Term project 1

Gabriella Zsiros

Data Engineering 1

1. About the Data:

https://www.mavenanalytics.io/data-playground

The relational dataset contains 6 tables, which contains primary and foreign keys. It contains: Movies (containing the titles and some details of the Harry Potter movies), Chapters (of movies), Characters, Places, Spells, and Dialogues (actual dialogues lines from the movies). Detailed description of the data is in Data_Dictionary.csv

The central dataset was proven to be the database of dialogues, in which characters, places and chapters are referenced with foreign keys. Interestingly, the 'spells' data table does not connect to the other tables.

2. "Layer 0":

Although some basic analysis could have been performed on the dataset, I personally to wanted to have an extra twist and thought it a nice touch to add a layer of Sentiment Analysis to the dialogues.

Before starting to analyze my dataset, I wrote a short script, using 'SentimentAnalysis' R package, which gives each text snippet (dialogue) a sentiment score* between -1 and 1 [1 being maximally positive, -1 being maximally negative, 0 being neutral].

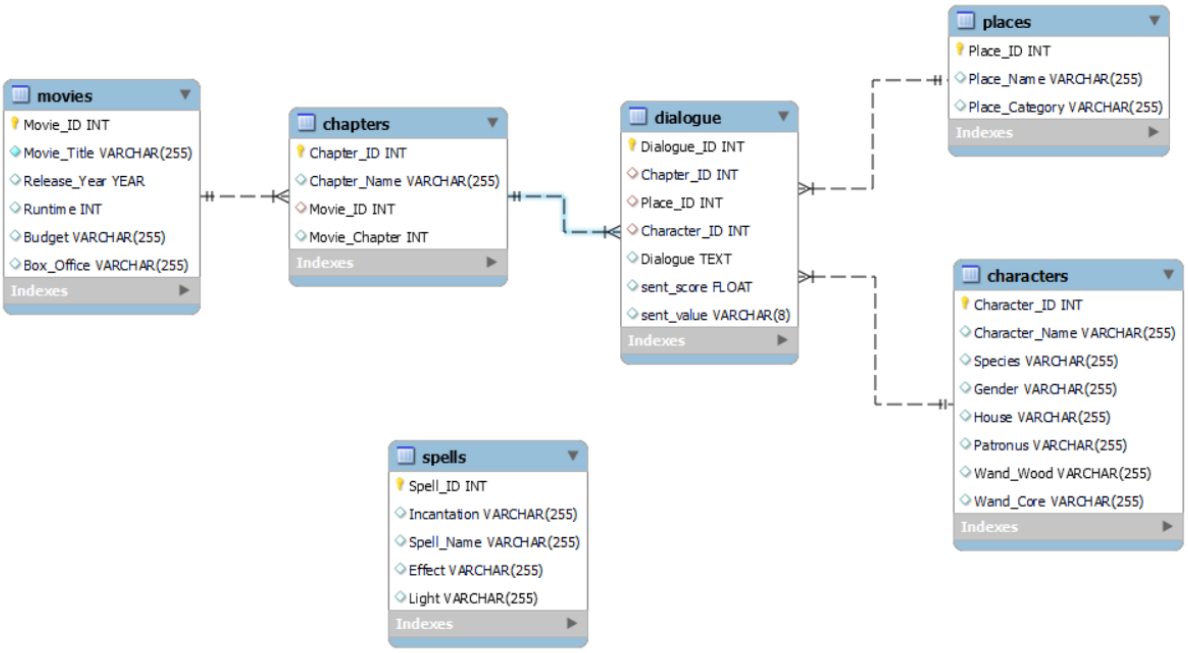
*The accuracy of the sentiment analysis is not subject to this term project

I modified my Dialogues database with the sentiment score (as well as a semantic evaluation for possible future use).

- Script used: hp_sentiment.R

3. Operational layer, modeling:

Hp_table create.sql loads the various csv files and sets up the relations between the datasets. (In some data table, I have already removed the header line in the source file, and in others, they were ignored).



- Model: hp_EER_model.mwb

3.1 Transformation

One data cleaning activity was necessary in the 'movies' database, before proceeding with the project. The observations for Budget and Box_Office variables were stored as text, comma separated at the thousand digits and marked with currency ($). As I was interested in the profit (i.e. the difference of box office and budget) for each movie, I added an extra column by transforming these two variables with the 'cast' and various text functions.

Scripts used

- hp_table_create.sql
- hp_dw.sql

4. Analytical Questions & Data marts

Given the applied sentiment analysis, I primarily approached the dataset from the Sentiment point of view, completing

- What are financial results for each movie? -> Box office view
- Which characters are the most positive -> Most positive characters view
  Taking the average sentiment score per character and limiting only to characters who have more than 50 lines in the movies.
- Which character are the most negative? -> Most negative characters view
  Note that by interpreting the results, one must consider that the sentiment analysis only focused

on the content of the text, therefore our antagonistic characters, who say misleadingly positive things, or pessimistic protagonists can distortive effect at first glance.

