Web Scraping_Selenium

November 21, 2020

[ ]:
```
!pip3 install selenium
!pip install chromedriver-binary
!pip3 install spacy
!python -m spacy download en_core_web_sm
```

[51]:
```
from bs4 import BeautifulSoup as bs
import requests
from selenium import webdriver
import chromedriver_binary
import spacy
import re
```

[52]:
```
driver = webdriver.Chrome()
driver.page_source
nlp = spacy.load('en_core_web_sm')
```

[69]:
```
def book_scraping(href, driver, nlp):
    driver.get(href)
    soup = bs(driver.page_source, features='lxml')
    bookTitle = soup.find_all('h1', id='bookTitle')[0].contents[0].strip()
    bookSeries = None
    try:
        bookSeries = soup.find_all('h2', id='bookSeries')[0].contents[1].
 ↪contents[0].strip()[1:-1]
    except Exception:
        pass
    bookAuthors = soup.find_all('span', itemprop='name')[0].contents[0].strip()
    descr = soup.find_all('div', id='description')[0].contents
    descr_fil= list(filter(lambda s: s!='\n', descr))
    if len(descr_fil) == 1:
        Plot = ''.join(descr_fil[0].contents[0])
    else:
        descr_fil = descr_fil[1:-1]
        x = [j for i in descr_fil for j in i.contents if (isinstance(j,␣
 ↪str)==True)]
        Plot = ''.join(x)
    NumberofPages = soup.find_all('span', itemprop='numberOfPages')[0].
↪contents[0].split()[0]
    ratingValue = soup.find_all('span', itemprop='ratingValue')[0].contents[0].
↪strip()
    ratings_reviews = soup.find_all('a', href='#other_reviews')
    for i in ratings_reviews:
        if i.find_all('meta',itemprop='ratingCount'):
            ratingCount = i.contents[2].split()[0]
        if i.find_all('meta',itemprop='reviewCount'):
            reviewCount = i.contents[2].split()[0]
    pub = soup.find_all('div', class_='row')[1].contents[0].split()[1:4]
    Published = ' '.join(pub)
    char = soup.find_all('a', href=re.compile('characters')) # find the regular␣
↪expression(re) 'characters' within the attribute href
    Characters = [i.contents[0] for i in char]
    sett = soup.find_all('a', href=re.compile('places')) # find the regular␣
↪expression(re) 'places' within the attribute href
    if len(sett) == 0:
        Setting = None
    else:
        Setting = [i.contents[0] for i in sett]
    doc = nlp(Plot)
    token_list = [token for token in doc]
    sentiment_analysis = [token for token in token_list if not token.is_stop␣
↪and not token.is_punct]
    return (bookTitle, bookSeries, bookAuthors, ratingValue, ratingCount,␣
↪reviewCount, Plot, NumberofPages, Published, Characters, Setting)
```

[70]:

```
href = 'https://www.goodreads.com/book/show/9222475-infernal-devices'
book_scraping(href, driver, nlp)
```

[70]:

```
('Infernal Devices',
 'Infernal Devices #1',
 'K.W. Jeter',
 '3.36',
 '2,698',
 '387',
 "HE INHERITED A WATCHMAKER'S STORE - AND A WHOLE HEAP OF TROUBLE. But idle
sometime-musician George has little talent for clockwork. And when a shadowy
figure tries to steal an old device from the premises, George finds himself
embroiled in a mystery of time travel, music and sexual intrigue. A genuine lost
classic, a steampunk original whose time has come.",
 '384',
 'April 7th 2011',
 [],
 None)
```

[71]:

```
href = 'https://www.goodreads.com/book/show/1137215.Boneshaker'
book_scraping(href, driver, nlp)
```

[71]:

```
('Boneshaker',
 'The Clockwork Century #1',
 'Cherie Priest',
 '3.51',
 '31,414',
 '4,073',
 'In the early days of the Civil War, rumors of gold in the frozen Klondike
brought hordes of newcomers to the Pacific Northwest. Anxious to compete,
Russian prospectors commissioned inventor Leviticus Blue to create a great
machine that could mine through Alaska's ice. Thus was Dr. Blue's Incredible
Bone-Shaking Drill Engine born.But on its first test run the Boneshaker went
terribly awry, destroying several blocks of downtown Seattle and unearthing a
subterranean vein of blight gas that turned anyone who breathed it into the
living dead.Now it is sixteen years later, and a wall has been built to enclose
the devastated and toxic city. Just beyond it lives Blue's widow, Briar Wilkes.
Life is hard with a ruined reputation and a teenaged boy to support, but she and
Ezekiel are managing. Until Ezekiel undertakes a secret crusade to rewrite
history.His quest will take him under the wall and into a city teeming with
ravenous undead, air pirates, criminal overlords, and heavily armed refugees.
And only Briar can bring him out alive.',
 '416',
 'September 29th 2009',
 ['Zombies',
  'Briar Wilkes',
  'Ezekiel (Zeke) Wilkes',
  'Croggon Beauregard Hainey'],
 ['Seattle, Washington'])
```

[72]:

```
href = 'https://www.goodreads.com/book/show/7082.
 ↪Do_Androids_Dream_of_Electric_Sheep_'
book_scraping(href, driver, nlp)
```

[72]:

```
('Do Androids Dream of Electric Sheep?',
 'Blade Runner #1',
 'Philip K. Dick',
 '4.08',
 '348,186',
 '13,281',
 'It was January 2021, and Rick Deckard had a license to kill.',
 '244',
 'June 1996 by',
 ['Rick Deckard',
  'John Isidore',
  'Roy Baty',
  'Rachael Rosen',
  'Iran Deckard',
  'Phil Resch',
  'Luba Luft',
  'Bill Barbour',
  'Pris Stratton',
  'Hannibal Sloat',
  'Irmgard Baty',
  'Inspector Garland',
  'Max Polokov',
  'Wilbur Mercer',
  'Buster Friendly',
  'Al Jarry'],
 ['San Francisco, California', 'California', 'United States of America'])
```

```
[73]: href = 'https://www.goodreads.com/book/show/24800.House_of_Leaves'
      book_scraping(href, driver, nlp)
```

```
[73]: ('House of Leaves',
       None,
       'Mark Z. Danielewski',
       '4.05',
       '132,654',
       '11,146',
       'Years ago, when  was first being passed around, it was nothing more than a
      badly bundled heap of paper, parts of which would occasionally surface on the
      Internet. No one could have anticipated the small but devoted following this
      terrifying story would soon command. Starting with an odd assortment of
      marginalized youth-musicians, tattoo artists, programmers, strippers,
      environmentalists, and adrenaline junkies-the book eventually made its way into
      the hands of older generations, who not only found themselves in those strangely
      arranged pages but also discovered a way back into the lives of their estranged
      children.Now, for the first time, this astonishing novel is made available in
      book form, complete with the original colored words, vertical footnotes, and
      newly added second and third appendices.The story remains unchanged, focusing on
      a young family that moves into a small home on Ash Tree Lane where they discover
      something is terribly wrong: their house is bigger on the inside than it is on
      the outside.Of course, neither Pulitzer Prize-winning photojournalist Will
      Navidson nor his companion Karen Green was prepared to face the consequences of
      that impossibility, until the day their two little children wandered off and
      their voices eerily began to return another story-of creature darkness, of an
      ever-growing abyss behind a closet door, and of that unholy growl which soon
      enough would tear through their walls and consume all their dreams.',
       '705',
       'March 7th 2000',
       ['Zampanò', 'Will Navidson', 'Karen Green', 'Johnny Truant'],
       None)
```