

# *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, DA*  
*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%

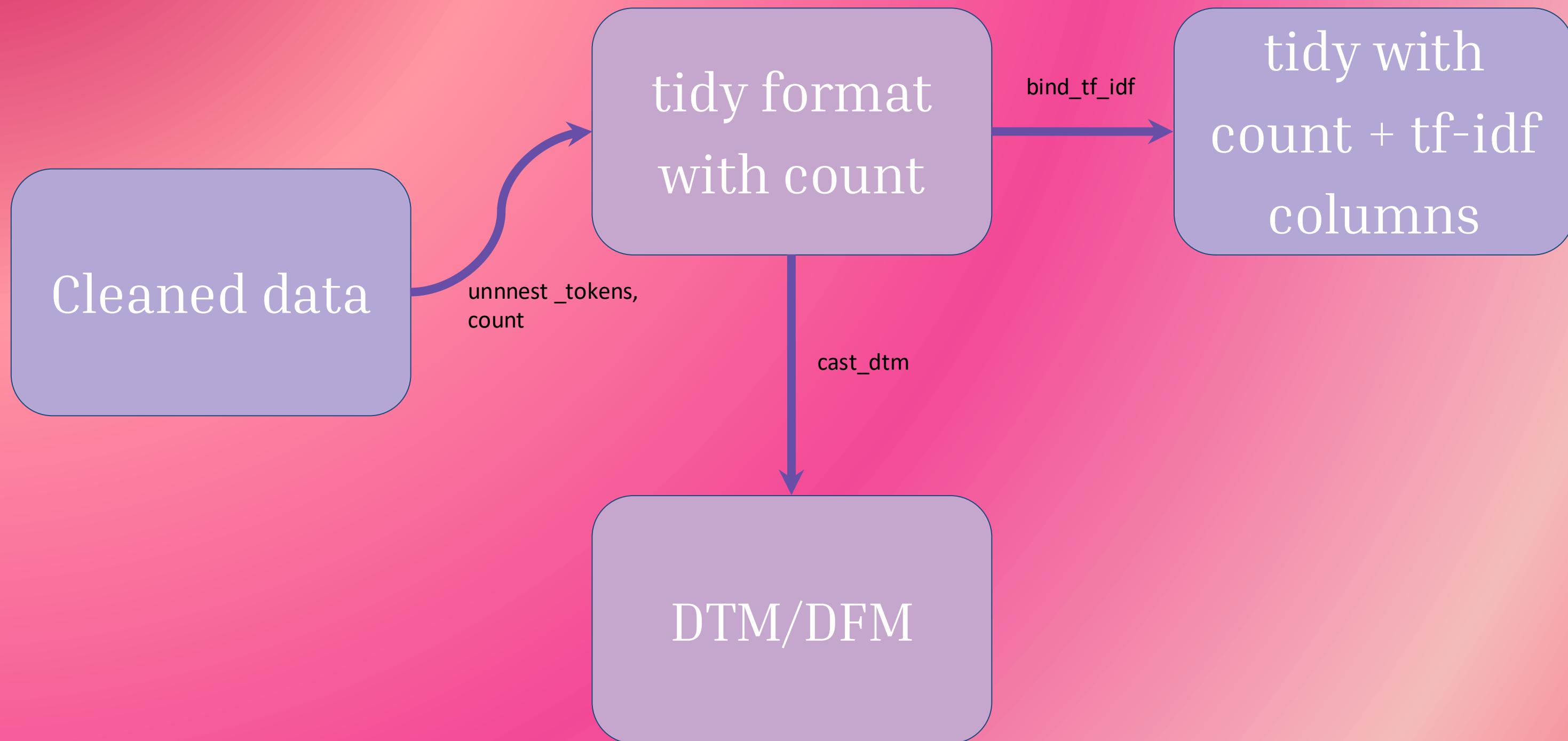
Sparsity

14

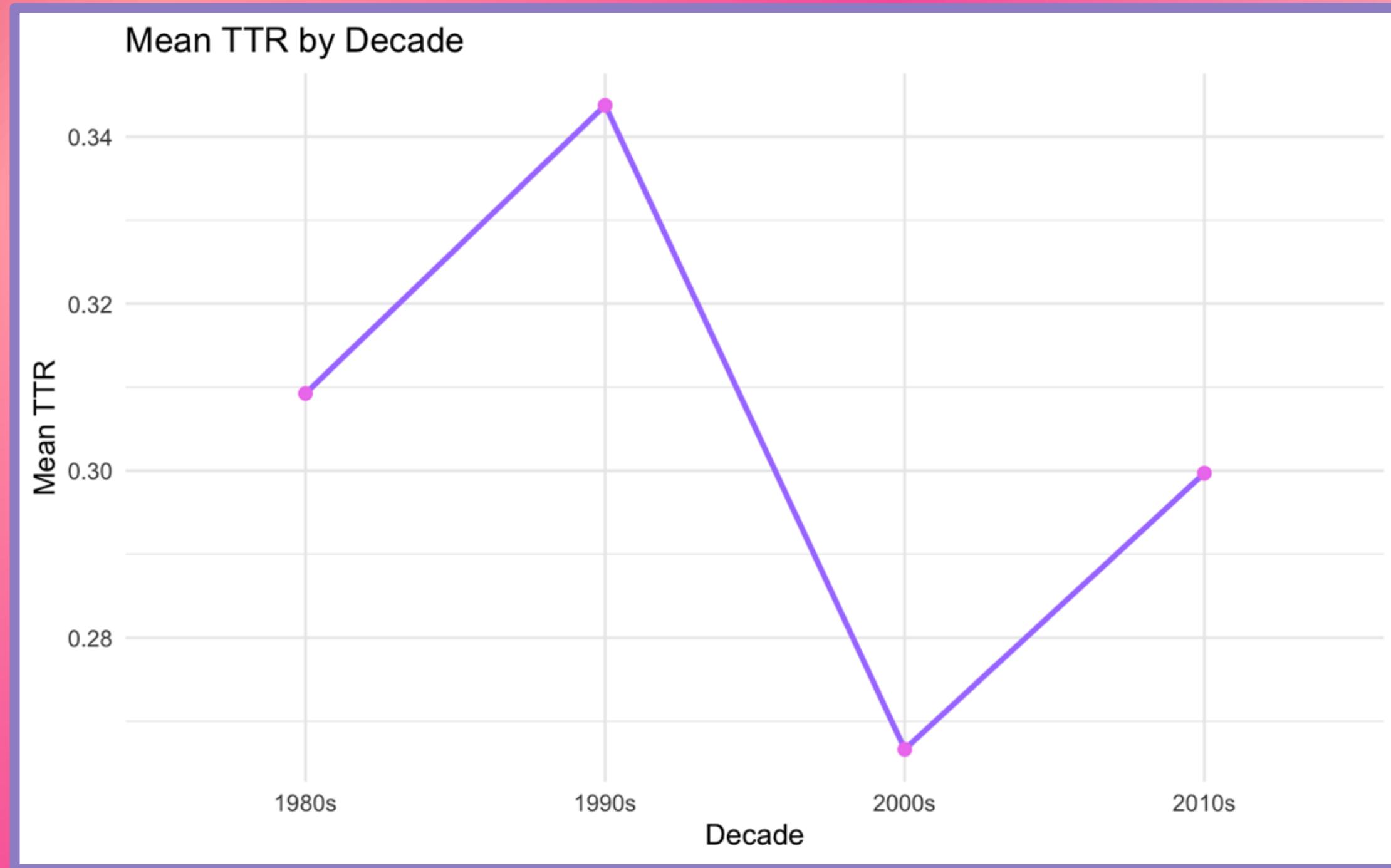
Maximal term length



# Data handling

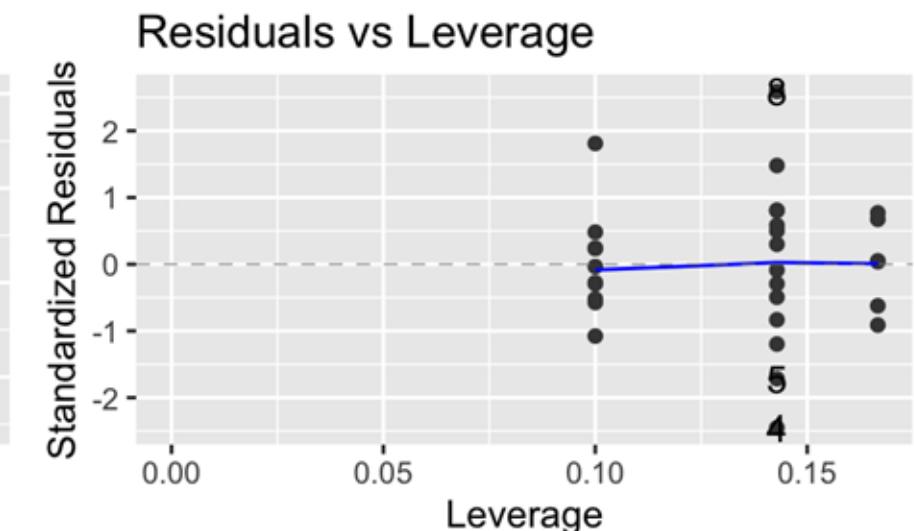
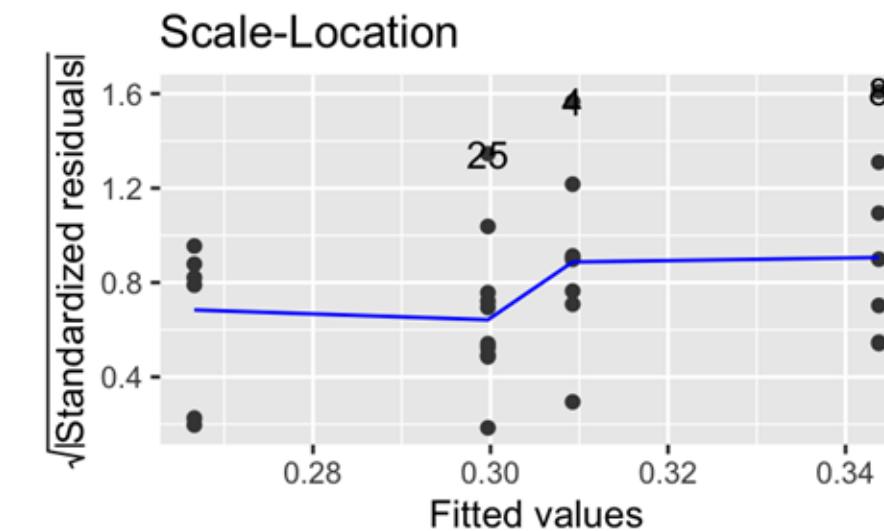
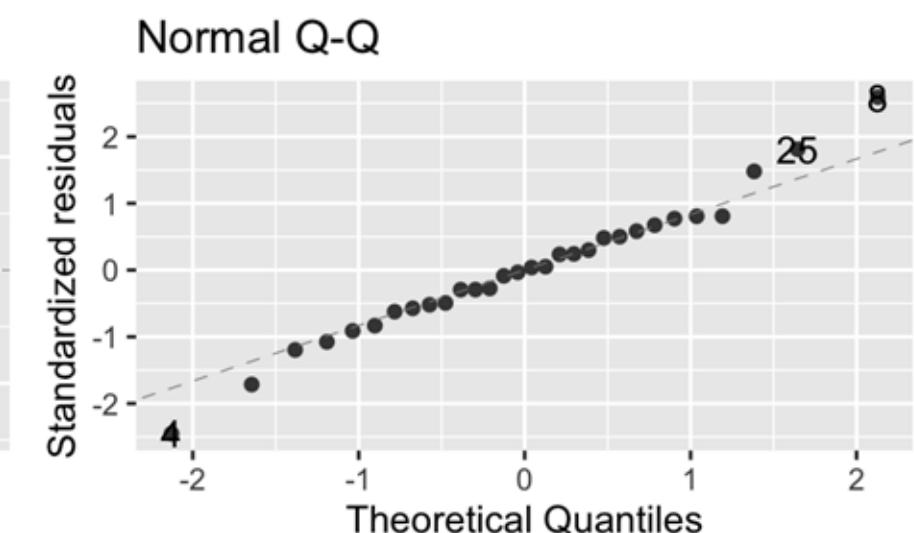
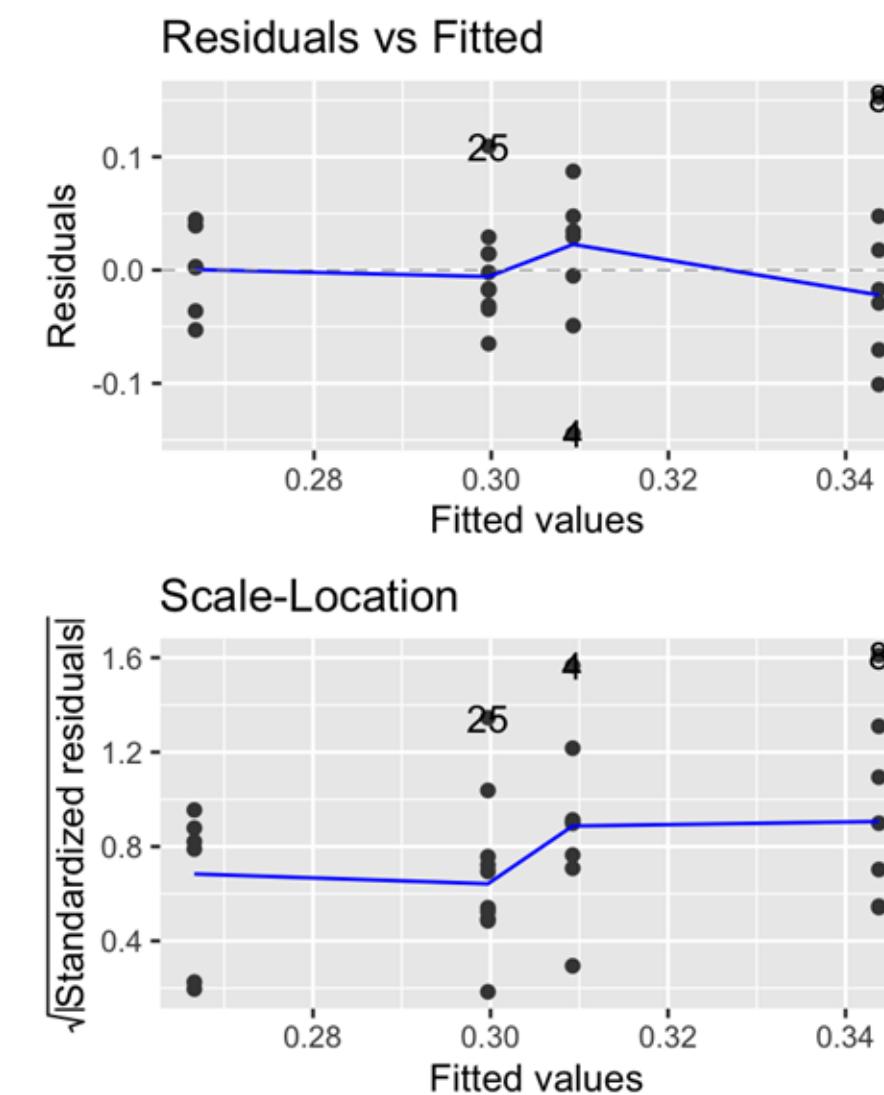


# How lexically diverse is each decade?



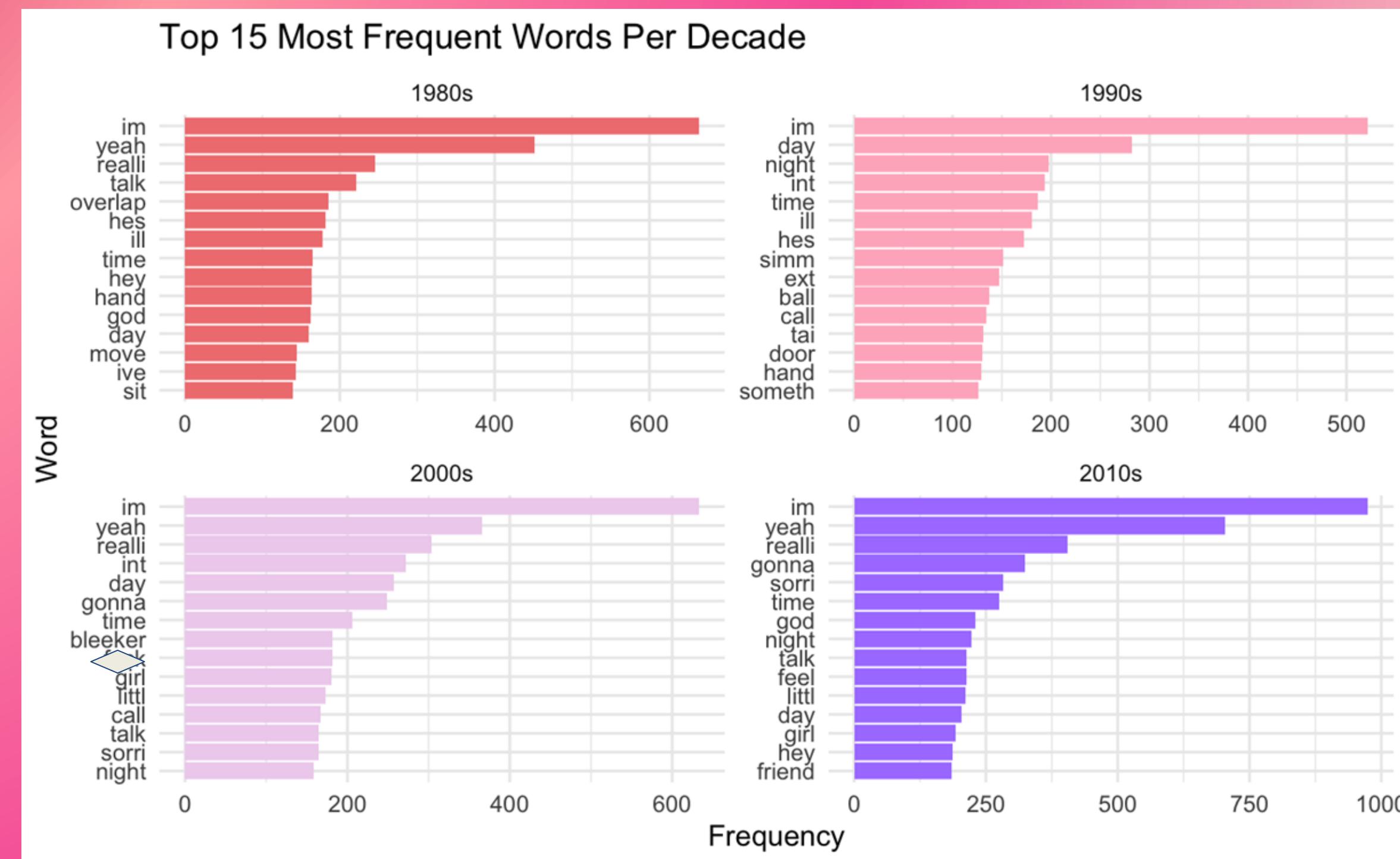
# Are these differences significant?

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
decade2010s	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

A word cloud for the 1980s featuring various slang terms. The words are colored in a gradient from blue to green. The most prominent words are "yeah" (yellow) and "im" (pink). Other visible words include "hey", "time", "overlap", "night", "hold", "stop", "sit", "shes", "talk", "uhlittl", "ive", "happened", "god", "move", "tbl", "illhes", "call", "someth", "feel", "hand", "didnt", "realli", "gonna", "start".

1990s

A word cloud for the 1990s featuring various slang terms. The words are colored in a gradient from blue to green. The most prominent words are "day" (orange), "time" (green), "ext", "lifedidnt", "someth", "talkshotint", "taipeopl", "ive", "head", "ill", "reallitri", "ballhes", "simm", "night" (purple), "feel", "hand", "littl", "watch", "start", "sit", "im" (pink), and "door".

