

## 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:

Dataset	Observations	Variables
basics	11,544,906	9
crew	11,544,906	3
principals	91,642,432	6
ratings	1,549,047	3

1. First, download the datasets from IMDb at <https://developer.imdb.com/non-commercial-datasets/>.
2. Then, place all the tsv files in a folder called `tsv` in the working directory (root) of this `import_save.Rmd` file.
3. Appropriately set and run the code chunks below in order to set the working directory and file directories. You can also set the number of threads working to hasten the importing process.

### Setup

```
# install.packages(data.table) # Uncomment to install if not yet installed
library(data.table)
gc() # Clear unused memory
setwd("~/Documents/06.sp25/04.stsci4100/stsci4100") # MODIFY TO CORRECT WD
setDTthreads(threads = 0) # Uncomment to set # of working CPUs: 0 for all
```

### Import

Note that the principals dataset will take longer.

```
# 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.

```
subset_data <- function(tsv_dir = "tsv/") {
  rating_tconst <- unique(ratings$tconst)
  basics <- basics[tconst %in% rating_tconst]
  crew <- crew[tconst %in% rating_tconst]
  principals <- principals[tconst %in% rating_tconst]
}
```

## Save

```
# Saving RDS from environment to directory
save_rds <- function(save_dir = "subset_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

```
import_tsv()
subset_data()
save_rds()
```