

```
In [124]: from bs4 import BeautifulSoup, NavigableString, Tag
from datascience import *
from collections import Counter
```

```
In [125]: data = Table.read_table('scripts_metadata.csv')
data.show(5)
```

title	Genres	Average user rating	IMSDb rating	IMSDb opinion	Script Date	Movie Release Date	Writers	Submi
10 Things I Hate About You Script	Comedy;Romance;	(8.76 out of 10)	(7 out of 10)	A better- than- most teen film.	: November 1997	nan	Karen McCullah Lutz;Kirsten Smith;William Shakespeare;	
12 Script	Comedy;Read "12" Script;	None available	Not available	None available	nan	nan	Lawrence Bridges;	
12 and Holding Script	Drama;	(7.00 out of 10)	Not available	None available	: April 2004	: May 2006	Anthony Cipriano;	
12 Monkeys Script	Drama;Sci- Fi;Thriller;	(9.25 out of 10)	Not available	None available	: June 1994	nan	David Peoples;Janet Peoples;	
12 Years a Slave Script	Drama;	None available	Not available	None available	nan	: November 2013	John Ridley;	: XXyTi

... (1166 rows omitted)

```
In [126]: data = data.where('title', are.not_equal_to('8 Mile Script'))
data.show(5)
```

title	Genres	Average user rating	IMSDb rating	IMSDb opinion	Script Date	Movie Release Date	Writers	Submi
10 Things I Hate About You Script	Comedy;Romance;	(8.76 out of 10)	(7 out of 10)	A better-than-most teen film.	: November 1997	nan	Karen McCullah Lutz;Kirsten Smith;William Shakespeare;	
12 Script	Comedy;Read "12" Script;	None available	Not available	None available	nan	nan	Lawrence Bridges;	
12 and Holding Script	Drama;	(7.00 out of 10)	Not available	None available	: April 2004	: May 2006	Anthony Cipriano;	
12 Monkeys Script	Drama;Sci-Fi;Thriller;	(9.25 out of 10)	Not available	None available	: June 1994	nan	David Peoples;Janet Peoples;	
12 Years a Slave Script	Drama;	None available	Not available	None available	nan	: November 2013	John Ridley; : XXyTi	

... (1165 rows omitted)

```
In [127]: data = data.where('script_path', are.not_equal_to('nan'))
data.show(5)
```

title	Genres	Average user rating	IMSDb rating	IMSDb opinion	Script Date	Movie Release Date	Writers	Submi
10 Things I Hate About You Script	Comedy;Romance;	(8.76 out of 10)	(7 out of 10)	A better-than-most teen film.	: November 1997	nan	Karen McCullah Lutz;Kirsten Smith;William Shakespeare;	
12 Script	Comedy;Read "12" Script;	None available	Not available	None available	nan	nan	Lawrence Bridges;	
12 and Holding Script	Drama;	(7.00 out of 10)	Not available	None available	: April 2004	: May 2006	Anthony Cipriano;	
12 Monkeys Script	Drama;Sci-Fi;Thriller;	(9.25 out of 10)	Not available	None available	: June 1994	nan	David Peoples;Janet Peoples;	
12 Years a Slave Script	Drama;	None available	Not available	None available	nan	: November 2013	John Ridley; : XXyTi	

... (1137 rows omitted)

```
In [128]: data = data.where('title', are.not_equal_to('Back to the Future Script'))
data.show(5)
```

title	Genres	Average user rating	IMSDb rating	IMSDb opinion	Script Date	Movie Release Date	Writers	Submit
10 Things I Hate About You Script	Comedy;Romance;	(8.76 out of 10)	(7 out of 10)	A better-than-most teen film.	: November 1997	nan	Karen McCullah Lutz;Kirsten Smith;William Shakespeare;	
12 Script	Comedy;Read "12" Script;	None available	Not available	None available	nan	nan	Lawrence Bridges;	
12 and Holding Script	Drama;	(7.00 out of 10)	Not available	None available	: April 2004	: May 2006	Anthony Cipriano;	
12 Monkeys Script	Drama;Sci-Fi;Thriller;	(9.25 out of 10)	Not available	None available	: June 1994	nan	David Peoples;Janet Peoples;	
12 Years a Slave Script	Drama;	None available	Not available	None available	nan	: November 2013	John Ridley; : XXyTi	

