HW1 Code:

Part A

In [37]:

```
""" Take a csv file line, returns a list of its values """
def split_commas(x):
  splitted = []
  values = x.split(',')
  i = 0
  while i < len(values):
    if values[i].startswith('"'):
      new_val = values[i]
      if not (values[i].endswith('"')):
        while (i < len(values)-1) and not (values[i+1].endswith('"')):
          i+=1
          new_val = new_val + ',' + (values[i])
        i+=1
        if i < len(values):
          new_val = new_val + ',' + values[i]
      splitted.append(new_val.split('"')[1])
    else:
      splitted.append(values[i])
    i += 1
  return splitted
```

In [38]:

```
csv_path = "/content/500k_daily_prog_data.csv"
txt_path = "500k_daily_prog_data.txt"
separator = '||'

""" Read a csv file, write it to txt file with || separators instead of commas """
with open(csv_path, 'r') as csv_file, open(txt_path, 'w') as txt_file:
    for line in csv_file:
      new_line = split_commas(line)
      txt_file.write(separator.join(new_line))
```

In [39]:

```
! pip install mrjob
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/publi
c/simple/
Requirement already satisfied: mrjob in /usr/local/lib/python3.10/dist-packages (0.7.4)
Requirement already satisfied: PyYAML>=3.10 in /usr/local/lib/python3.10/dist-packages (f
rom mrjob) (6.0)
```

In [40]:

```
%%file hw1.py

from mrjob.job import MRJob
from mrjob.step import MRStep

class hw1(MRJob):
  # Condition 1:
  def filter_airtime(self, time):
    return int(time) >= 70000 and int(time) < 90000

  # Condition 2:
  def filter_genre(self, genre):
    flag = False
    for g in ['Talk', 'Politics', 'Spanish', 'Community', 'Martial arts']:
```

```
      if g in genre.split(","):
          flag = True
    return flag

  # Condition 3:
  def filter_title(self, title):
    flag = False
    for let in ['j', 'q', 'z']:
      if let in title:
        flag = True
    return flag

  def mapper(self, _, txt_row):
    row = txt_row.split("||")
    if row[0] != 'prog_code':
      if self.filter_airtime(row[4]) and self.filter_genre(row[2]) and self.filter_title
(row[1].lower()):
        yield (row[1], row[2]), row[3]

  def reducer(self, title_genre, dates):
    yield title_genre, len(set(dates))

if __name__ == '__main__':
  hw1.run()
```

```
Overwriting hw1.py
```

In [41]:

```
! python hw1.py < "500k_daily_prog_data.txt" > output.txt
```

```
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/hw1.root.20230517.092923.241692
Running step 1 of 1...
reading from STDIN
job output is in /tmp/hw1.root.20230517.092923.241692/output
Streaming final output from /tmp/hw1.root.20230517.092923.241692/output...
Removing temp directory /tmp/hw1.root.20230517.092923.241692...
```

