Stemmed all words
- Removed sparse terms

35162/122068

Non-/sparse entries

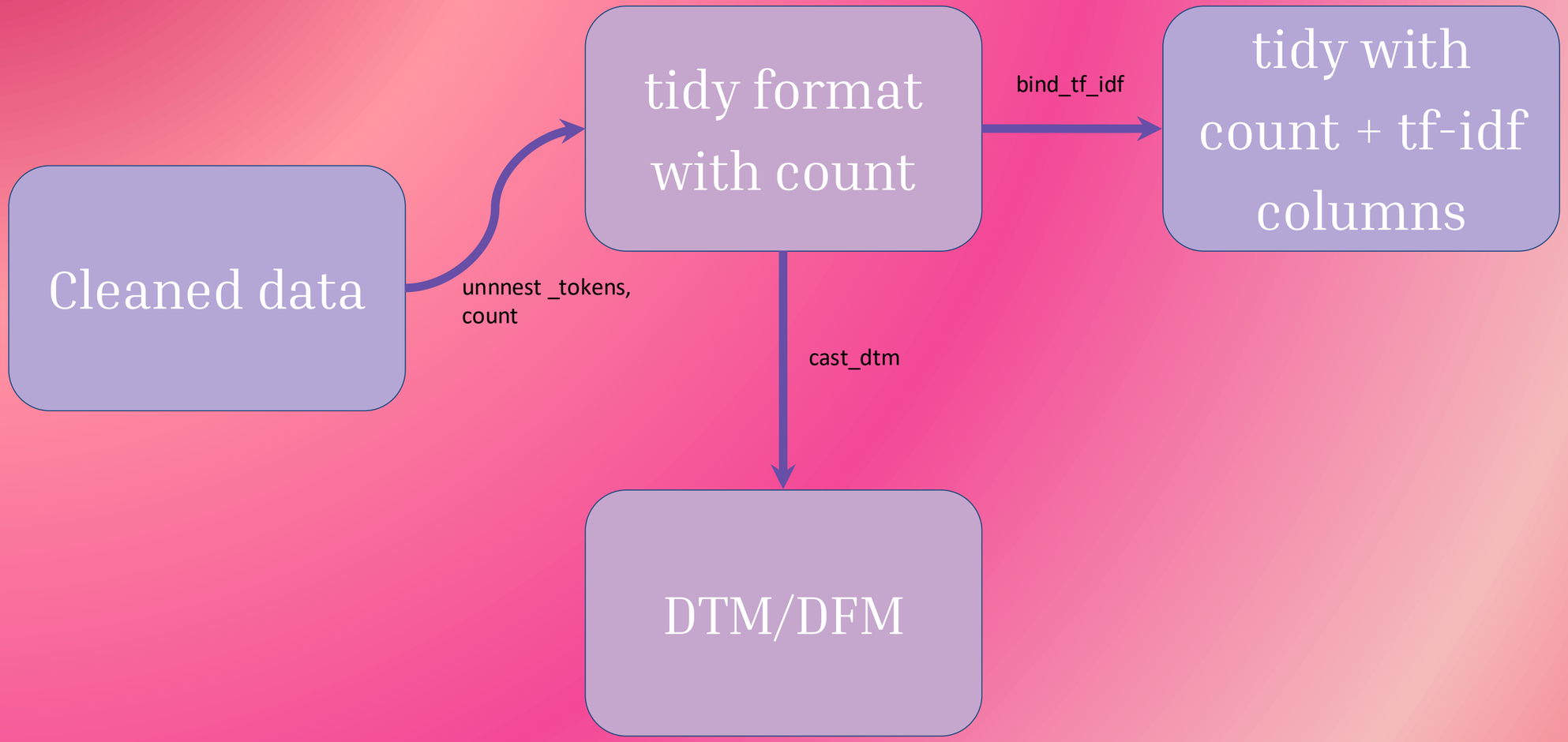
78%

Sparsity

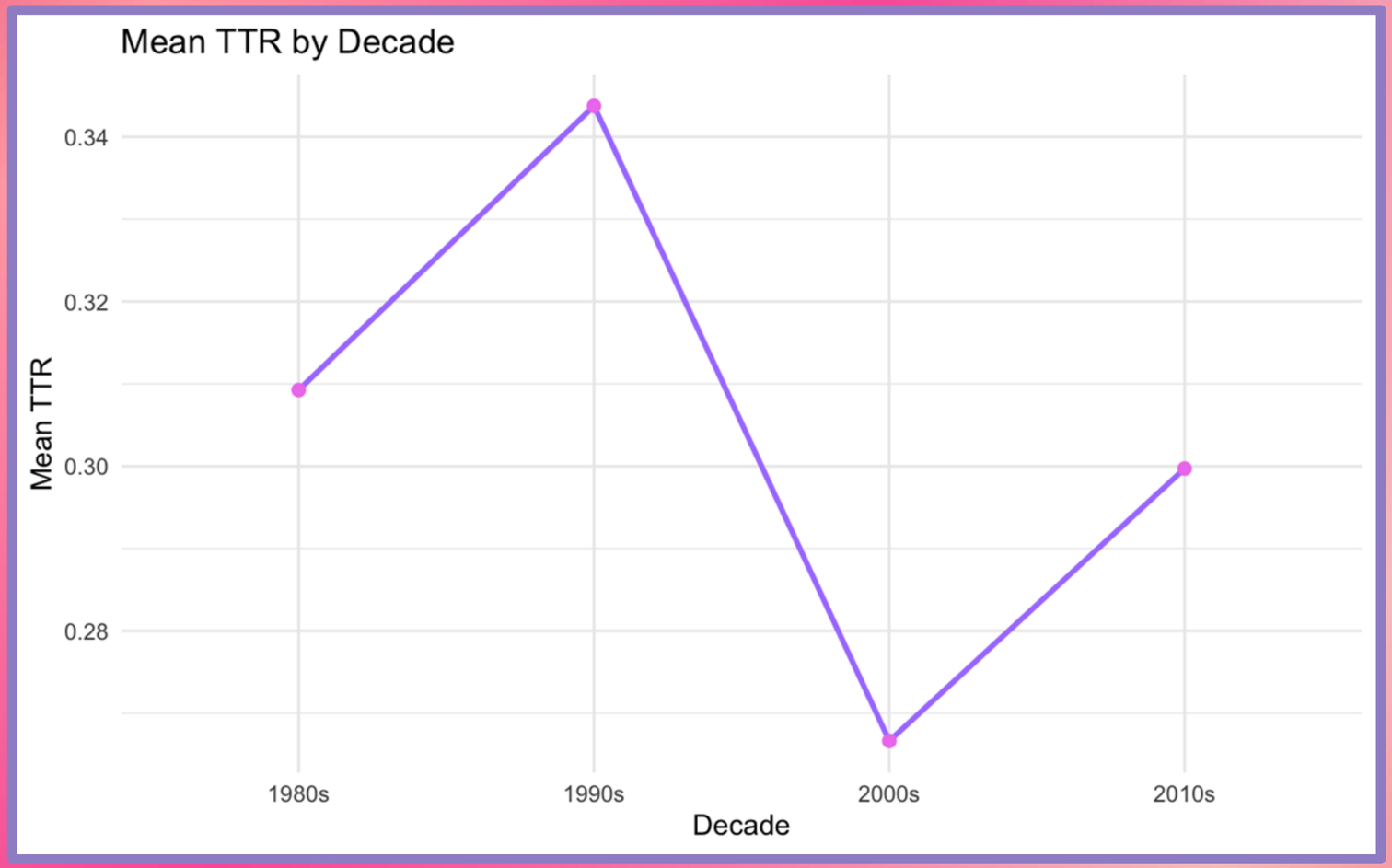
14

Maximal term length

Data handling



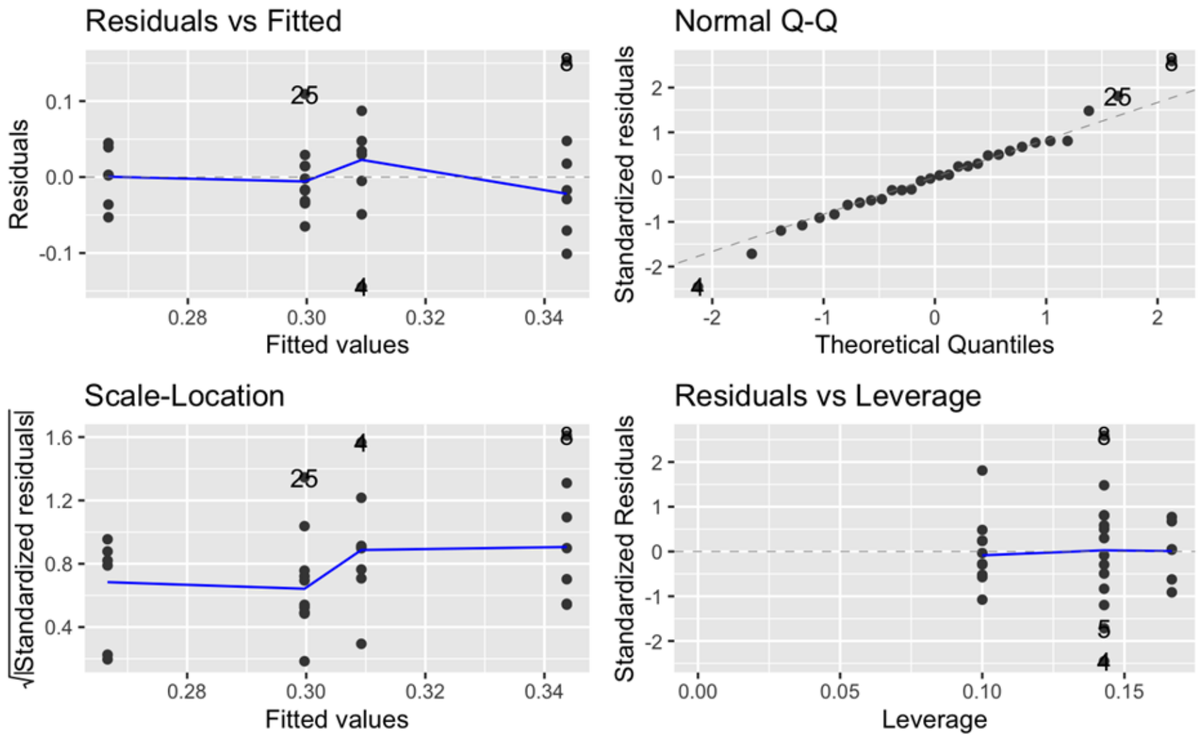
How lexically diverse is each decade?



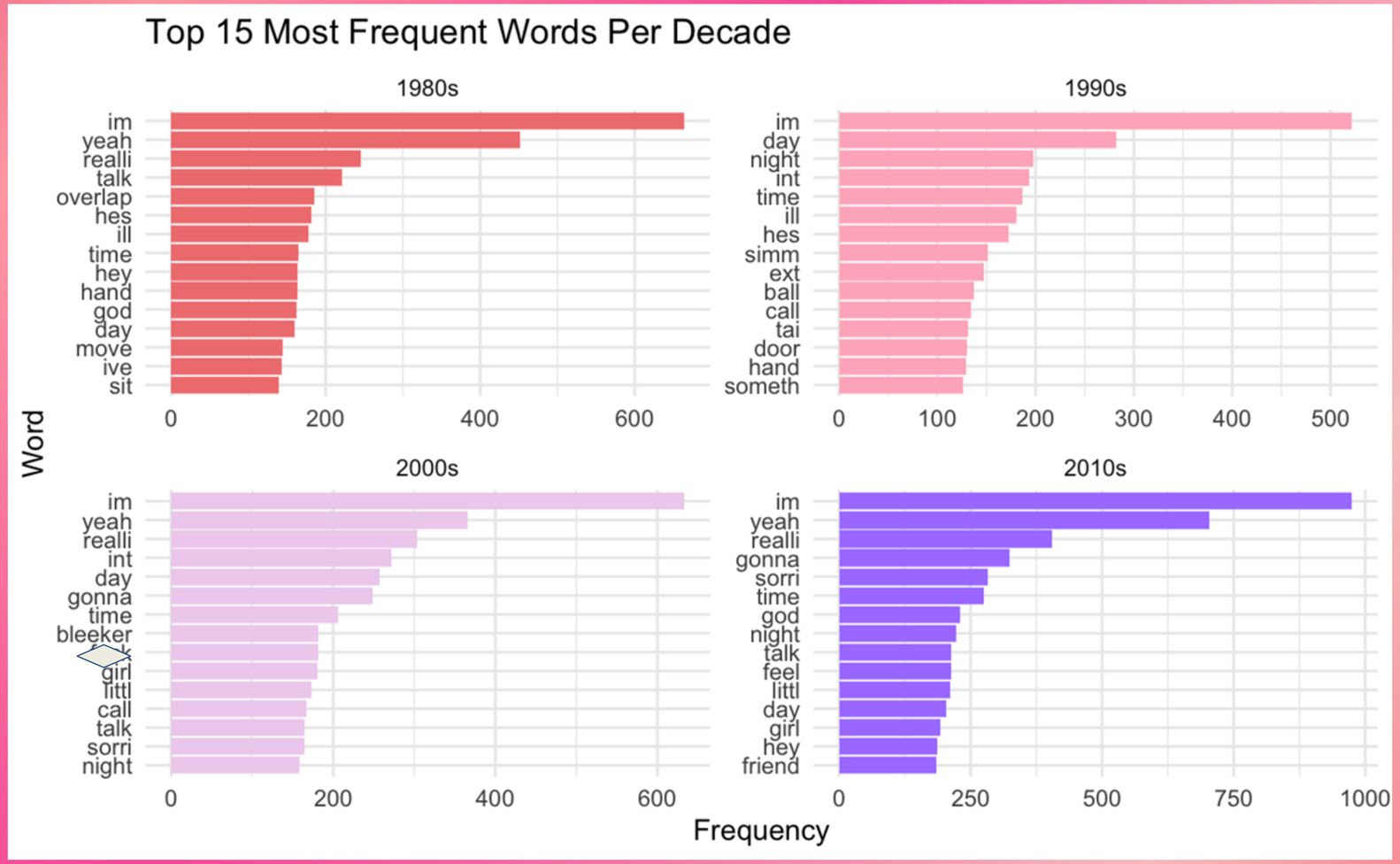
Are these differences significant?