2000s

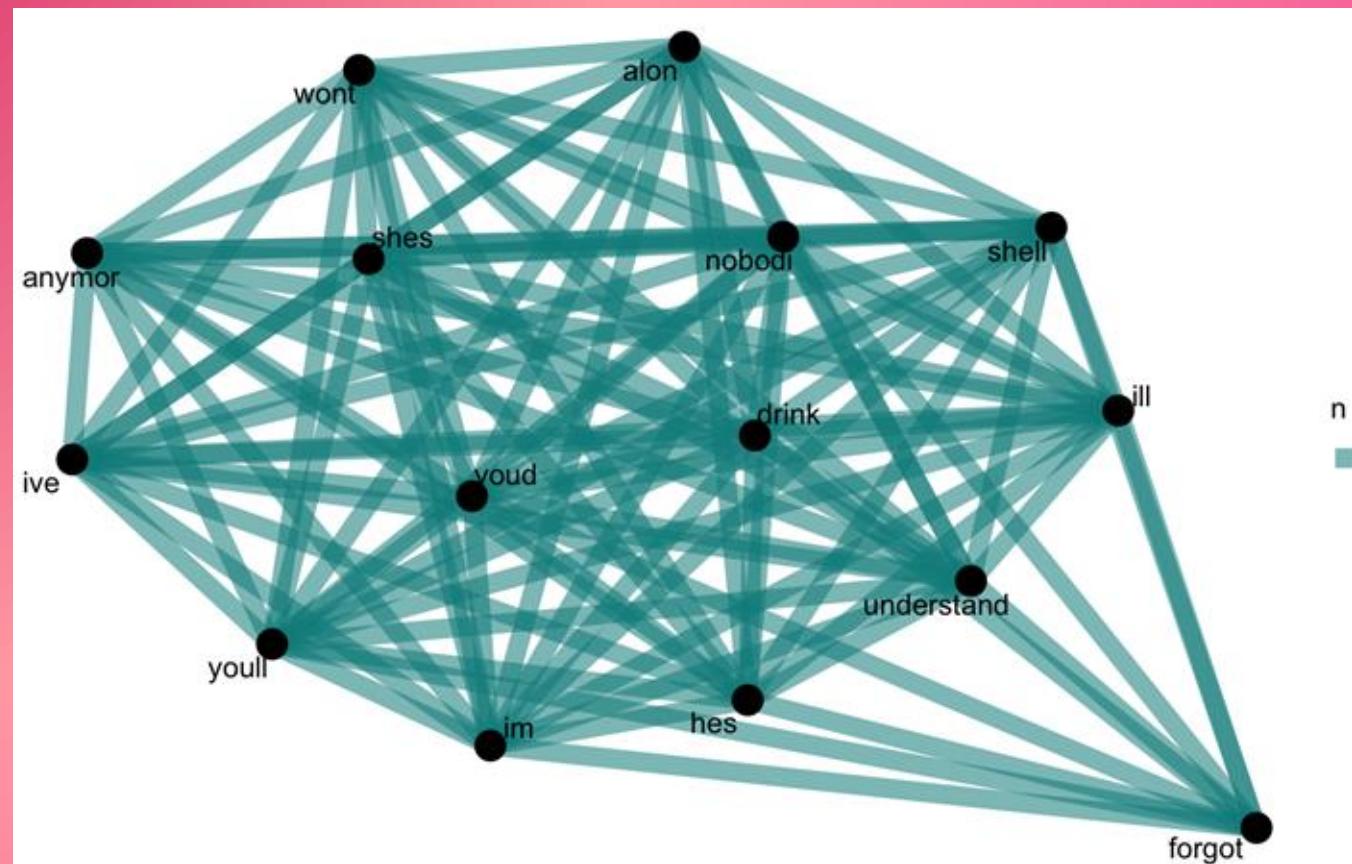
A word cloud for the 2000s featuring various slang terms. The words are colored in a gradient from blue to purple. The most prominent words are "im" (pink), "bleeker", "pm", "hes", "babi", "gonna", "girl", "car", "shes", "time", "didnt", "sex", "shes", "littl", "final", "night", "someth", "int", "fuk", "mayb", "real", "god", "talk", "sorri", "product", "hey", "call". A black diamond symbol is located near the bottom center of the word cloud.

2010s

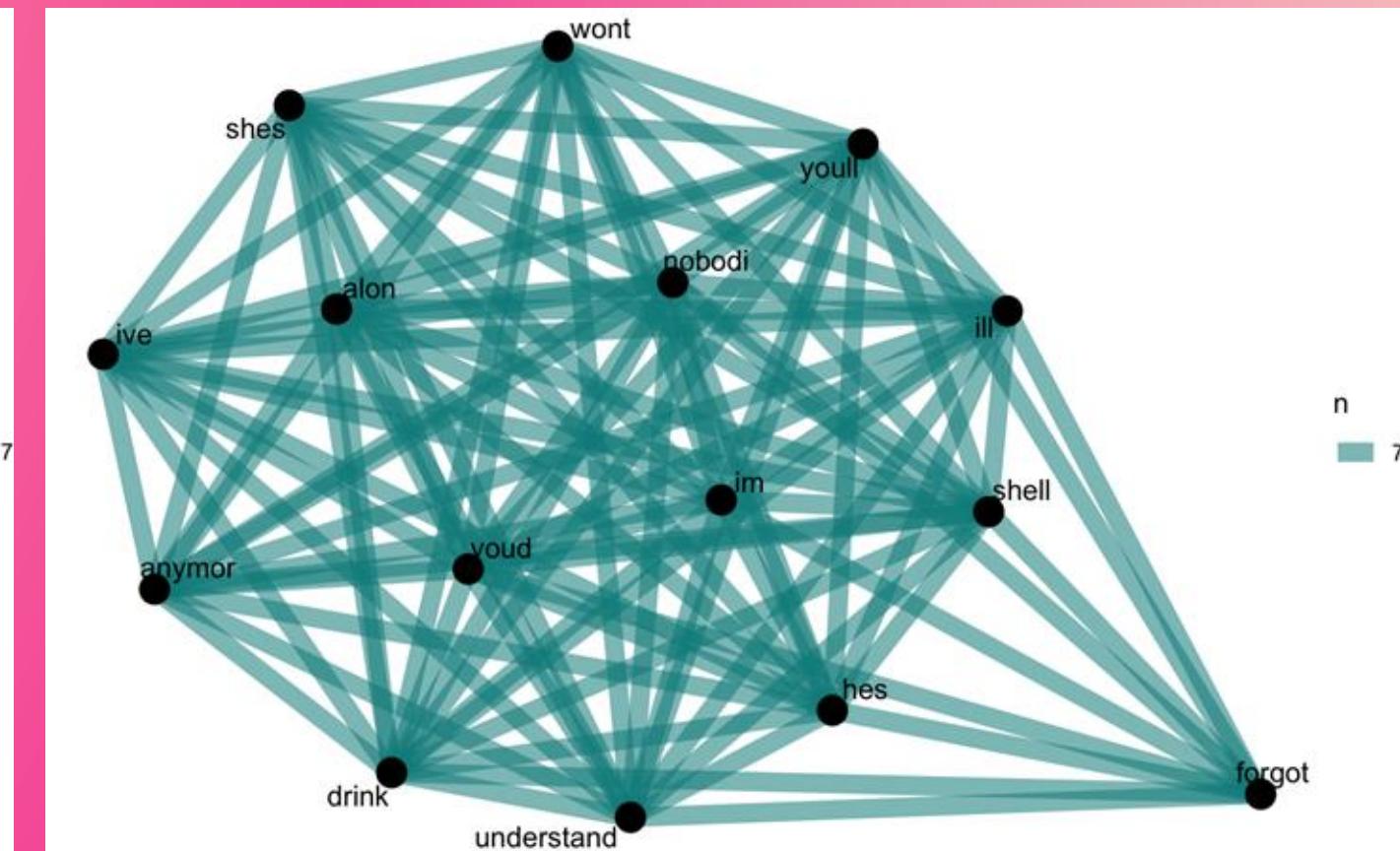
A word cloud for the 2010s featuring various slang terms. The words are colored in a gradient from blue to green. The most prominent words are "yeah" (yellow), "alway", "littl", "didnt", "shes", "uh", "night", "friend", "time", "god", "girl", "im" (pink), "talk", "um", "ill", "int", "feel", "day", "sorri", "life", "hes", "hey", "ive", "tri", "real", "gonna", "someth", "peik", "nice".