In [42]:

```
""" Reads the output text and prints it according to the required format """
with open("output.txt", 'r') as txt_file:
  for line in txt_file:
      new_line = line.split("\t")
      sum_dates = new_line[1].strip()
      title_genre = new_line[0].split('"')
      title = title_genre[1]
      genre = title_genre[3]
      print(f'({title}, {genre}), ({sum_dates}, {len(genre.split(","))})')
```

```
(The Gossip Queens, Talk,Entertainment), (1, 2)
(The Hundred-Foot Journey, Spanish,Comedy-drama), (1, 2)
(The Josh Wolf Show, Talk,Comedy), (3, 2)
(The Late Late Show With James Corden, Talk,Comedy), (52, 2)
(The Queen Latifah Show, Talk,Variety), (48, 2)
(The Quest, Spanish,Action), (1, 2)
(The Tonight Show Starring Jimmy Fallon, Talk,Comedy), (9, 2)
(Todo Lo Que T\u00fa Quieras, Spanish,Drama), (1, 2)
(Town Square, Community), (1, 1)
(Un Balazo para Quintana, Spanish,Action), (1, 2)
(Una Mujer Para los S\u00e1bados, Spanish,Drama), (1, 2)
(Una Mujer Sin Amor, Spanish,Drama), (1, 2)
(Viaje Redondo, Spanish,Drama), (1, 2)
(WLJC Spring Telethon, Special,Community), (1, 2)
(Walking Tall: Lone Justice, Spanish,Action), (1, 2)
(What Would Julieanna Do?, Talk,Cooking), (1, 2)
(\u00bfQui\u00e9n Paga la Cuenta?, Spanish,Comedy), (1, 2)
(q, Talk,Entertainment,Variety), (8, 3)
(Jimmy Kimmel Live, Talk,Comedy), (9, 2)
(Jonathan Last on The Dadly Virtues, Special,Talk), (1, 2)
```

```
(Juan sin Miedo, Spanish,Drama), (2, 2)
(Judo Budapest Grand Prix 2014 Highlights, Special,Sports non-event,Martial arts), (1, 3)
(Justice With Judge Jeanine, Talk,News), (4, 2)
(La Maldici\u00f3n de la Momia Azteca, Spanish,Horror), (2, 2)
(La Masacre de los P\u00e9rez, Spanish,Drama), (1, 2)
(La Monja Alf\u00e9rez, Spanish,Drama), (1, 2)
(La Otra Mujer, Spanish,Drama), (1, 2)
(La Oveja Negra, Spanish,Drama), (1, 2)
(La Vida Dif\u00edcil de una Mujer F\u00e1cil, Spanish,Drama), (1, 2)
(La visita que no toc\u00f3 el timbre, Spanish,Comedy), (1, 2)
(Lamberto Quintero, Spanish,Drama), (1, 2)
(Last Week Tonight With John Oliver, Talk,Comedy), (8, 2)
(Last Week Tonight With John Oliver, Talk,Comedy,Interview), (2, 3)
(Late Night Joy, Talk), (1, 1)
(Lejos del Mundo, Spanish,Suspense,Drama), (1, 3)
(Life Today With James Robison, Talk,Religious), (5, 2)
(Liquidation Channel, Community,Consumer), (12, 2)
(Liquidation Channel, Special,Community,Consumer), (1, 3)
(Lo Azul del Cielo, Spanish,Drama,Romance,Suspense), (1, 4)
(Lo Mejor de Caso Cerrado, Law,Reality,Talk), (1, 3)
(Lo Mejor de la Madre Ang\u00e9lica, Talk,Religious), (15, 2)
(Los Campeones Justicieros, Spanish,Action), (1, 2)
(Los Fern\u00e1ndez de Peralvillo, Spanish,Drama), (1, 2)
(Los Hijos de Peralvillo, Spanish,Drama), (1, 2)
(Los Maistros: El D\u00eda de la Santa Cruz, Spanish,Comedy), (1, 2)
(Los de Abajo, Spanish,Historical drama), (1, 2)
(M\u00e1s Vale P\u00e1jaro en Mano, Spanish,Comedy), (1, 2)
(MediaBuzz, News,Talk,Public affairs,Politics), (2, 4)
(Mejor Estar Solo, Spanish,Comedy), (1, 2)
(Mojados, Spanish,Drama), (1, 2)
(Mojoe, Entertainment,Talk,Newsmagazine), (1, 3)
(Music for Change: The Global Citizen, Special,Music,Community), (1, 3)
(Operaci\u00f3n Jaque, Spanish,Drama), (1, 2)
(Para Usted, Jefa, Spanish,Drama), (1, 2)
(Programa do J\u00f4, Talk,Interview), (12, 2)
(Q & A, News,Talk,Interview), (1, 3)
(Q, Talk,Entertainment,Variety), (2, 3)
(Quadriga - The International Talk Show, Talk,Public affairs,Newsmagazine), (1, 3)