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	0.309252	0.024055	12.856	8.96e-13
decade1990s	0.034504	0.034019	1.014	0.320
decade2000s	-0.042619	0.035408	-1.204	0.240
decade				763

Multiple R Squared	Adjusted R Squared	F-Statistic	p-value
0.1579	0.0607	1.625 (3 and 26 DF)	0.2078



What words appear most in each decade?



1980s

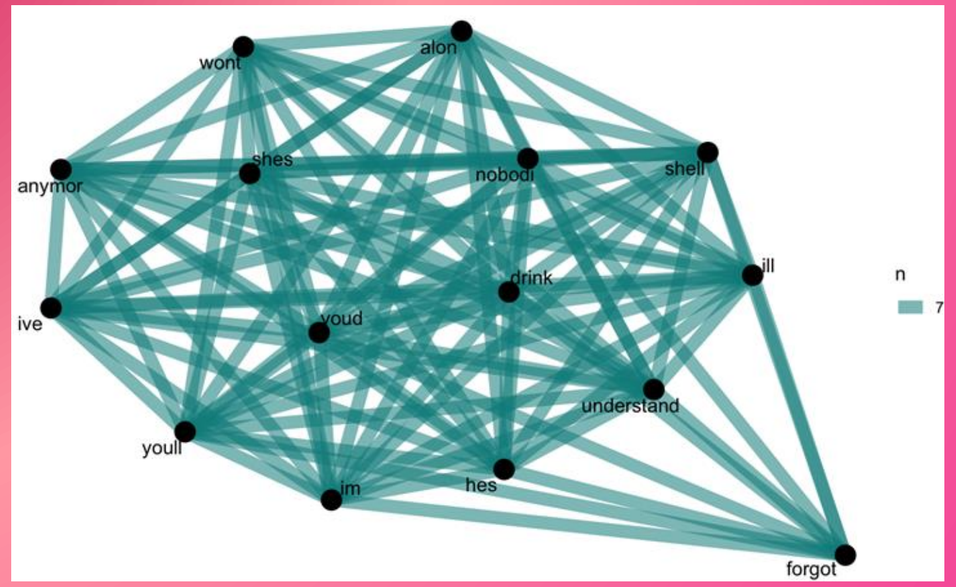
1990s

2000s

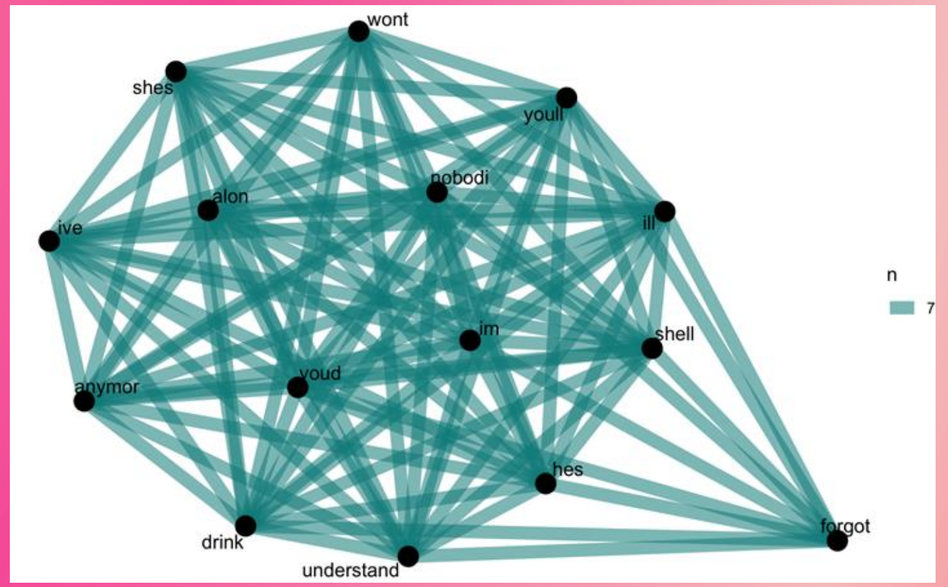
2010s