# Word co-occurrences by decade

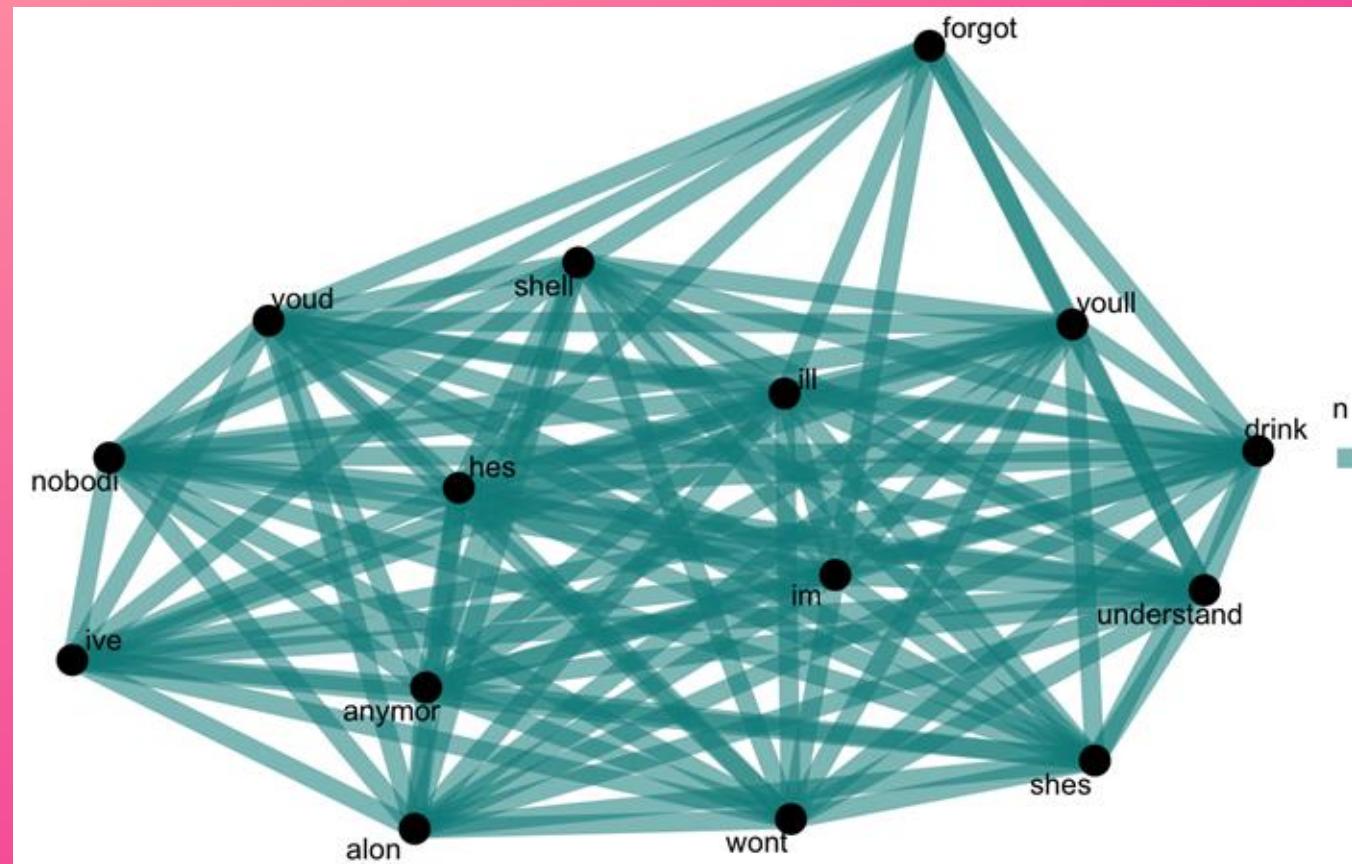
1980s



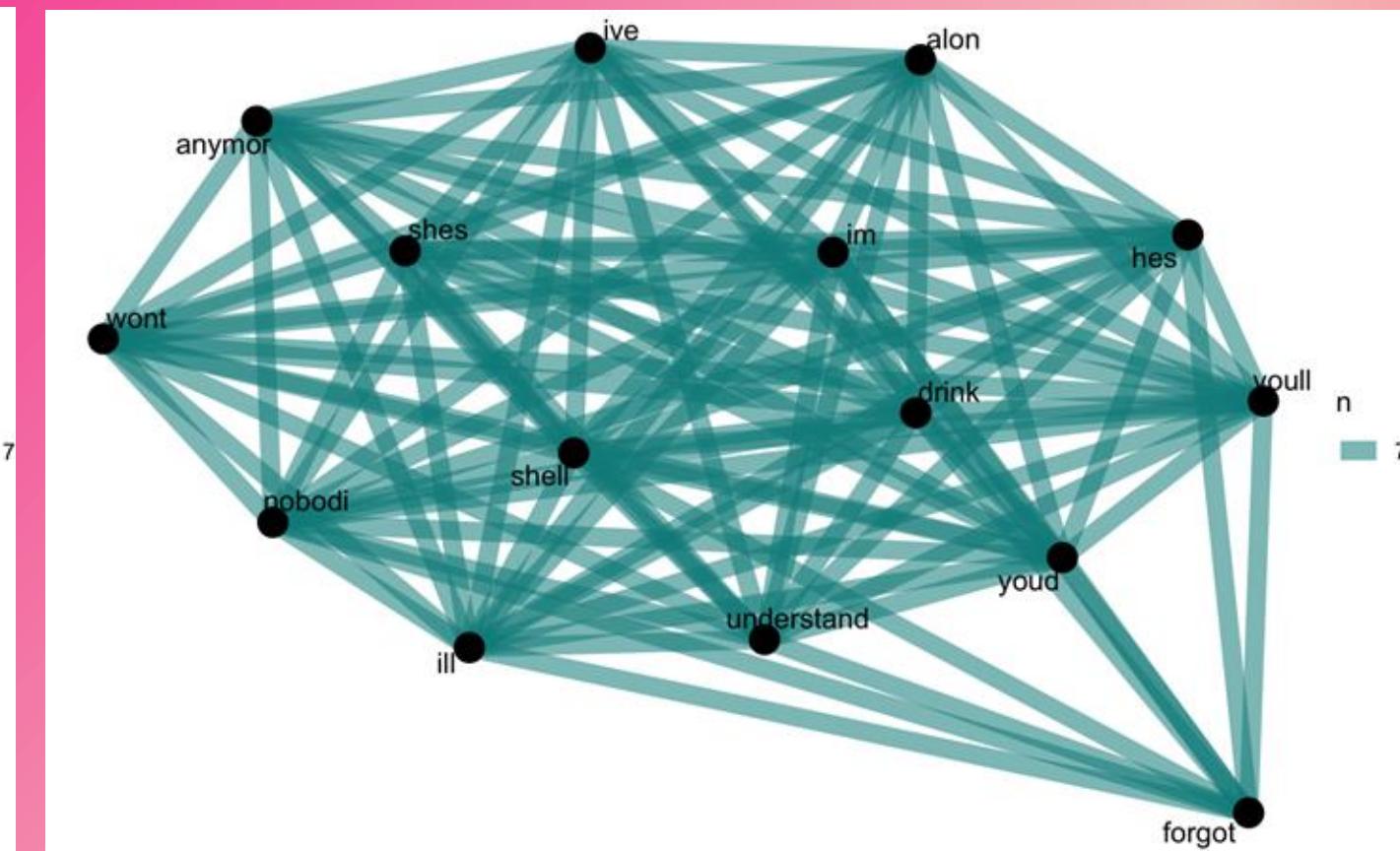
1990s



2000s

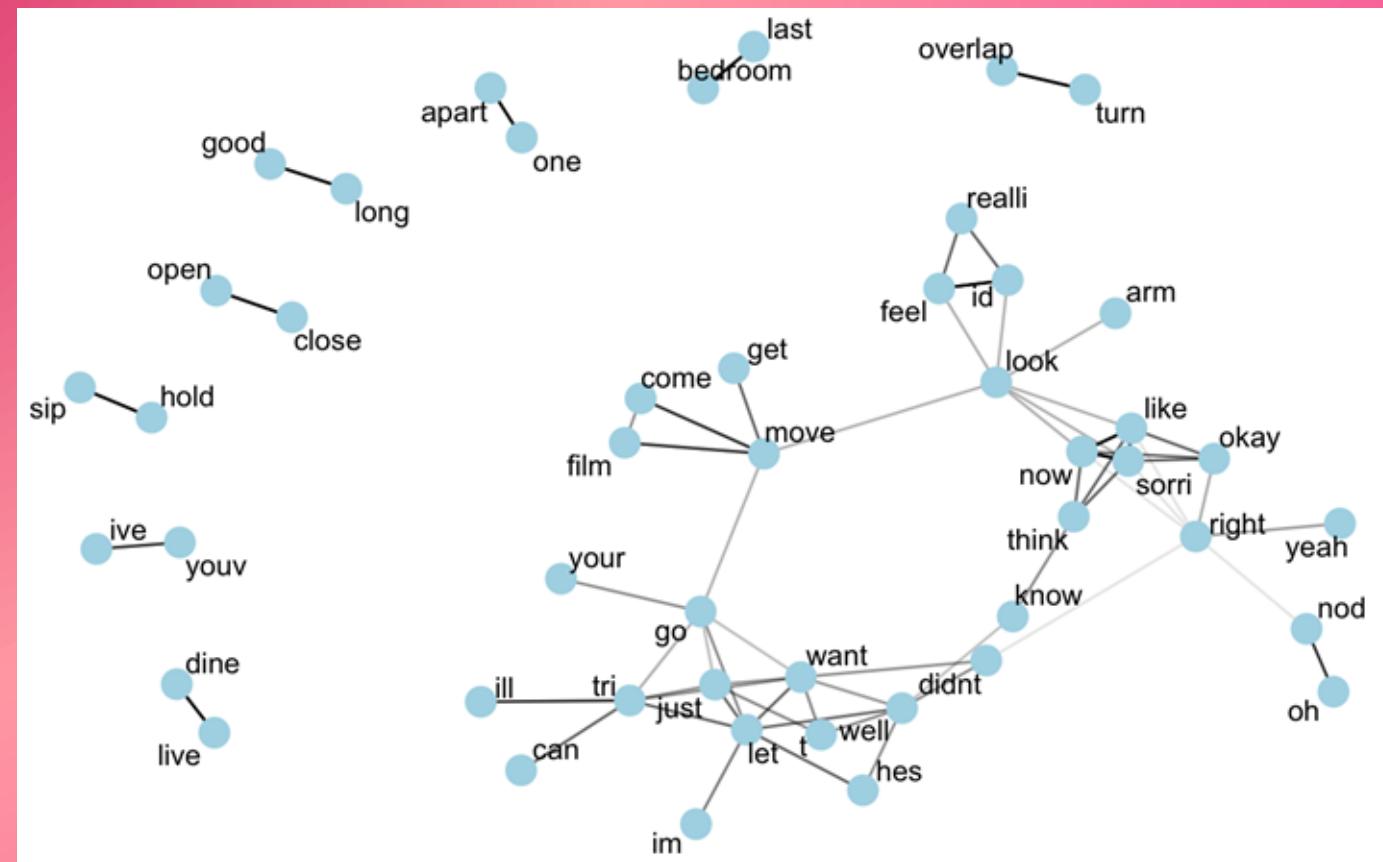


2010s

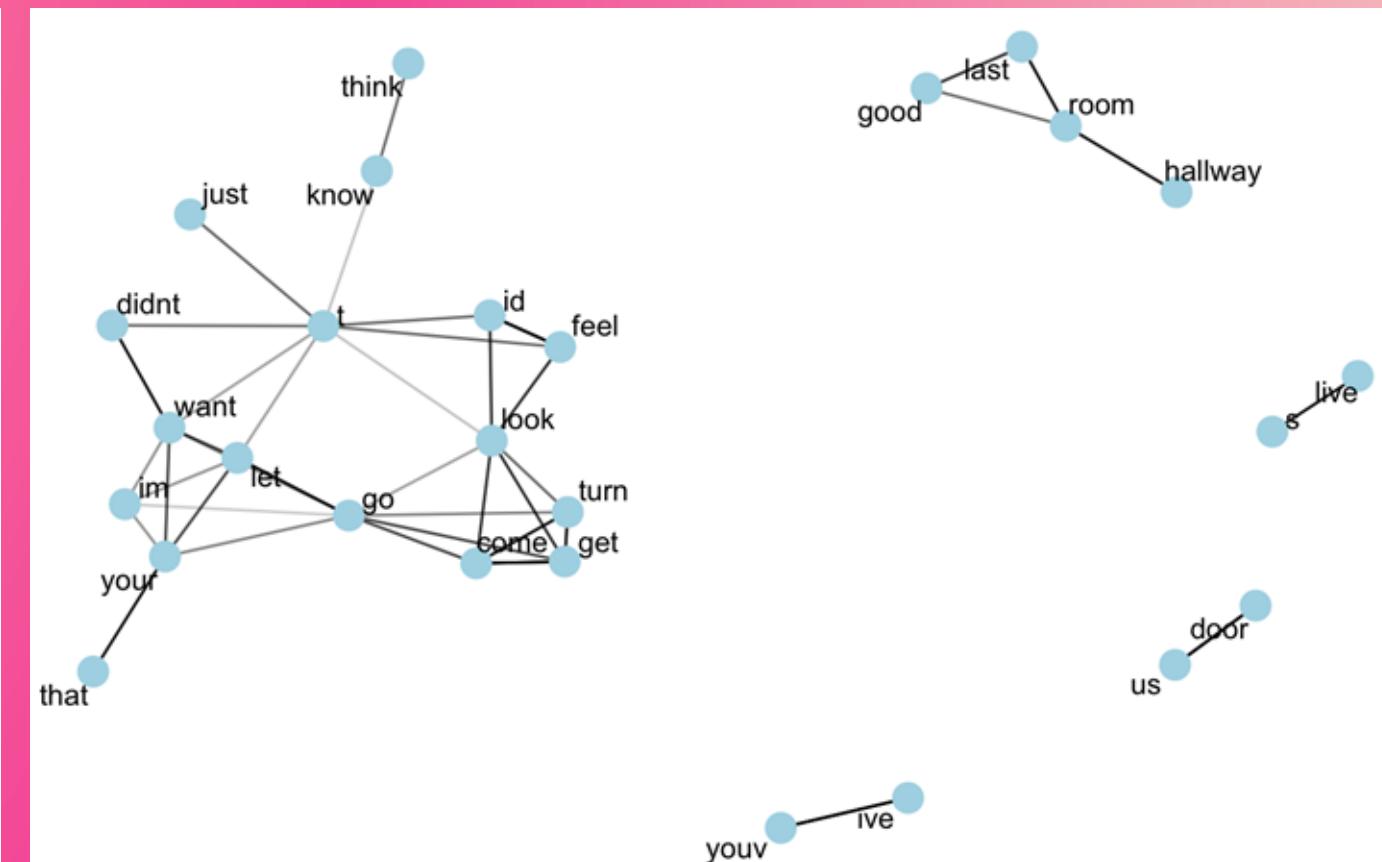


# Correlation analysis with bigrams

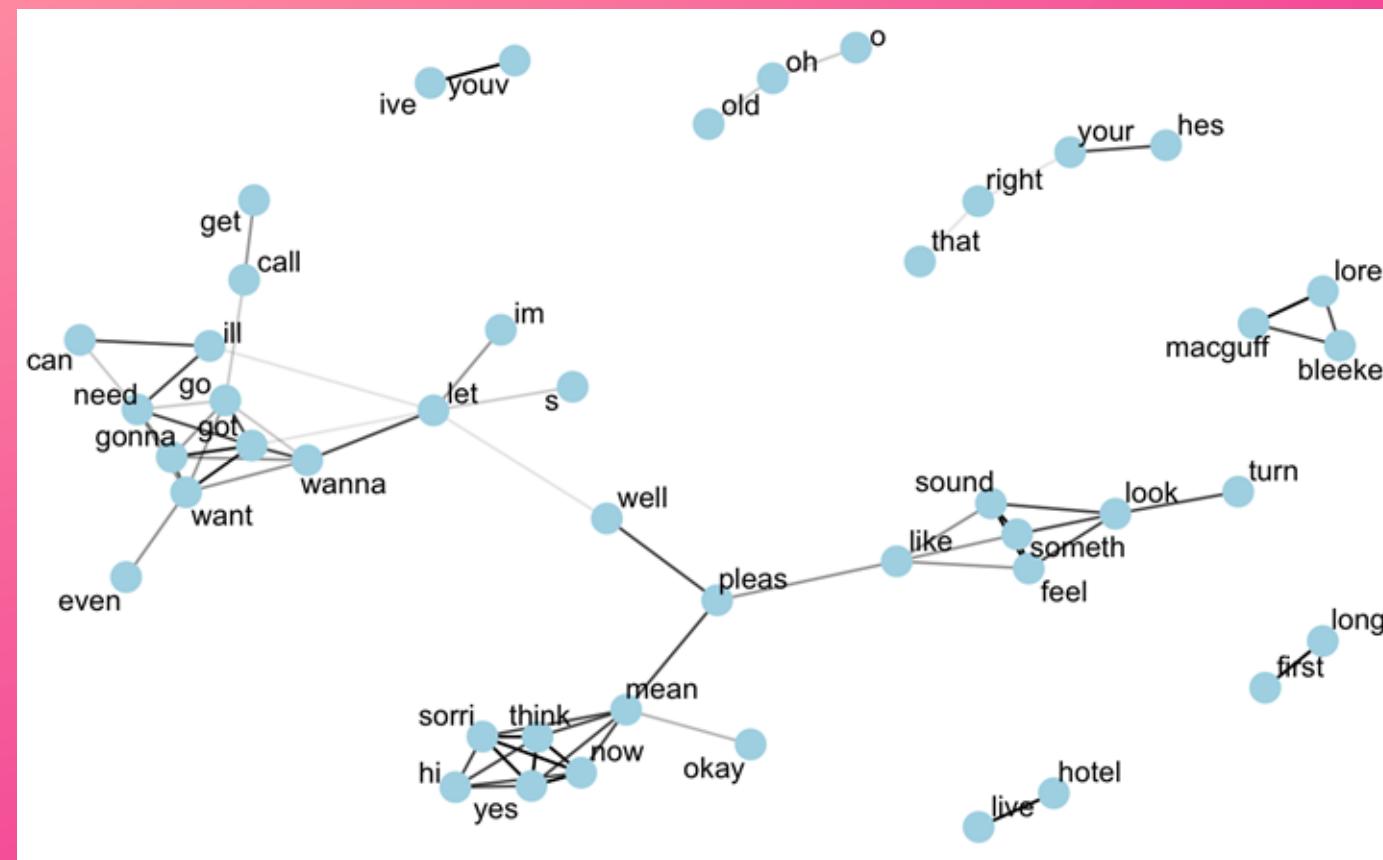
# 1980s



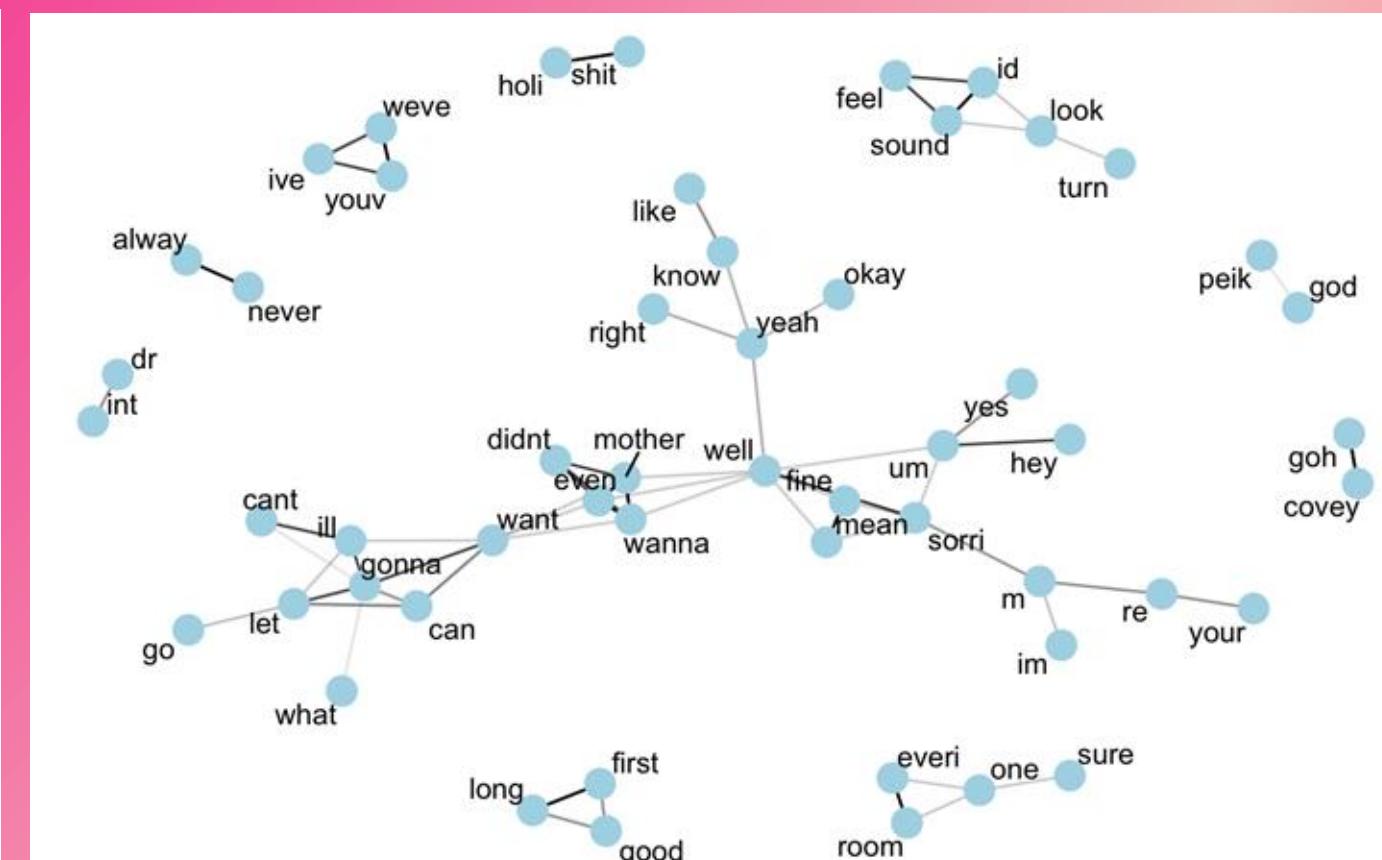
