SI 206 Final Project:

Common Words and Sentiment Analysis in Various Forms of Media Titles

Carolyn Cullen, Eugene Cousino

University of Michigan, School of Information

SI 206: Data-Oriented Programming

Dr. Barbara Ericson

December 12, 2023

GitHub link: https://github.com/genocous/cousino_cullen_final.git

I. Project Goal

Originally, we planned to use the Spotify API, Apple Music API, and Google Books API. We planned on gathering title and genre information in order to find the most common words in titles grouped by genre.

II. Achieved Goals

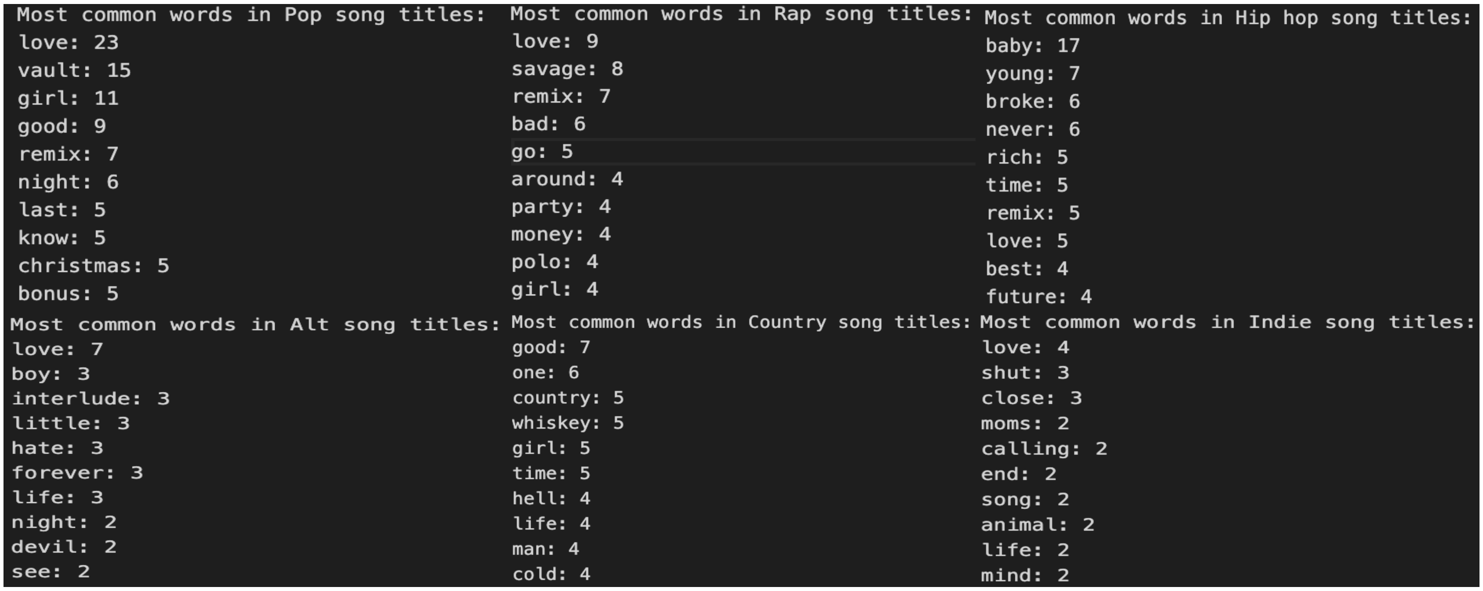
During our initial attempts at gathering data, we realized that the APIs we chose would not be sufficient for all calculations. So, we instead used the Spotify API, IMDb API, and Open Library API. We successfully found the most common words in titles by genre and decided to take our project a step further by performing sentiment analysis on all titles by genre.

III. The problems that you faced

We faced a number of problems, including in our initial stages when we discovered a few things. One, the Apple Music API required $99.99 developer "membership" in order to pull data from the API. We had to find a different API and ended up using the IMDb API instead. We also found that the Google Books API was hard to work with, so we used the Open Library API instead, and it was much easier to use. We also dealt with significant runtime issues of data collection from our API's, specifically the IMDb API. We considered switching to another database, but ultimately decided against it, because we only had to run the code a few times to get the data that we needed. Once we had the data, we had to deal with cleaning it. We created a list of excluded words for each source of media, because they skewed the results and made it harder to analyze the content. We also had to ensure that we were using English titles, because particularly in the Spotify database, there were many foreign tracks, which were not helpful to us.

IV. Calculations

Music Results:



```
Most common words in Pop song titles:
love: 23
vault: 15
girl: 11
good: 9
remix: 7
night: 6
last: 5
know: 5
christmas: 5
bonus: 5
```

```
Most common words in Rap song titles:
love: 9
savage: 8
remix: 7
bad: 6
go: 5
around: 4
party: 4
money: 4
polo: 4
girl: 4
```