Word co-occurrences by decade

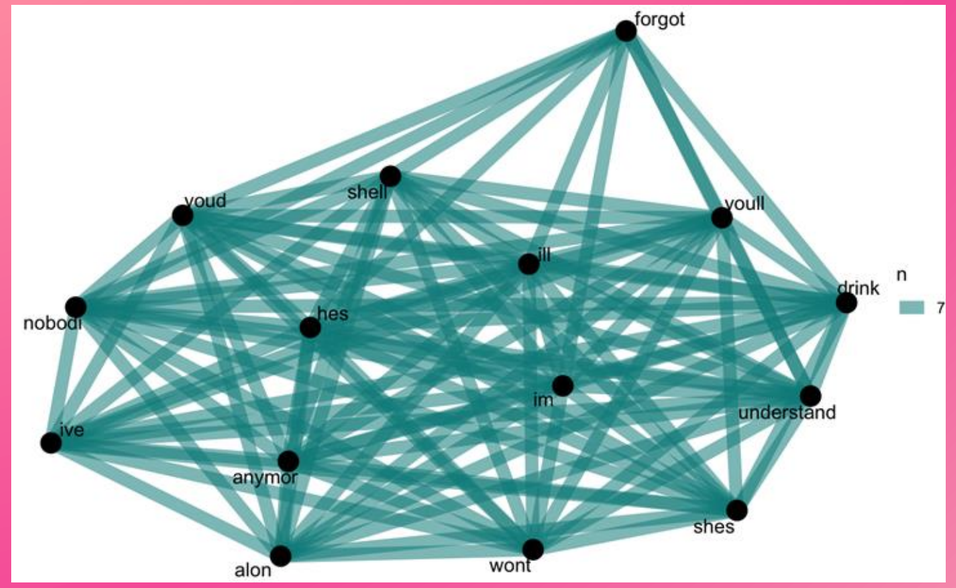
1980s



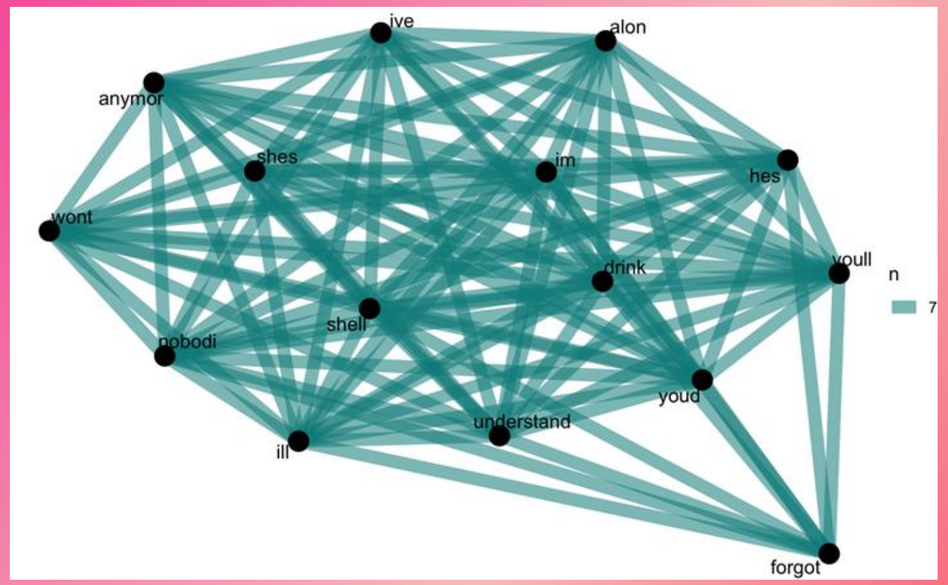
1990s



2000s

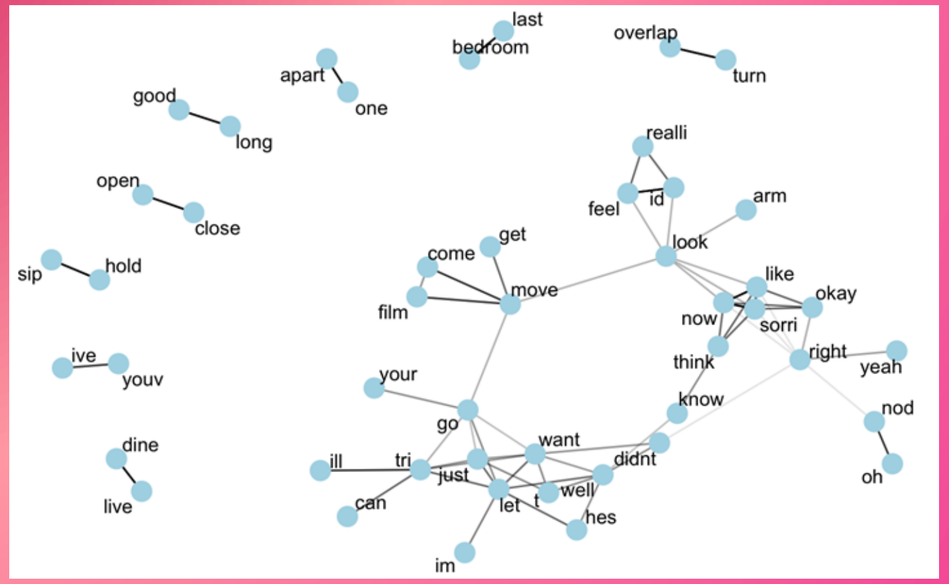


2010s

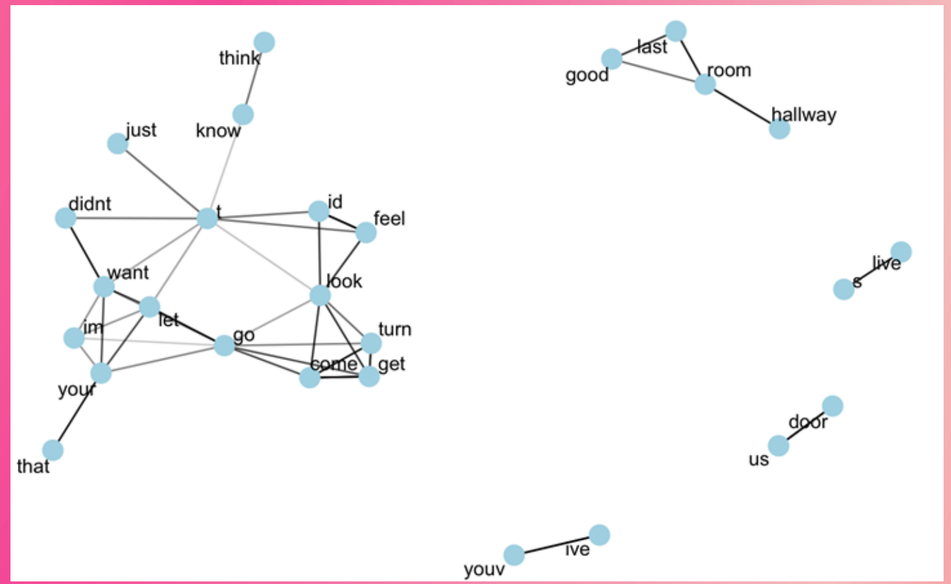


Correlation analysis with bigrams

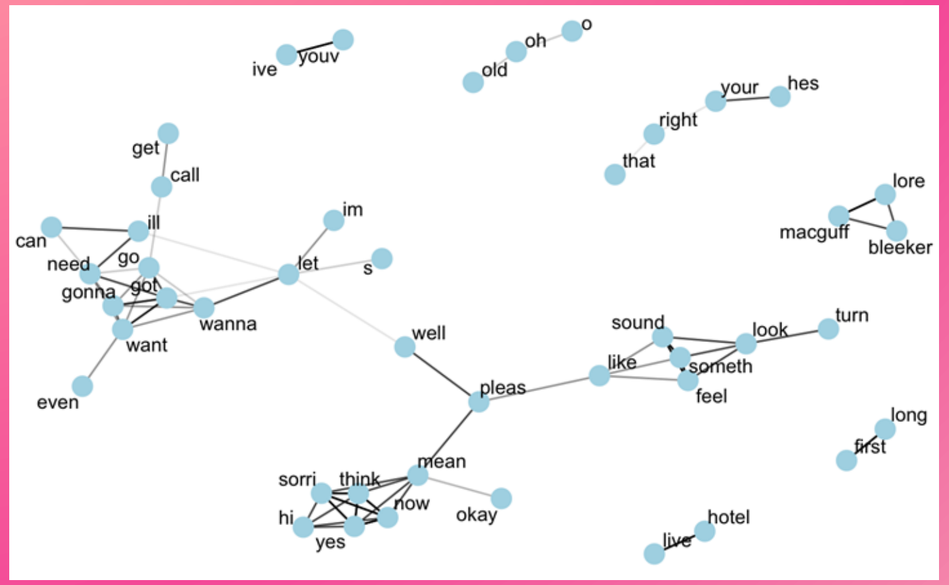
1980s



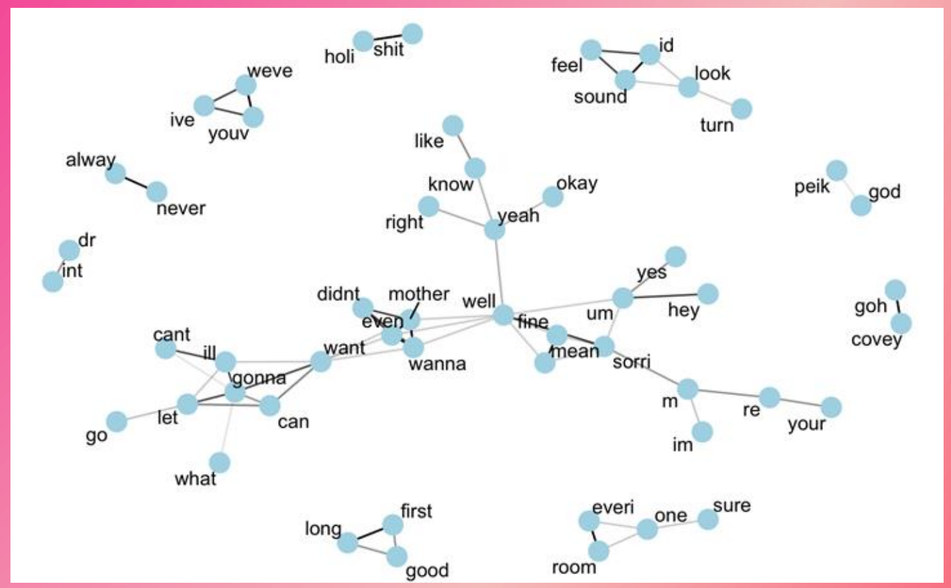
1990s



2000s

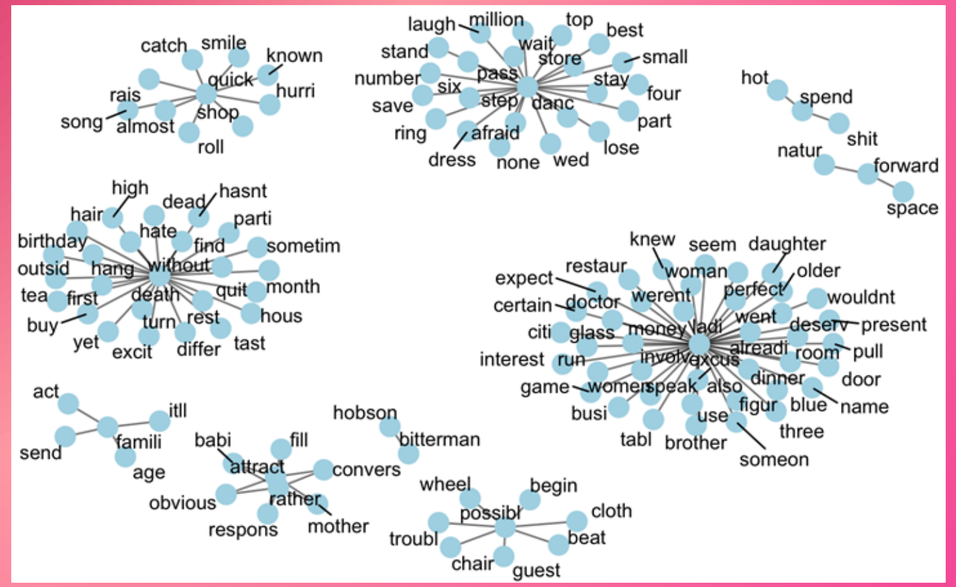


2010s

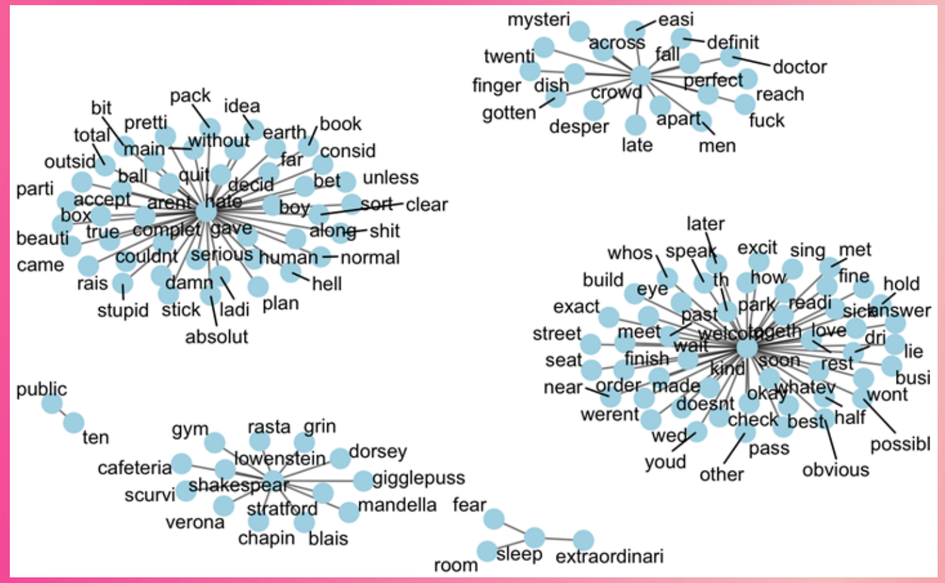


