Web Scraping Stargate Scripts

https://imsdb.com/TV/Stargate%20SG1.html (https://imsdb.com/TV/Stargate%20SG1.html)

In [1]:
```
import requests
from bs4 import BeautifulSoup
from bs4 import element
import re
import pandas as pd
import pickle
import json
```

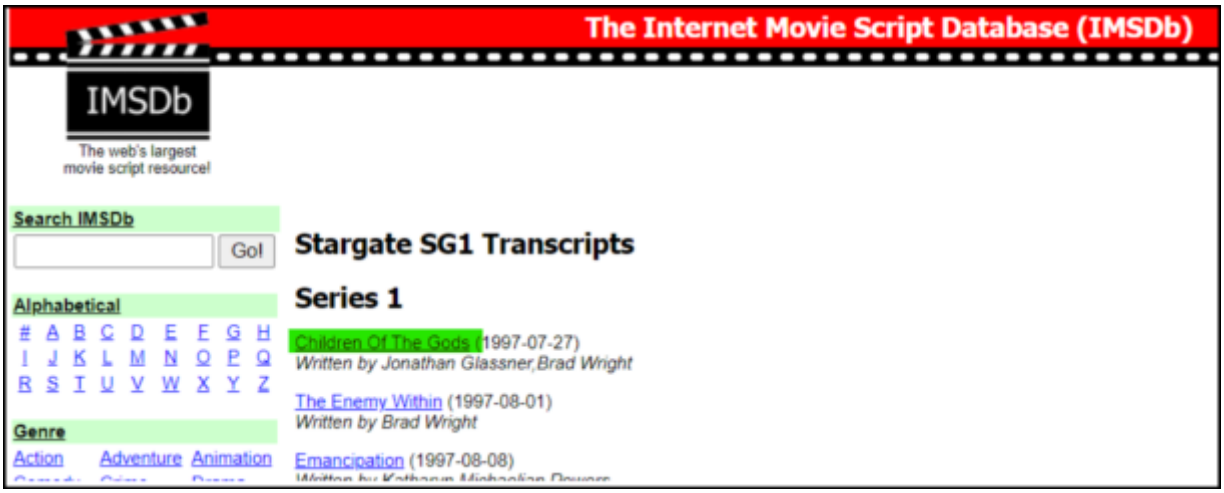
...

In [2]:
```
pd.set_option('display.max_columns', None)  # or 1000
pd.set_option('display.max_rows', None)  # or 1000
pd.set_option('display.max_colwidth', None)  # or 199
```