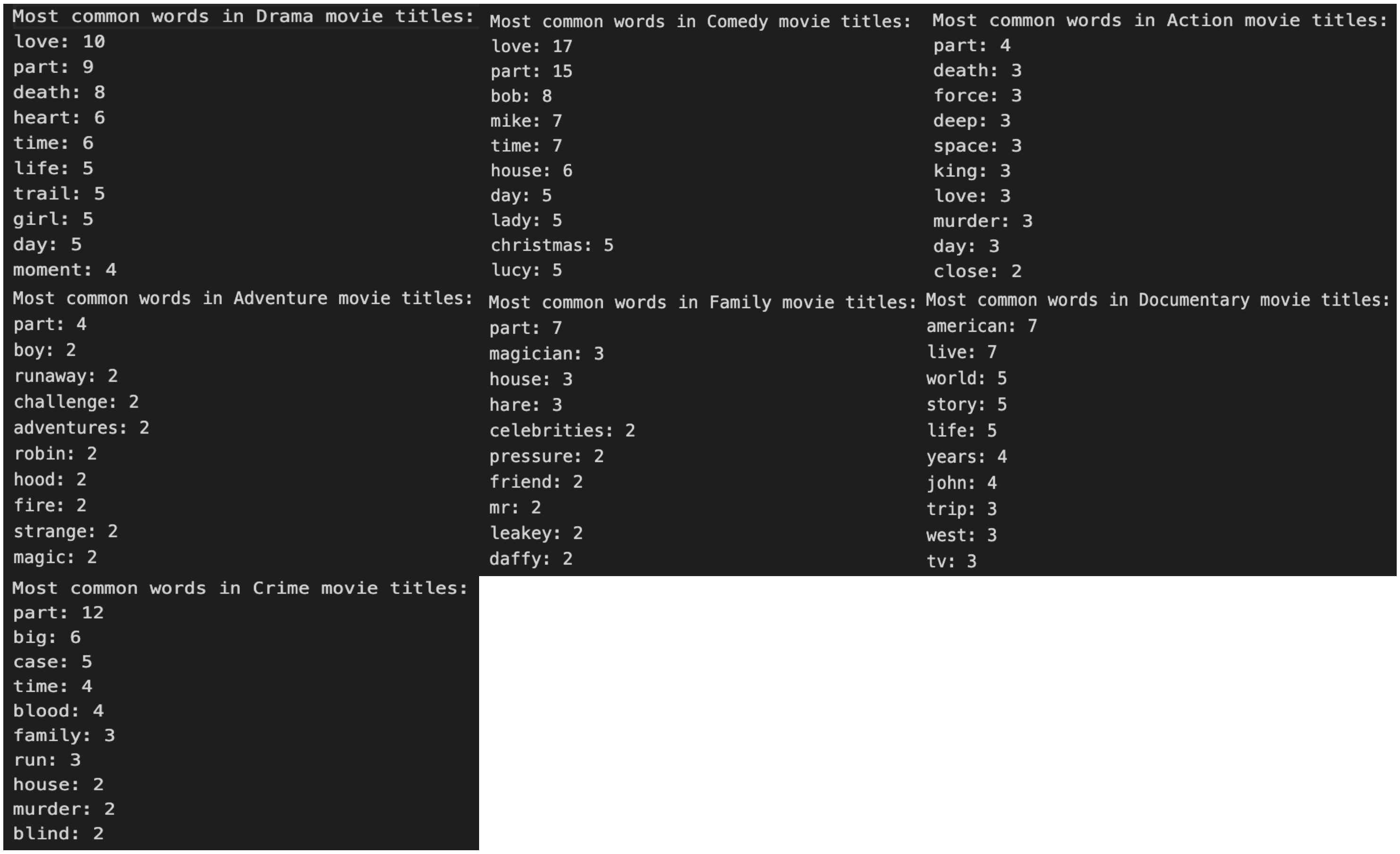
```
Most common words in Hip hop song titles:
baby: 17
young: 7
broke: 6
never: 6
rich: 5
time: 5
remix: 5
love: 5
best: 4
future: 4
```

```
Most common words in Alt song titles:
love: 7
boy: 3
interlude: 3
little: 3
hate: 3
forever: 3
life: 3
night: 2
devil: 2
see: 2
```

```
Most common words in Country song titles:
good: 7
one: 6
country: 5
whiskey: 5
girl: 5
time: 5
hell: 4
life: 4
man: 4
cold: 4
```

```
Most common words in Indie song titles:
love: 4
shut: 3
close: 3
moms: 2
calling: 2
end: 2
song: 2
animal: 2
life: 2
mind: 2
```

Movie Results:



```
Most common words in Drama movie titles:
love: 10
part: 9
death: 8
heart: 6
time: 6
life: 5
trail: 5
girl: 5
day: 5
moment: 4
```

```
Most common words in Comedy movie titles:
love: 17
part: 15
bob: 8
mike: 7
time: 7
house: 6
day: 5
lady: 5
christmas: 5
lucy: 5
```

```
Most common words in Action movie titles:
part: 4
death: 3
force: 3
deep: 3
space: 3
king: 3
love: 3
murder: 3
day: 3
close: 2
```