Correlation analysis with trigrams

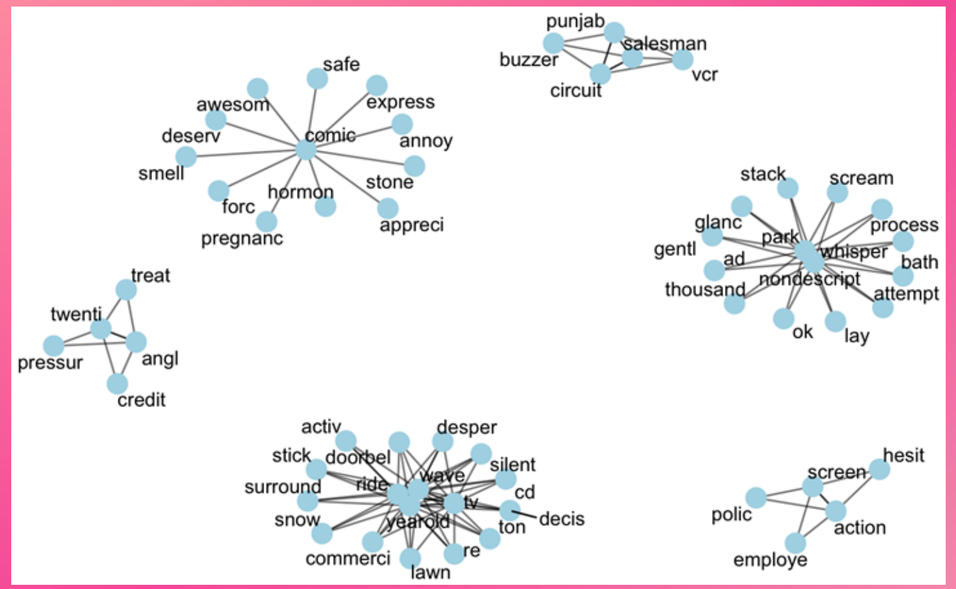
1980s



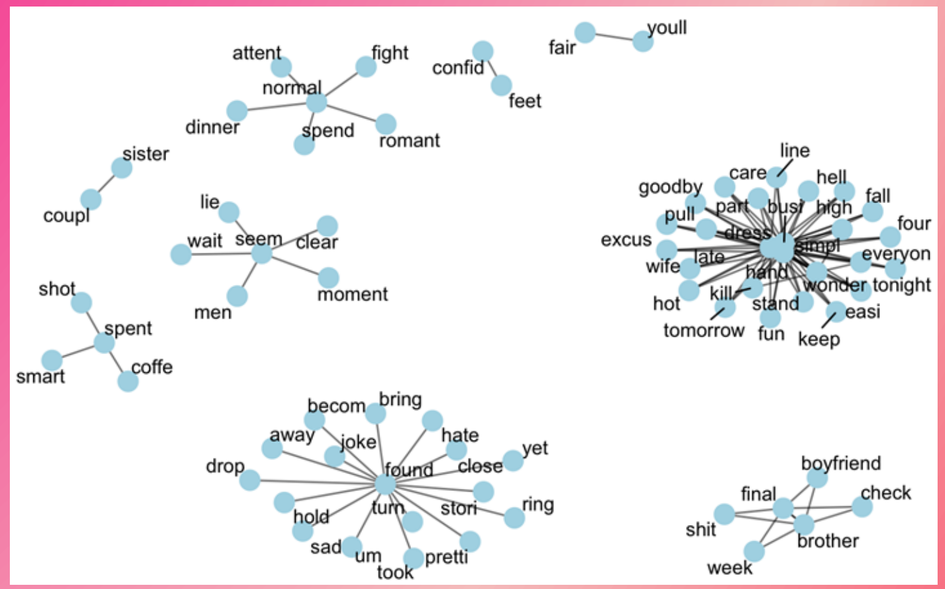
1990s

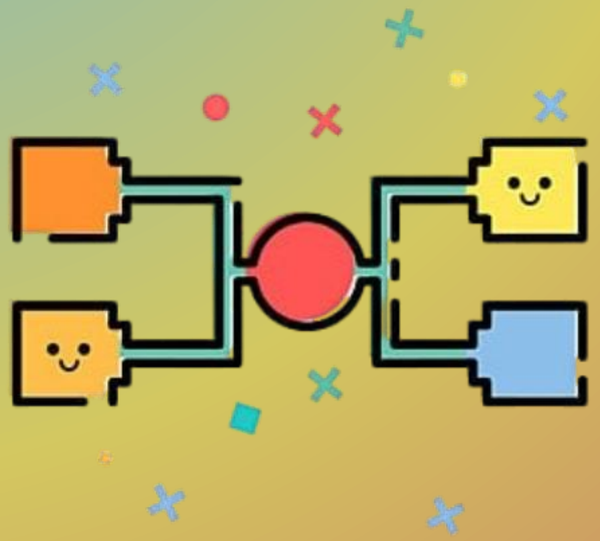


2000s



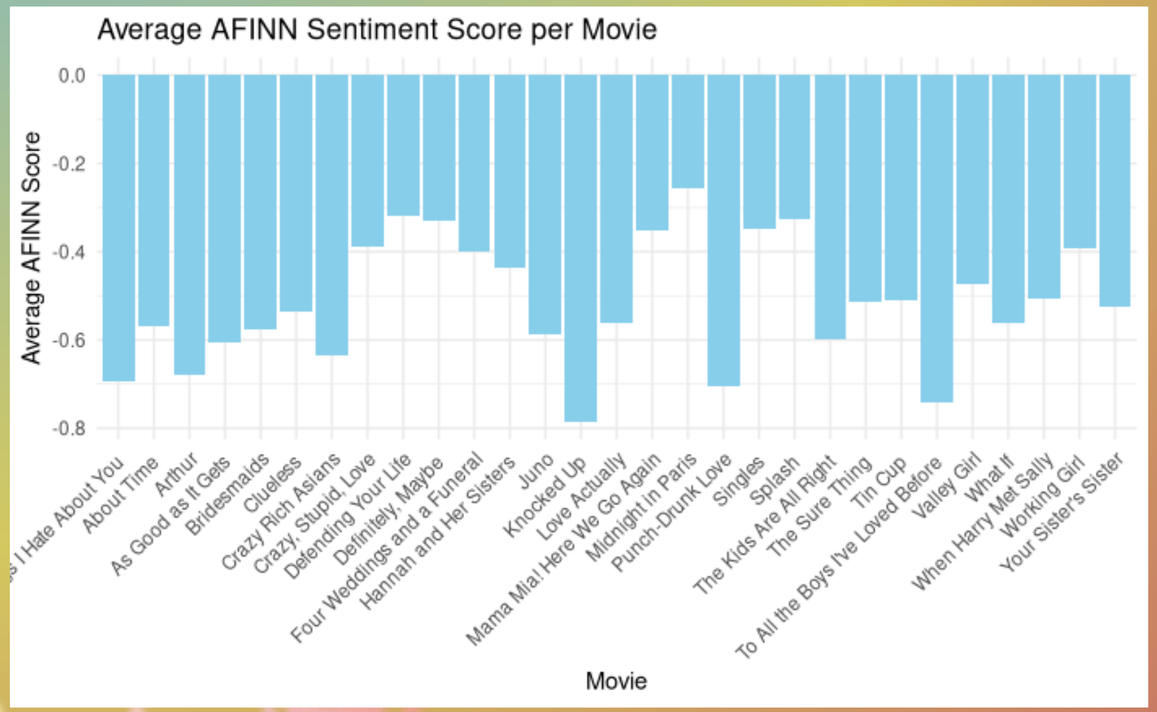
2010s

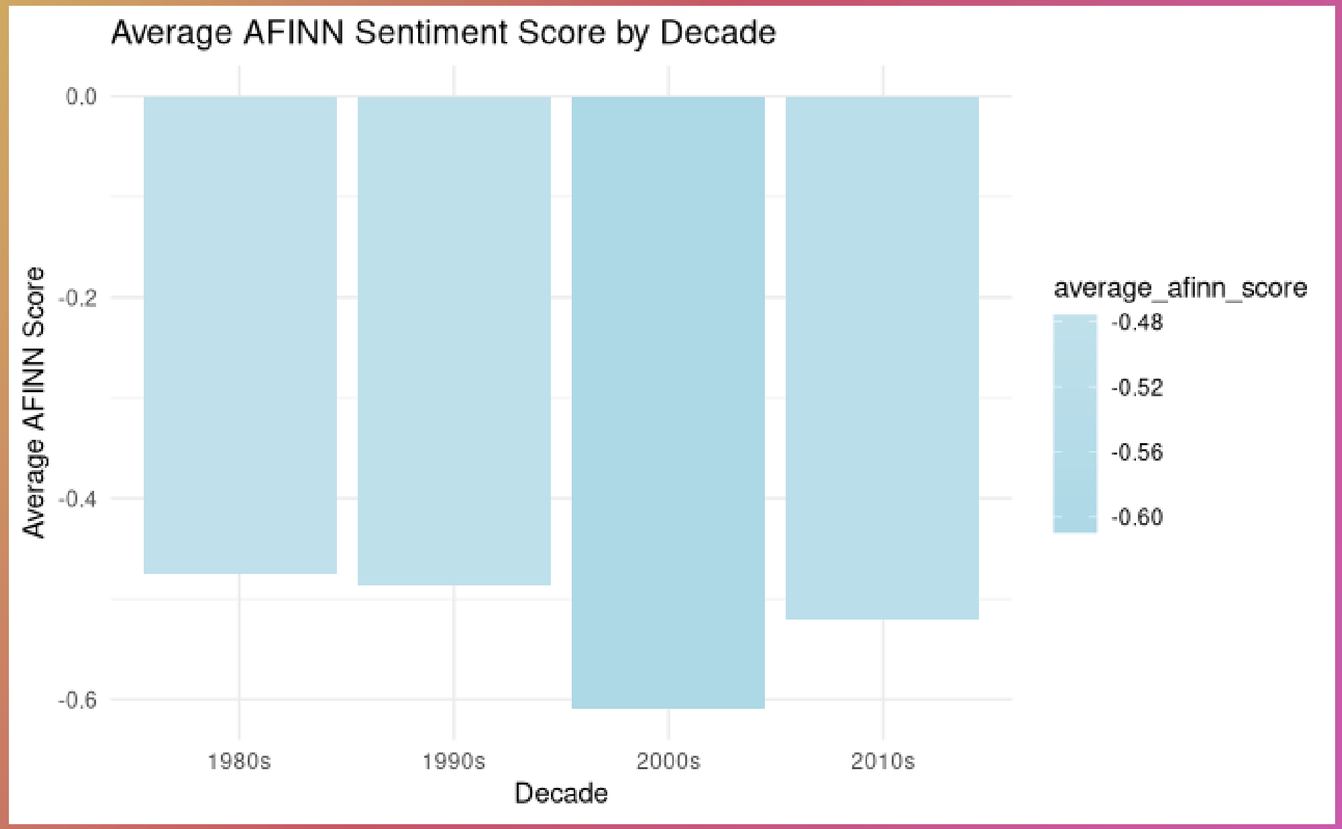




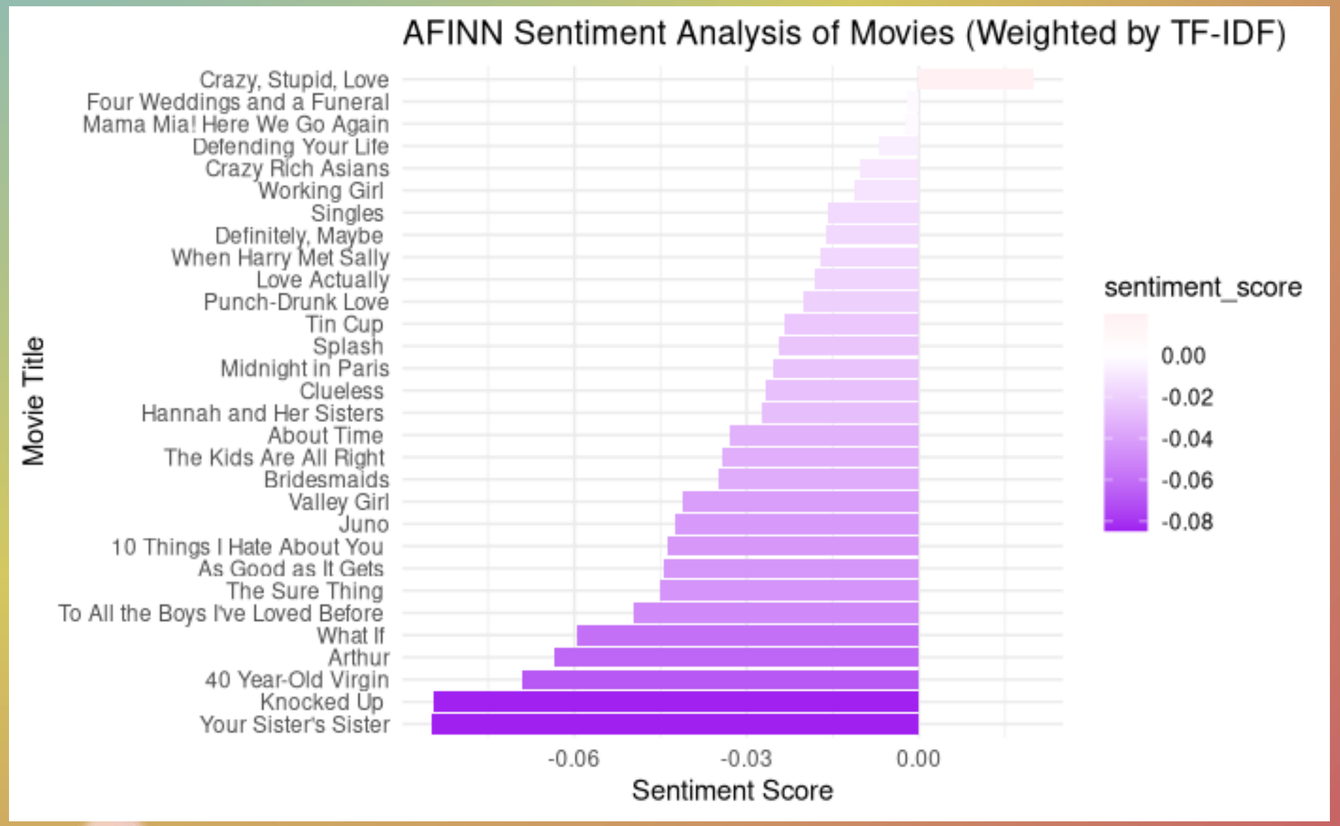
Sentiment Analysis and Clustering

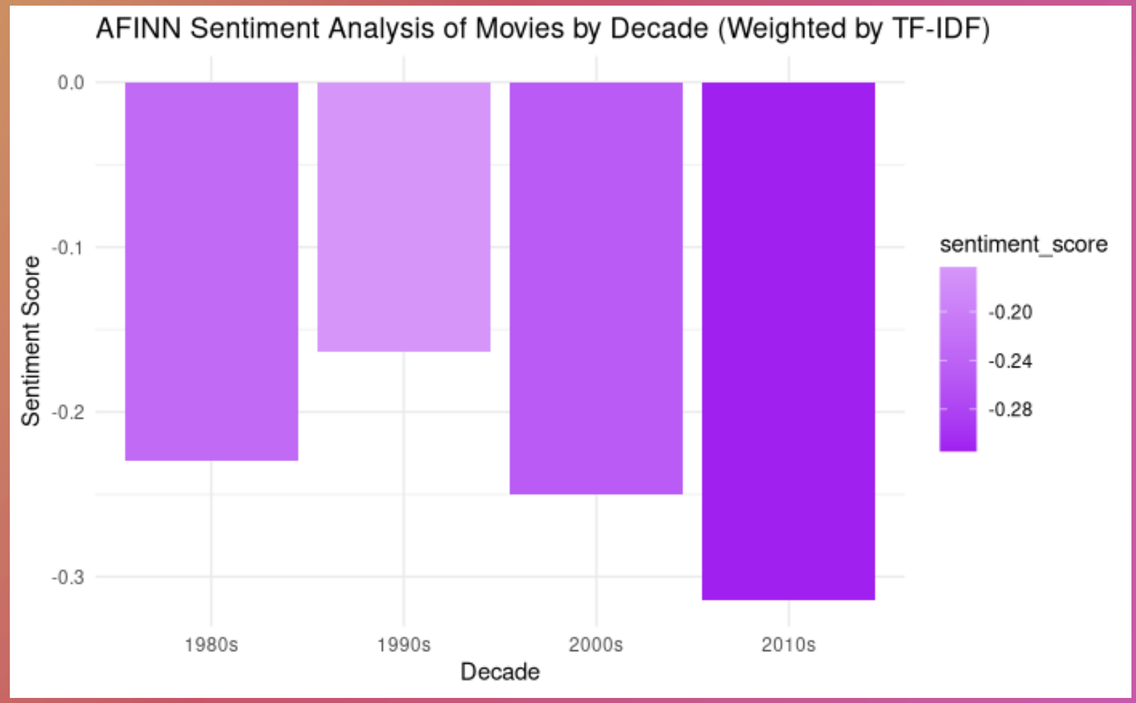
How Does Emotional Tone Change Across Decades?



