# lec01

August 9, 2021

[1]:
```
# Stuff that will appear at the top of notebooks;
# You don't have to understand how this works or change it for now.

from datascience import *
import numpy as np
%matplotlib inline
import matplotlib.pyplot as plots
plots.style.use('fivethirtyeight')
import warnings
warnings.simplefilter(action="ignore", category=FutureWarning)

from urllib.request import urlopen
import re
def read_url(url):
    return re.sub('\\s+', ' ', urlopen(url).read().decode())
```

[2]:
```
2+3
```

[2]:
```
5
```

[3]:
```
# Read two books, fast!

huck_finn_url = 'https://www.inferentialthinking.com/data/huck_finn.txt'
huck_finn_text = read_url(huck_finn_url)
huck_finn_chapters = huck_finn_text.split('CHAPTER ')[44:]

little_women_url = 'https://www.inferentialthinking.com/data/little_women.txt'
little_women_text = read_url(little_women_url)
little_women_chapters = little_women_text.split('CHAPTER ')[1:]
```

[ ]:
```
huck_finn_chapters
```

[4]:
```
Table().with_column('Chapters', huck_finn_chapters)
```

[4]:
```
Chapters
I. YOU don't know about me without you have read a book  …
II. WE went tiptoeing along a path amongst the trees bac …
```

```
III. WELL, I got a good going-over in the morning from o …
IV. WELL, three or four months run along, and it was wel …
V. I had shut the door to. Then I turned around and ther …
VI. WELL, pretty soon the old man was up and around agai …
VII. "GIT up! What you 'bout?" I opened my eyes and look …
VIII. THE sun was up so high when I waked that I judged  …
IX. I wanted to go and look at a place right about the m …
X. AFTER breakfast I wanted to talk about the dead man a …
… (33 rows omitted)
```

[5]:
```
np.char.count(huck_finn_chapters, 'Tom')
```

[5]:
```
array([ 6, 24,  5,  0,  0,  0,  2,  2,  0,  0,  2,  3,  1,  0,  0,  0,  3,
        5,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  1,  4, 19, 15,
       14, 18,  9, 32, 11, 11,  8, 30,  6])
```

[6]:
```
np.char.count(huck_finn_chapters, 'Jim')
```

[6]:
```
array([ 0, 16,  0,  8,  0,  0,  0, 22, 11, 19,  4, 20,  9,  6, 16, 28,  0,
       10, 13, 18,  1,  0,  9,  5,  0,  0,  0,  1,  3,  5, 17,  0,  5, 17,
       18, 23,  4, 27, 10, 13,  0, 12,  6])
```

[7]:
```
counts = Table().with_columns([
    'Tom', np.char.count(huck_finn_chapters, 'Tom'),
    'Jim', np.char.count(huck_finn_chapters, 'Jim'),
    'Huck', np.char.count(huck_finn_chapters, 'Huck'),
])
counts
```

[7]:
```
Tom  | Jim  | Huck
6    | 0    | 3
24   | 16   | 2
5    | 0    | 2
0    | 8    | 1
0    | 0    | 0
0    | 0    | 2
2    | 0    | 0
2    | 22   | 5
0    | 11   | 1
0    | 19   | 0
… (33 rows omitted)
```