# 1990s



# 2000s

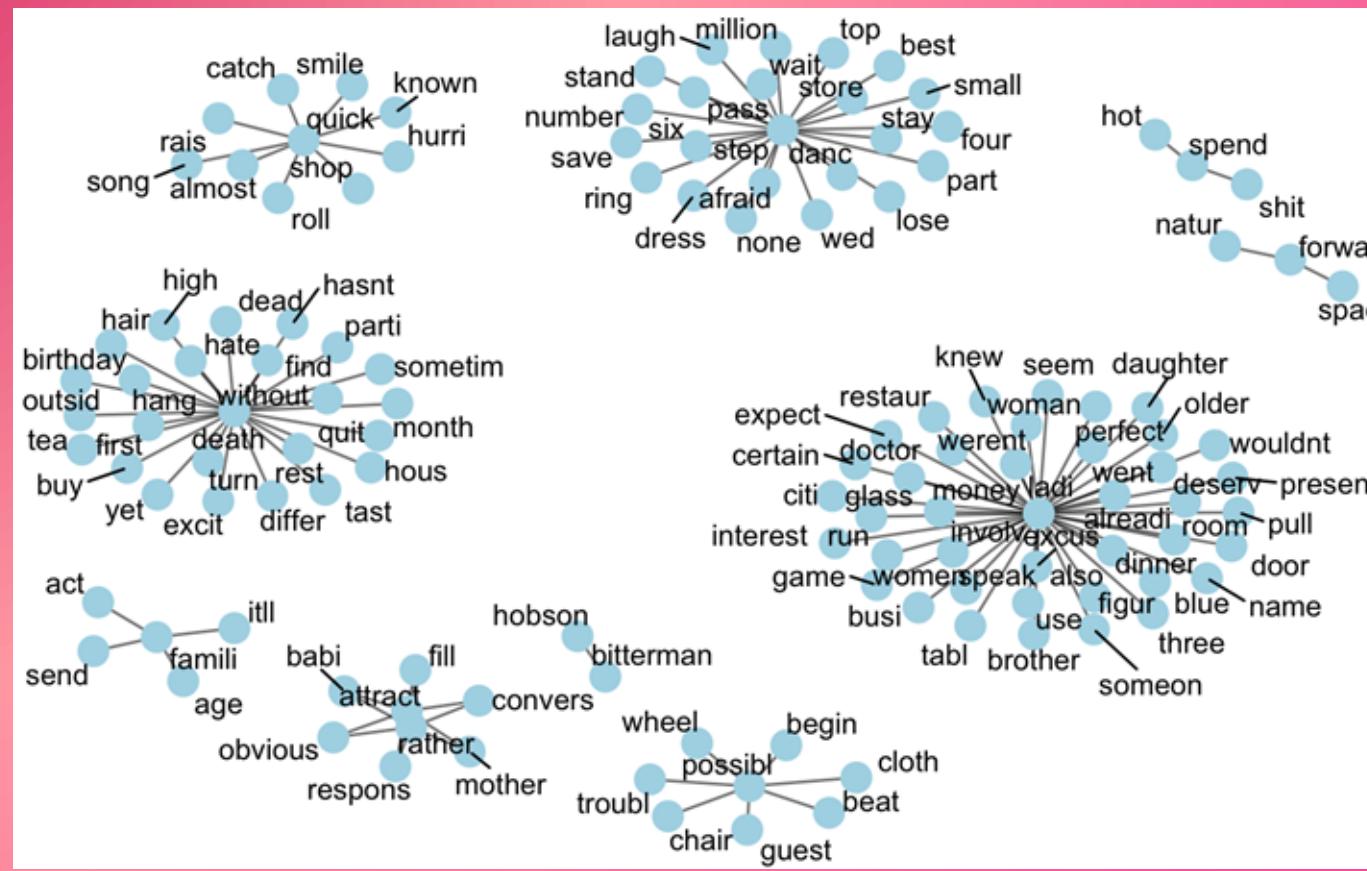


2010s

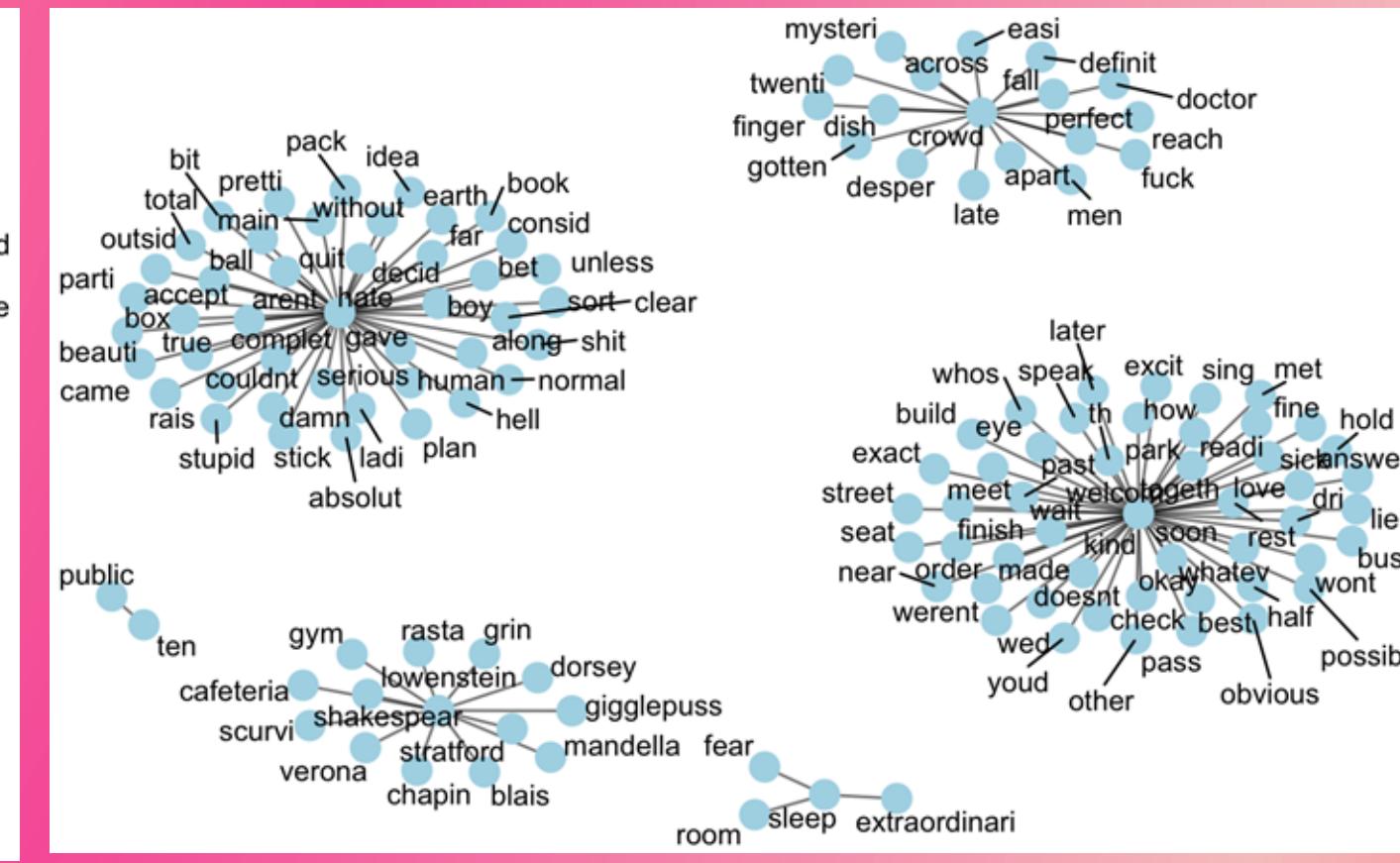


# Correlation analysis with trigrams

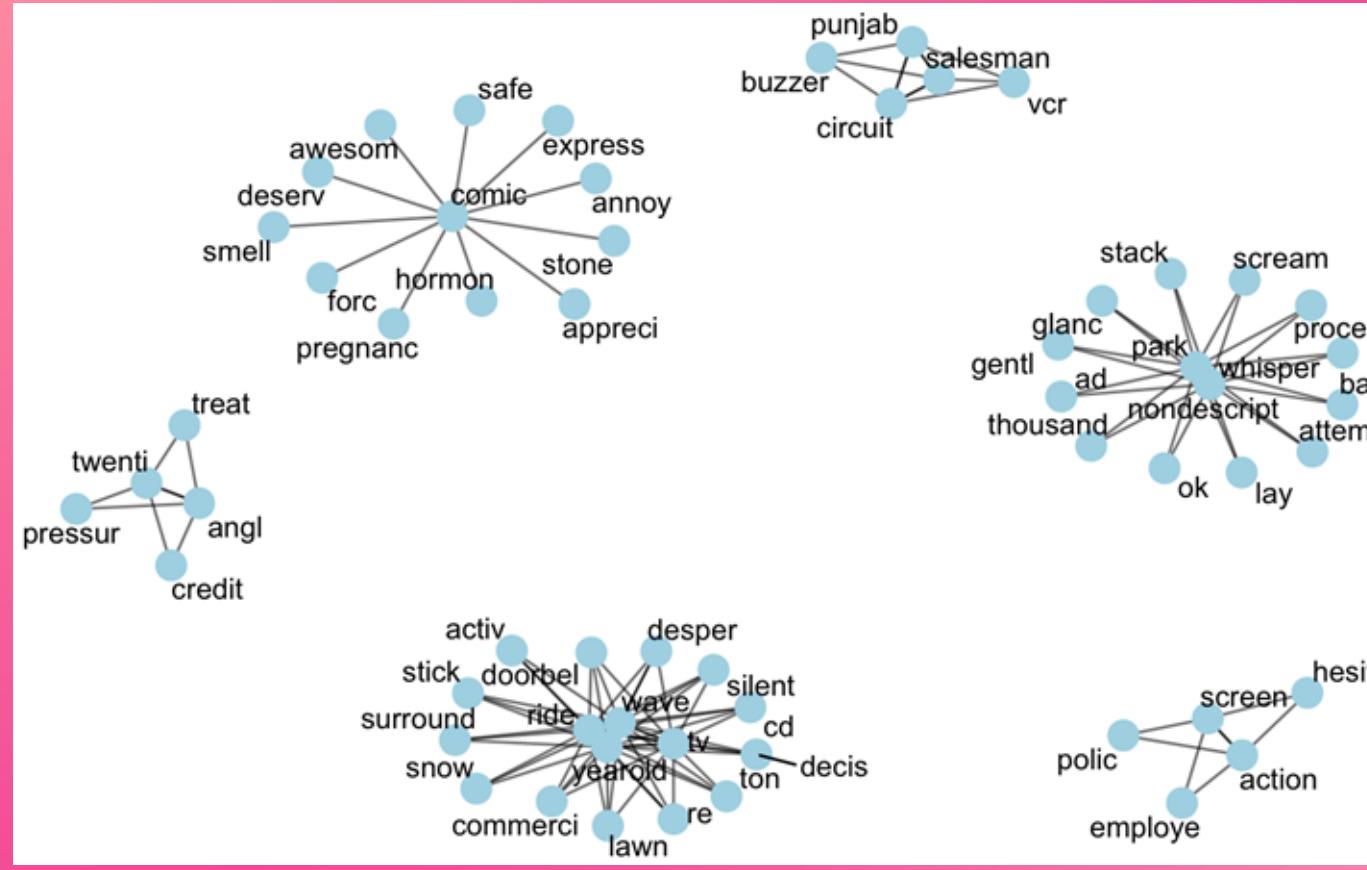
# 1980s



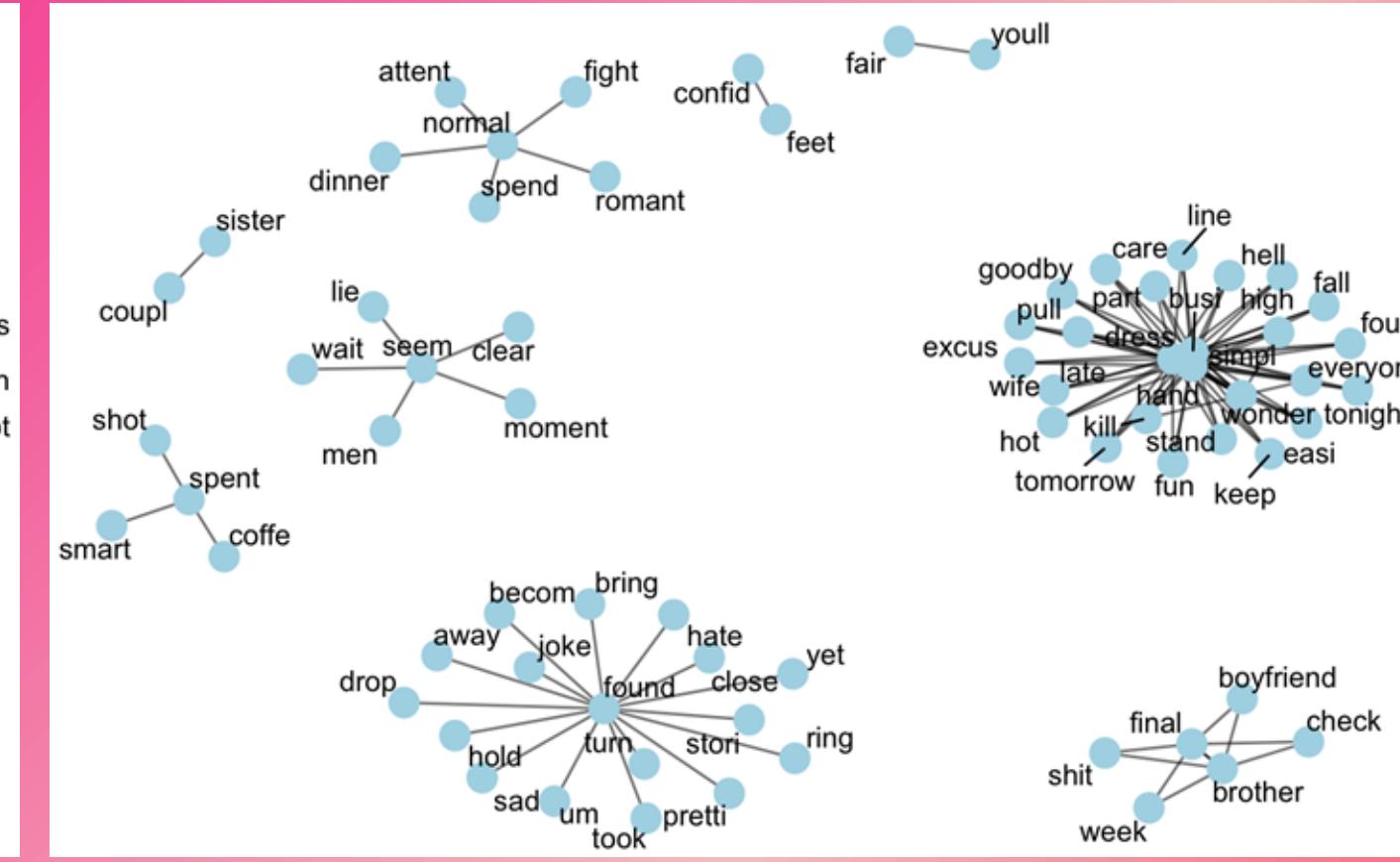
# 1990s

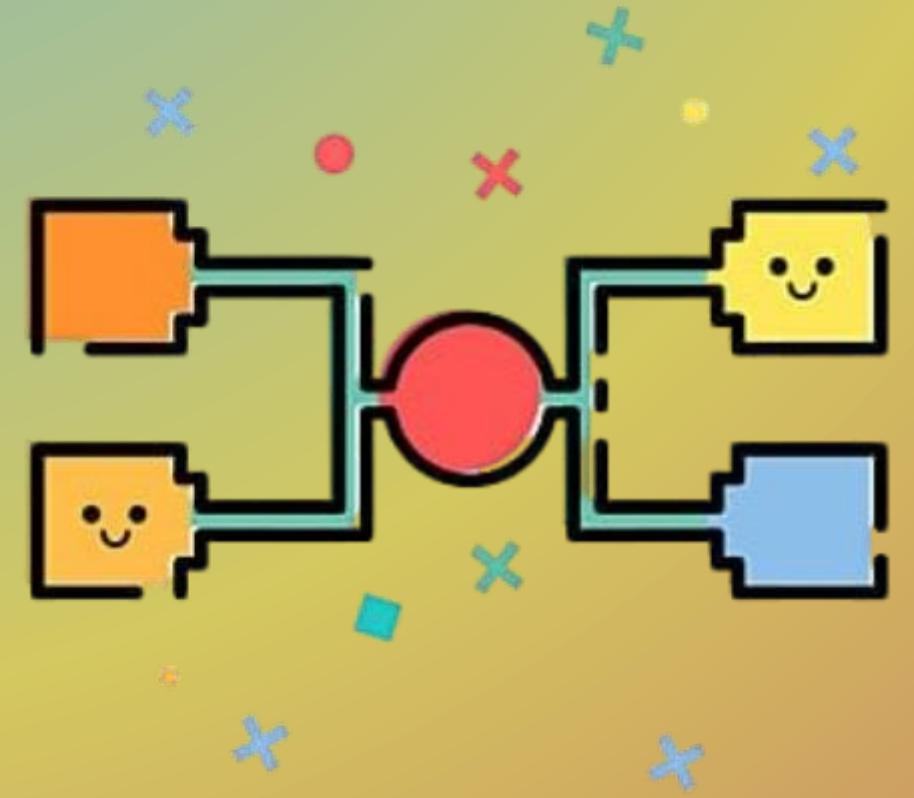


# 2000s



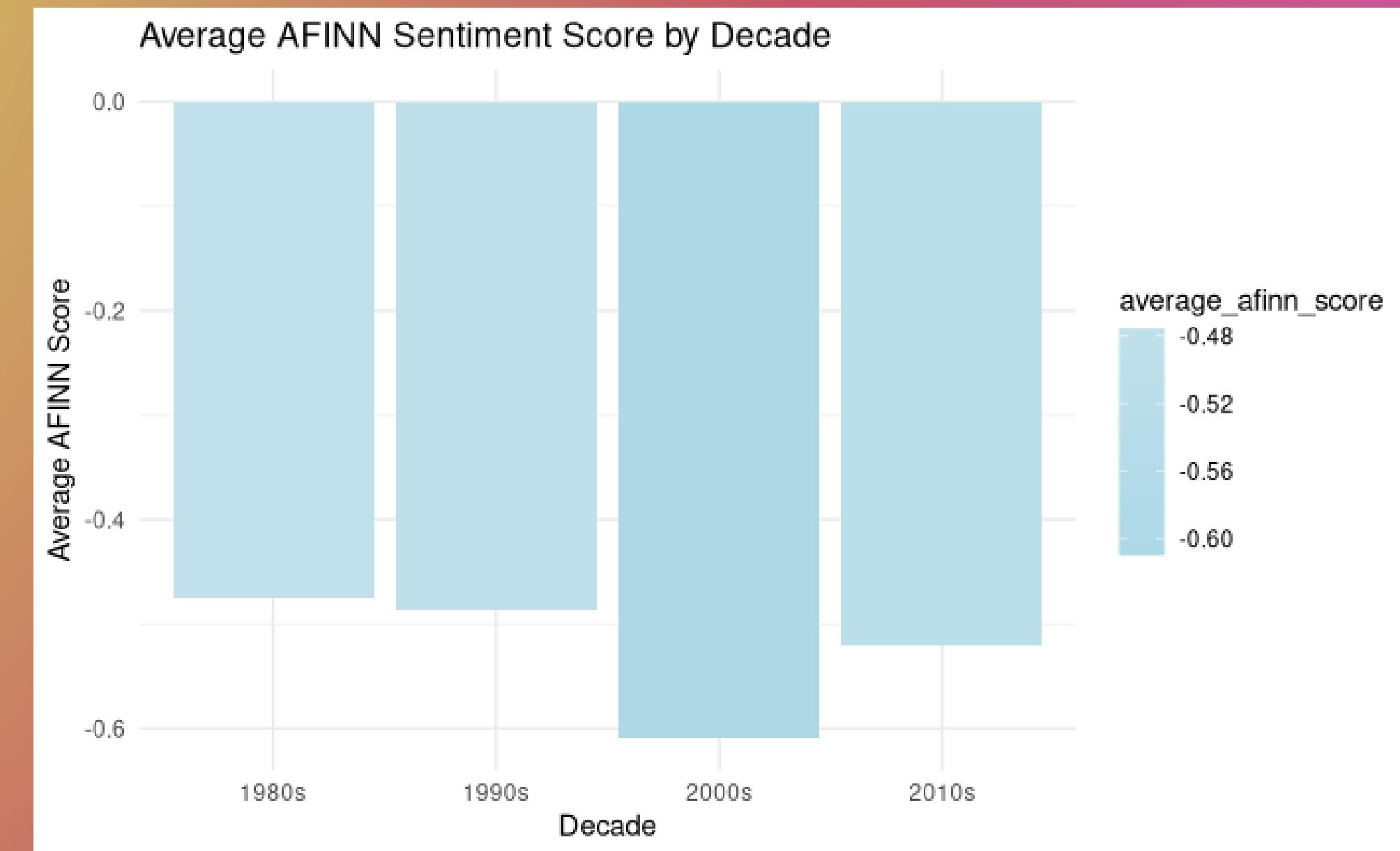
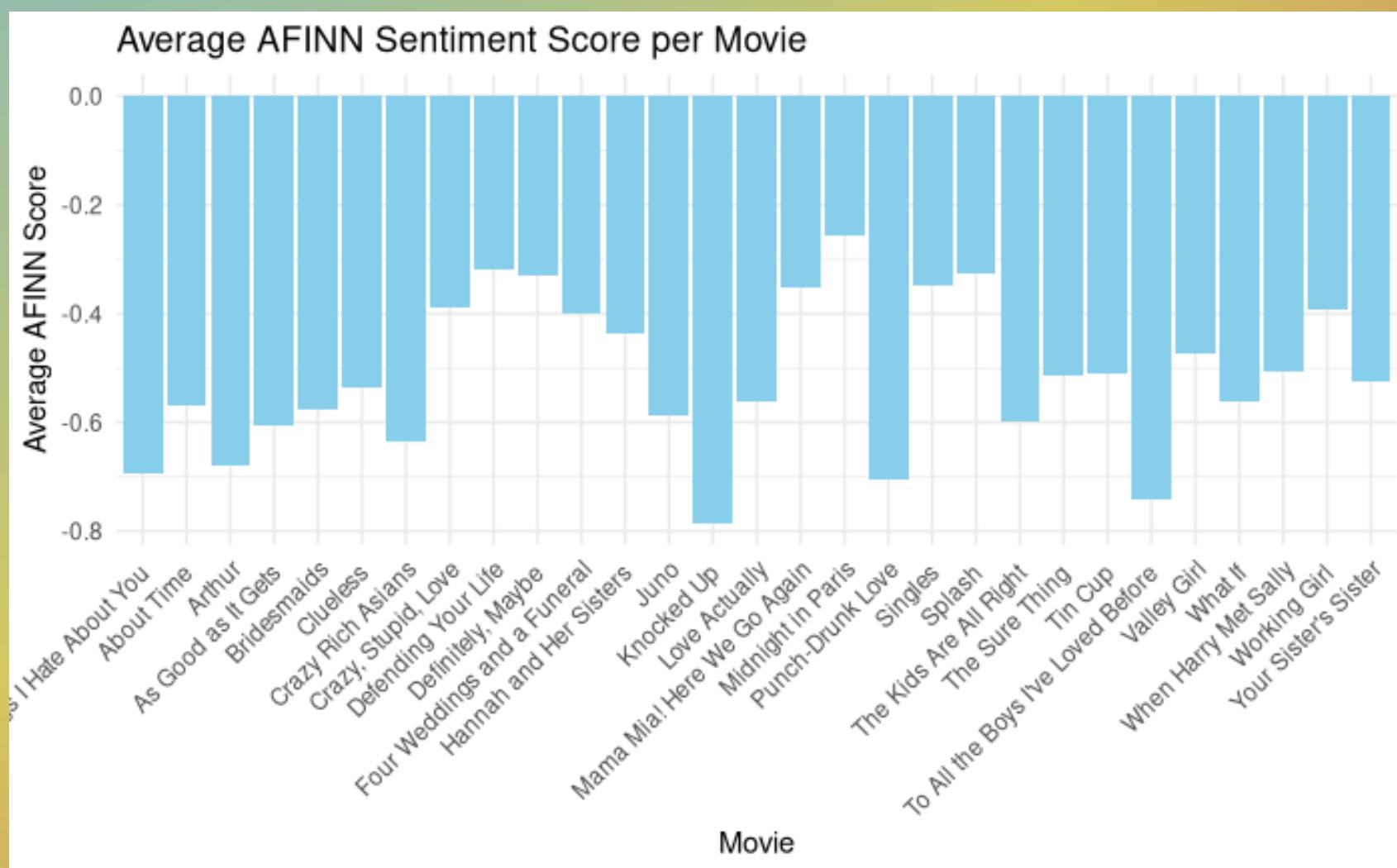
# 2010s