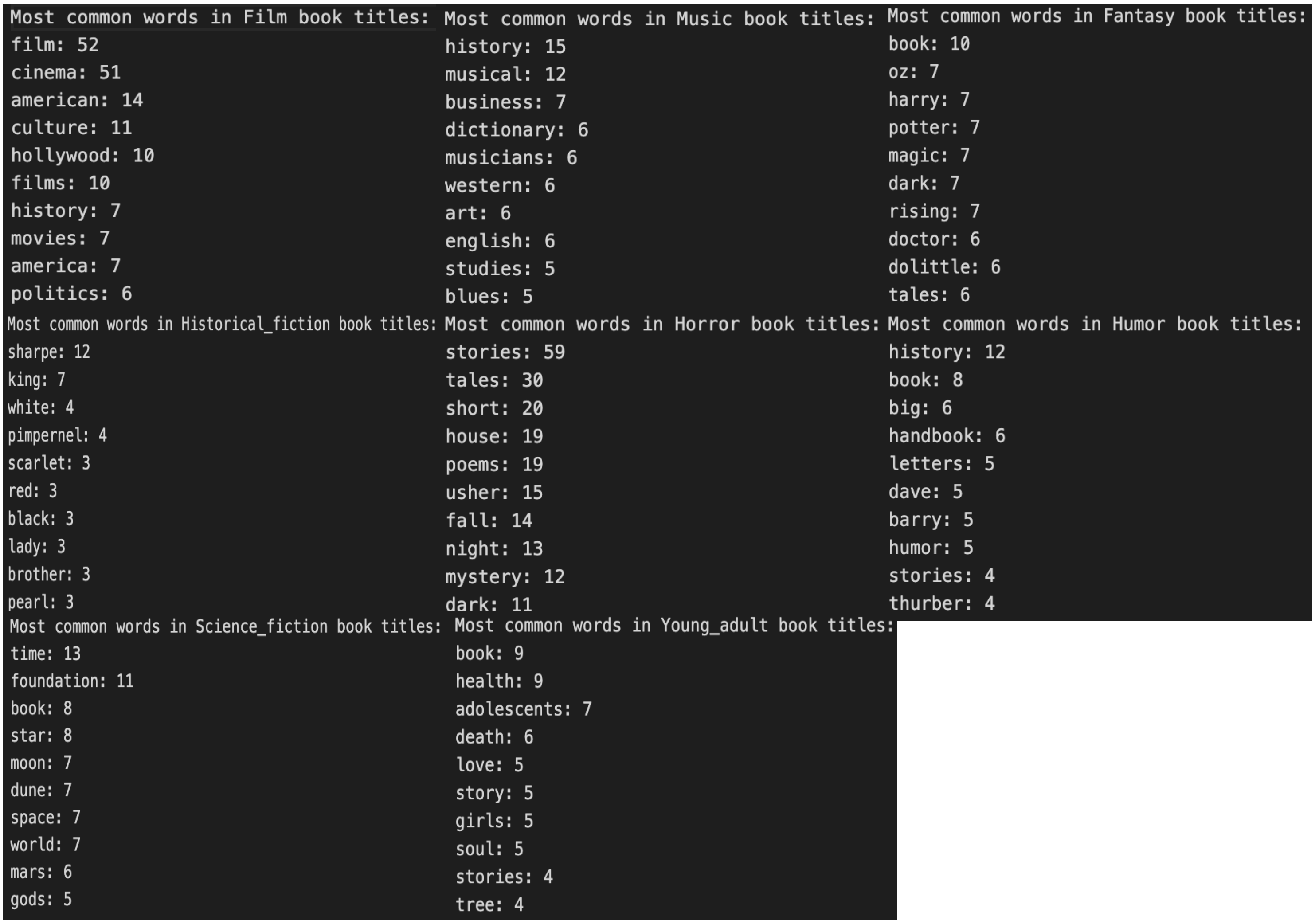
```
Most common words in Adventure movie titles:
part: 4
boy: 2
runaway: 2
challenge: 2
adventures: 2
robin: 2
hood: 2
fire: 2
strange: 2
magic: 2
```

```
Most common words in Family movie titles:
part: 7
magician: 3
house: 3
hare: 3
celebrities: 2
pressure: 2
friend: 2
mr: 2
leakey: 2
daffy: 2
```

```
Most common words in Documentary movie titles:
american: 7
live: 7
world: 5
story: 5
life: 5
years: 4
john: 4
trip: 3
west: 3
tv: 3
```

```
Most common words in Crime movie titles:
part: 12
big: 6
case: 5
time: 4
blood: 4
family: 3
run: 3
house: 2
murder: 2
blind: 2
```

Book Results:



```
Most common words in Film book titles:
film: 52
cinema: 51
american: 14
culture: 11
hollywood: 10
films: 10
history: 7
movies: 7
america: 7
politics: 6
```

```
Most common words in Music book titles:
history: 15
musical: 12
business: 7
dictionary: 6
musicians: 6
western: 6
art: 6
english: 6
studies: 5
blues: 5
```

```
Most common words in Fantasy book titles:
book: 10
oz: 7
harry: 7
potter: 7
magic: 7
dark: 7
rising: 7
doctor: 6
dolittle: 6
tales: 6
```

```
Most common words in Historical_fiction book titles:
sharpe: 12
king: 7
white: 4
pimpernel: 4
scarlet: 3
red: 3
black: 3
lady: 3
brother: 3
pearl: 3
```