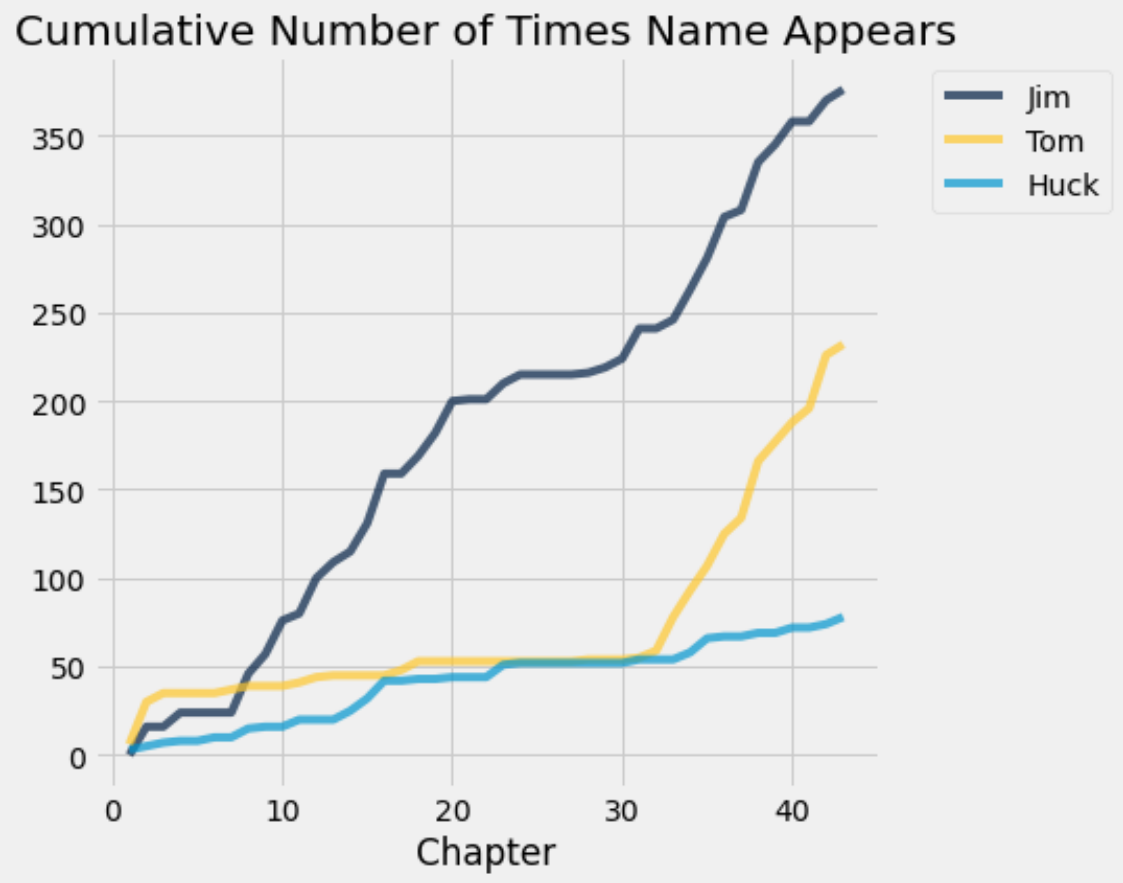
[8]:
```
# Count how many times the names Jim, Tom, and Huck appear in each chapter

counts = Table().with_columns([
        'Jim', np.char.count(huck_finn_chapters, 'Jim'),
        'Tom', np.char.count(huck_finn_chapters, 'Tom'),
```

```
        'Huck', np.char.count(huck_finn_chapters, 'Huck')
    ])

# Plot the cumulative counts:
# how many times in Chapter 1, how many times in Chapters 1 and 2, and so on.

cum_counts = counts.cumsum().with_column('Chapter', np.arange(1, 44, 1))
cum_counts.plot(column_for_xticks=3)
plots.title('Cumulative Number of Times Name Appears');
```




[9]:
```
# The chapters of Little Women

Table().with_column('Chapters', little_women_chapters)
```

[9]:
```
Chapters
ONE PLAYING PILGRIMS "Christmas won't be Christmas witho …
TWO A MERRY CHRISTMAS Jo was the first to wake in the gr …
THREE THE LAURENCE BOY "Jo! Jo! Where are you?" cried Me …
FOUR BURDENS "Oh, dear, how hard it does seem to take up …
```

```
FIVE BEING NEIGHBORLY "What in the world are you going t …
SIX BETH FINDS THE PALACE BEAUTIFUL The big house did pr …
SEVEN AMY'S VALLEY OF HUMILIATION "That boy is a perfect …
EIGHT JO MEETS APOLLYON "Girls, where are you going?" as …
NINE MEG GOES TO VANITY FAIR "I do think it was the most …
TEN THE P.C. AND P.O. As spring came on, a new set of am …
… (37 rows omitted)
```

