# IMDB Data Processing

Download datset from IMDB website (https://www.imdb.com/interfaces/ (https://www.imdb.com/interfaces/)) so that we can check if the word (or phrase) we found is an actor, an actress, a director or a title

All imports go here

In [1]:

```
import csv
import pickle
```

Load raw data
name_data.tsv --> name.basics.tsv.gz (https://datasets.imdbws.com/name.basics.tsv.gz (https://datasets.imdbws.com/name.basics.tsv.gz))
title_data.tsv --> title.akas.tsv.gz (https://datasets.imdbws.com/title.akas.tsv.gz (https://datasets.imdbws.com/title.akas.tsv.gz))

In [2]:

```
actor = []
actress = []
director = []
person = []
title = []
data = []
with open('name_data.tsv', 'r', encoding = 'utf-8') as data_file:
    reader = csv.reader(data_file, delimiter="\t")
    for row in reader:
        data.append(row)
for i in range(1, len(data)):
    row = data[i]
    job_titles = row[4].split(',')
    is_famous = False
    if 'actor' in job_titles:
        actor.append(row[1])
        is_famous = True
    if 'actress' in job_titles:
        actress.append(row[1])
        is_famous = True
    if 'director' in job_titles:
        director.append(row[1])
        is_famous = True
    if is_famous:
        person.append(row[1])
print(actor[:10])
with open('title_data.tsv', 'r', encoding = 'utf-8') as data_file:
    reader = csv.reader(data_file, delimiter="\t")
    for row in reader:
        data.append(row)

for i in range(1, len(data)):
    row = data[i]
    if row[3]=='US':
        title.append(row[2])
print(title[:10])
```

```
['Fred Astaire', 'John Belushi', 'Ingmar Bergman', 'Humphrey Bogart', 'Marlon Brand
o', 'Richard Burton', 'James Cagney', 'Gary Cooper', 'James Dean', 'Kirk Douglas']
['Carmencita', 'The Clown and His Dogs', 'Blacksmithing Scene', 'Blacksmith Scene #
1', 'Blacksmithing', 'Blacksmith Scene', 'Chinese Opium Den', 'Corbett and Courtney
Before the Kinetograph', 'The Corbett-Courtney Fight', 'Jim Corbett vs. Peter Courtn
ey']
```

Convert each list to a set for fast indexing

In [3]:

```
actor = set(actor)
actress = set(actress)
director = set(director)
person = set(person)
title = set(title)
```

Check if each set works as expected

In [4]:

```
print('Daniel Craig' in actor)
```

True

In [5]:

```
print('Scarlett Johansson' in actress)
```

True

In [6]:

```
print('The Godfather' in title)
```

True

In [7]:

```
print('James Cameron' in director)
```

True

In [8]:

```
print('Chris Evans' in person)
```

True

Save the dictionarys to a pickle file

In [9]:

```
dataset = [actor,actress,director,person,title]
with open('imdb_data.pkl', 'wb') as save_file:
    pickle.dump(dataset, save_file, protocol=pickle.HIGHEST_PROTOCOL)
```