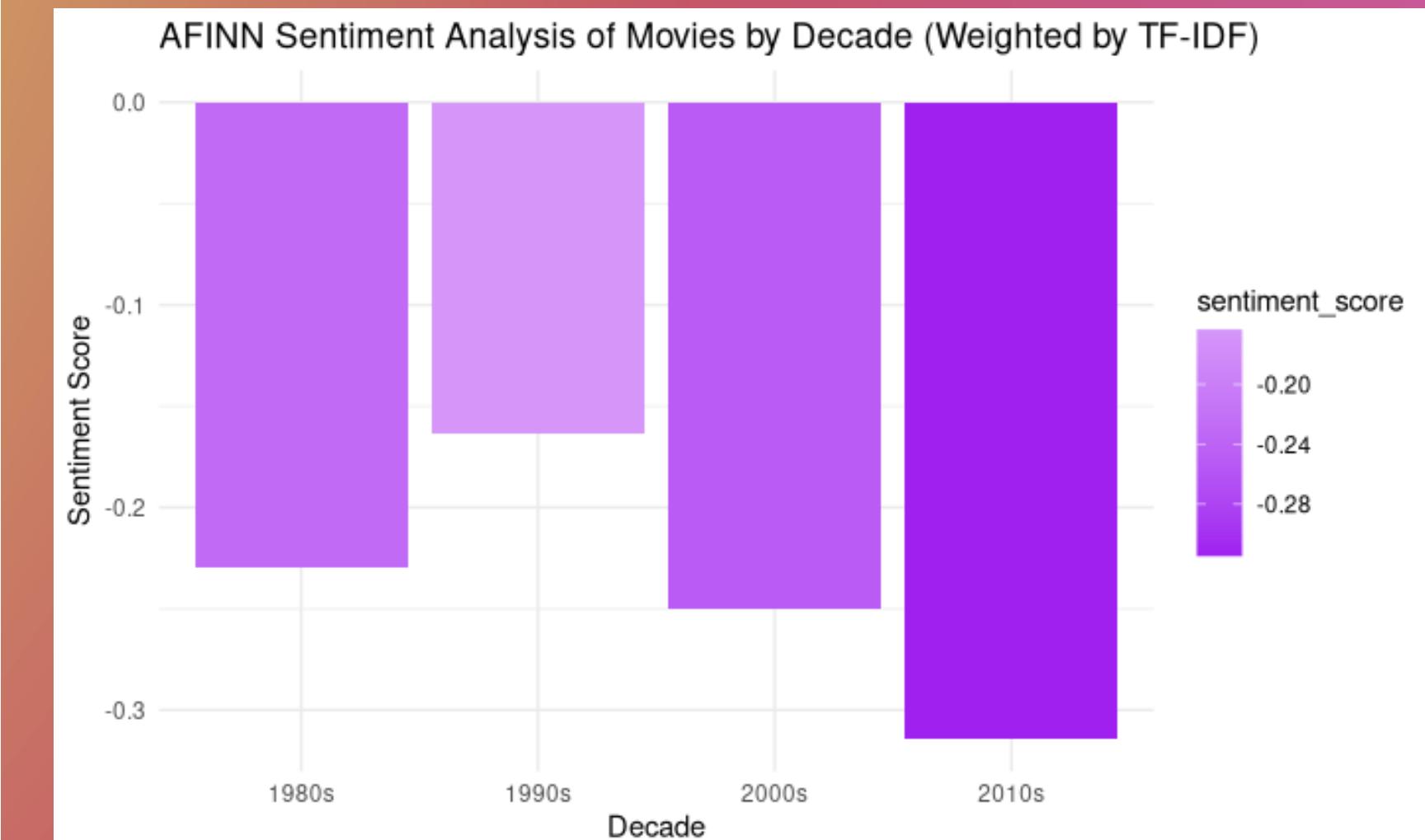
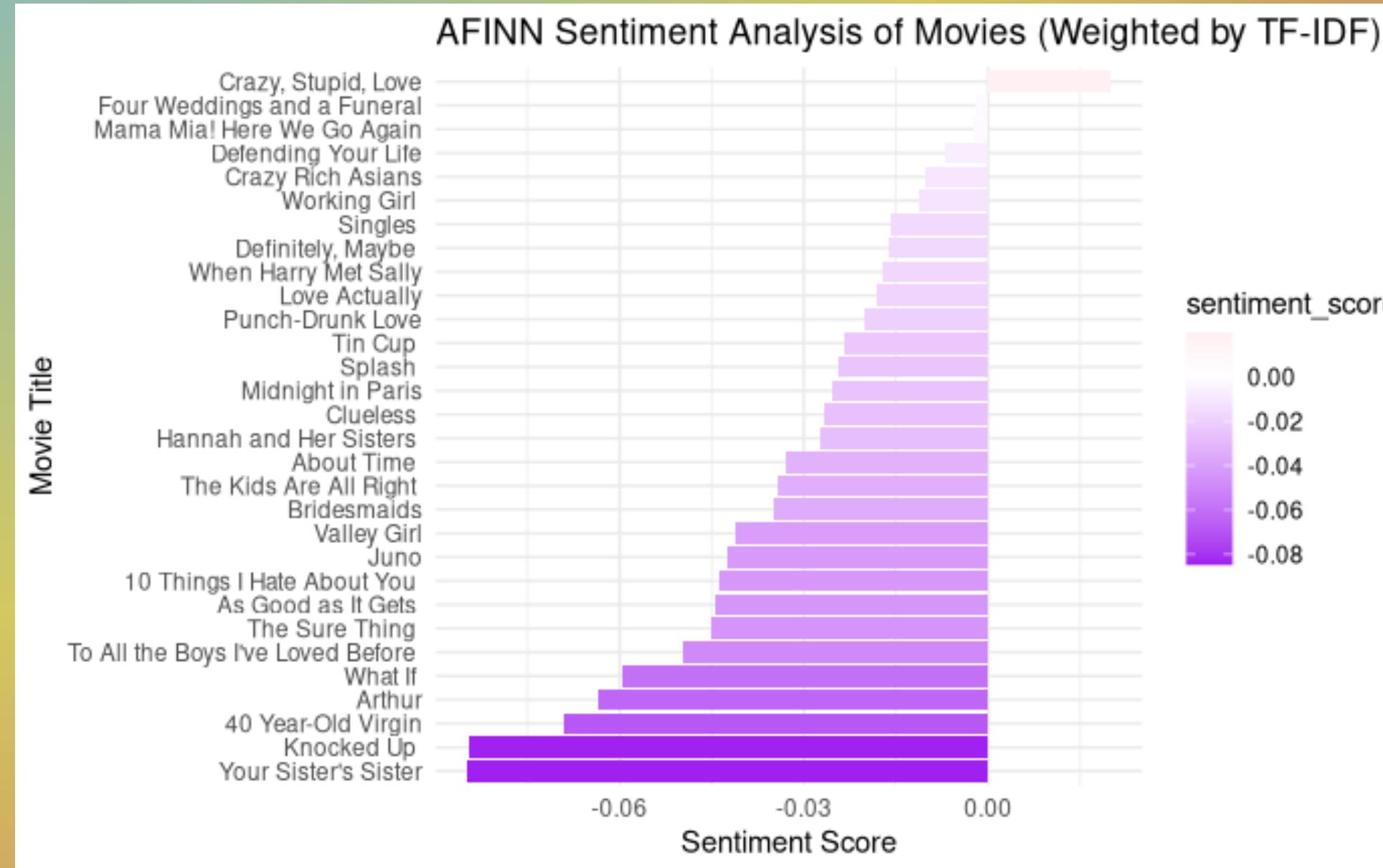


# Sentiment Analysis and Clustering

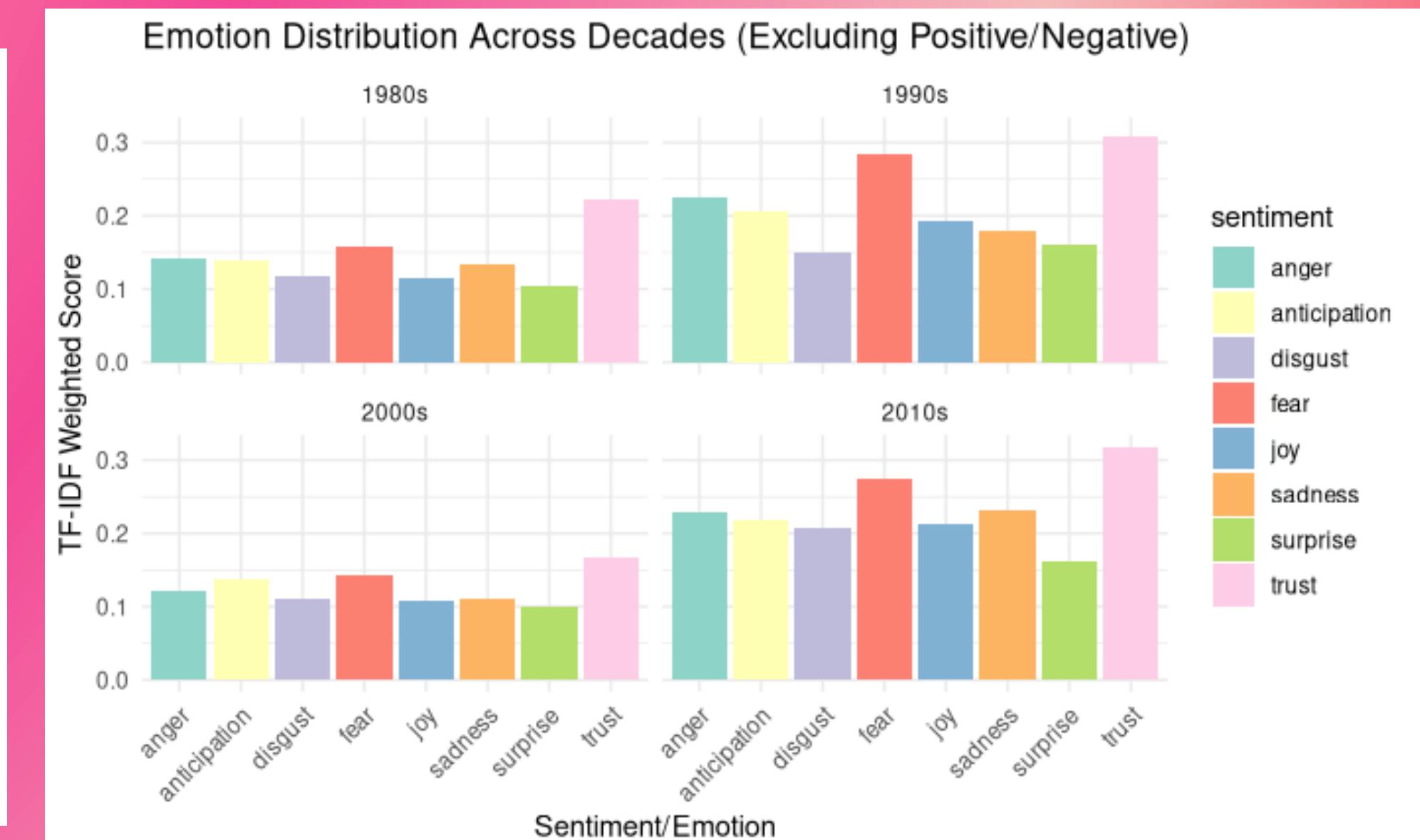
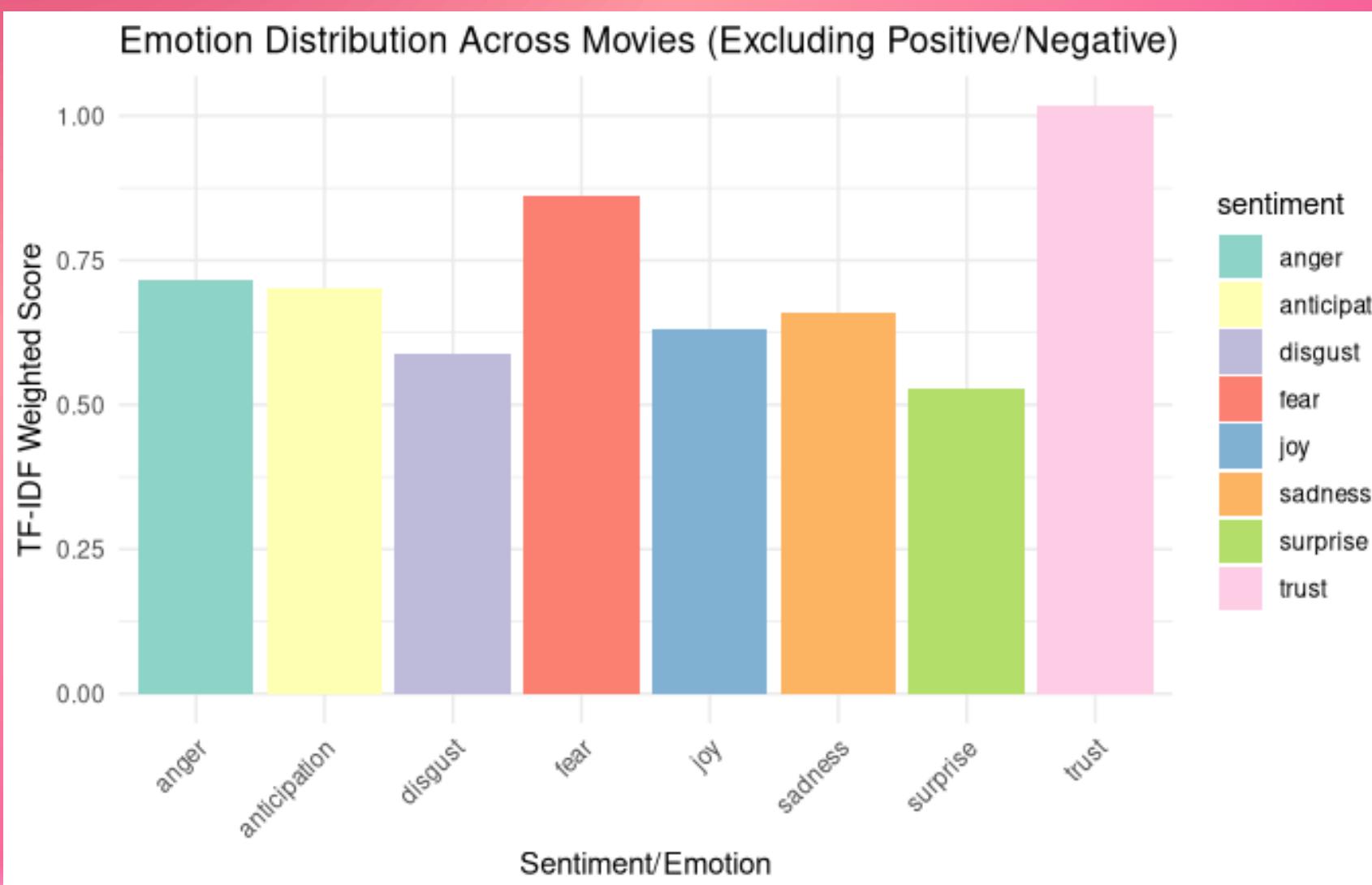
# How Does Emotional Tone Change Across Decades?



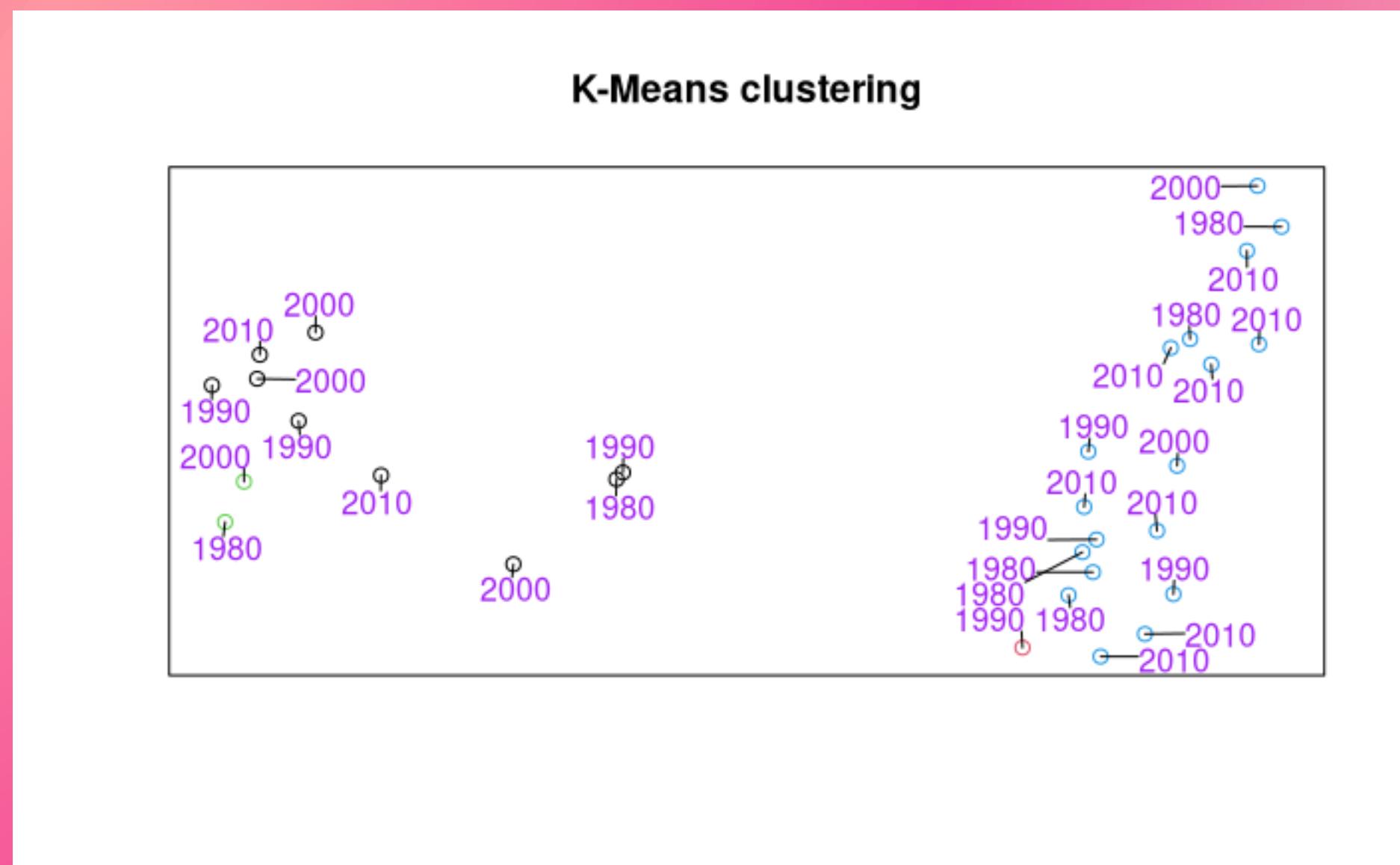
# Are the sentiments of common words changing over time?



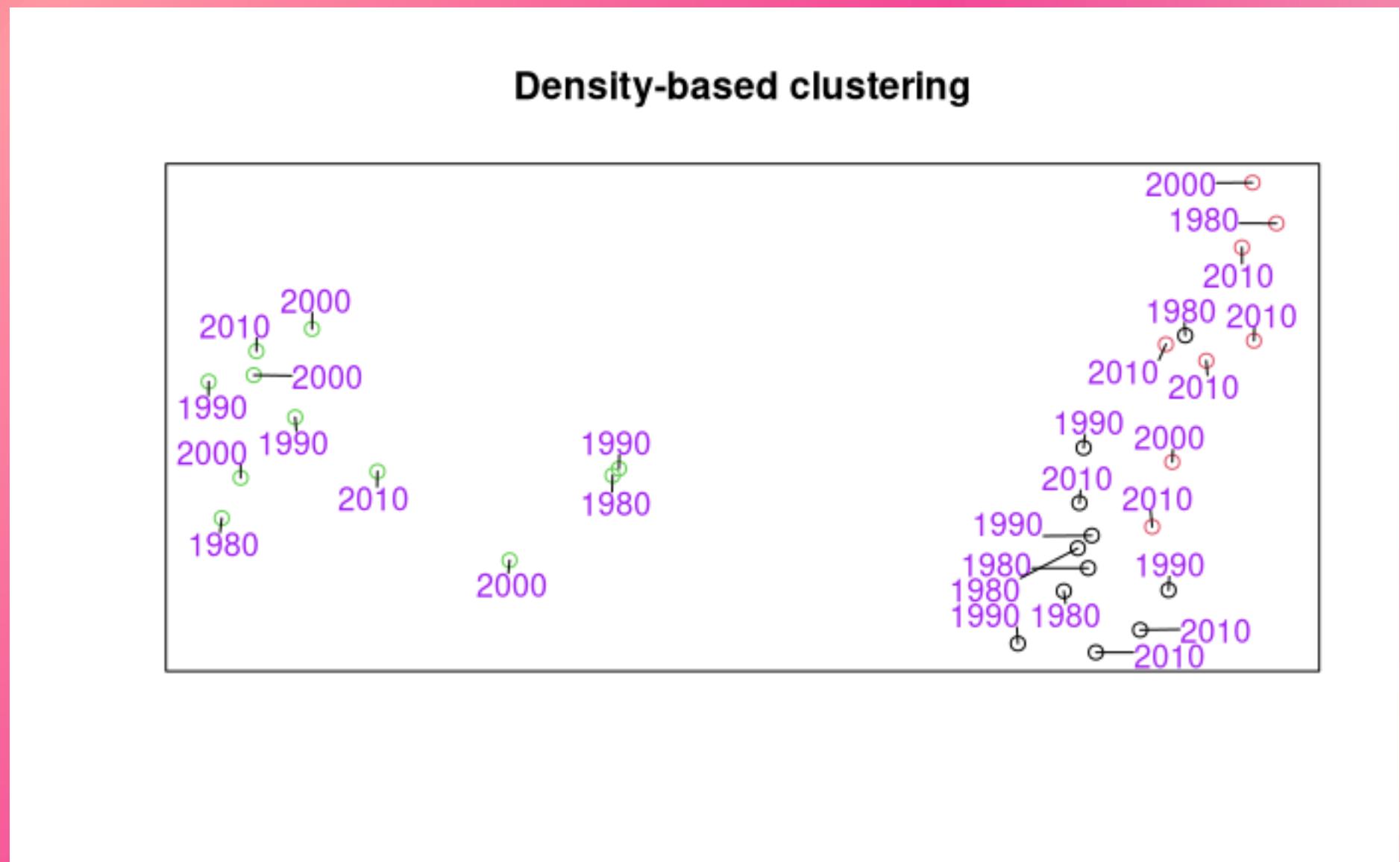
# Which emotions dominate, and how do they change over time?



