Web Scrapping Project (Harsh)

In [2]:

```
# Importing All Required Libraries.
import requests
from bs4 import BeautifulSoup
import re
import pandas as pd
headers = {'User-Agent': 'Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, li
```

In [3]:

```
# Code for all pages in one List.
url_list = []
base_url = 'https://www.themoviedb.org/movie?page='
for item in range(1,51):
    url_list.append(base_url+str(item))
```

In [4]:

```
# Code for Individual Movie Link in one list
preffix_url = "https://www.themoviedb.org"
movieurl_list = []
for url in url_list:
    src_data = requests.get(url,headers = headers).text
    soup_data = BeautifulSoup(src_data,"lxml")
    all_divs = soup_data.find_all('div',class_="card style_1")
    for anchor in all_divs:
        each_url = anchor.find('a')['href']
        movieurl_list.append(preffix_url+str(each_url))
```

In [5]:

```
All_Movie_Data = []
#For Each Movie URL.
for url in movieurl_list:
    src_data = requests.get(url,headers = headers).text
    soup_data = BeautifulSoup(src_data,"lxml")
    all_content = soup_data.find('div',class_="single_column")
    # For Each Movie Name.
    title_unclean = all_content.find('h2').text
    title = re.sub("[\(].*?[\)]", "",title_unclean).strip()
    # For Each Movie Rating.
    rating = all_content.find('div',class_="user_score_chart")['data-percent']
    # For Each Movie Genre.
    genre = all_content.find('span',class_="genres").text.strip()
    # For Each Movie Release Date.
    unclean_release_date = all_content.find('span',class_="release").text
    release_date = re.sub("[\(].*?[\)]", "", unclean_release_date).strip()
    #For Each Movie Runtime.
    def RUN():
        try:
            run_time = all_content.find('span',class_="runtime").text.strip()
            return(run_time)
        except AttributeError:
            return("N/A")
    runtime = RUN()
    # For Each Movie Director.
    movie_character = []
    movie_dir_lst = []
    src_director = all_content.find_all('li',class_="profile")
    for i in src_director:
        a = i.find('p',class_="character").text.strip()
        b = i.find('p').text.strip()
        movie_character.append(a)
        movie_dir_lst.append(b)
    c = 0
    for char in movie_character:
        if("Director" in char):
            c = movie_character.index(char)
    def DIR():
        try:
            return(movie_dir_lst[c])
        except IndexError:
            return("N/A")
    director = DIR()
    #For all Data to be put in Dictionary inside List.
    data ={"Name":title,
        "Rating":rating,
        "Genre":genre,
        "Release date":release_date,
        "Runtime":runtime,
        "Director":director,
        "Url":url}
    All_Movie_Data.append(data)
```

In [6]:

```
#Converting Data to Excel format using Pandas.
df = pd.DataFrame(All_Movie_Data)
df.to_excel("WebScrapping_excel.xlsx")
```

In [ ]:

```
#Converting Data to CSV format
import csv
keys = All_Movie_Data[0].keys()
with open('WebScrapping_CSV.csv','w',newline='') as op_file:
    dict_writer = csv.DictWriter(op_file,keys)
    dict_writer.writeheader()
    dict_writer.writerows(All_Movie_Data)
```