(Rosario Tijeras, Spanish,Crime drama,Romance), (1, 3)
(Santo y Mantequilla N\u00e1poles, Spanish,Action), (1, 2)
(Serpiente Azteca, Spanish,Drama), (1, 2)
(Soy el Hijo del Tah\u00far, Spanish,Action,Drama), (1, 3)
(St. Joe Live Presents, Community), (1, 1)
(State of Mine: Jim Hunt Story, Special,Community), (1, 2)
(The Daily Show With Jon Stewart, Talk,Interview,Comedy), (11, 3)
(The Dr. Oz Show, Talk,Health), (94, 2)
(2014 LBJ Civil Rights Summit, Community), (2, 1)
(7 Cajas, Spanish,Action,Suspense), (1, 3)
(Adventures of Johnny Tao: Rock, Action,Adventure,Martial arts), (1, 3)
(Al Rojo Vivo, Talk,Newsmagazine), (4, 2)
(Alicia Menendez Tonight, Talk,Politics), (1, 2)
(Amor y Frijoles, Spanish,Comedy-drama), (2, 2)
(Antiques Roadshow: In Bismarck, Collectibles,Community), (1, 2)
(Aqu\u00ed Nos Toc\u00f3 Vivir, Community,Travel), (1, 2)
(Arizona Horizon, Community), (6, 1)
(Arizona Horizon, Community,News), (19, 2)
(Around the Corner With John McGivern, Community), (2, 1)
(Arquitectos de lo Imposible, Community), (1, 1)
(Art Basel Design District Magazine, Community,Public affairs), (5, 2)
(Big Morning Buzz Live, Talk,Entertainment,News), (2, 3)
(Blazing Saddles, Spanish,Comedy,Western), (1, 3)
(Check Please! Arizona, Community), (2, 1)
(Choque de Opiniones, Talk,News,Debate), (3, 3)
(Cilantro y Perejil, Spanish,Romance-comedy), (1, 2)
(Cool Jobs, Community,Educational), (8, 2)
(Coruj\u00e3o do Esporte, Sports non-event,Talk), (1, 2)
(Crazy People, Spanish,Comedy), (1, 2)
(Crazy Talk, Comedy,Talk), (30, 2)
(Crimenes De Lujuria, Spanish,Drama,Suspense), (1, 3)
(Di\u00e1logos en Confianza, Talk), (1, 1)
(Dos Mojados En Apuros, Spanish,Comedy), (1, 2)
(Dulces Navajas, Spanish,Drama), (1, 2)
```

```
(Duro y Parejo en la Casita del Pecado, Spanish,Comedy), (1, 2)
(Ek The Raja Ek Thi Rani, Community), (4, 1)
(El Baile de San Juan, Spanish,Historical drama), (1, 2)
(El Cuerpazo del Delito, Spanish,Comedy-drama), (1, 2)
(El Efecto Tequila, Spanish,Comedy-drama), (1, 2)
(El Esqueleto de la Se\u00f1ora Morales, Spanish,Comedy-drama), (1, 2)
(El Joven Ju\u00e1rez, Spanish,Biography), (1, 2)
(El Joven del Carrito, Spanish,Comedy), (1, 2)
(El Oreja Rajada, Spanish,Drama), (1, 2)
(El Palenque, Talk), (1, 1)
(El Rediezcubrimiento de M\u00e9xico, Spanish,Comedy-drama), (1, 2)
(El Santos vs la Tetona Mendoza, Spanish,Comedy,Animated), (1, 3)
(El Santos vs. la T...a Mendoza, Spanish,Comedy,Animated), (1, 3)
(El Tejedor de Milagros, Spanish,Drama), (1, 2)
(El Tigre de Guanajuato, Spanish,Adventure), (1, 2)
(El Vizconde de Montecristo, Spanish,Comedy), (1, 2)
(El mejor, Spanish,Drama), (1, 2)
(Esos de P\u00e9njamo, Spanish,Drama), (1, 2)
(Estrella Sin Luz, Spanish,Drama), (1, 2)
(Fallaste Coraz\u00f3n, Spanish,Drama), (1, 2)
(Fashionably Late With Rachel Zoe, Talk,Fashion), (1, 2)
(Godzilla, Spanish,Science fiction,Action,Adventure), (1, 4)
(Hagit - Designer Jewelry, Shopping,Talk), (1, 2)
(Israel Tour June 2015, Community), (5, 1)
(Izrail' Plyus Predstavlyaet, Community), (1, 1)
(J. Edgar, Spanish,Biography,Historical drama), (1, 3)
(Jack Holt At The River, Religious,Community), (1, 2)
(Jade Warrior, Spanish,Action,Adventure,Martial arts), (1, 4)
(Jagged Edge, Spanish,Crime drama), (1, 2)
(Jayne Mansfield's Car, Spanish,Drama), (1, 2)
(Jeepers Creepers 2, Spanish,Horror), (1, 2)
(Jerry Springer, Talk), (109, 1)
```

