**The Movie Lens dataset**

**Data cleaning (Please refer to “movies.ipynb” file)**

1. Removing the rows with no genres which reduces the shape of data frame from (62423, 3) to (57361, 3) (Cell 4).
2. Converting the type of genres column from data frame to list for different genres of movies (Cell 5).
3. Adding rows for a title of movie according to its list of genres (Cell 5). The result of the first ten rows is like below:

Table

Description automatically generated

1. Grouping the titles of movies according to genres column using “group by” (Cell 6). The result is as follows:

Table

Description automatically generated

1. To remove numbers, punctuations, conjunctions, auxiliary verbs, prepositions, and nonsense words using NLTK library, the function “str\_per\_genre” is created (Cell 7). The result is as follows:

A picture containing calendar

Description automatically generated

The results are stored in a dictionary named “dict\_titles\_per\_genre”.

**MapReduce job**

1. To use the titles of each genre as input to MapReduce job, text files called “titles\_i.txt” (0 <= i <= 18) are created (Cell 11).
2. To save the outputs of MapReduce job, a text file called “top10.txt” is created.
3. The implementation of MapReduce job (counting the words and giving the top ten frequent ones) is written into a python file called “map\_reduce.py”.
4. The outputs of top 10 frequent words are saved in “top10.txt” file.

**Commands to be executed in terminal:**

1. echo "genre\_name" >> “your\_path”/top10.txt
2. echo "==============================" >> “your\_path”/top10.txt
3. python “your\_path”\map\_reduce.py “your\_path”\”titles\_i.txt” >> “your\_path”\top10.txt