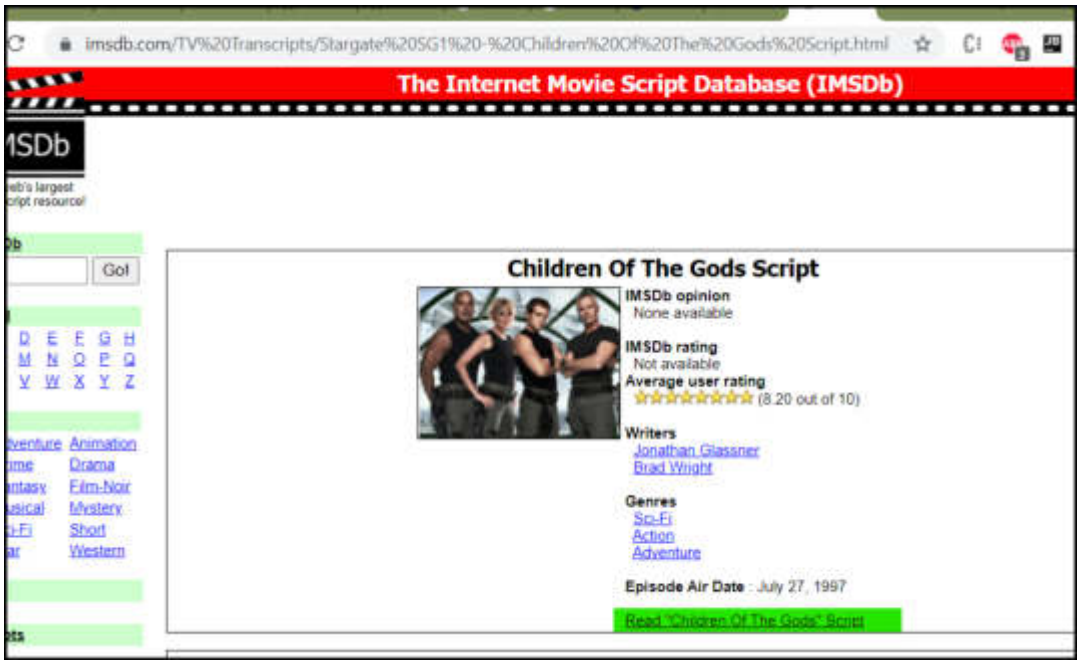
Obtaining the Links to the Scripts

https://imsdb.com/TV/Stargate%20SG1.html (https://imsdb.com/TV/Stargate%20SG1.html)

1) Obtain all episode links



2) Obtain link for script



In [3]:
```
# Obtain HTML code of Main Page
# source: https://stackoverflow.com/questions/45204152/beautifulsoup-pull-p-tag-while-between-defined-h2-tags
URL = "https://imsdb.com/TV/Stargate%20SG1.html"
page = requests.get(URL)
#page.text
```

In [4]:
```
# Parse HTML page with beautiful soup
soup = BeautifulSoup(page.content, "html.parser")

# h2 tags with season name
seasons = [t.text for t in soup.find_all("h2")]

# Extractiong the links for all episodes
# --List comprehension with python dictionary
# --Dictionary will store the episode linkes sorted by season
episode_links = {season_nr:{} for season_nr in range(1,len(seasons)+1,1)}
season_count = 1
episode_count = 1

# Populate the episode link dictionary
for season_tag in soup.find_all("h2"):
    episode_count = 1
    # Get all episodes that follow a "season headline" (see image 1 above)
    for tag in season_tag.next_siblings:
        if isinstance(tag, element.Tag):

            # go to next season
            if season_count < 9 and tag.text == seasons[season_count]: # check if the current tag contains next "season headline"
                    season_count+=1
                    break

            # go to next episode
            if  tag.find("a"):
                episode_links[season_count][episode_count] = "https://imsdb.com" + tag.find("a")['href']
                episode_count+=1

del episode_links[6][14] # episode is missing
#episode_links
```

Obtaining the Scripts' HTML → Creating Pandas Dataframes

In [5]:
```
def get_script(soup:BeautifulSoup)->element.Tag:
    """Obtain a script's HTML code"""
    for tag in soup.find_all('a'):
        if re.match(r"Read\s\".*\"\sScript", tag.text) and isinstance(tag, element.Tag):
            URL = "https://imsdb.com"+tag['href']
            page = requests.get(URL)
            soup = BeautifulSoup(page.content, 'html.parser')
            script = soup.find("pre").find("body")
            return script
    return None
```

In [6]:
```
def string_empty(string:str)->bool:
    """Check if string is empty or empty space characters"""
    # source: https://thispointer.com/python-check-if-string-is-empty-or-blank-or-contain-spaces-only/
    return not string.text.strip() or string.text.isspace()

def is_tag(tag)->bool:
    return isinstance(tag, element.Tag)

def is_text(text)->bool:
    return isinstance(text, element.NavigableString)

# Very rudimental data cleaning
def remove_newline(string:str)->str:
    return string.replace("\n", "").replace("\r", "")
```