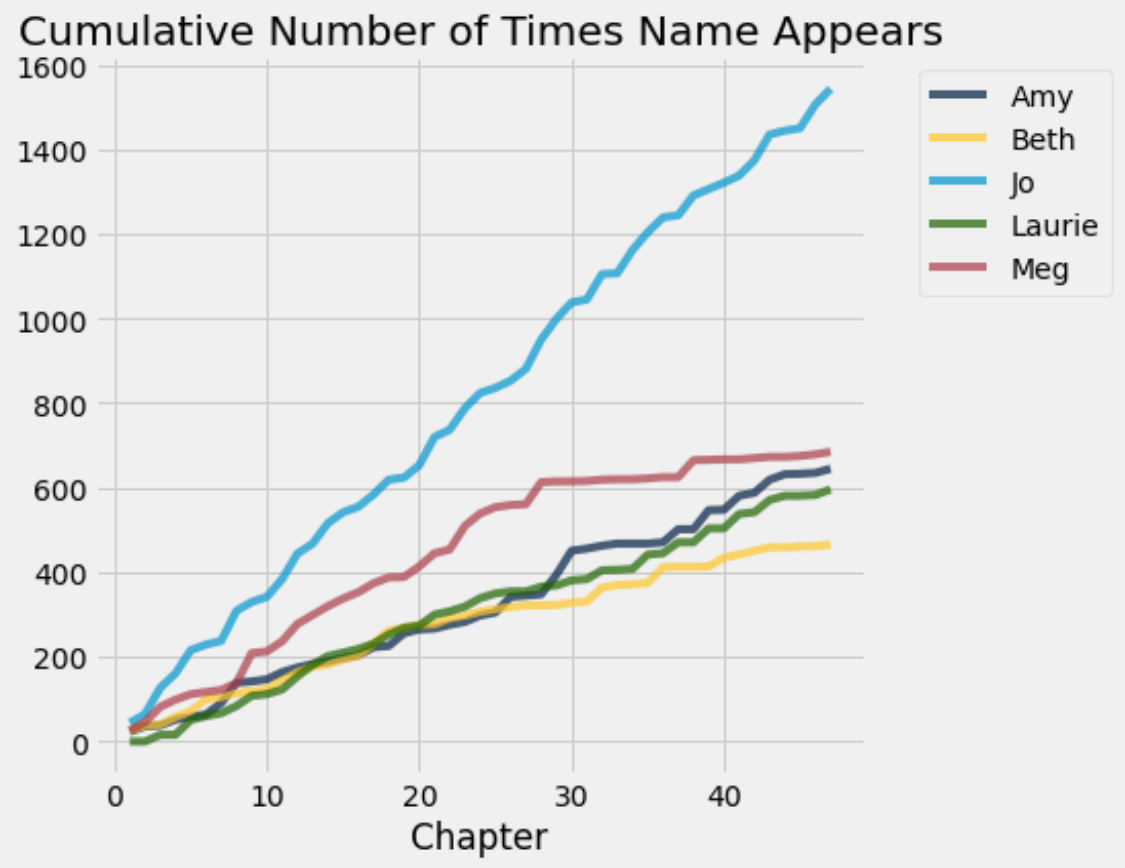
[10]:
```
# Counts of names in the chapters of Little Women

people = ['Amy', 'Beth', 'Jo', 'Laurie', 'Meg']
people_counts = {pp: np.char.count(little_women_chapters, pp) for pp in people}

counts = Table().with_columns([
        'Amy', people_counts['Amy'],
        'Beth', people_counts['Beth'],
        'Jo', people_counts['Jo'],
        'Laurie', people_counts['Laurie'],
        'Meg', people_counts['Meg']
    ])
```

[11]:
```
# Plot the cumulative counts

cum_counts = counts.cumsum().with_column('Chapter', np.arange(1, 48, 1))
cum_counts.plot(column_for_xticks=5)
plots.title('Cumulative Number of Times Name Appears');
```




[12]:
```
# In each chapter, count the number of all characters;
# call this the "length" of the chapter.
# Also count the number of periods.

chars_periods_hf = Table().with_columns([
        'HF Chapter Length', [len(s) for s in huck_finn_chapters],
        'Number of Periods', np.char.count(huck_finn_chapters, '.')
    ])
chars_periods_lw = Table().with_columns([
        'LW Chapter Length', [len(s) for s in little_women_chapters],
        'Number of Periods', np.char.count(little_women_chapters, '.')
    ])
```

[13]:
```
# The counts for Huckleberry Finn

chars_periods_hf
```

[13]:
```
HF Chapter Length | Number of Periods
7026              | 66
11982             | 117
```