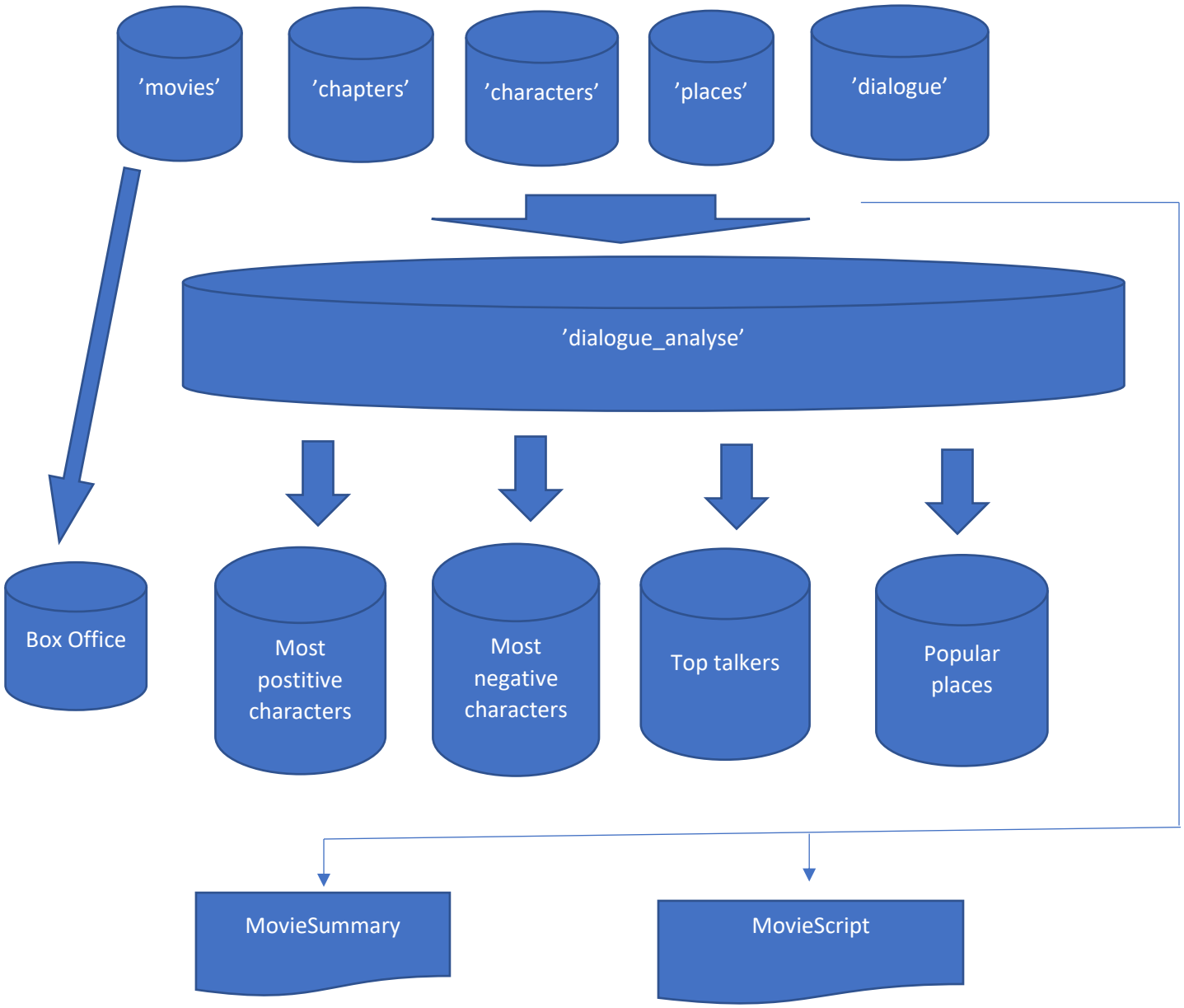
- Who talks the most? -> Top talkers view
- What are the most popular places -> Popular places view

- Script used: hp_datamarts.sql

5. Analytical layer

Given that the *dialogue* database was the one with most observations, and had foreign key to other tables as well, this is what I used as the basis of joining transformations. Mapping the actual values to the IDs from relational databases enables us to get a clear picture and interpretable data views.

5.1 Overview



6. Stored Procedures

Both of the stored procedures take an ‘x’ variable where ‘x’ is the movie number (1-8). MovieSummary creates a materialized view (which mostly has benefit on much larger datasets) with the main quantifiable characteristic of the data (number of characters, chapters, places,

- Script used: hp_stored_procedures.sql

7. Final thoughts on the analysis

Data marts coming from a static database might pose the of usability and operational impact, especially with a dataset like Harry Potter Dialogues. (Even though with a more refined sentiment analysis might open statistical possibilities to examine sentiments and financial performance for instance).

However, the outline of the process can easily be applied to such environment, where text and sentiments can have business impact. Let’s thinks about switching dialogues to call center conversations, characters to agents, places to teams, movies to departments and we might have a solid basis to gain significant operation insight in a business.

Summary of files and scripts:

- HP_movies / Chapters.csv, Characters.csv, Data_dictionary.csv, Dialogues.csv, Places.csv, Spells.csv, Movies.csv
- hp_sentiment.R
- hp_table_create.sql
- hp_dw.sql
- hp_datamarts.sql
- hp_stored_procedures.sql
- hp_EER_model.mwb