CS5200 Fall 2020: Practicum 2

Chandra Davis, Evan Douglass

Overview

We've decided to work with SQLite for this practicum. As such, to work with these files you will need SQLite installed on your machine. The data we are using is provided by IMDB at https://datasets.imdbws.com/

Download all the compressed tsv files and extract them into a folder called `data/`. The extracted files will all be called `data.tsv` so you will have to rename each. Rename them based on the compressed file name as follows in order to ensure they work with the R-notebook:

`<first_part>-<second_part>-data.tsv`

for example, `name-basics-tsv` and `title-ratings-data.tsv`

Setup

Before working with the data we need to create a place to store it. This section will set up any constants needed later and connect to a SQLite database.

```r
# TSV file path constants
tsv.prefix <- "data/"
tsv.postfix <- "-data.tsv"
NAME_BASICS <- paste(tsv.prefix, "name-basics", tsv.postfix, sep="")
TITLE_BASICS <- paste(tsv.prefix, "title-basics", tsv.postfix, sep="")
TITLE_EPISODE <- paste(tsv.prefix, "title-episode", tsv.postfix, sep="")
TITLE_RATINGS <- paste(tsv.prefix, "title-ratings", tsv.postfix, sep="")
TITLE_AKAS <- paste(tsv.prefix, "title-akas", tsv.postfix, sep="")
TITLE_CREW_DATA <- paste(tsv.prefix, "title-crew-data", tsv.postfix, sep="")
TITLE_PRINCIPLES <- paste(tsv.prefix, "title-principals", tsv.postfix, sep="")

# Setup SQLite database
library(RSQLite)
DB_NAME <- "imdb-data.db"
conn <- dbConnect(RSQLite::SQLite(), DB_NAME)
```