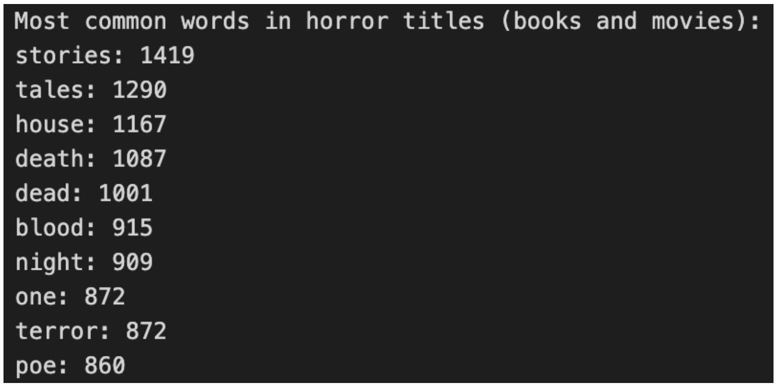
```
Most common words in Horror book titles:
stories: 59
tales: 30
short: 20
house: 19
poems: 19
usher: 15
fall: 14
night: 13
mystery: 12
dark: 11
```

```
Most common words in Humor book titles:
history: 12
book: 8
big: 6
handbook: 6
letters: 5
dave: 5
barry: 5
humor: 5
stories: 4
thurber: 4
```

```
Most common words in Science_fiction book titles:
time: 13
foundation: 11
book: 8
star: 8
moon: 7
dune: 7
space: 7
world: 7
mars: 6
gods: 5
```

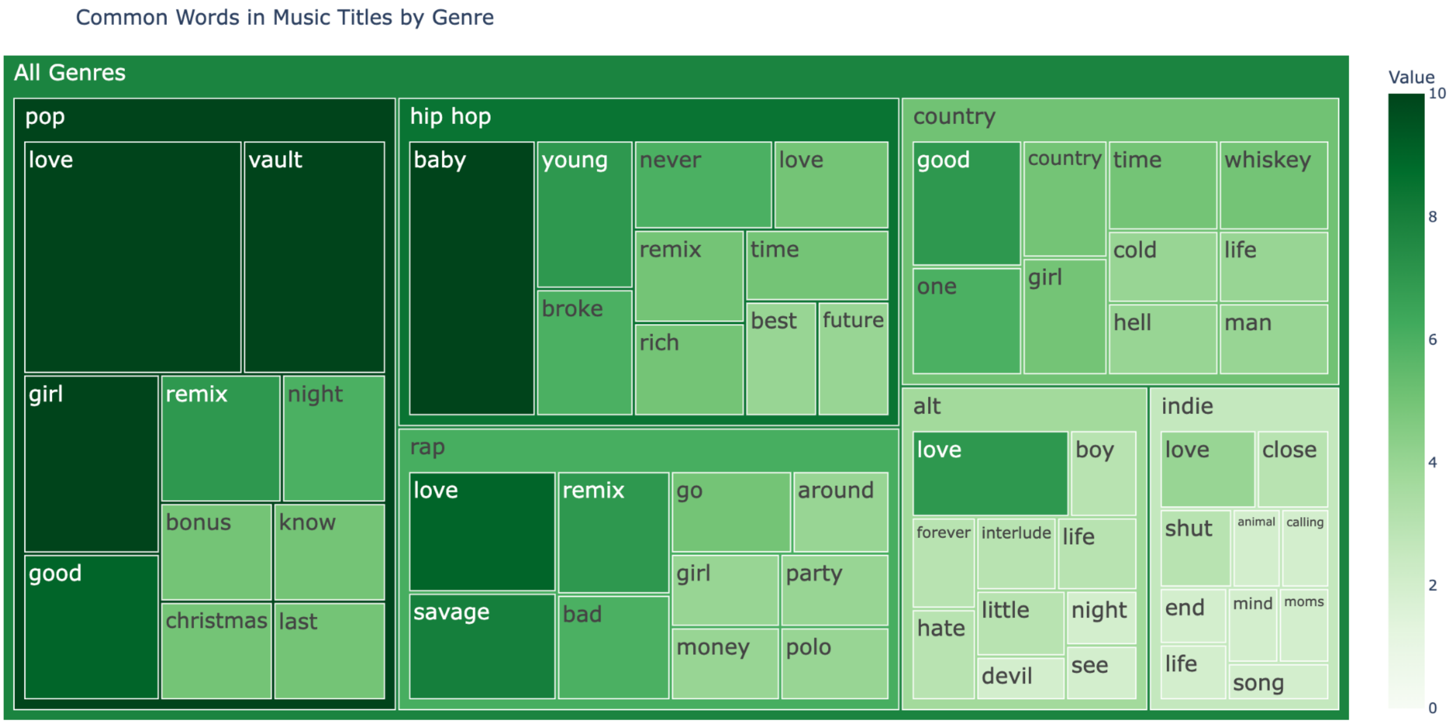
```
Most common words in Young_adult book titles:
book: 9
health: 9
adolescents: 7
death: 6
love: 5
story: 5
girls: 5
soul: 5
stories: 4
tree: 4
```

