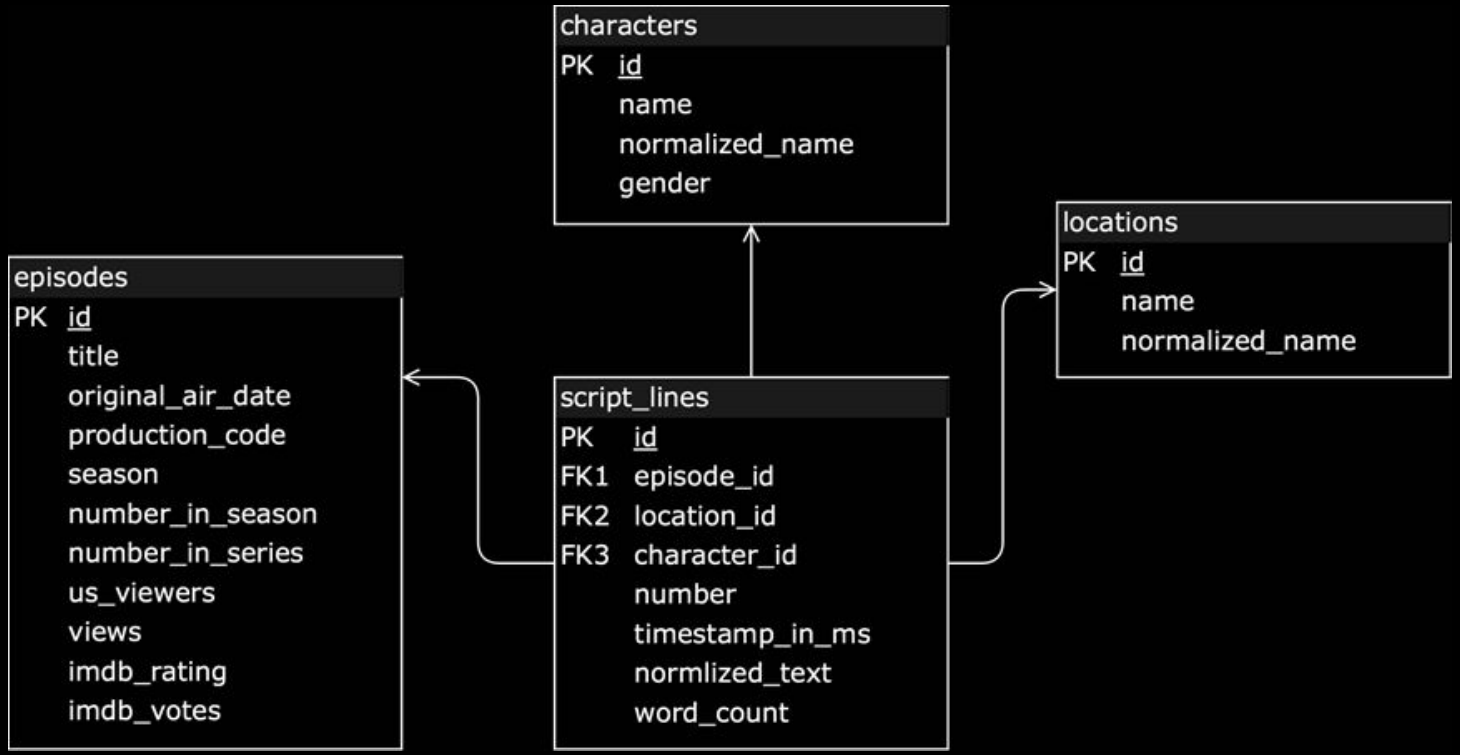
with R

data/simpsons


characters
PK id
name
normalized_name
gender
locations
PK id
name
normalized_name
episodes
PK id
title
original_air_date
production_code
season
number_in_season
number_in_series
us_viewers
views
imdb_rating
imdb_votes
script_lines
PK id
FK1 episode_id
FK2 location_id
FK3 character_id
number
timestamp_in_ms
normlized_text
word_count

INTRODUCTION
EXPLORE THE DATA SET

Q1: What does a record in each table represent?

Q2: What columns exist in each table and what do they contain?

Q3: How many records are in each table?

Q4: How are the tables connected?

Q5: Which columns contain text data?

GROUP WORK PART 1
TRANSFORM AND VISUALIZE

Q1: How many episodes debuted each year in the dataset?

Q2: What is the proportion of female to male characters?

Q3: What is the distribution of IMDB ratings?

Q4: How many records are in each table?

Q5: Which columns contain text data?

GROUP WORK PART 2
SEARCH

Q1: How often is Barack Obama mentioned in all episodes?

Q2: How often do characters mention Trump?

Q3: How often are *any* US presidents mentioned?

Q4: In which episodes do characters swear?

Q5: Which episodes reference Star Wars?

GROUP WORK PART 3
JOINS

Q1: How often do characters other than Homer say "Donut"?

Q2: In which location does Homer speak the most?

Q3: Who explains to Homer the German word "Schadenfreude"?

Q4: Is there an episode in which Lisa does not speak at all?

Q5: Which character speaks the most?

Q6: Who is the most frequent guest in the Simpson's home?

GROUP WORK PART 4
TOKENIZATION

Q1: Apply the five steps to tokenize the script lines into atomic words!

1. Filter the data to include only episodes from the first season

2. Apply transformations to clean and normalize the scripts

3. Split the script lines into words (or tokens)

4. Remove common english stop words

5. Add part of speech tags to your data

GROUP WORK PART 5
RULE-BASED CURSE IDENTIFICATION

Q1: Create a top 10 list of the characters who curse the most!

1. Create a theory-driven dictionary for curse words

2. Apply the dictionary to the tokenized script lines

3. Decide on a metric and aggregate keyword matches

4. Assign "curse" or "no curse" to each line based on the metric

5. Review the result and refine your dictionary

GROUP WORK PART 6
DO LLMS KNOW SIMPSONS?

Q1: Sample 5 subsequent lines from different episodes! Can a state-of-the-art LLM recognize the exact season and episode based on the script lines?

Q2: Sample 10 random lines from the same character and present them to a state-of-the-art LLM. Can it predict the character's name?