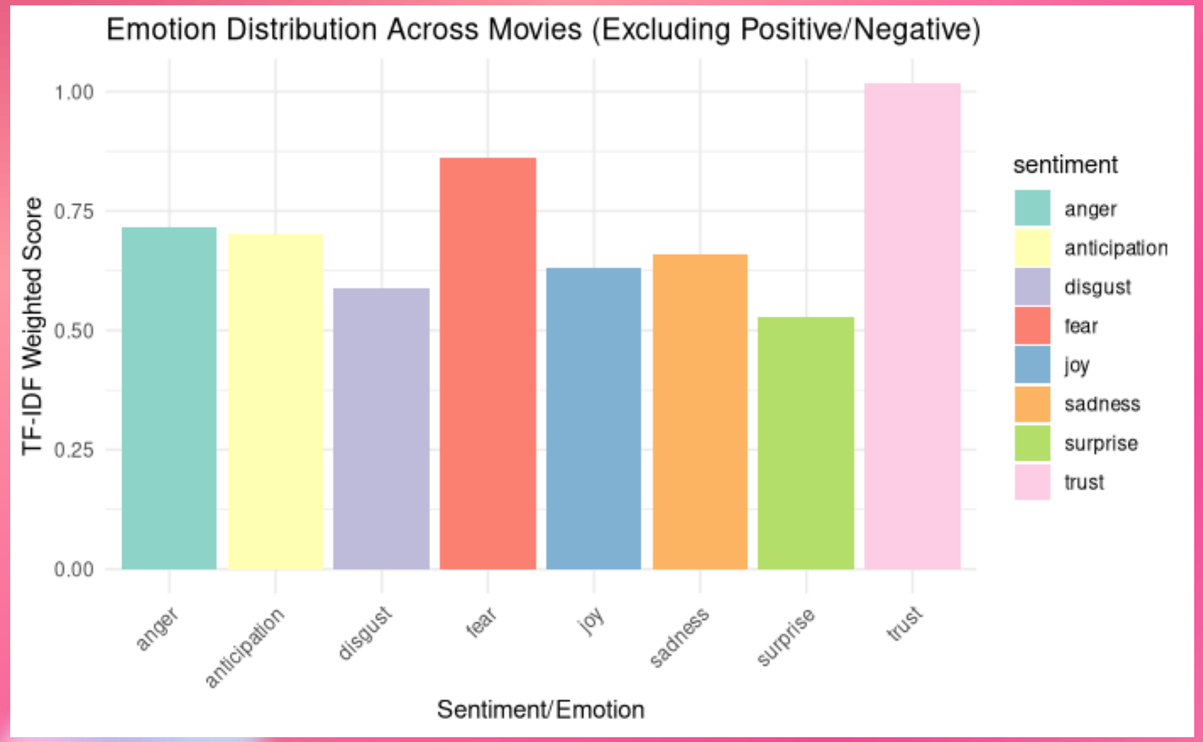


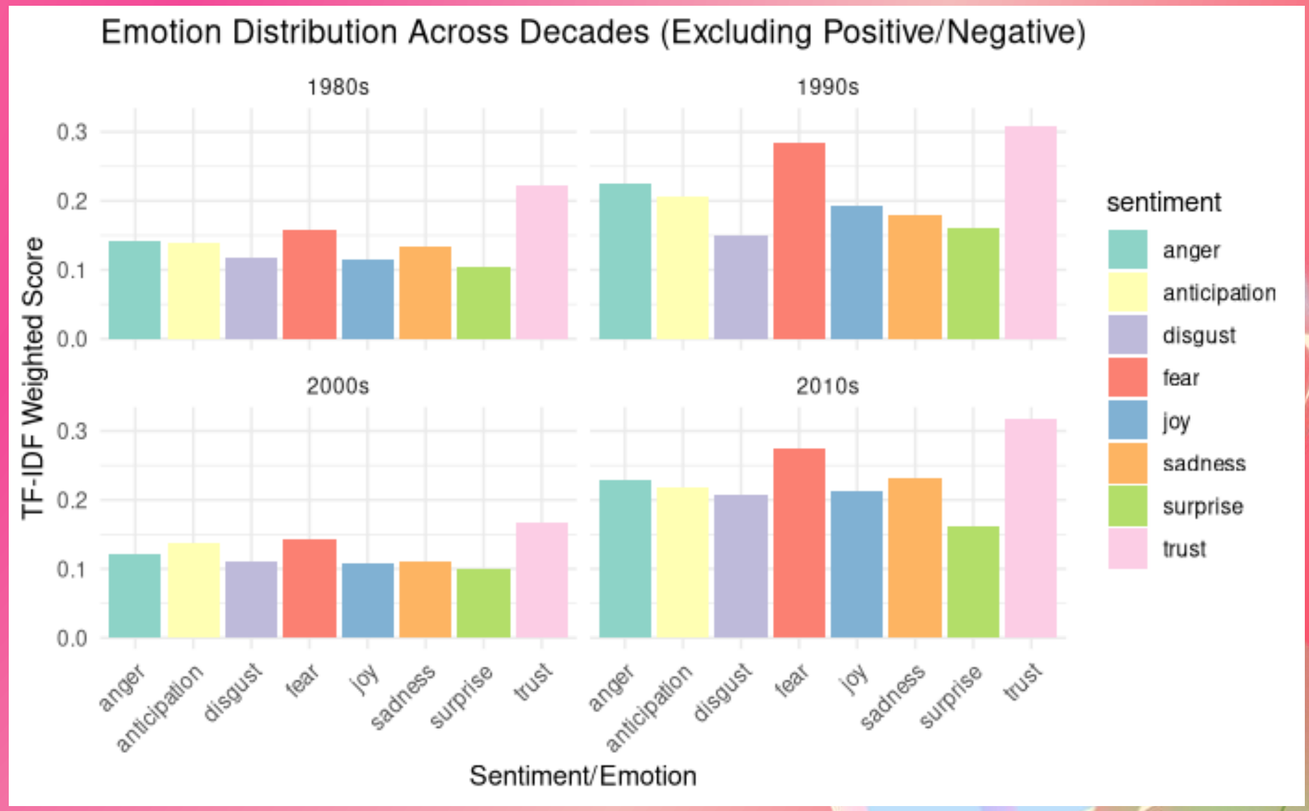
Are the sentiments of common words changing over time?



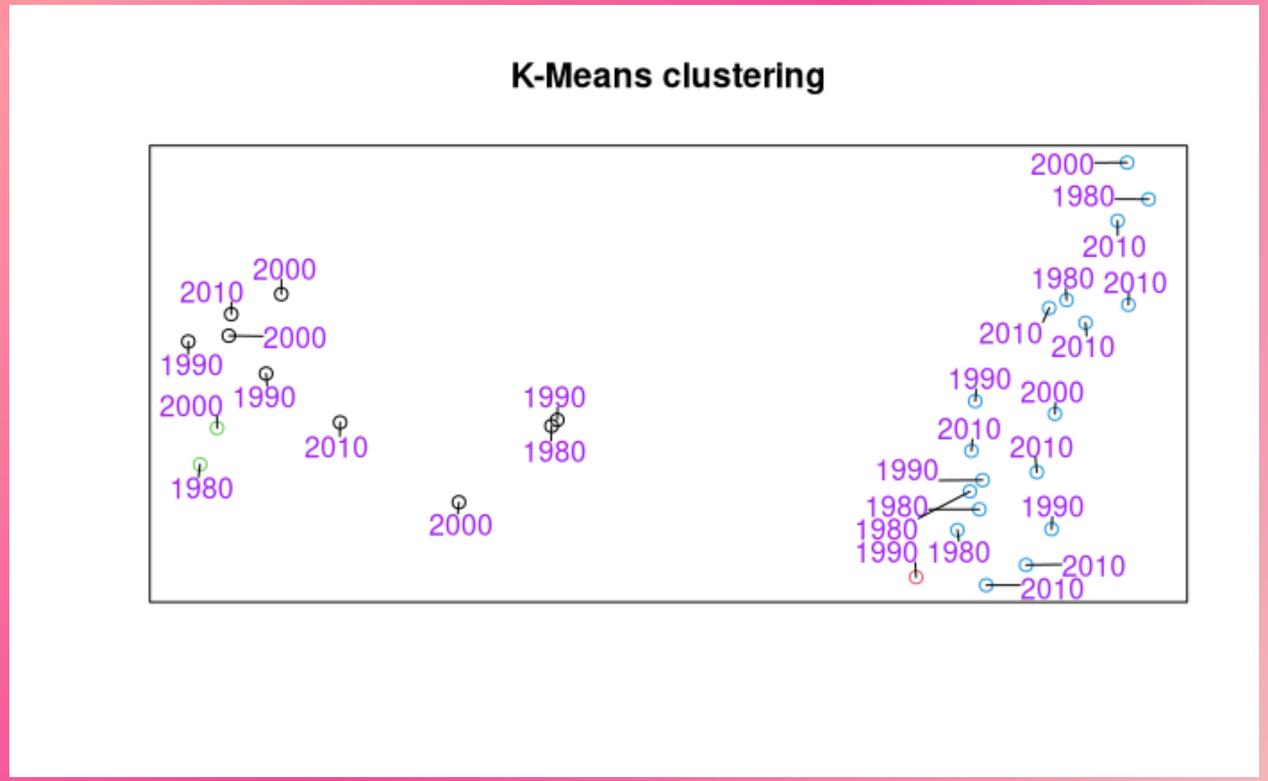


Which emotions dominate, and how do they change over time?