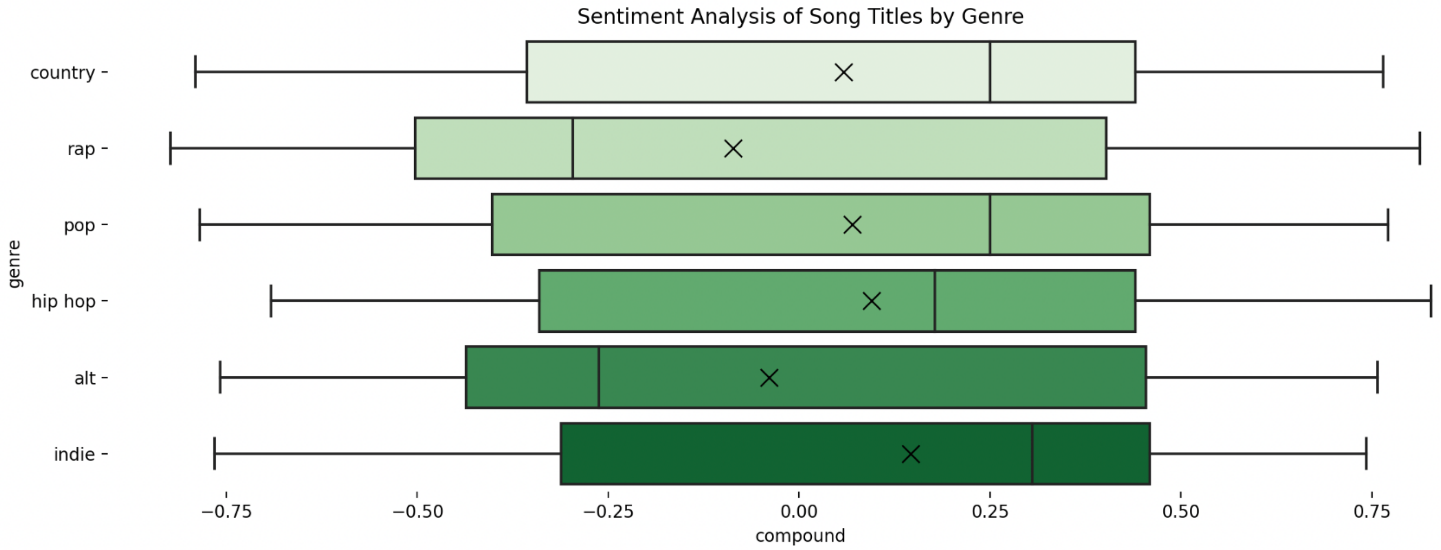
Book/Movie Horror Results:

