Práctica Netflix

Tecnologías para el Análisis de Datos Masivos

Daniel Ramos & Sergi Fornés

Importación y limpieza de los datos

Información sobre el Raw data

Vamos a realizar el análisis a partir de datos sobre valoraciones de películas de Netflix. Los datos que nos pueden resultar más interesantes son las notas de las valoraciones, la cantidad de valoraciones que tiene cada película y la identificación de quien hace estas valoraciones.

Obtenemos los datos a partir de 5 ficheros `.txt` y un `.csv`.

`combined_data_x.txt`

Tenemos a nuestra disposición 4 ficheros de este tipo: `combined_data_1.txt`, `combined_data_2.txt`, `combined_data_3.txt` y `combined_data_4.txt`. En estos ficheros tenemos información sobre valoraciones numéricas puestas por usuarios a películas de Netflix. Cada bloque de valoraciones está precedido por un número que indica la película (ID de la película). Para cada película existe un conjunto de valoraciones, de las cuales tenemos información sobre la fecha de la valoración, la nota (del 1 al 5) y un identificador del usuario.

`filas_ID_combined_all.txt`

En el fichero tenemos la localización de los bloques de valoraciones por películas en los ficheros `combined_data_x.txt`.

Columnas del fichero:

- `X1`: Identificación del bloque. Es el mismo nombre que aparece en los ficheros `combined_data_x.txt`.
- `fila`: Número de fila en el que se encuentra la identificación del bloque.
- `ID`: ID de la película asociada al bloque.
- `fila_final`: Última fila que contiene una valoración de la película `ID`.
- `data`: Número de fichero `combined_data_x.txt` al que pertenece la película `ID`. Tiene 4 valores posibles.

`movie_titles.csv`

En este archivo podemos encontrar información sobre las películas. Tenemos el ID de la película, su año de estreno y su nombre en inglés.

Obtención de las peliculas de interés

Decidimos las películas que vamos a analizar usando como semilla aleatoria nuestras dos últimas cifras del DNI.

```
set.seed(3793)
rand_pelis <- sort(sample(1:17770, 250, replace = FALSE))
```

Obtenemos del fichero `filas_ID_combined_all.txt` únicamente la información de las películas que nos ha tocado analizar.

```
loc_pelis <- read_csv("../data/filas_ID_combined_all.txt", col_types = "ciiii") %>%
  filter(ID %in% rand_pelis)
```

Creación del data frame

De cada uno de los archivos `combined_data_x.txt`, buscamos los datos de las películas que analizaremos y los metemos en un data frame, especificando en cada observación cual es el ID de la película.

```
i <- 1
df <- tibble()
for(comb_data in 1:4){
  file <- paste("../data/combined_data_", comb_data, ".txt", sep = "")
  while(loc_pelis[i,"data"] == comb_data & i <= 250){
    df_aux <- read_csv(file,
                       col_names = c("User","Score","Date"),
                       col_types = "iiD",
                       skip = loc_pelis[[i,"fila"]],
                       n_max = loc_pelis[[i,"fila_final"]] - loc_pelis[[i,"fila"]]) %>%
      mutate(ID_film = rand_pelis[i])
    df <- rbind(df, df_aux)
    i <- i + 1
  }
}
```

Cargamos el fichero con los nombres de las películas, tenemos en cuenta que el nombre de la película puede llevar , y arreglamos los valores `NA`. Este nuevo data frame cuenta con el ID de cada película, por lo que podemos unirlo con el data frame de las valoraciones de los usuarios para ampliarlo con información sobre las películas. Después lo guardamos en un fichero `.csv` para poder cargarlo y analizarlo en la siguiente sección.

```
names_pelis <- read_tsv("../data/movie_titles.csv",
                        locale = readr::locale(encoding = "ISO-8859-1"),
                        col_names = FALSE) %>%
  separate(col = X1,
           sep = ",",
           into = c("ID_film","Release_Year","Title"),
           extra = "merge",
           convert = TRUE) %>%
  mutate(Release_Year = ifelse(Release_Year == "NULL", NA, as.integer(Release_Year)))

df <- inner_join(df, names_pelis)

write_csv(df, "../data/pelis.csv")
```

Análisis exploratorio de los datos