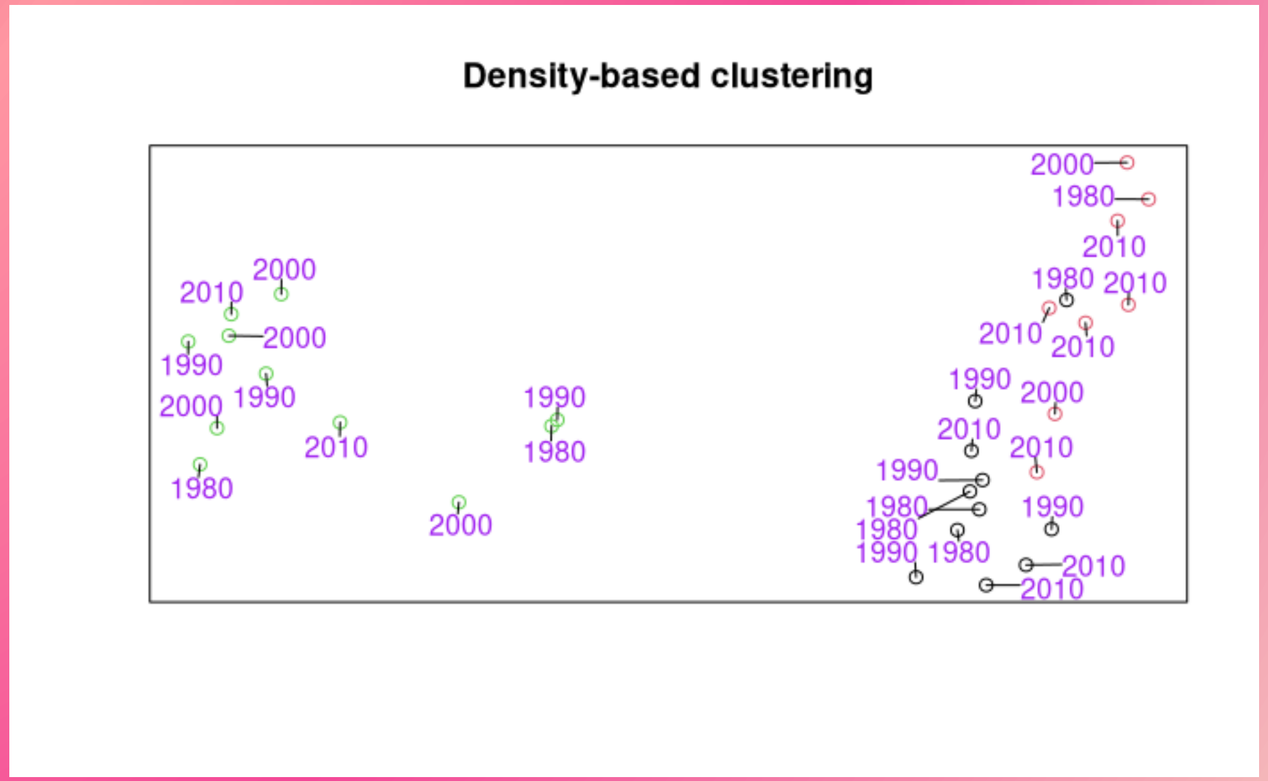




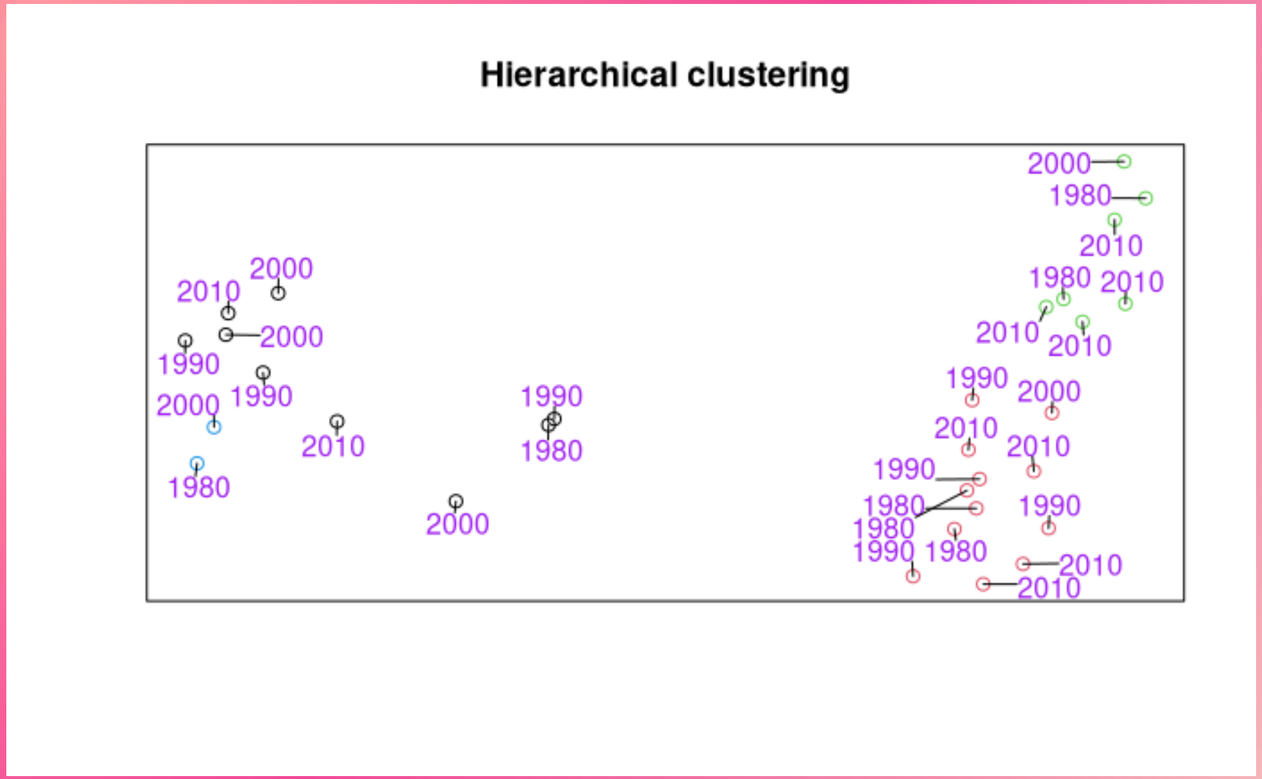
Can we separate the movies by decade?



Can we separate the movies by decade?

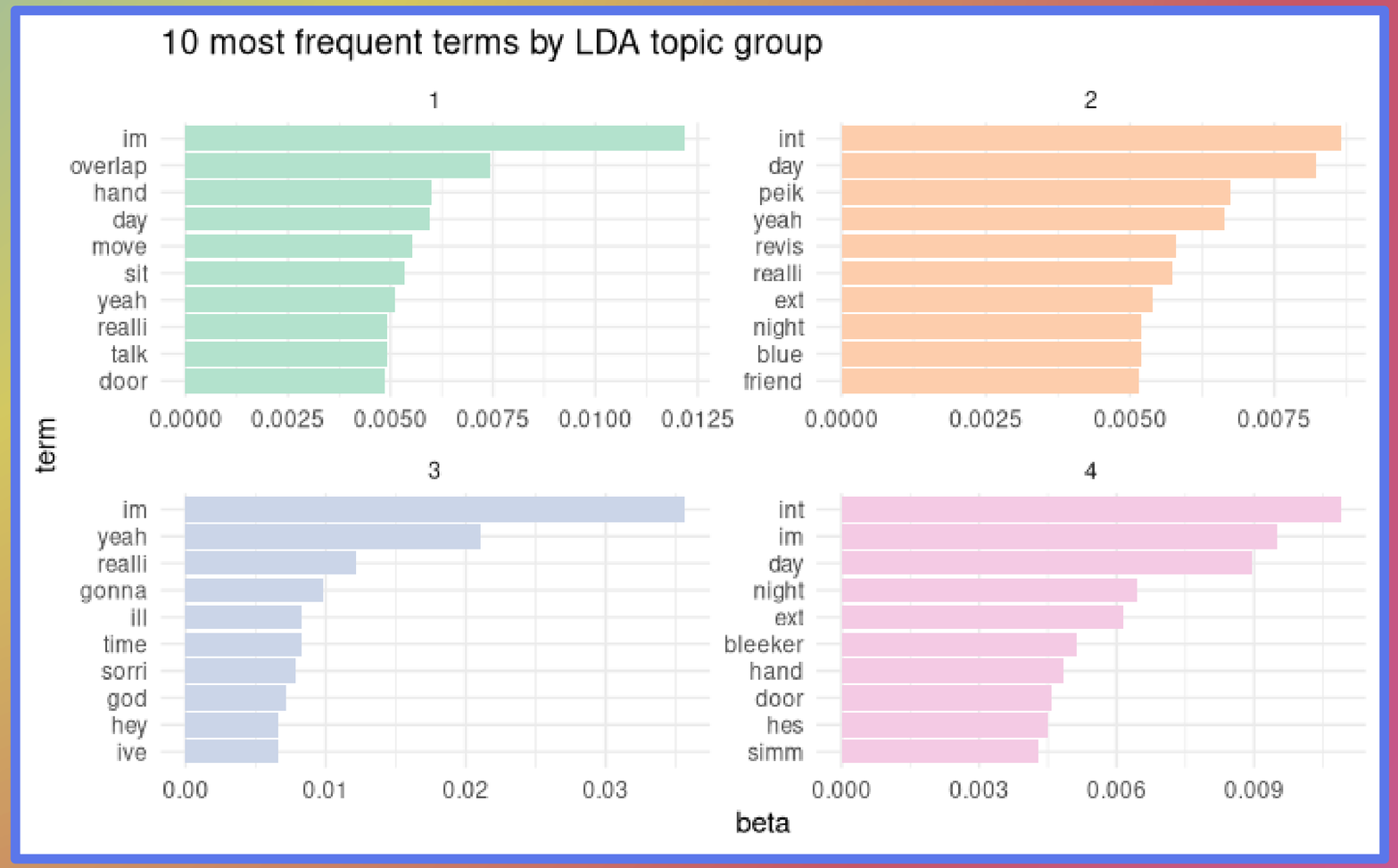


Can we separate the movies by decade?

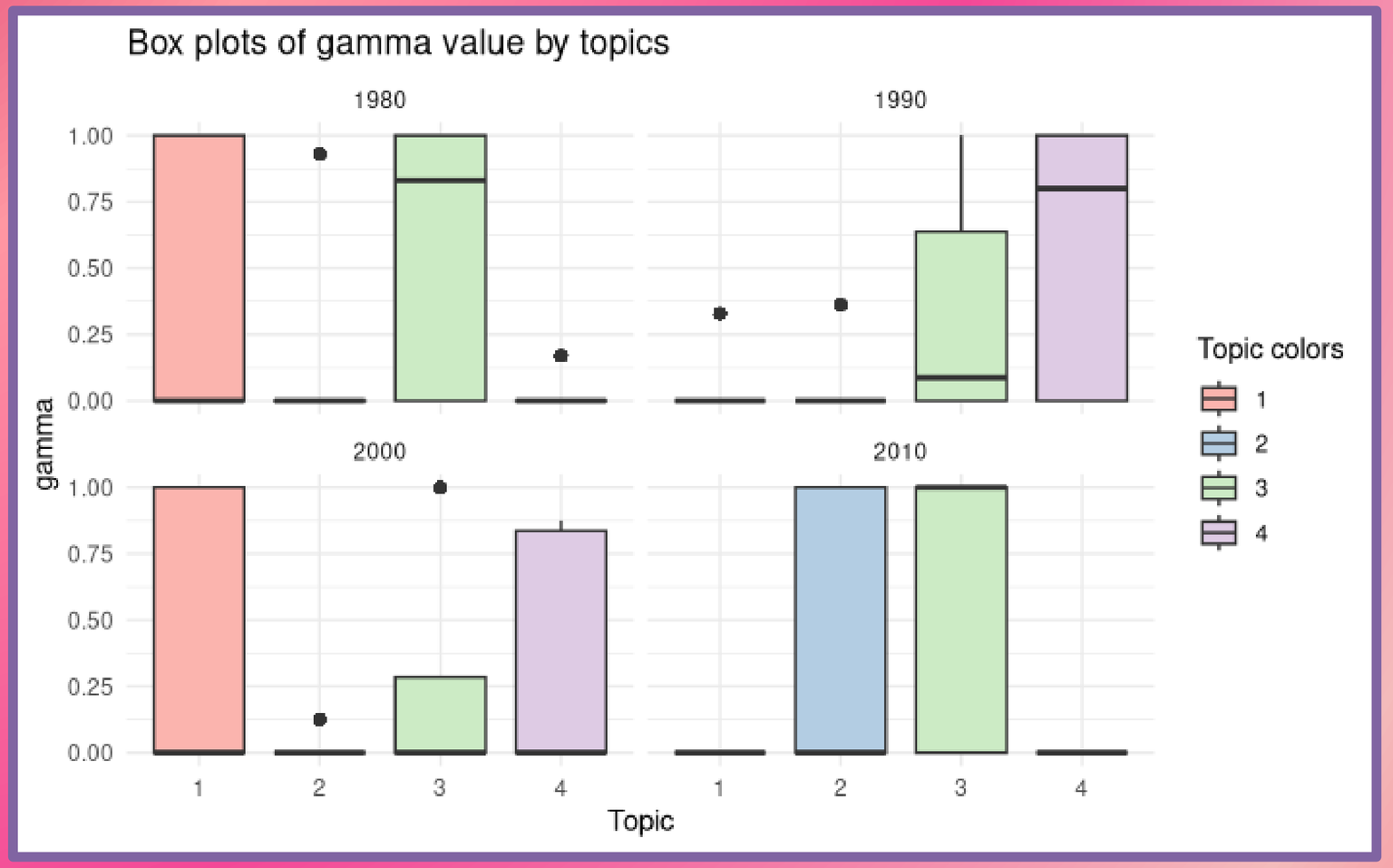


Topic Modeling, LLM Integration, and Prediction

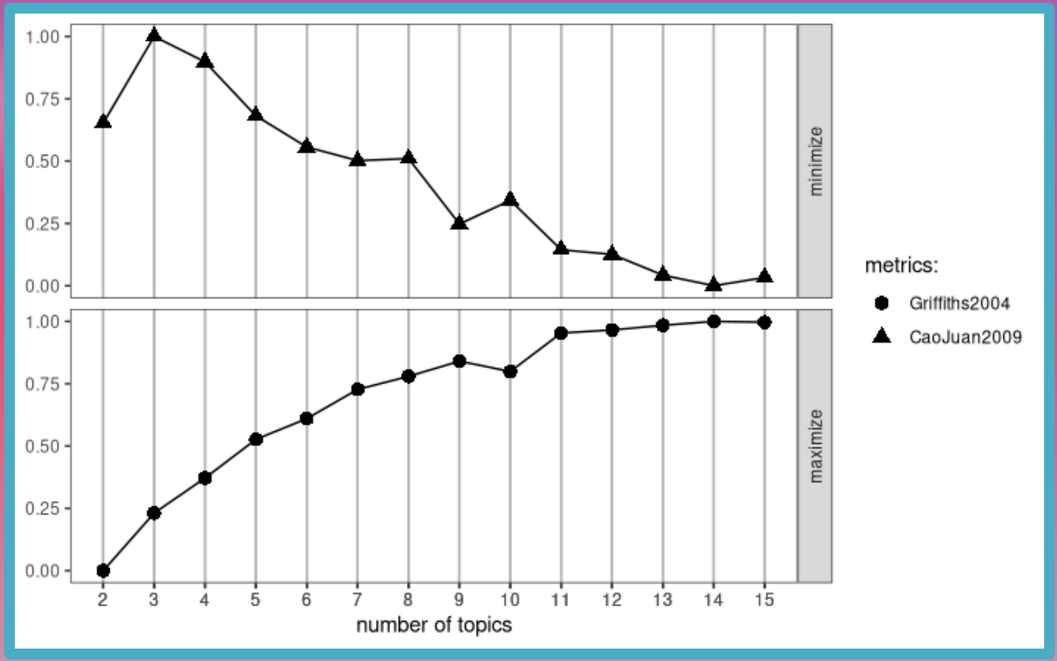
What are the most popular terms in each topic group?