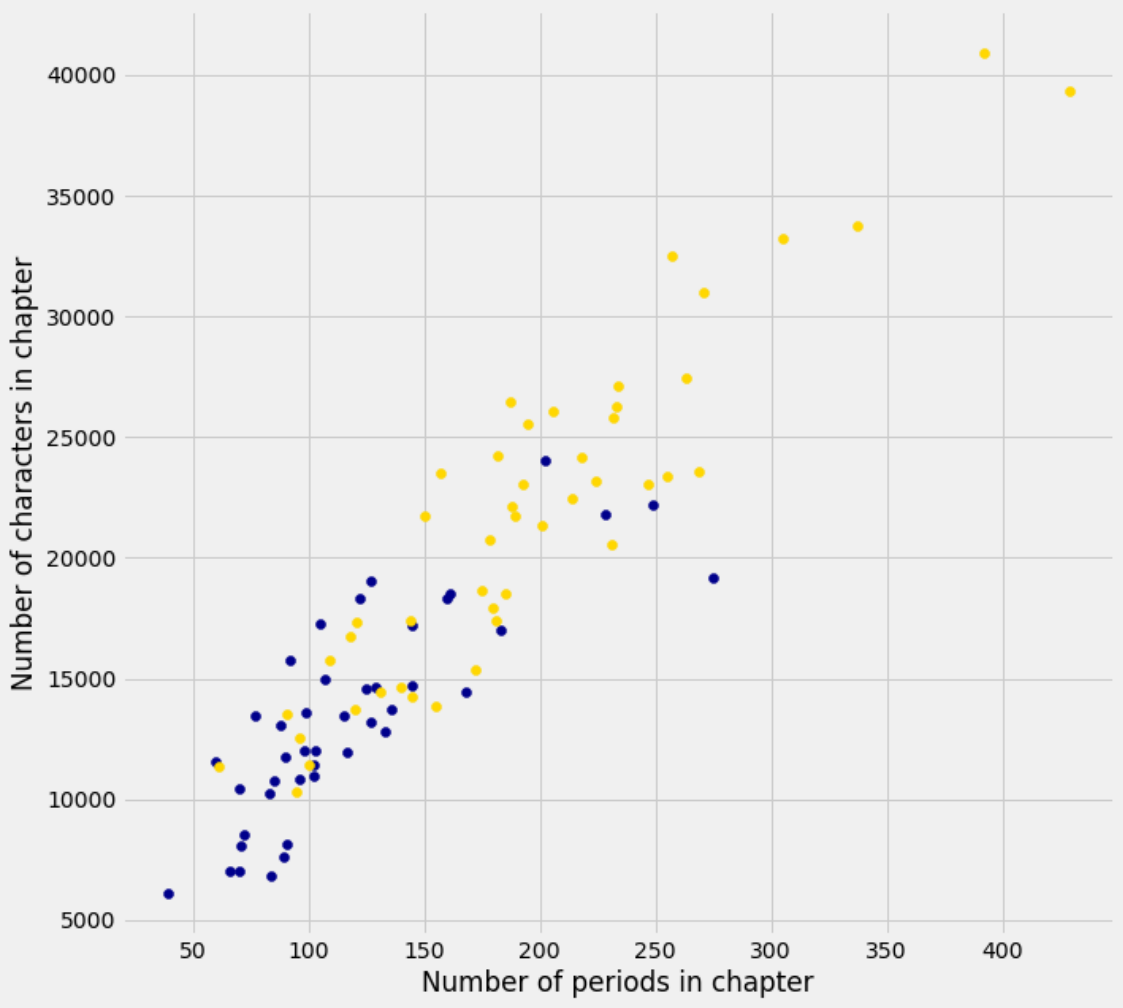
```
8529              | 72
6799              | 84
8166              | 91
14550             | 125
13218             | 127
22208             | 249
8081              | 71
7036              | 70
… (33 rows omitted)
```

[14]:
```
# The counts for Little Women

chars_periods_lw
```

[14]:
```
LW Chapter Length | Number of Periods
21759             | 189
22148             | 188
20558             | 231
25526             | 195
23395             | 255
14622             | 140
14431             | 131
22476             | 214
33767             | 337
18508             | 185
… (37 rows omitted)
```

[15]:
```
plots.figure(figsize=(10,10))
plots.scatter(chars_periods_hf[1], chars_periods_hf[0], color='darkblue')
plots.scatter(chars_periods_lw[1], chars_periods_lw[0], color='gold')
plots.xlabel('Number of periods in chapter')
plots.ylabel('Number of characters in chapter');
```


Number of characters in chapter
Number of periods in chapter
40000
35000
30000
25000
20000
15000
10000
5000
50
100
150
200
250
300
350
400