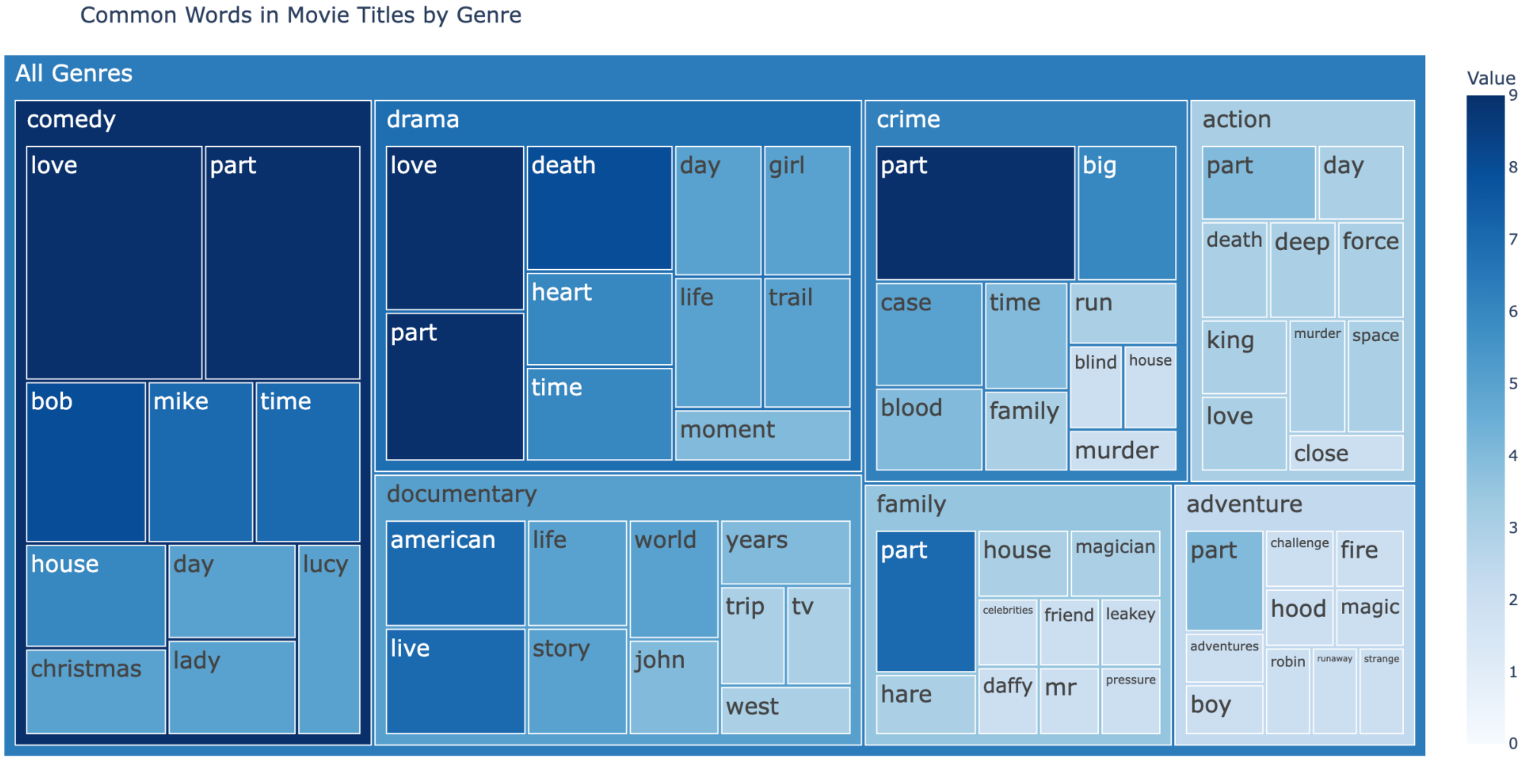


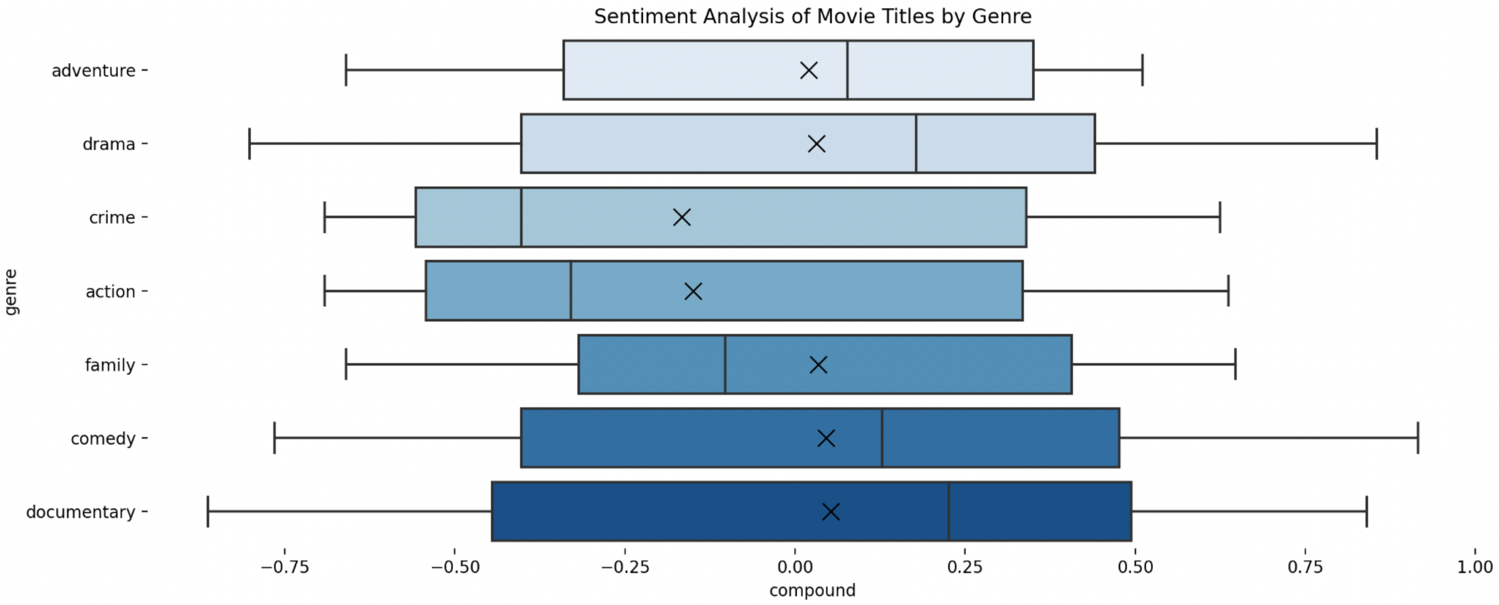
```
Most common words in horror titles (books and movies):
stories: 1419
tales: 1290
house: 1167
death: 1087
dead: 1001
blood: 915
night: 909
one: 872
terror: 872
poe: 860
```

