Vorstellung des Datensatzes

Der ausgewählte Datensatz, der vor allem die Skripte der 600 Folgen in den Staffeln 1 bis 28 (bis Folge 4) der Fernsehserie „The Simpsons“ umfasst, besteht aus 4 CSV Dateien. Die Struktur unserer Daten ist also tabellenartig und könnte deshalb nicht nur durch interpretative Suche, sondern auch durch Datenbankabfragen erforscht werden. Die Details der einzelnen Dateien sind in folgender Tabelle aufgeführt:

Dateiname	**Zeilen**	**Spalten (*Anzahl*)**
simpsons_script_lines.csv	157.462	id, episode_id, number, raw_text, timestamp_in_ms, speaking_line, character_id, location_id, raw_character_text, war_location_text, spoken_words, normalized_text, word_count (*13*)
simpsons_characters.csv	6.177	id, name, normalized_name, gender (*3*)
simpsons_episodes.csv	600	id, title, original_air_date, production_code, season, number_in_season, number_in_series, us_viewers_in_millions, views, imdb_rating, imdb_votes, image_url, video_url (*13*)
simpsons_locations.csv	3.144	id, name, normalized_name (*4*)

Der klare Fokus für dieses Projekt liegt auf der script_lines Datei, die die gesprochenen Worte eines jeden Charakters pro Redeanteil in den einzelnen Folgen umfasst und auf der sich somit inhaltlich nach bestimmten Topics suchen lässt. In dieser Datei sind allerdings auch Regieanweisungen verzeichnet, in diesem Fall ist der Wert der Spalte speaking_line „FALSE“, sodass über diesen Wert nicht in Regieanweisungen, sondern in tatsächlich gesprochenen Worten gesucht werden kann. Getan wird dies in der Spalte „normalized_text“. Zur Präsentation werden über die IDs der Charaktere, Orte und Folgen jedoch auch deren Werte erfasst. Als Dokumente des Datensatzes hätten einzelne Redeanteile, ganze Folgen oder Szenen deklariert werden können. Dabei erschien uns Letzteres als das beste Maß, sodass weder nur ein einzelner Redebeitrag noch eine Folge in ganzer Länge präsentiert wird. Die Unterteilung in Szenen erfolgt dabei durch den bereits erwähnten „FALSE“-Wert der Spalte speaking_line, durch die das aktuelle Gespräch beendet wird. Somit erhielten wir insgesamt 26.164 Szenen als Dokumente.