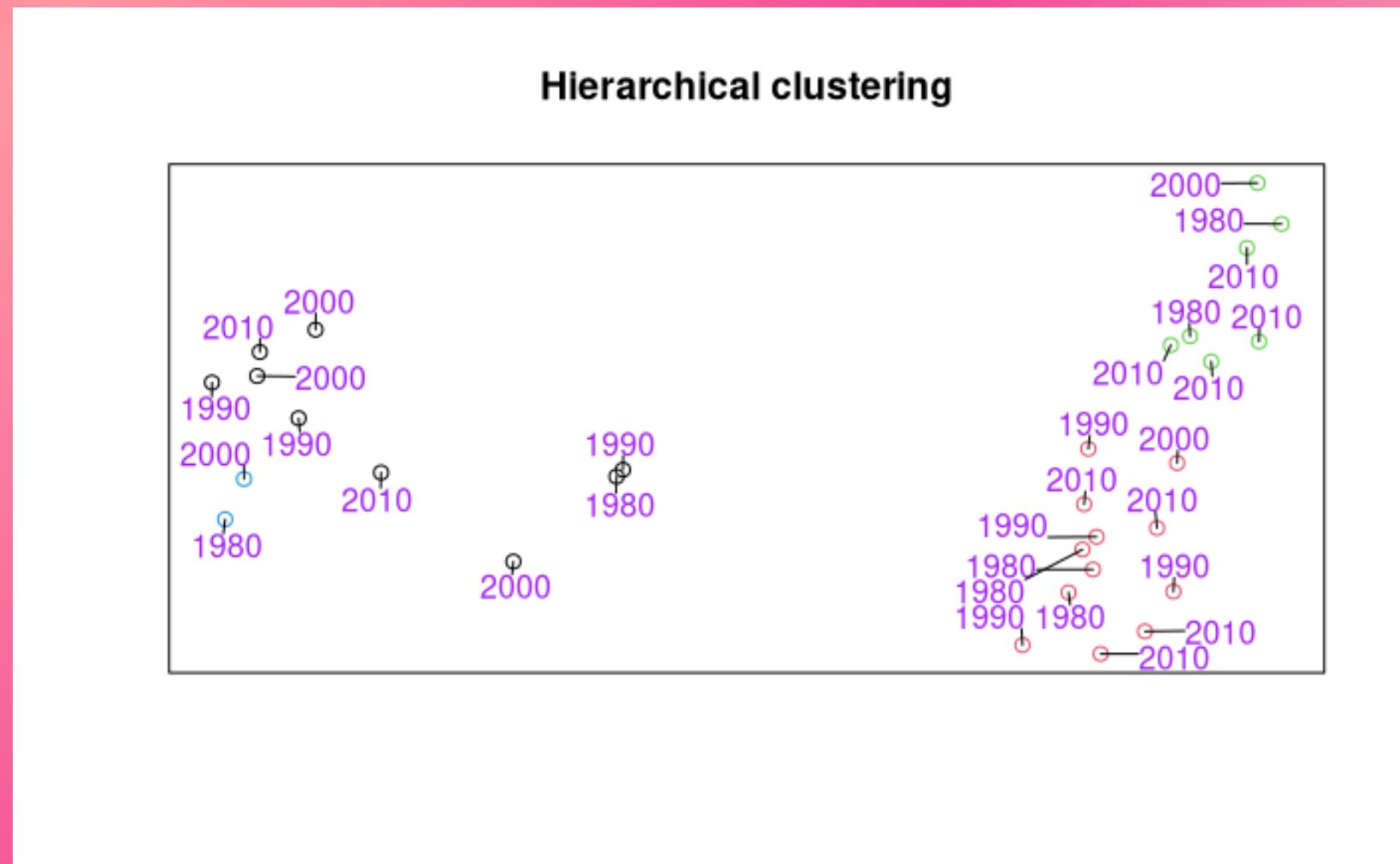
# Can we separate the movies by decade?



# Can we separate the movies by decade?



# Can we separate the movies by decade?

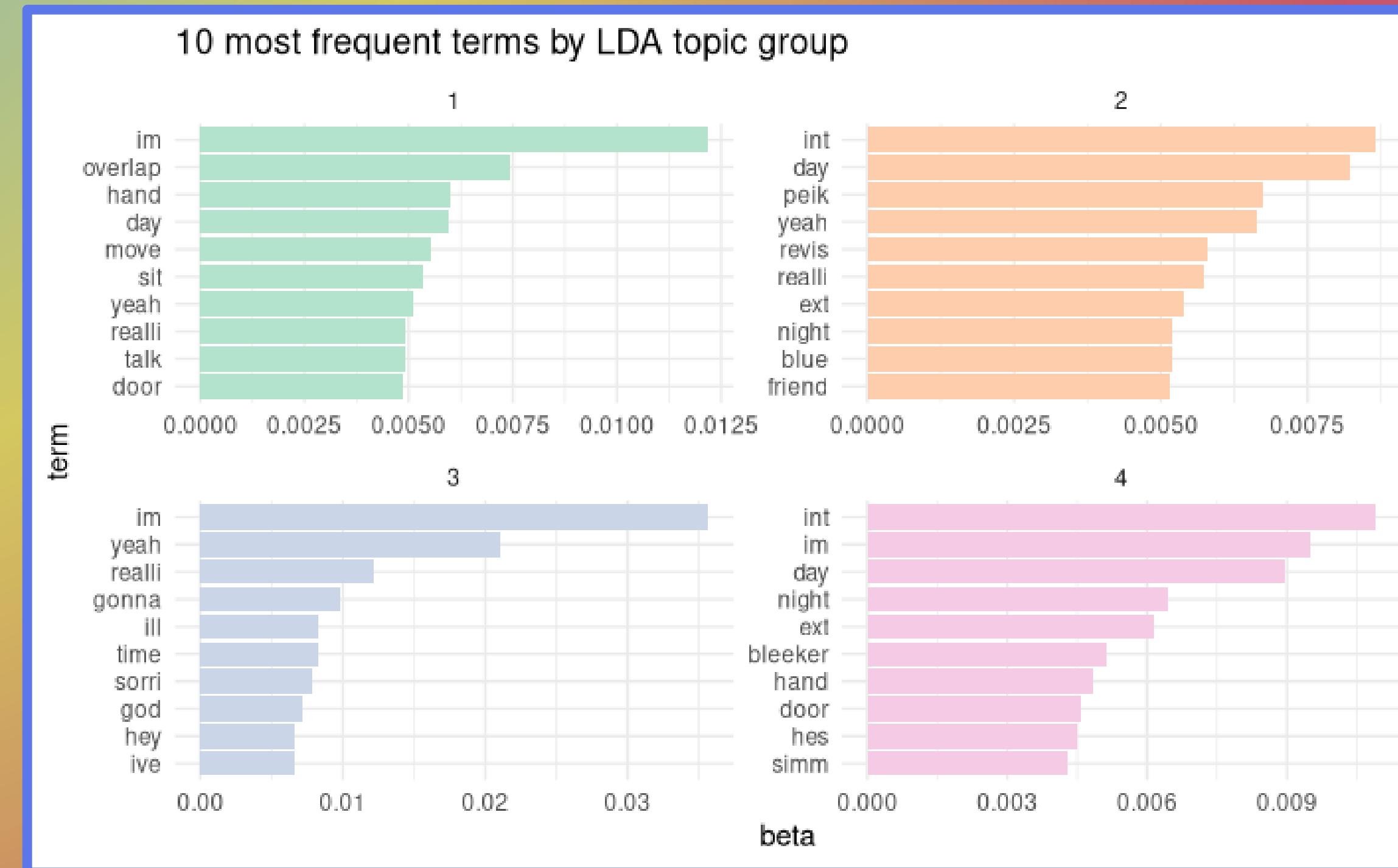


# Topic Modeling, LM Integration, and Prediction

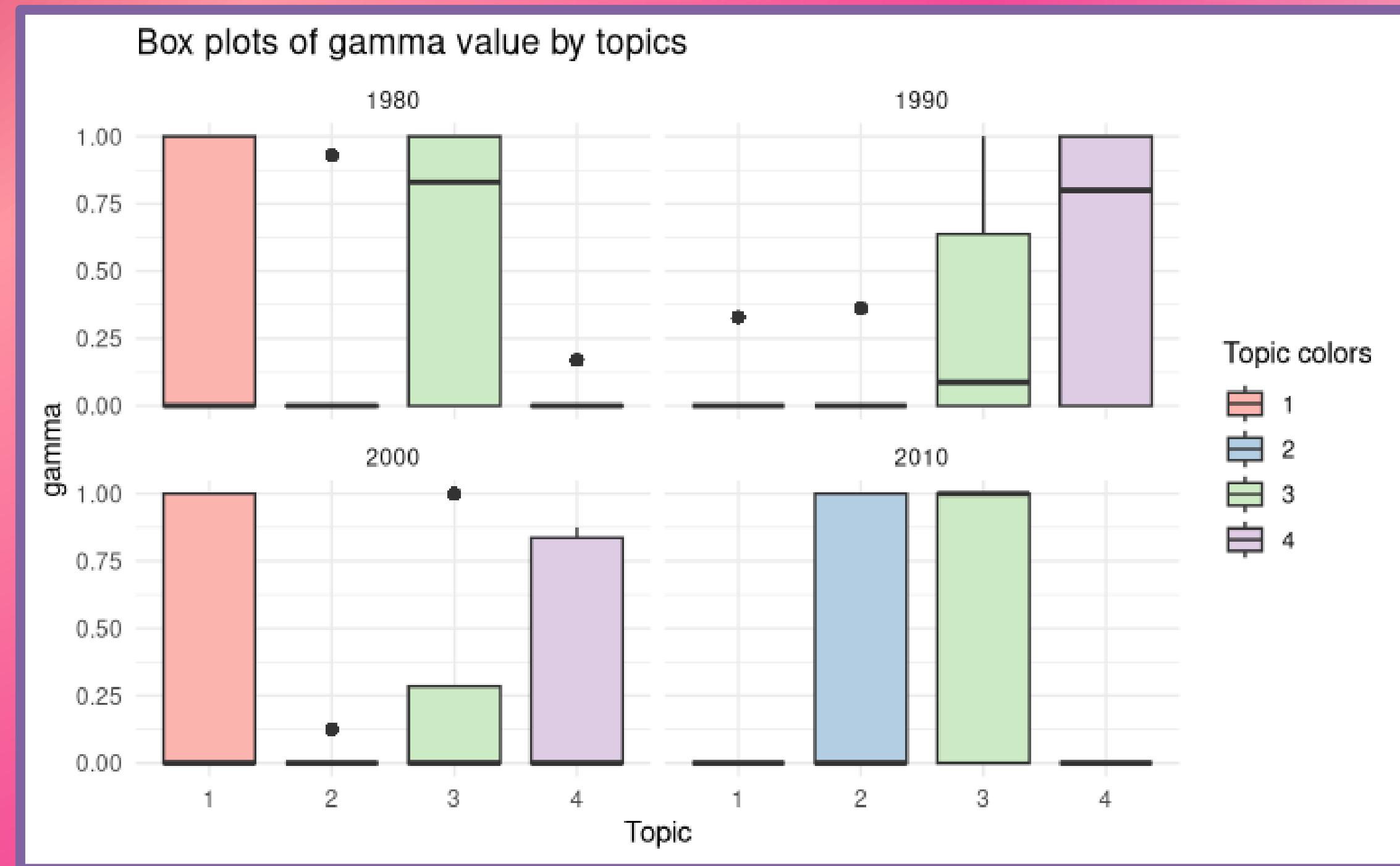


# W

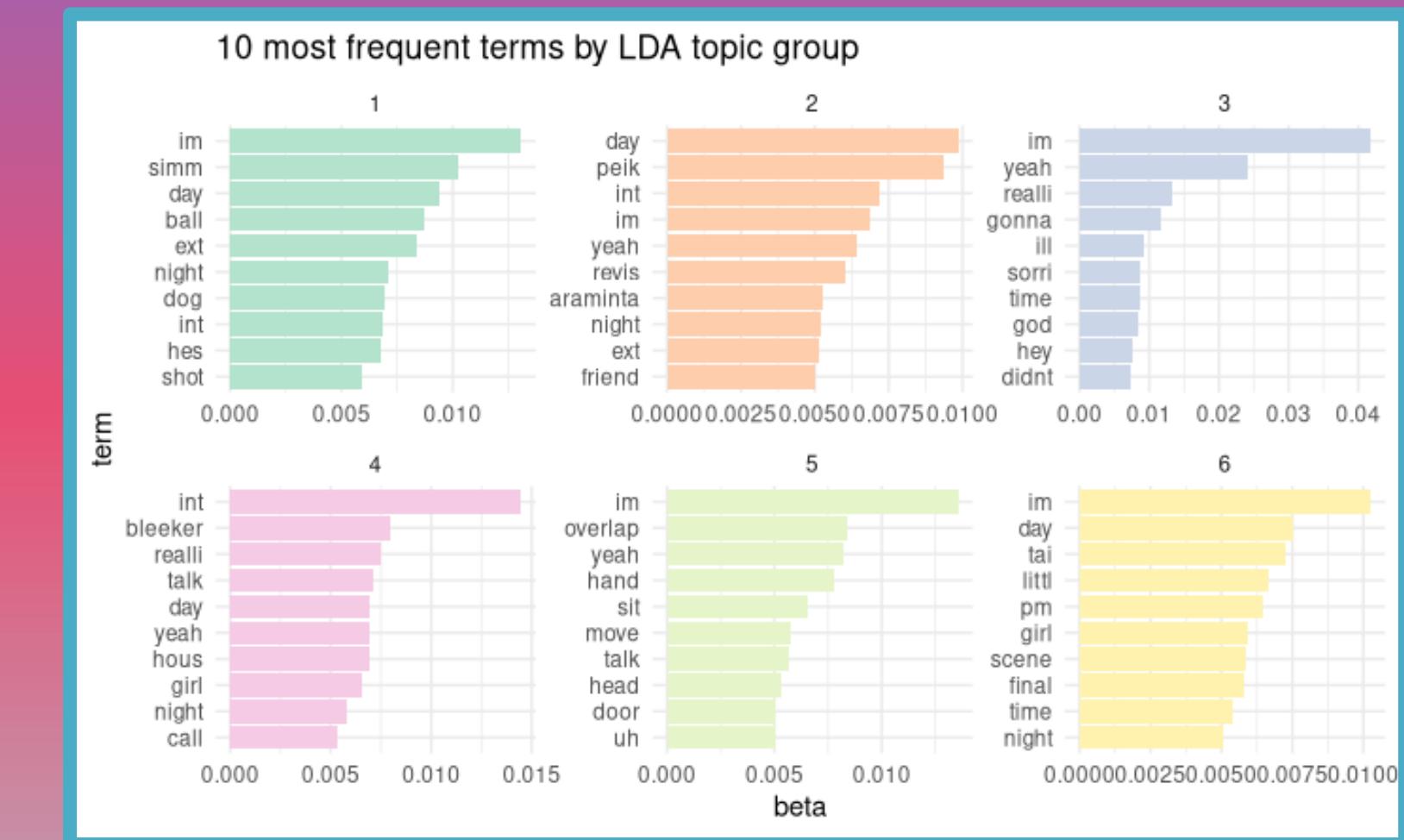
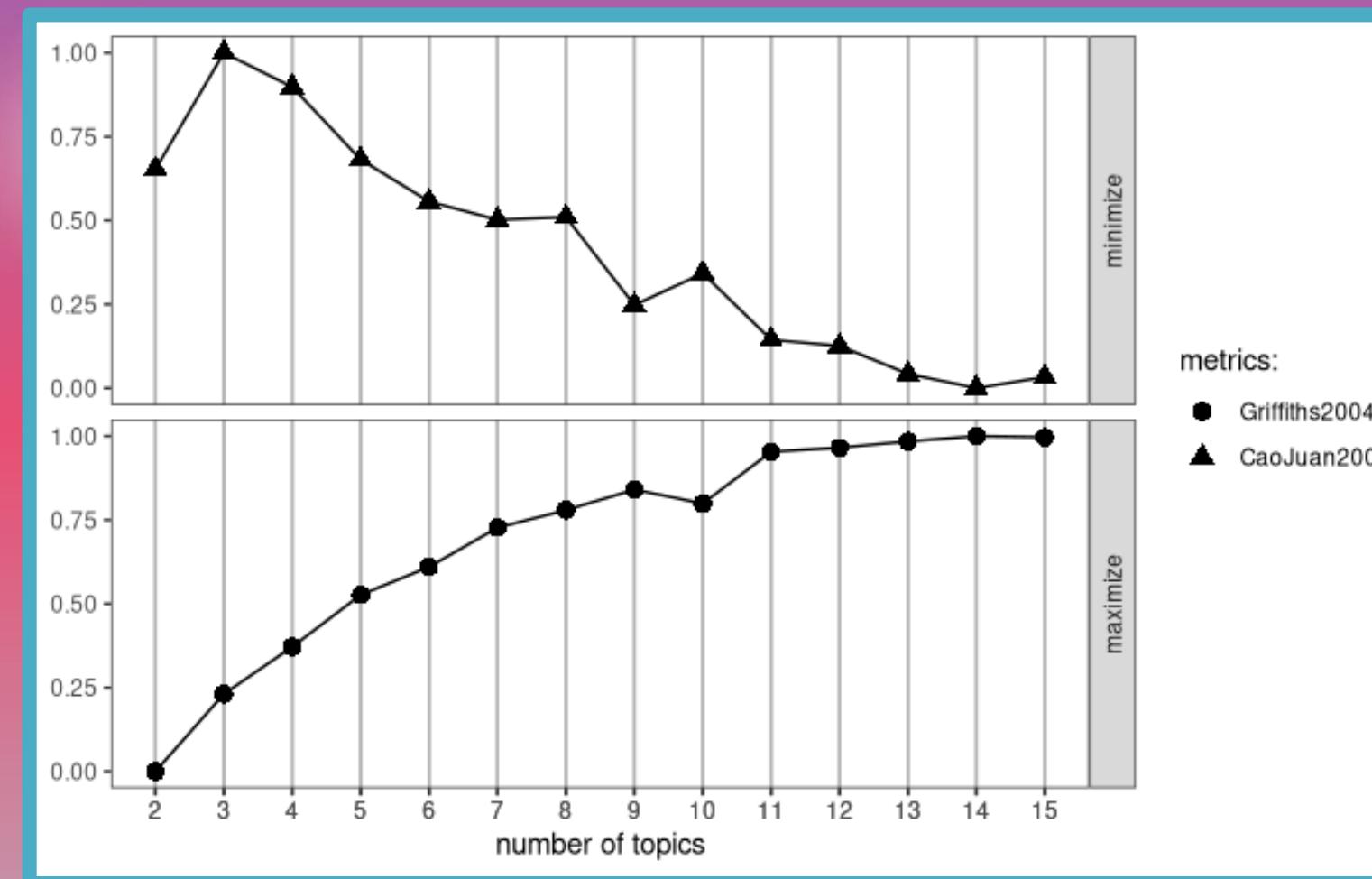
# What are the most popular terms in each topic group?