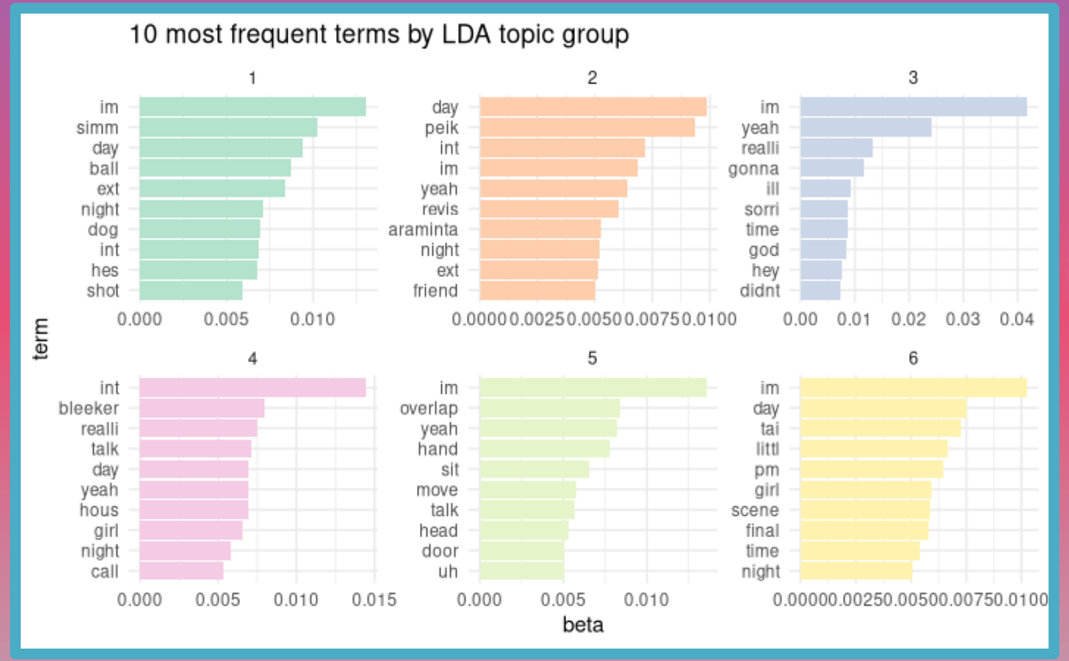


Do our topics correspond with our decade splits?

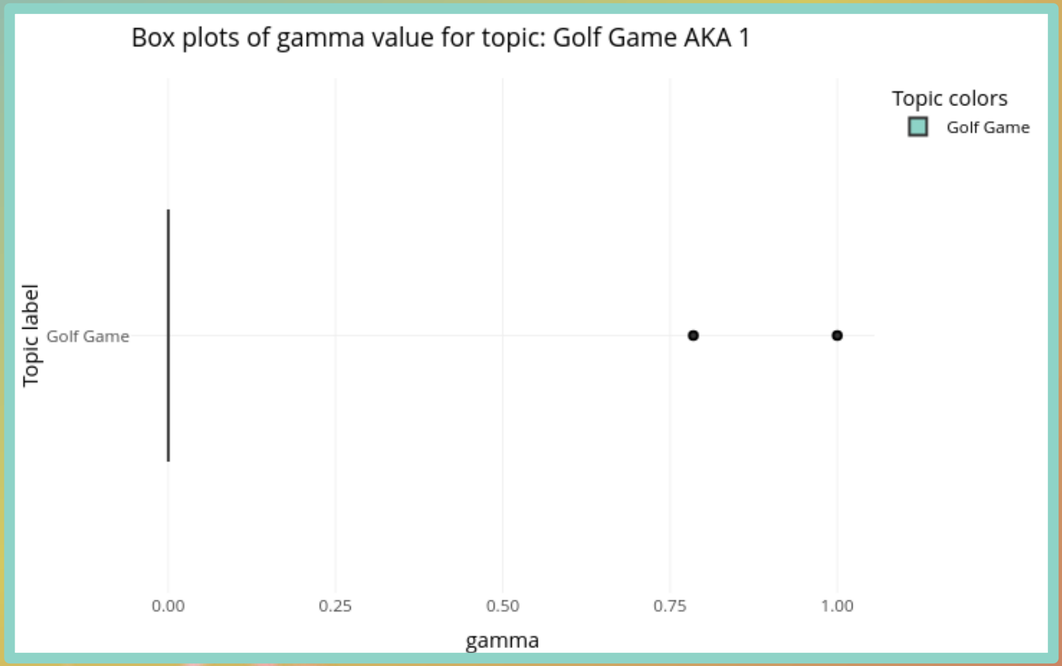


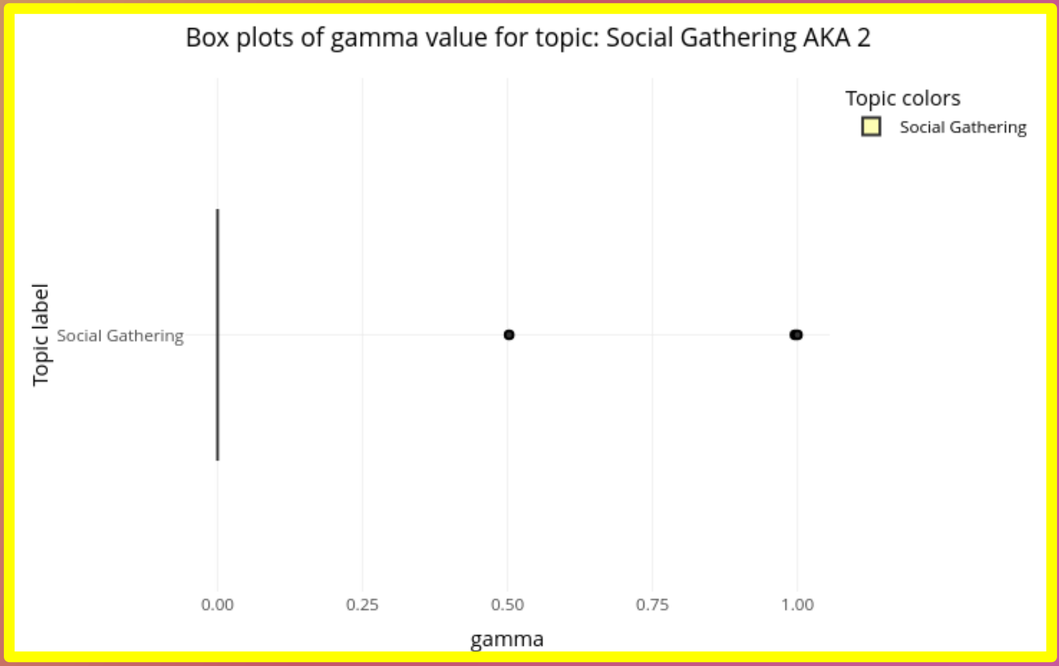
What are the true topics found in the data?



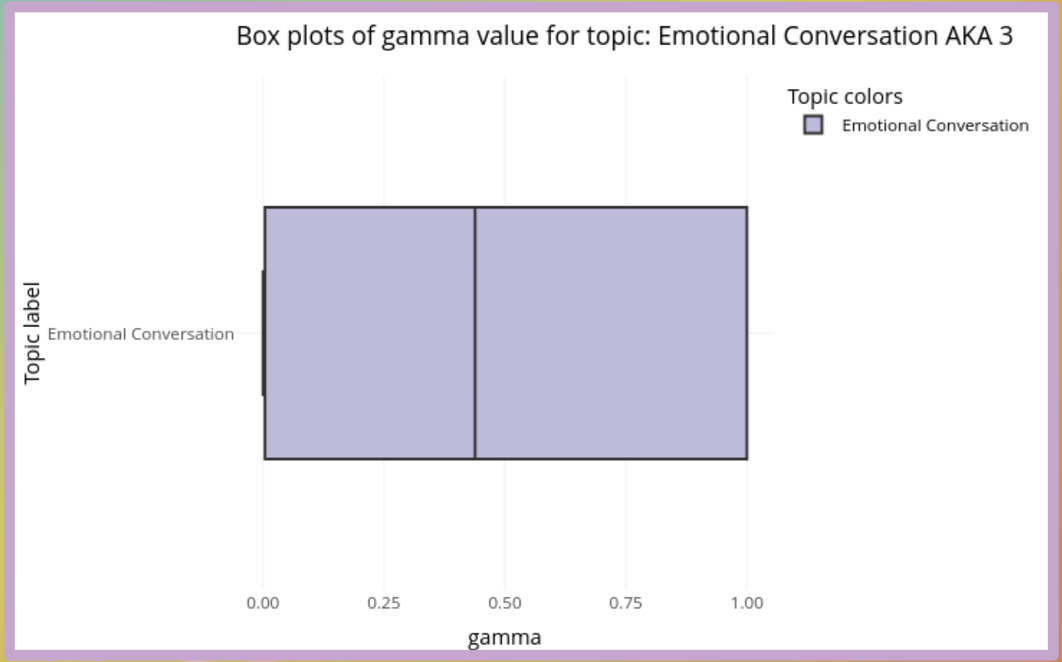


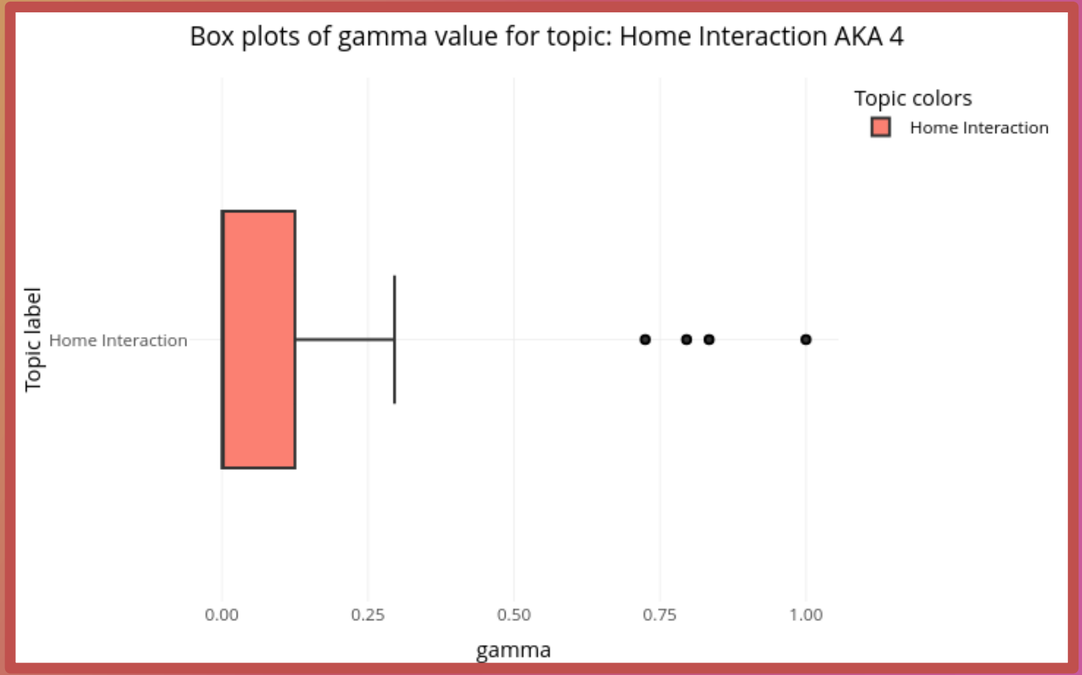
How are documents associated with these topics (labeled by ChatGPT)?