V. Visualizations

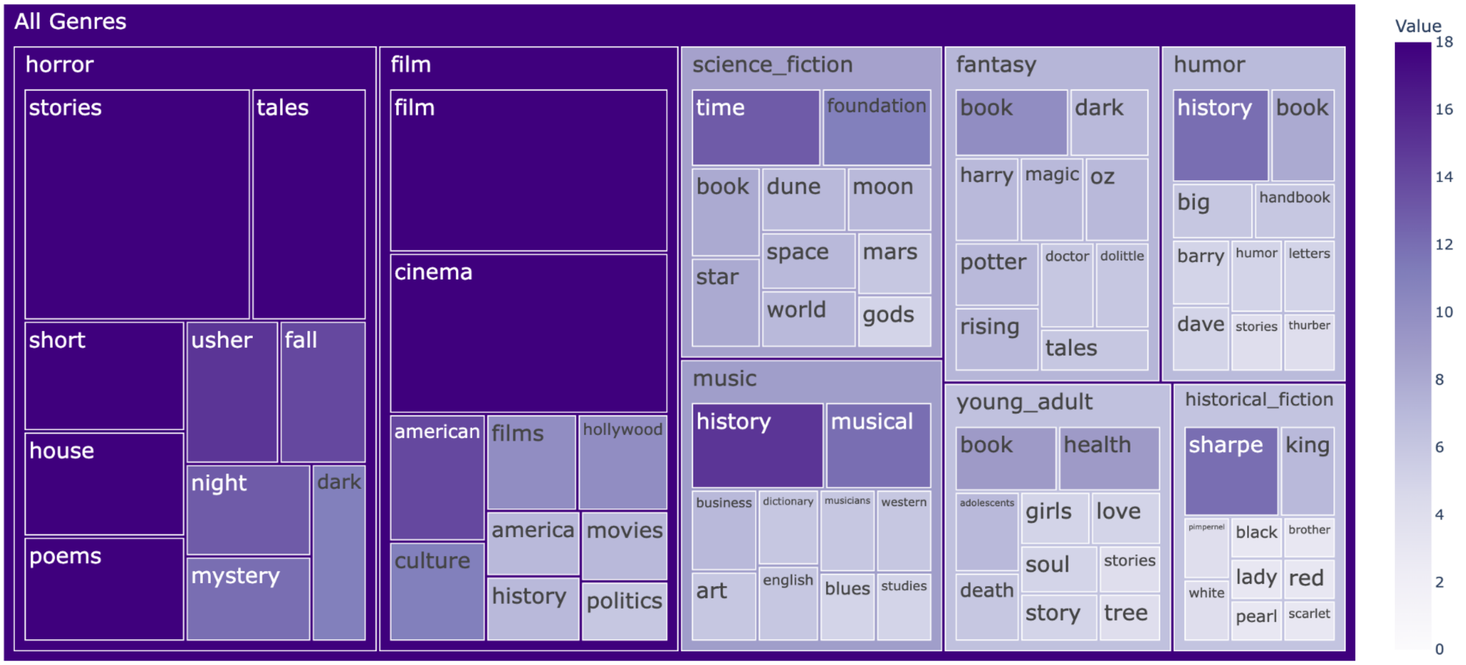




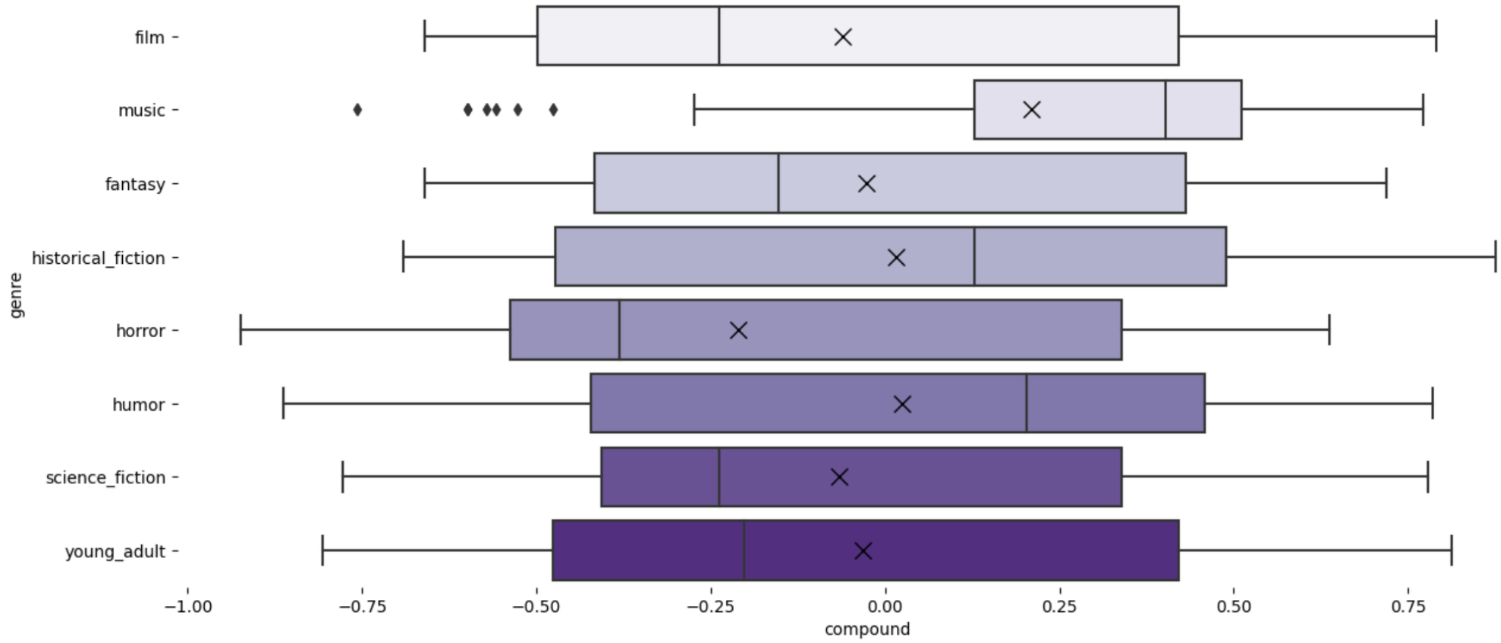

Common Words in Movie Titles by Genre
All Genres
comedy
love
part
bob
mike
time
house
day
lucy
christmas
lady
drama
love
death
day
girl
heart
life
trail
part
time
moment
documentary
american
life
world
years
trip
tv
live
story
john
west
crime
part
big
case
time
run
blind
house
blood
family
murder
family
part
house
magician
celebrities
friend
leakey
hare
daffy
mr
pressure
action
part
day
death
deep
force
king
murder
space
love
close
adventure
part
challenge
fire
hood
magic
adventures
robin
runaway
strange
boy
Value
9
8
7
6
5
4
3
2
1
0


