stat29000project01solutions

January 30, 2020

1 STAT29000 Project 1 Solutions

1.1 Question 1

[1]:

```
import csv
from block_timer.timer import Timer
import requests
from collections import defaultdict
```

[ ]:

```
# open a file from a link
file = requests.get("https://raw.githubusercontent.com/zygmuntz/goodbooks-10k/
 ↪master/books.csv").text

# strip whitespace and split line by line
file = file.strip().split('\n')

# solution
reader = csv.reader(file)
for idx, row in enumerate(reader):
    if idx % 10 == 0:
        print(row)
```

[ ]:

```
# refresh
reader = csv.reader(file)

# skip the header row
next(reader)

for row in reader:
    print(f'{row[9]}: {row[12]}')
```

1.2 Question 2

[4]:

```
keywords = ('magic', 'game', 'mystery')

# stored keywords in a tuple as we don't need to modify
# or change the size of our keywords

# refresh reader
reader = csv.reader(file)

for row in reader:
    if keywords[0] in row[9] or keywords[1] in row[9] or keywords[2] in row[9]:
        print(row)
```

```
['332', '65605', '65605', '1031537', '312', '60764902', '9.78006076491e+12',
'C.S. Lewis', '1953.0', "The magician's nephew", "The Magician's Nephew
(Chronicles of Narnia, #6)", 'eng', '4.01', '268355', '303570', '8690', '3168',
'14602', '69051', '105375', '111374', 'https://images.gr-
assets.com/books/1308814770m/65605.jpg', 'https://images.gr-
assets.com/books/1308814770s/65605.jpg']
['9633', '20510241', '20510241', '35441994', '56', '62332589',
'9.78006233258e+12', 'James Frey, Nils Johnson-Shelton', '2014.0', 'Endgame: The
Calling', 'The Calling (Endgame, #1)', 'eng', '3.77', '10900', '12934', '1900',
'538', '1078', '2900', '4668', '3750', 'https://images.gr-
assets.com/books/1411272821m/20510241.jpg', 'https://images.gr-
assets.com/books/1411272821s/20510241.jpg']
```

[5]:

```
good_books = []
ehh_books = []
books = []

# refresh reader
reader = csv.reader(file)

# skip header row
next(reader)

for row in reader:
    if float(row[12]) >= 4.02:
        good_books.append(row[7])
    else:
        ehh_books.append(row[7])

    # regardless of the book rating, store the subset
    books.append(row[7])
```

[17]:

```python
both = []
only_ehh = []

for ebook in ehh_books:
    for gbook in good_books:

        if ebook == gbook:
            # author has written both a good and ehh book
            if gbook not in both:
                both.append(gbook)


    # author has only written an ehh book
    # we've checked for this author in all good books
    if ebook not in both:
        if ebook not in only_ehh:
            only_ehh.append(ebook)
```

[23]:

```python
both = {}
only_ehh = {}

both = set.intersection(set(good_books), set(ehh_books))
only_ehh = set(ehh_books) - set(good_books)

print(len(both))
print(len(only_ehh))
```

```
590
2175
```

[8]:

```python
with Timer(title="2c"):
    both = []
    only_ehh = []

    for ebook in ehh_books:
        for gbook in good_books:

            if ebook == gbook:
                # author has written both a good and ehh book
                if gbook not in both:
                    both.append(gbook)


        # author has only written an ehh book
        # we've checked for this author in all good books
        if ebook not in both:
            if ebook not in only_ehh:
                only_ehh.append(ebook)

with Timer(title="2d"):
    both = {}
    only_ehh = {}

    both = set.intersection(set(good_books), set(ehh_books))
    only_ehh = set(ehh_books) - set(good_books)
```

```
[2c] Total time 1.54169 seconds.
[2d] Total time 0.00100 seconds.
```

1.3 Question 3

[24]:

```
good_books = []
ehh_books = []
books = []

# refresh reader
reader = csv.reader(file)

# skip header row
next(reader)

books = [row[7] for row in reader]
reader = csv.reader(file)
next(reader)
good_books = [row[7] for row in reader if float(row[12]) >= 4.02]
reader = csv.reader(file)
next(reader)
ehh_books = [row[7] for row in reader if float(row[12]) < 4.02]

both = []
only_ehh = []

both = {gbook for ebook in ehh_books for gbook in good_books if ebook==gbook␣
 ↪and gbook not in both}
only_ehh = {ebook for ebook in ehh_books if ebook not in both}

print(len(both))
print(len(only_ehh))
```

```
590
2175
```

[10]:

```
# refresh reader
reader = csv.reader(file)

# skip header row
next(reader)

langs = defaultdict(int)
for row in reader:
    langs[row[11]] += int(row[13])

print(langs)

# the most popular are: English, Spanish, Arabic
```

```
defaultdict(<class 'int'>, {'eng': 387206346, 'en-US': 106950932, 'en-CA':
4855486, '': 26964645, 'spa': 1583857, 'en-GB': 9485353, 'fre': 724719, 'nl':
83695, 'ara': 1043827, 'por': 100415, 'ger': 265740, 'nor': 41107, 'jpn':
127161, 'en': 86604, 'vie': 24907, 'ind': 239921, 'pol': 67378, 'tur': 19806,
'dan': 24039, 'fil': 21497, 'ita': 23731, 'per': 33161, 'swe': 6435, 'rum':
11079, 'mul': 11223, 'rus': 9287})
```

[ ]:

```
# refresh
reader = csv.reader(file)

# skip the header row
next(reader)

for row in reader:
    authors = [a.strip() for a in row[7].split(',')]
    print(f'{row[9]} by {" and ".join(authors)}: {row[12]}')
```