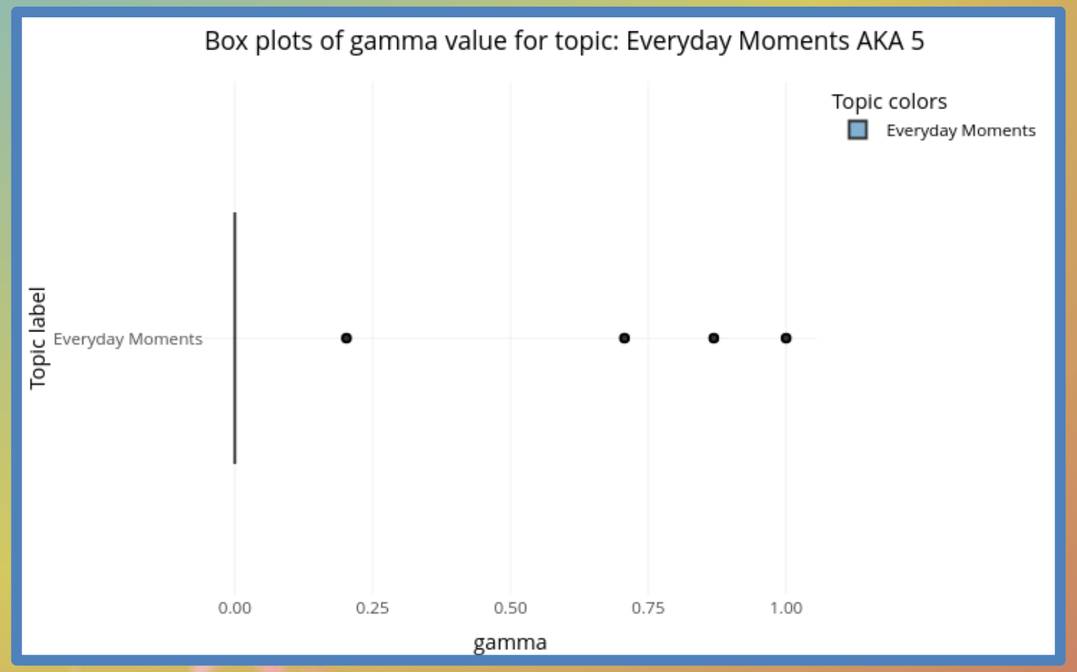


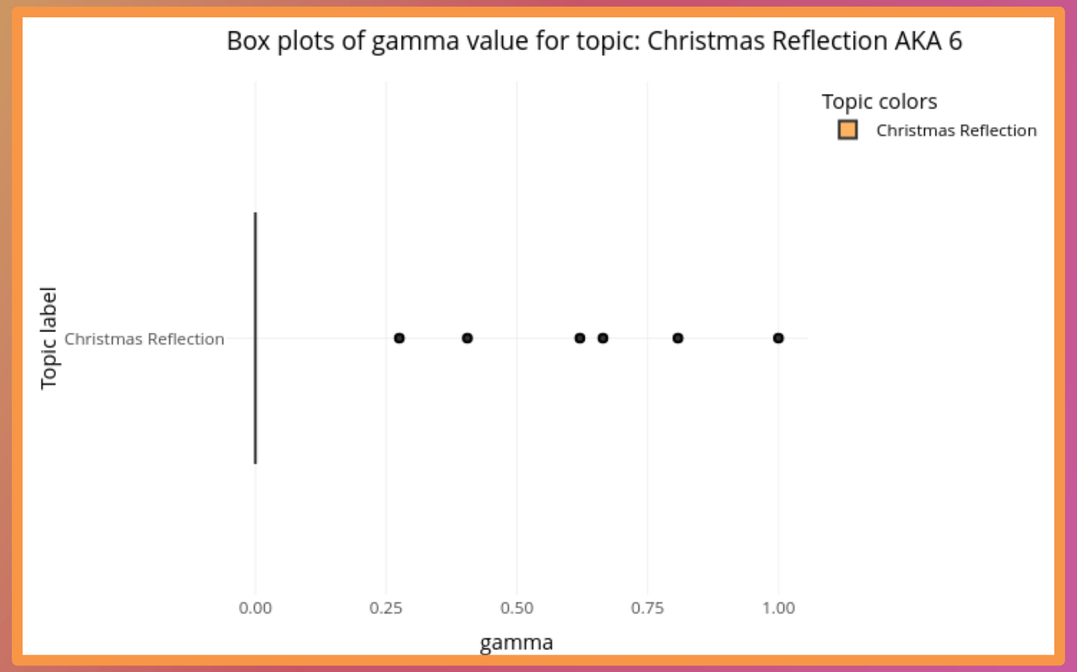
How are documents associated with these topics (labeled by ChatGPT)?



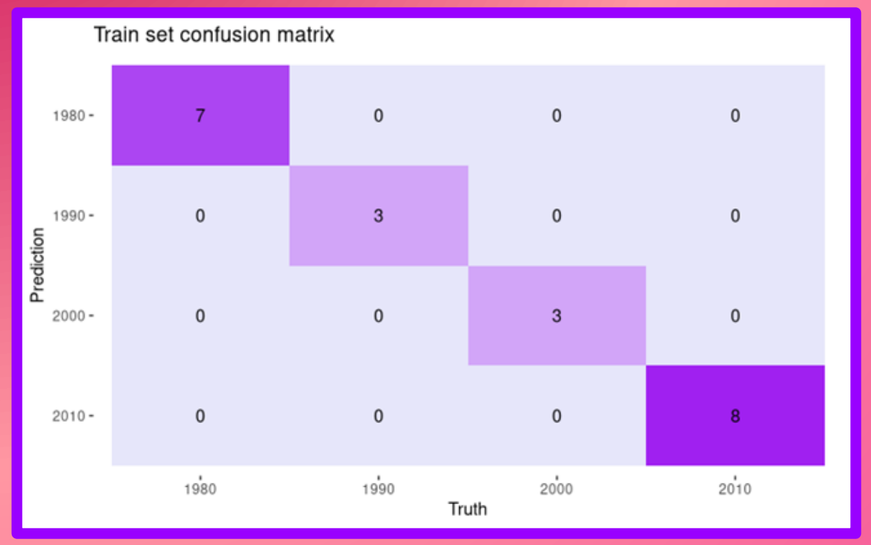


How are documents associated with these topics (labeled by ChatGPT)?

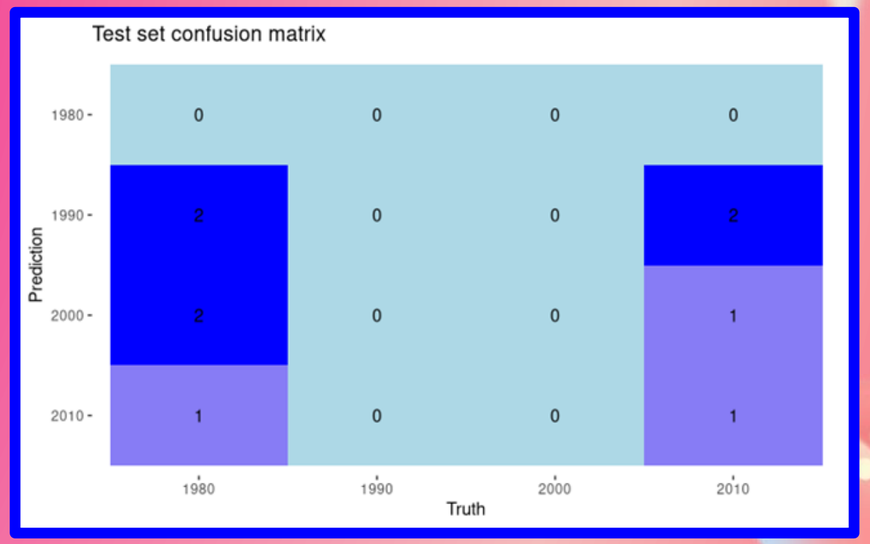




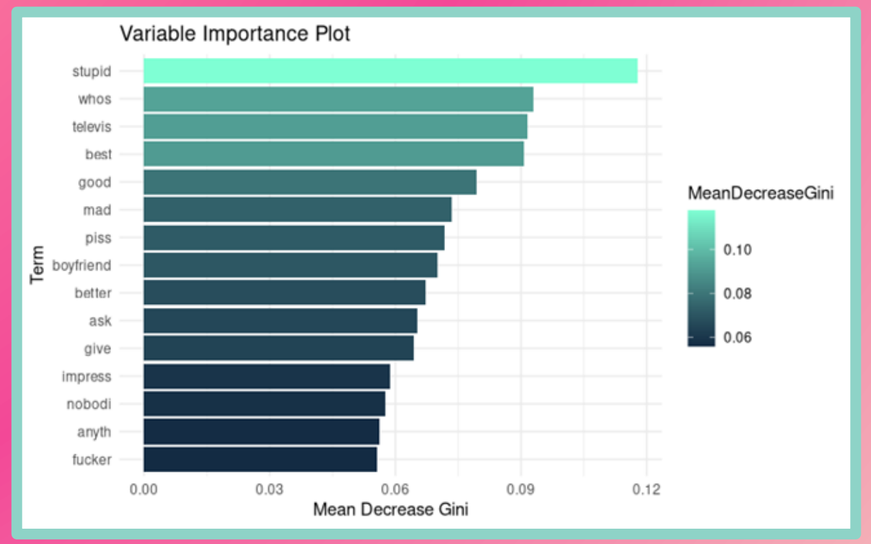
Can we predict the decade of a rom-com from its text?



Accuracy: 100%

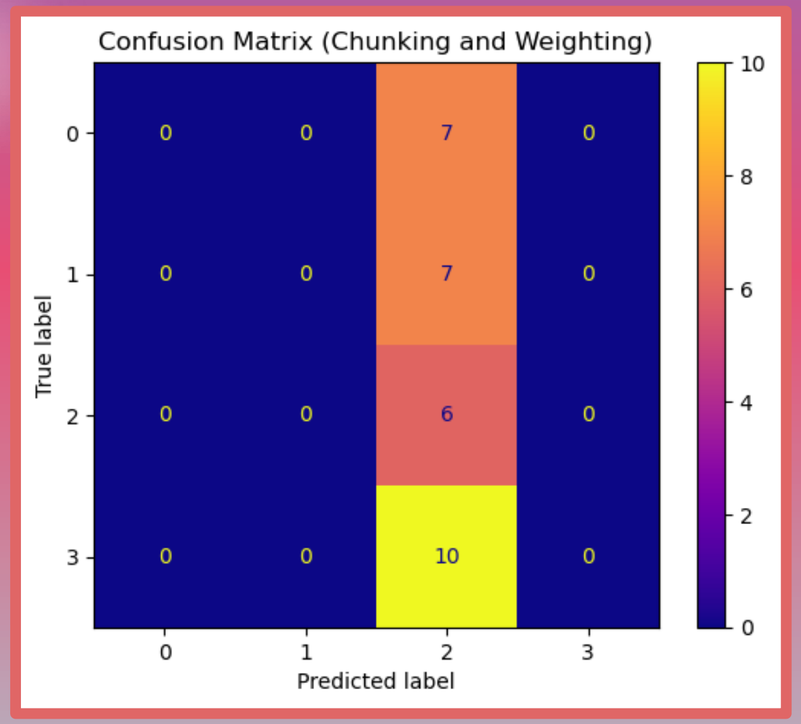


Accuracy: 11.11%

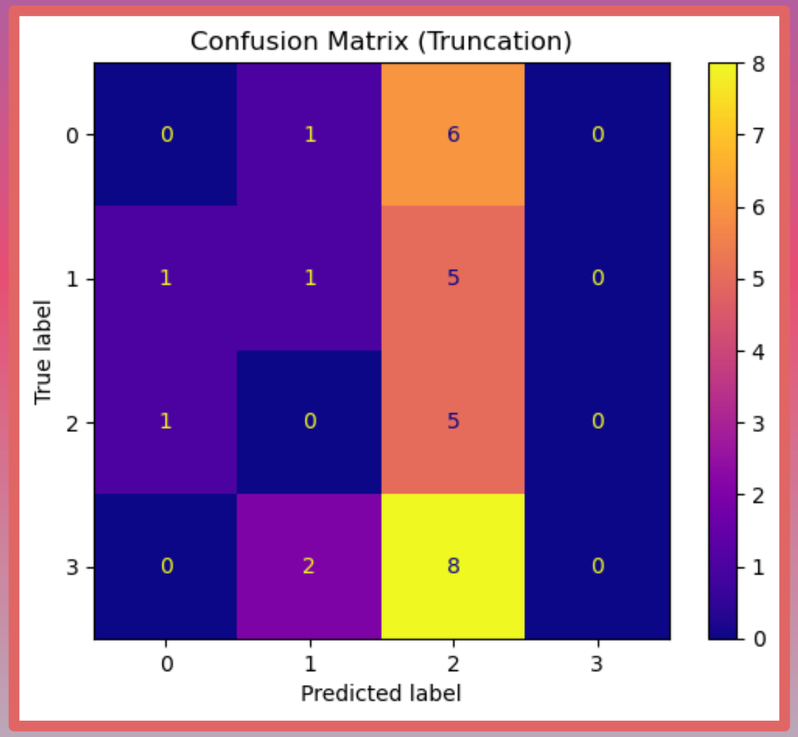


Can we predict the decade of a rom-com from its text? (LLM)

Using a fine tuned DistilBERT model trained on a custom dataset of 2,859 movie scripts



Accuracy: 20%



Accuracy: 20%

Our Questions Answered

- Are recent romantic comedies more correlated with each other than those of the past?
- Have romantic comedy scripts become less complex (in terms of vocabulary)?
- Can the era of a romantic comedy be easily predicted based on aspects of the script?