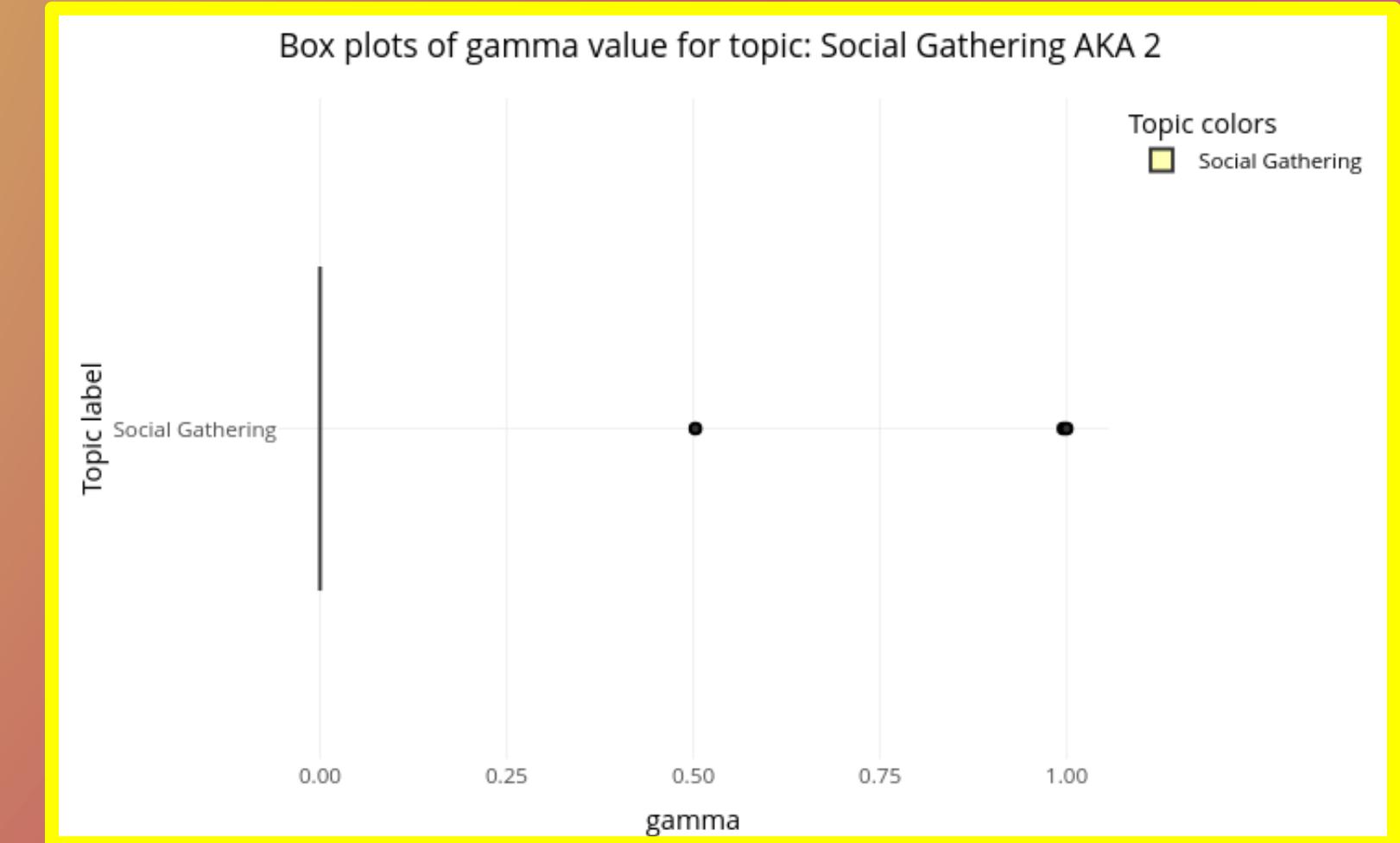
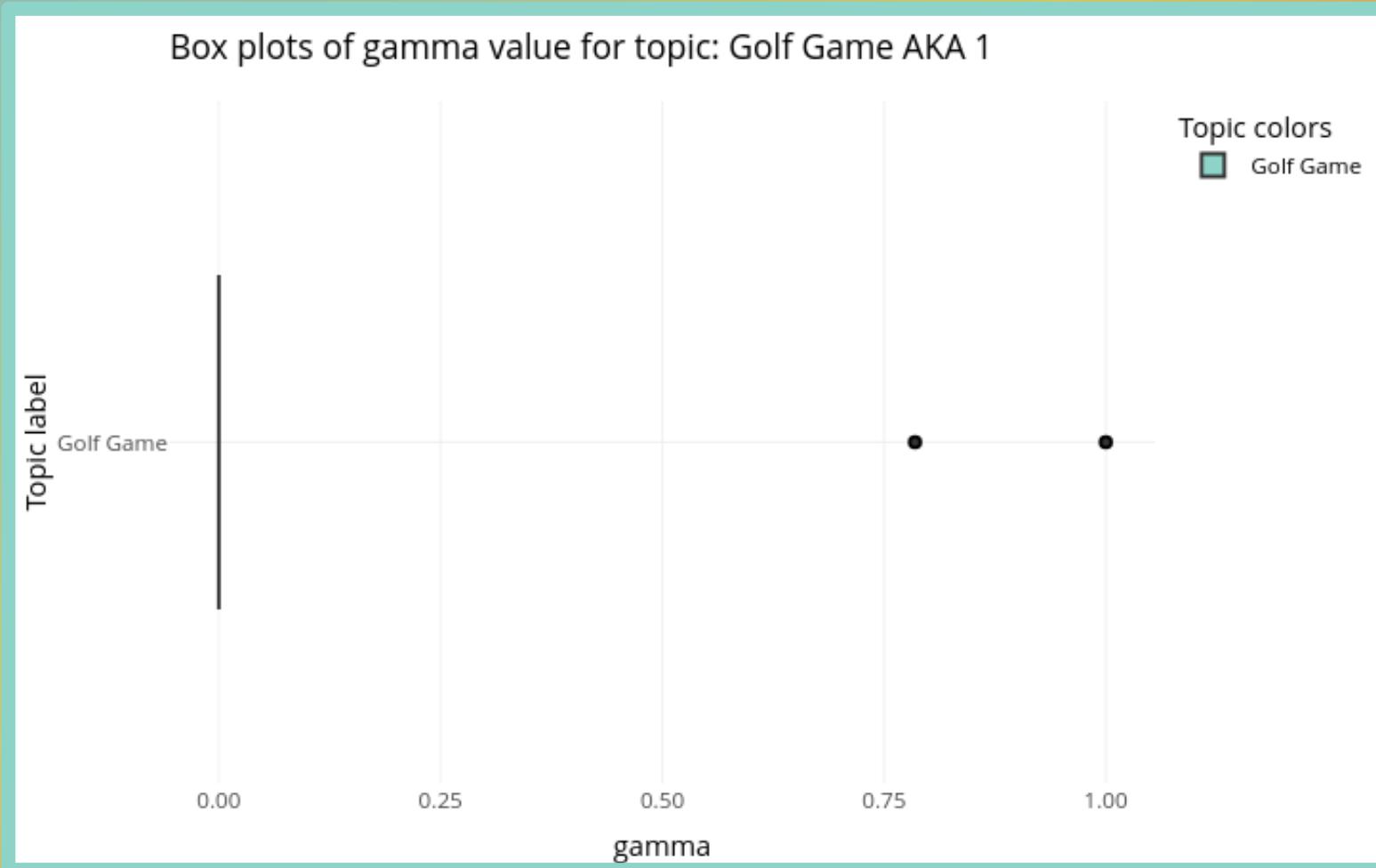
# Do our topics correspond with our decade splits?



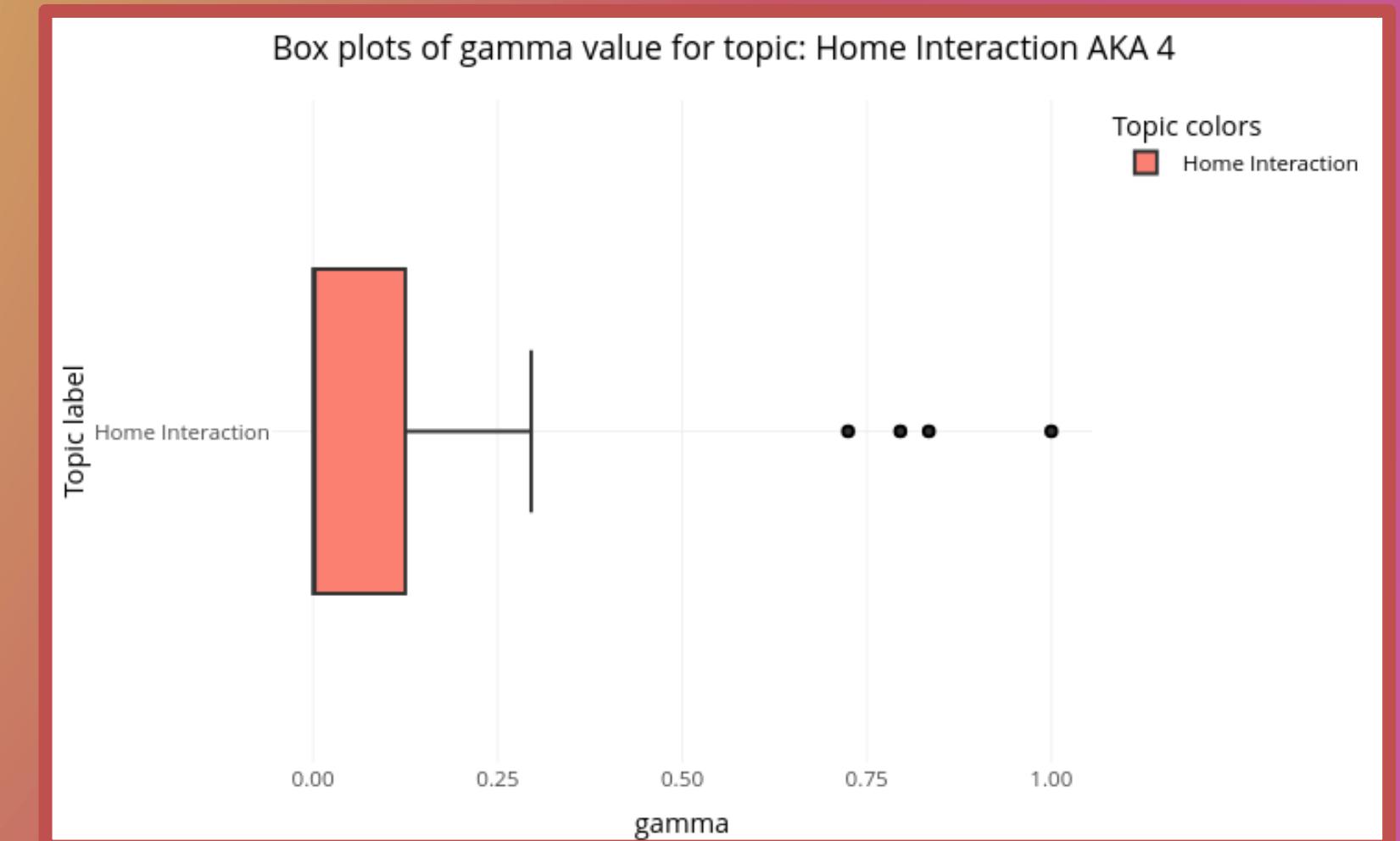
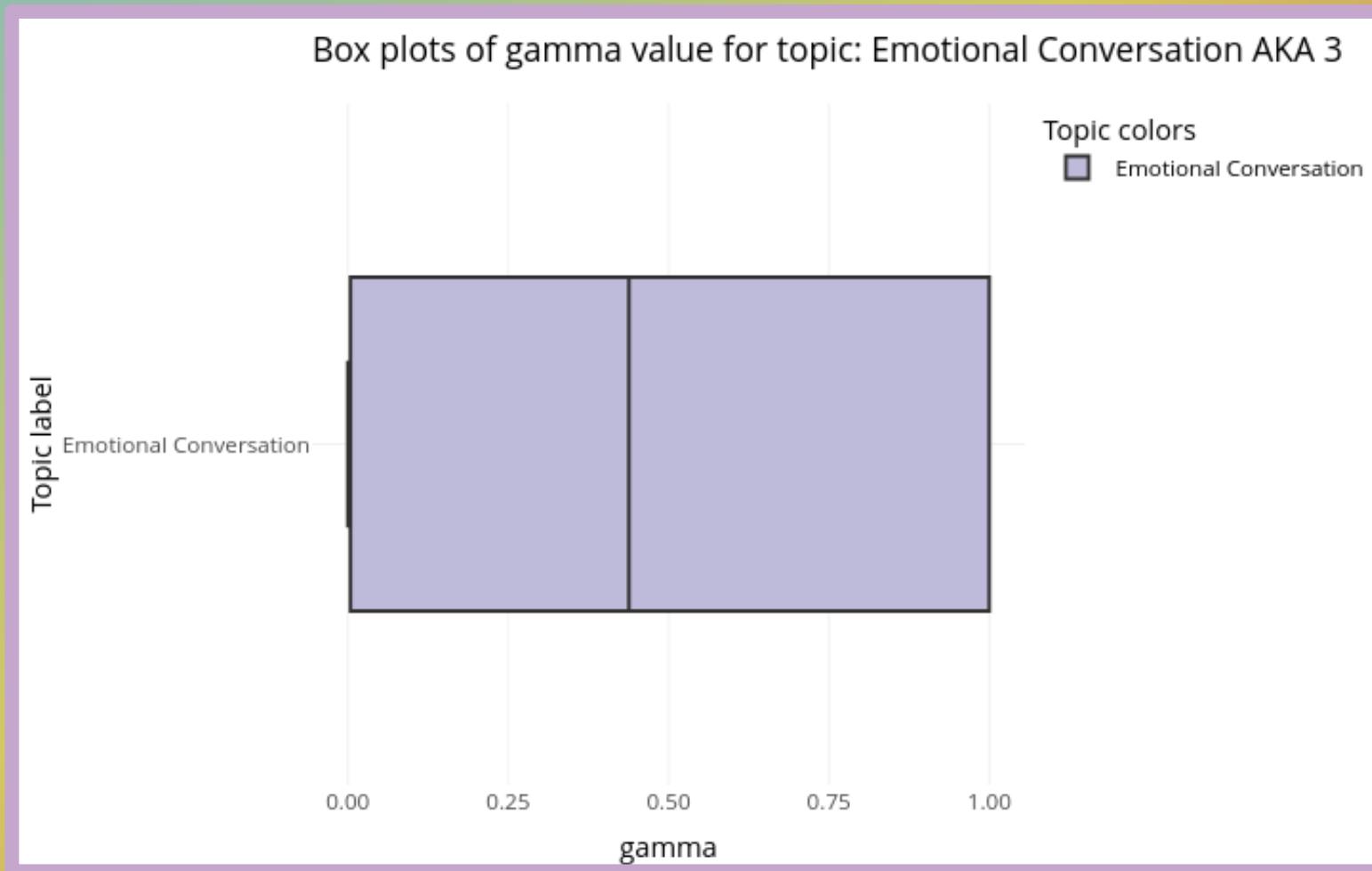
# What are the true topics found in the data?



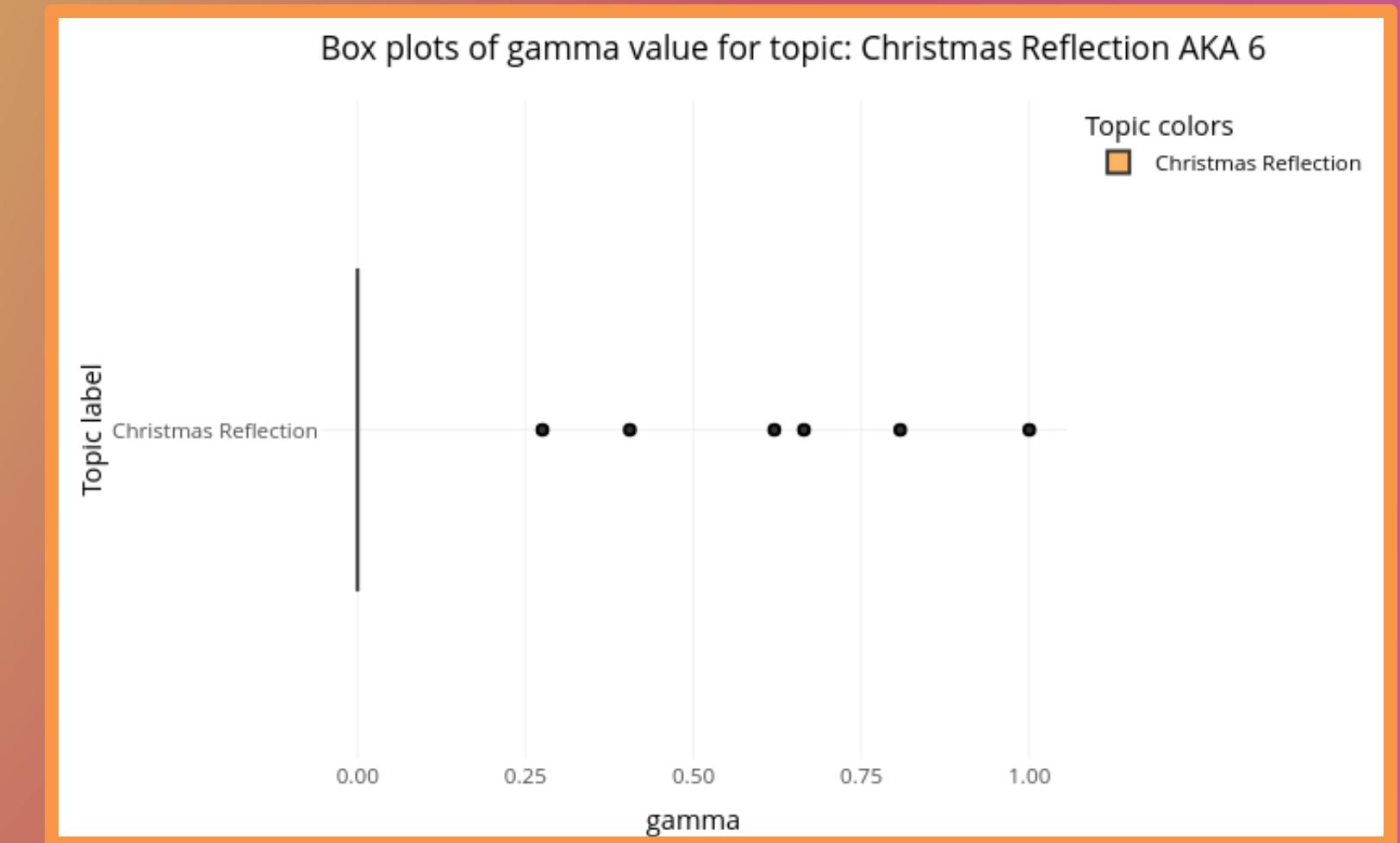
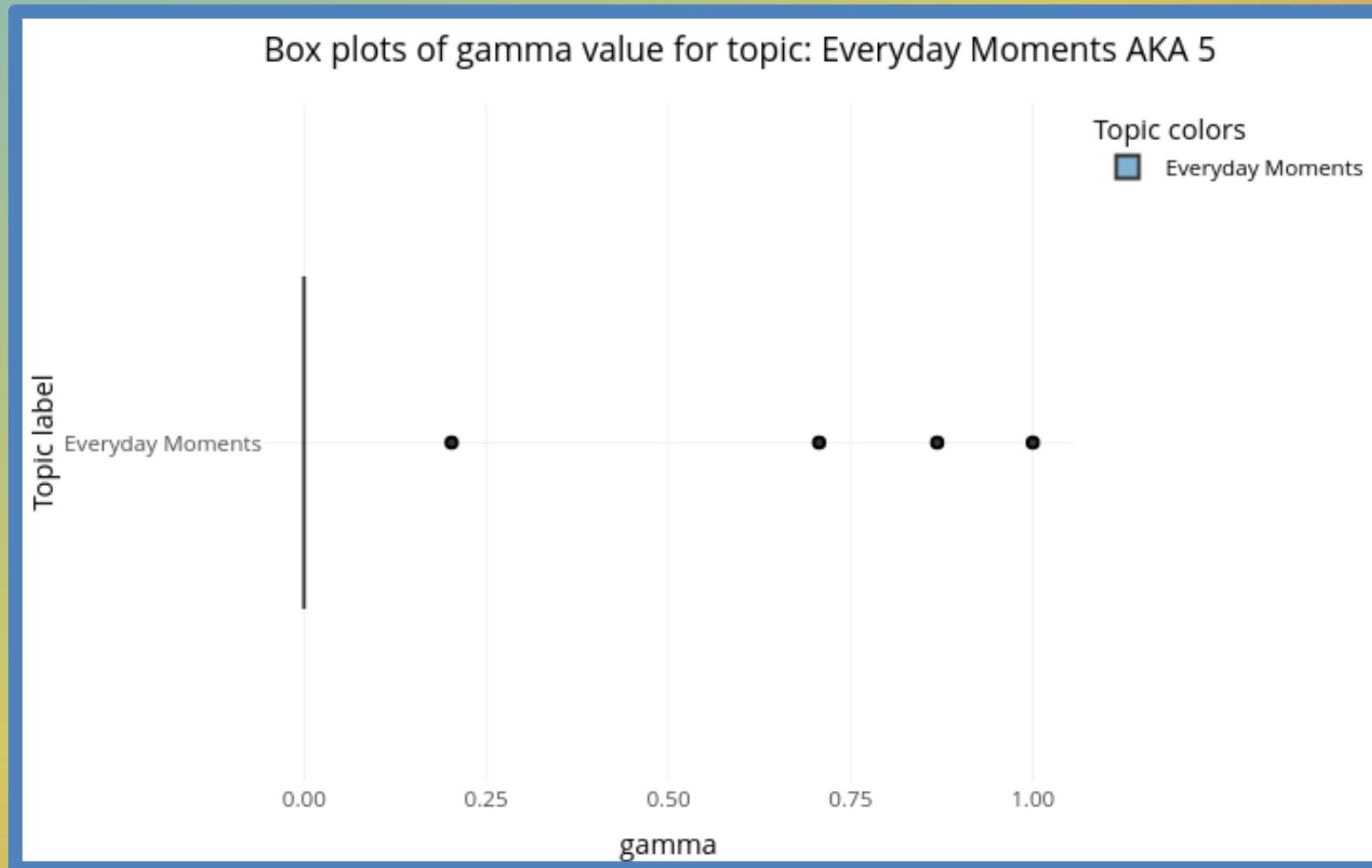
# How are documents associated with these topics (labeled by ChatGPT)?



