Import and Saving IMDb Datasets

This page includes instructions on how to import and save the IMDb datasets. The datasets we will be observing have the dimensions below. Note that these numbers will change as IMDb adds more data into their publicly available datasets.

Dataset	Observations	Variables
`basics`	11,654,015	9
`crew`	11,655,944	3
`principals`	92,540,008	6
`ratings`	1,568,336	3

1. First, download the datasets from IDMb at https://developer.imdb.com/non-commercial-datasets/.
2. Then, place all the tsv.gz files in the folder `tsv` in the working directory (root) of this project.
3. Uncompress all four `tsv.gz` files into `tsv` files.
4. Open this file (`import_save.Rmd`) and set your working directory to the root of this project.
5. Appropriately set and run the code chunks below in order to set the working directory and file directories if there are any mismatching directories. You can also set the number of threads working to hasten the importing process.

Setup

```r
# install.packages(data.table) # Uncomment to install if not yet installed
library(data.table)
gc() # Clear unused memory
setDTthreads(threads = 0) # Uncomment to set # of working CPUs: 0 for all
```

Import

Note that the principals dataset will take longer.

```r
# Set limit to number of rows imported: -1 for no limit
number.of.rows = -1
# Importing tsv files into environment
import_tsv <- function(tsv_dir = "tsv/") {
  basics <<- fread(paste0(tsv_dir, "title.basics.tsv"), sep = "\t",
                   na.strings = "\\N", nrows = number.of.rows, quote = "")
  crew <<- fread(paste0(tsv_dir, "title.crew.tsv"), sep = "\t",
                 na.strings = "\\N", nrows = number.of.rows, quote = "")
  principals <<- fread(paste0(tsv_dir, "title.principals.tsv"), sep = "\t",
                       na.strings = "\\N", nrows = number.of.rows, quote = "")
  ratings <<- fread(paste0(tsv_dir, "title.ratings.tsv"), sep = "\t",
                    na.strings = "\\N", nrows = number.of.rows, quote = "")
}
```

Subset by unique tconst

Since not all titles have available ratings, we will subset all datasets to the size of ratings by unique tconst values in `ratings`.

After the subsetting of datasets, the data will have the dimensions below:

Dataset	Observations	Variables
basics	1,568,336	9
crew	1,568,333	3
principals	21,693,507	6
ratings	1,568,336	3

```r
subset_data <- function() {
  rating_tconsts <- unique(ratings$tconst)
  basics <<- basics[tconst %in% rating_tconsts]
  crew <<- crew[tconst %in% rating_tconsts]
  principals <<- principals[tconst %in% rating_tconsts]
}
```

Save

```r
# Saving RDS from environment to directory
save_rds <- function(save_dir = "rds/") {
  saveRDS(basics, paste0(save_dir, "basics.rds"))
  saveRDS(crew, paste0(save_dir, "crew.rds"))
  saveRDS(principals, paste0(save_dir, "principals.rds"))
  saveRDS(ratings, paste0(save_dir, "ratings.rds"))
}
```

Run

```r
import_tsv(tsv_dir = "tsv/")
subset_data()
save_rds(save_dir = "rds/")
```

full_df

```
full_df
```