Part B

In [43]:

```
! pip install pyspark
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/publi
c/simple/
Requirement already satisfied: pyspark in /usr/local/lib/python3.10/dist-packages (3.4.0)
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.10/dist-packages
(from pyspark) (0.10.9.7)
```

In [44]:

```
import pyspark
from pyspark.sql import SparkSession
from pyspark.mllib.random import RandomRDDs
from pyspark.sql.types import*
```

In [45]:

```
if 'spark' in dir():
    print("spark context is already created for you!")
else: print("You need to create your own SparkSession object")
```

```
spark context is already created for you!
```

In [46]:

```
spark = SparkSession.builder.appName('hw1').getOrCreate()
sc = spark.sparkContext
```

In [47]:

```
path = "/content/500k_daily_prog_data.csv"
data_rdd = sc.textFile(path)
```

In [48]:

```
""" Removing header and turns every line separated with commas to a line separated with |
| """
header = data_rdd.first()
new_data_rdd = data_rdd.filter(lambda row: row != header)\
                        .map(lambda x: split_commas(x))
```

In [49]:

```
""" Adding points to score by required conditions """
def score_calc(genre, duration):
  score = 0
  if len(genre) == 1:
    if genre[0] == "Sitcom":
      score += 5
    elif genre[0] == "Action" or genre[0] == "Documentary":
      score += 90
  return score + float(duration)/10

""" Filtering lines by required conditions """
def title_filter(title):
  wont_watch = ["big", "the", "bang", "theory", "community", "almanac", "met", "mother",
                "your", "city", "anatomy", "game", "thrones", "guy", "family", "friends"
, "senate"]
  for t in wont_watch:
    if t in title.split():
      return False
  return True
```

In [50]:

```
title_genre_result = new_data_rdd.filter(lambda x: title_filter(x[1].lower()))\
                                  .map(lambda x: ( (x[1], x[2]), score_calc(x[2].split("
,"), x[5]) ))\
                                  .reduceByKey(lambda  x, y: x + y)\
                                  .sortBy(lambda t: t[1], ascending=False)\
                                  .take(25)
```

In [51]:

```
""" Printing output according to required format """
for tg in title_genre_result:
  print("{", end='')
  print(tg[0][0], end='')
  for g in tg[0][1].split(","):
    print(',', end=' ')
    print(g, end='')
  print('} |', end=' ')
  print(tg[1])
```

```
{SIGN OFF, Special} | 30957.19999999996
{Documentary, Documentary} | 13776.100000000013
{Two and a Half Men, Sitcom} | 11316.400000000001
{Everybody Loves Raymond, Sitcom} | 11220.800000000005
{ABC World News Now, News} | 10671.000000000033
{Mike & Molly, Sitcom} | 9951.899999999998
{Weather Radar, Weather} | 9504.0
{Paid Programming, Shopping} | 9086.500000000022
{NHL Hockey, Sports event, Hockey} | 9072.0
{Hot in Cleveland, Sitcom} | 8790.5
{Anger Management, Sitcom} | 8503.0
{Seinfeld, Sitcom} | 8286.7
{Rules of Engagement, Sitcom} | 8276.600000000002
{MLB Baseball, Sports event, Baseball} | 8269.5
{Un Mundo Maravilloso, Documentary} | 8013.0
{Strange Inheritance, Documentary} | 7998.0
{Drug Wars, Documentary} | 7812.0
{NBA Basketball, Sports event, Basketball} | 7507.5
{Classic Arts Showcase, Art} | 7459.0
{Local Weather, Weather} | 7098.0
{Cougar Town, Sitcom} | 6910.700000000001
```

```
{Cougar Town, Sitcom} | 6919.700000000001
{Smooth Jazz, Music} | 6354.0
{Urban Beat, Music} | 6054.0
{True Life, Documentary} | 5969.199999999999
{MC Rap, Music} | 5913.0
```

In [ ]:

```
```