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

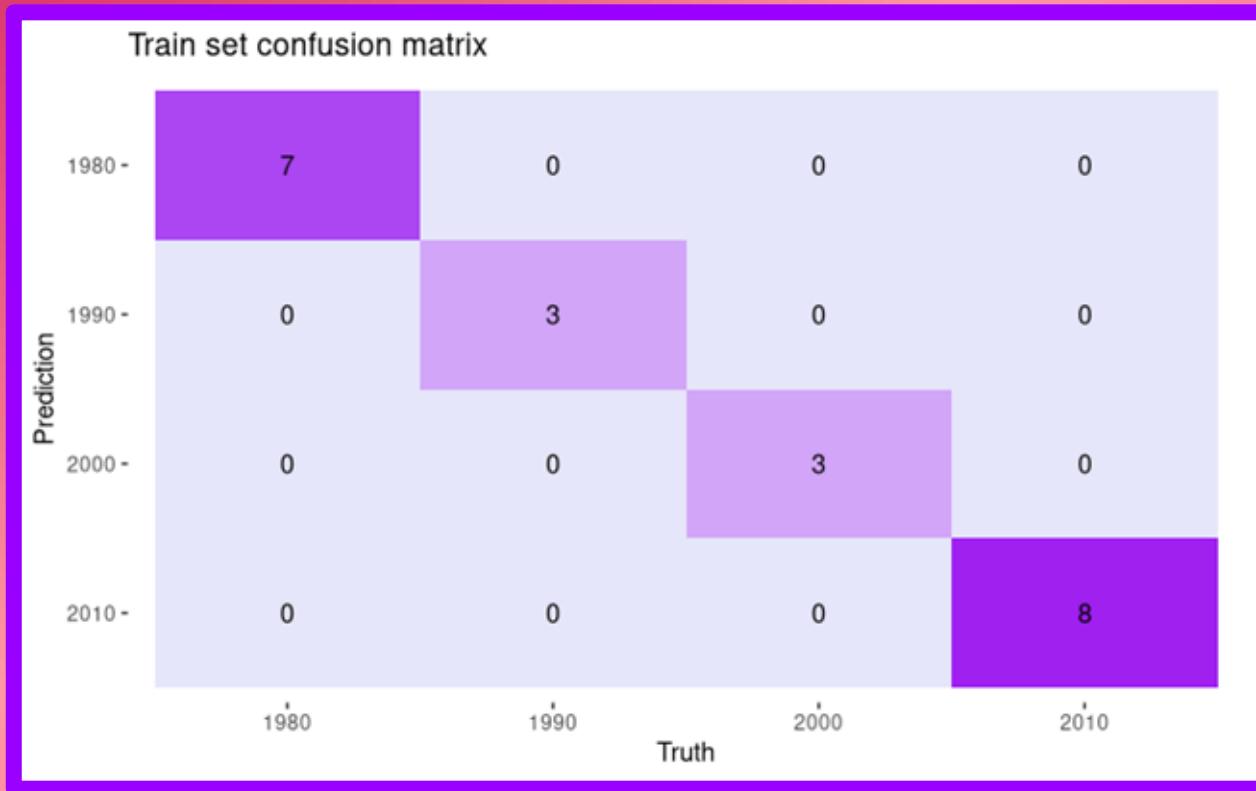


# How are documents associated with these topics (labeled by ChatGPT)?

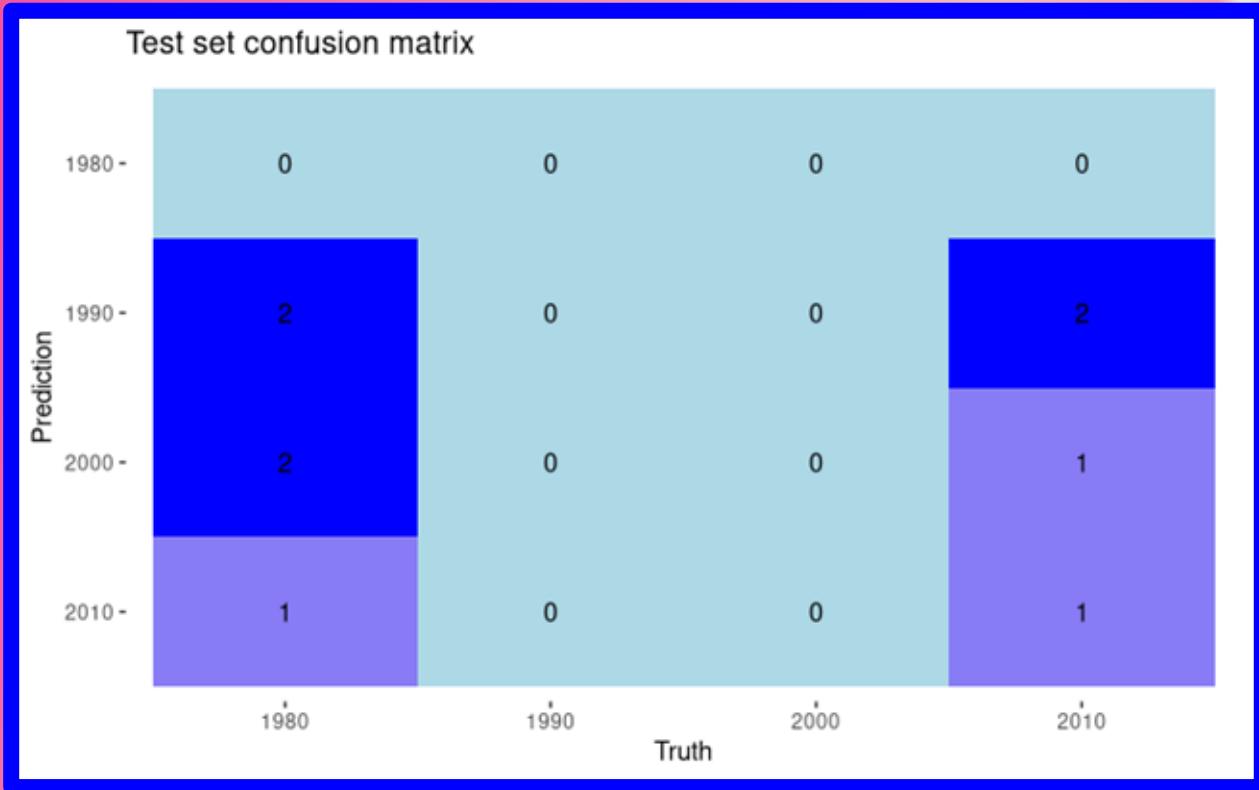




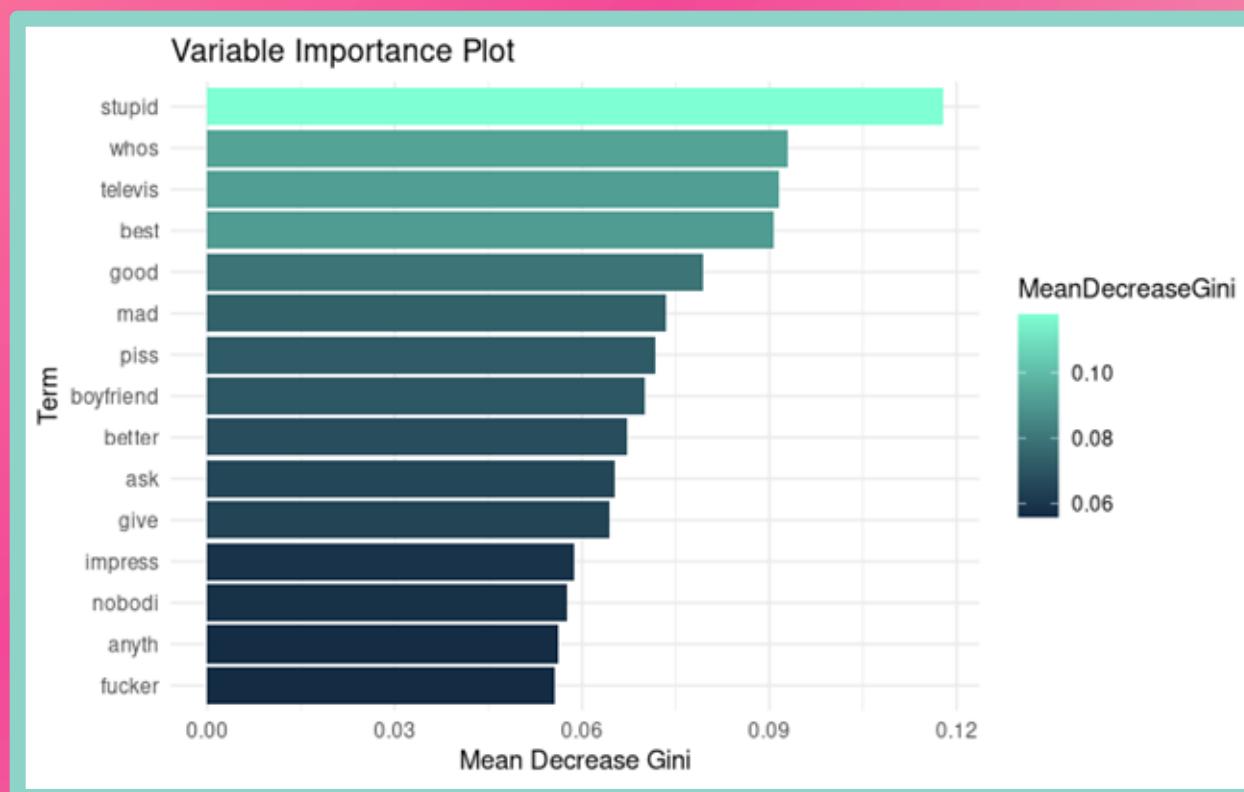
# Can we predict the decade of a rom-com from its text?



Accuracy: 100%



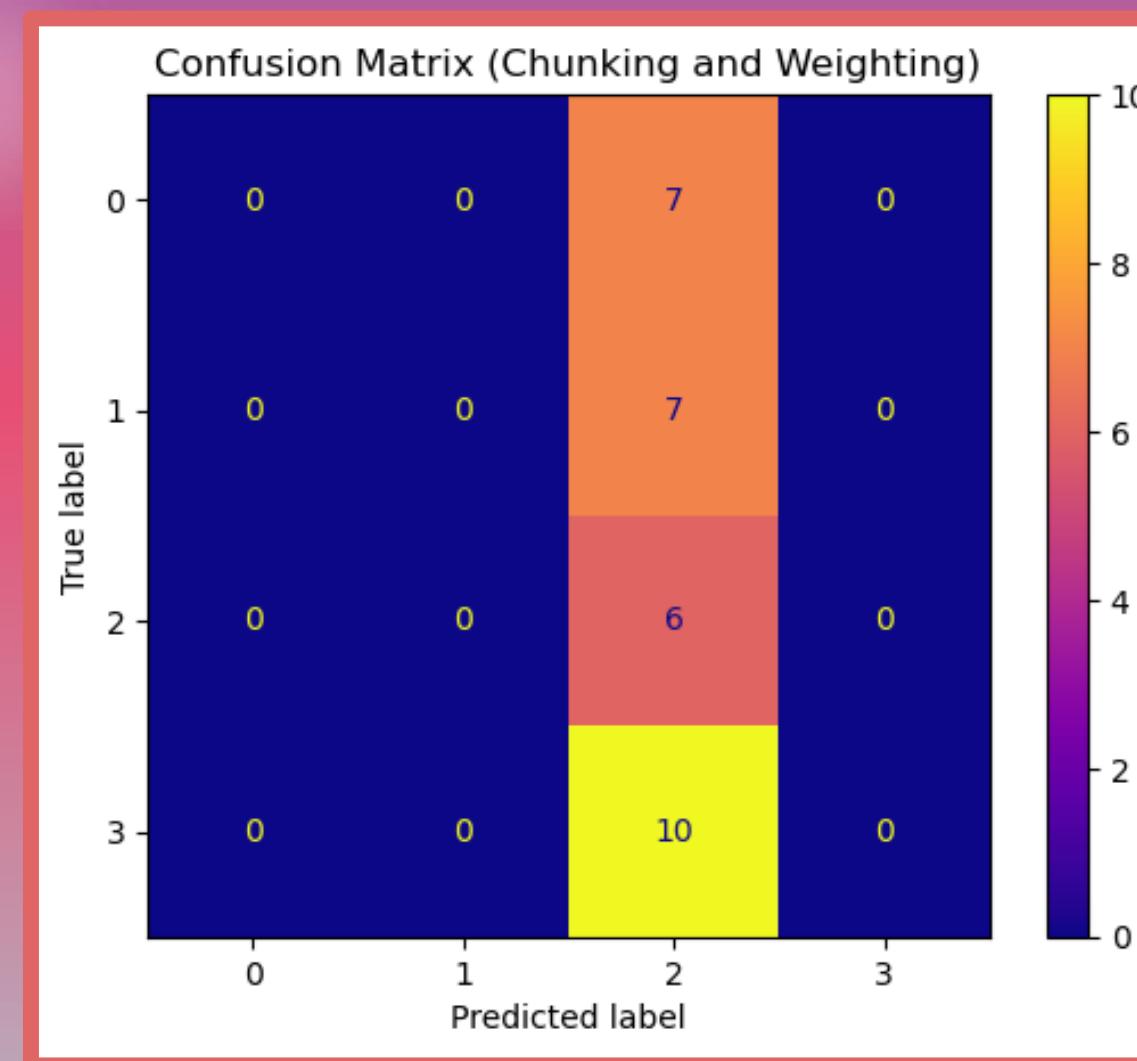
Accuracy: 11.11%



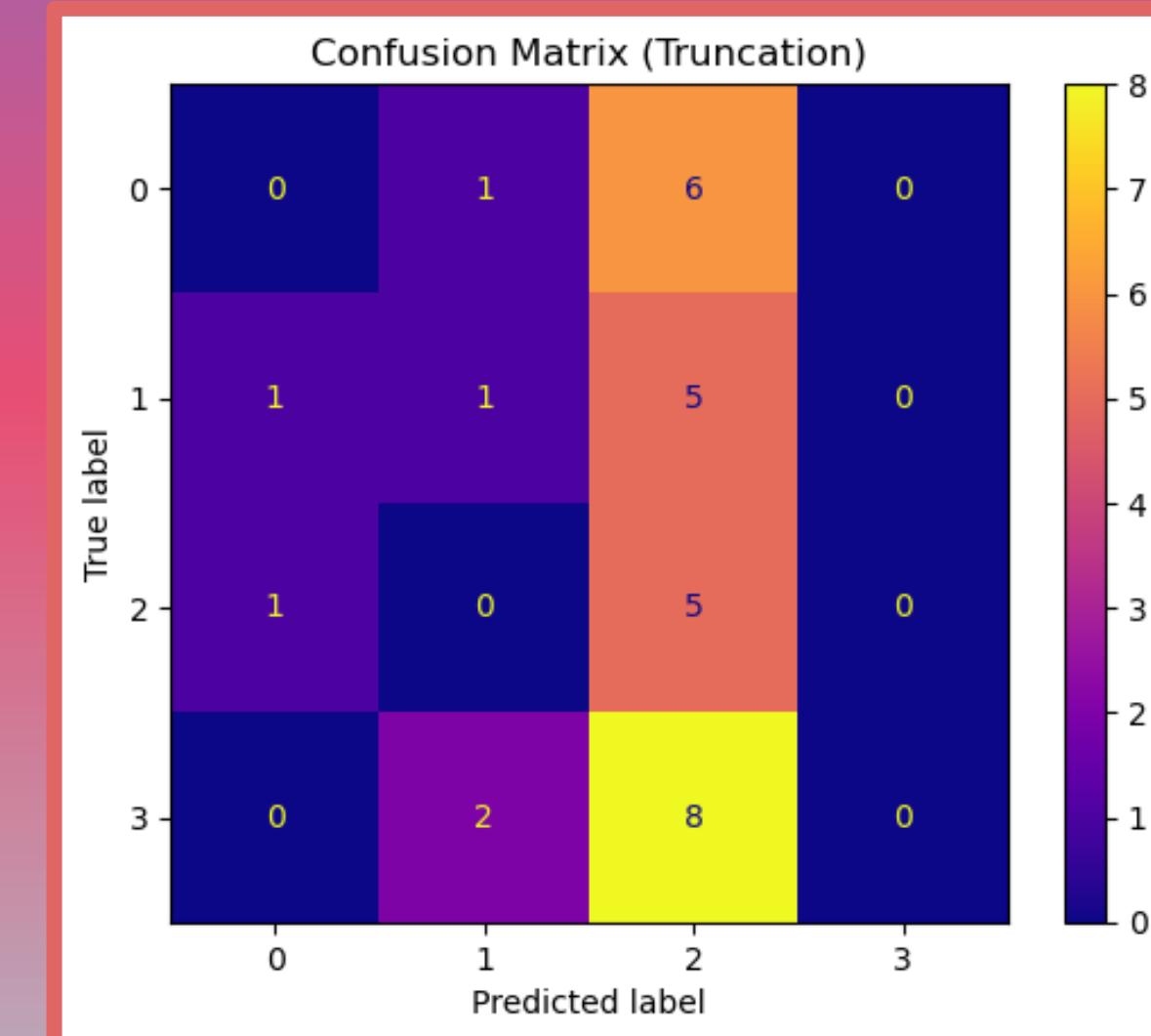


# Can we predict the decade of a rom-com from its text? (LLM)

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?