Sentiment Analysis of Movie Titles by Genre
genre
adventure
drama
crime
action
family
comedy
documentary
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
compound

Common Words in Book Titles by Genre



Sentiment Analysis of Book Titles by Genre



VII. Instructions for running code

The first step is collecting all of the data and placing it in a database. There are three separate files for collecting data from each of the API's that we used. The 'spotify_api' is the file that you should start with, and collect as much data as you want. We have it limited at 25 tracks per runtime, but you can change that number to whatever you want it to be, and that is the case for all of these files as well. The Spotify file puts its data in a table called tracks within the database 'title_analysis'. You should then run the IMDb file (imdb_api), which will take around 6-7 minutes to collect 25 titles and genres. This happened every time we ran the code, so expect to wait for the data to load. That file puts movie titles in a table called movies within the title_analysis database. Then, run the open_library_api function, which puts 25 book titles in a table called books within the title_analysis database. After you run these files, ensure that the data is loaded correctly by viewing the tables in SQLite, or a related database server.

Your next step is to run the 'visualization_functions' file, which will generate treemap visuals for each category of data (tracks, movies, books). Within this file, you can specify certain genres that you want to see in the main() function, as well as edit words that you want excluded from the analysis. This file also writes the top 10 most common words by genre into a text file. Then, you can run the 'join_analysis' function, which looks at the most common words in horror titles across both books and movies, and writes the result into a text file.

Finally, you should run the analysis_function file, which will perform the sentiment analysis of the various titles. You can also change the genres that you want to see results for in the main() function. The results of the analysis_function file are visualizations for each media type. We also added another file called join_analysis, where we joined together titles from horror movies and books to compare them.

VIII. Documentation for each function that you wrote. This includes describing the input and output for each function

spotify_api:

create_table():

Description: Creates a table called *tracks* to store track titles and genres

Parameters:

- None

Returns:

- None

Example:

create_table()

fetch_tracks():

Description: Stores 25 tracks and their corresponding genres to the *tracks* table

Parameters:

- None

Returns:

- None

Example:

fetch_tracks()

imdb_api:

create_table():

Description: Creates a table called *movies* to store movie titles and genres

Parameters:

- None

Returns:

- None

Example:

create_table()

fetch_movies():

Description: Stores 25 movies and their corresponding genres to the *movies* table

Parameters:

- None

Returns:

- None

Example:

fetch_movies()

open_library_api:

create_table():

Description: Creates a table called books to store book titles and genres

Parameters:

- None

Returns:

- None

Example:

create_table()

fetch_books():

Description: Stores 25 books and their corresponding genres to the books table

Parameters:

- None

Returns:

- None

Example:

fetch_books()

visualization_functions:

common_words(db_name, file_name, genre_keywords, excluded_words, data_row):

Description: Create a list of the 10 most common words by genre

Parameters:

- db_name (table): name of the table that is providing data
- file_name (string) : file name for the results file
- genre_keywords (list): keywords to use as genres stored within a list
- excluded_words (list): words that should be excluded from analysis, stored in a list
- data_row (list): list of columns in the given table

Returns:

Dictionary containing the 10 most common words by genre

Example:

common_words(tracks, ‘spotify_output.txt’, gernre_keywords, excluded_words, [‘id’, ‘name’, ‘genre’])

visualizations(common_words_dict, color_scheme, color_range, title)

Description: Create a visualization of common words per genre

Parameters:

- common_words_dict (dict): dictionary containing the 10 most common words by genre
- color_scheme (string): indicates which color scheme the visualization should be
- color_range (tuple): indicates the range of the color scheme for the visualization
- title (string): title of the visualization

Returns:

Treemap visualization of the most common words per genre

Example:

visualizations(common_words_result, 'Purples', [0,18], 'Common Words in Book Titles by Genre')

analysis_function:

visualize_sentiment_analysis(db_name, genre_keywords, data_row, color_scheme, g_title):

Description: Create a visualization of common words per genre

Parameters:

- common_words_dict (dict): dictionary containing the 10 most common words by genre
- color_scheme (string): indicates which color scheme the visualization should be
- color_range (tuple): indicates the range of the color scheme for the visualization
- title (string): title of the visualization

Returns:

Treemap visualization of the most common words per genre

Example:

visualizations(common_words_result, 'Purples', [0,18], 'Common Words in Book Titles by Genre')

IX. Resources used

Date	**Issue Description**	**Location of Resource**	**Result (did it solve the issue?)**
12/2/2023	IMDb API running slow	chat.openai.com	Issue solved - prepped statement to insert data into the tables. Runtime improved by 2-3 minutes per run.
12/3/2023	General issues with sentiment analysis	https://www.kaggle.com/code/robikscube/sentiment-analysis-python-youtube-tutorial	Issue solved - learned how to use nltk SentimentIntensityAnalysis