In [7]:
```
def script_to_df(script: element.Tag):
    """Extract the df frrom a html movie script.
        Puts the content into a two column dataframe with one column marking the speaking character
        and another columb which contains the spoken text."""
    script_table = []
    for tag in script.find_all("b"):
        if is_tag(tag) and string_empty(tag):
            sibling = tag.next_sibling
            # get character dialouge
            if is_tag(sibling) and not string_empty(sibling):
                script_table.append([remove_newline(str(sibling.text)), remove_newline(str(sibling.next_sibling))])
            # get intermediate text (all text that is not a character speaking)
            elif is_text(sibling):
                script_table.append(["interlude", remove_newline(str(sibling))])
    script_df = pd.DataFrame(script_table)
    script_df.columns = ["character", "text"]
    return script_df
```

In [8]:
```
# Dictionary that contains the links to the dataframes sorted by season and epispde number
corpus_dict = {season_nr:{} for season_nr in range(1,len(seasons)+1,1)}
#corpus_dict
```

In [9]:
```
raw_data_folder = "../raw_data/"
```

In [10]:
```
# Create data
for season_nr, episodes in episode_links.items():
    for episode_nr, episode_link in episodes.items():

            # Create DataFrame
            page = requests.get(episode_link)
            soup = BeautifulSoup(page.content, 'html.parser')
            script = get_script(soup)
            script_df = script_to_df(script)

            # Save DataFrame pickle
            filename = "S"+str(season_nr)+"-E"+str(episode_nr)+".pkl"
            script_df.to_pickle(raw_data_folder+filename)
            corpus_dict[season_nr][episode_nr] = filename
```

In [11]:
```
# pickle corpus (store)
# source: https://stackoverflow.com/questions/11218477/how-can-i-use-pickle-to-save-a-dict
with open(raw_data_folder+'/corpus_dict.pkl', 'wb') as handle:
    pickle.dump(corpus_dict, handle, protocol=pickle.HIGHEST_PROTOCOL)
```

In [12]:
```
# Load data (deserialize)
with open(raw_data_folder+'/corpus_dict.pkl', 'rb') as handle:
    corpus_dict = pickle.load(handle)

json.loads(json.dumps(corpus_dict))
```

Out[12]:

```
{'1': {'1': 'S1-E1.pkl',
  '2': 'S1-E2.pkl',
  '3': 'S1-E3.pkl',
  '4': 'S1-E4.pkl',
  '5': 'S1-E5.pkl',
  '6': 'S1-E6.pkl',
  '7': 'S1-E7.pkl',
  '8': 'S1-E8.pkl',
  '9': 'S1-E9.pkl',
  '10': 'S1-E10.pkl',
  '11': 'S1-E11.pkl',
  '12': 'S1-E12.pkl',
  '13': 'S1-E13.pkl',
  '14': 'S1-E14.pkl',
  '15': 'S1-E15.pkl',
  '16': 'S1-E16.pkl',
  '17': 'S1-E17.pkl',
  '18': 'S1-E18.pkl',
  '19': 'S1-E19.pkl',
```


In [13]:
```
# Example df
with open(raw_data_folder+corpus_dict[2][3], 'rb') as handle:
    unserialized_data = pickle.load(handle)

pickle.loads(pickle.dumps(unserialized_data))
```

Out[13]:

	character	text
0	JACK	All right kids. We re due back, unless you can tell me you ve discovered something Earth shattering, I m read to bag this one.
1	DANIEL	We practically just got here. We have no idea what this planet has to offer.
2	JACK	Trees and Moss.
3	DANIEL	Well, a few miles from the Stargate, granted, but...
4	JACK	Captain any signs of radio traffic in the last twenty-four hours? Any chemical traces in the air to indicate civilization?
5	SAM	None, sir.
6	interlude	Daniel tries to speak and puts his finger in the air. Jack points his own finger
7	JACK	Ah! Ah! We ll flag it for an aerial survey. As much as I love a good rain forest..
8	TEAL'C	O Neill
9	interlude	A man comes out of the tree's and falls to his knees
10	MAN	Help me. They find me. Taldor.
11	JACK	Who?