... (1136 rows omitted)

```
In [129]: data = data.where('title', are.not_equal_to('Back to the Future II & III Script'))
data.show(5)
```

title	Genres	Average user rating	IMSDb rating	IMSDb opinion	Script Date	Movie Release Date	Writers	Submit
10 Things I Hate About You Script	Comedy;Romance;	(8.76 out of 10)	(7 out of 10)	A better-than-most teen film.	: November 1997	nan	Karen McCullah Lutz;Kirsten Smith;William Shakespeare;	
12 Script	Comedy;Read "12" Script;	None available	Not available	None available	nan	nan	Lawrence Bridges;	
12 and Holding Script	Drama;	(7.00 out of 10)	Not available	None available	: April 2004	: May 2006	Anthony Cipriano;	
12 Monkeys Script	Drama;Sci-Fi;Thriller;	(9.25 out of 10)	Not available	None available	: June 1994	nan	David Peoples;Janet Peoples;	
12 Years a Slave Script	Drama;	None available	Not available	None available	nan	: November 2013	John Ridley; : XXyTi	

... (1135 rows omitted)

In [ ]:

```

In [130]: ## make an empty ditionary then append everthing to it
all_scripts = {}

for fname in data['script_path']:

    print(fname)
    with open(fname, 'r') as f:
        raw = f.read()
        soup = BeautifulSoup(raw, 'html5lib')

    try:
        bolded = soup.find('td', {'class': 'scrtext'}) .find_all('b') #find
        text = soup.find('td', {'class': 'scrtext'}) .text
        b_text = [b.text.strip() for b in bolded]
        bolded_text = [b for b in b_text if len(b) > 0]
        sift_out = ['INT.', "EXT.", "-"] #differenetiaste between scene cues
        characters = []
        scenes = []
        for c in bolded_text:
            character = True
            for s in sift_out:
                if s in c:
                    character = False
            if character == True:
                characters.append(c)
            elif len(c) > 4:
                scenes.append(c)

        characters = [c[0] for c in Counter(characters).most_common() if c[1]
        scenes.extend([c[0] for c in Counter(characters).most_common() if c[1]

        movie_name = fname.split('/')[ -1 ][ :-5 ].replace(' Script', '')

        all_scripts[movie_name] = {}
        all_scripts[movie_name]['cast'] = characters
        all_scripts[movie_name]['scenes'] = scenes
        all_scripts[movie_name]['text'] = text

    except:
        pass

```

```

scripts/10 Things I Hate About You Script.html
scripts/12 Script.html
scripts/12 and Holding Script.html
scripts/12 Monkeys Script.html
scripts/12 Years a Slave Script.html
scripts/127 Hours Script.html
scripts/1492: Conquest of Paradise Script.html
scripts/15 Minutes Script.html
scripts/17 Again Script.html
scripts/187 Script.html
scripts/2001: A Space Odyssey Script.html
scripts/2012 Script.html
scripts/25th Hour Script.html
scripts/30 Minutes or Less Script.html

```

```

-----
--
KeyboardInterrupt                                Traceback (most recent call las
t)
<ipython-input-130-2a6a504be4e6> in <module>()
      7     with open(fname, 'r') as f:
      8         raw = f.read()
----> 9     soup = BeautifulSoup(raw, 'html5lib')
     10
     11     try:

/srv/app/venv/lib/python3.6/site-packages/bs4/__init__.py in __init__(sel
f, markup, features, builder, parse_only, from_encoding, exclude_encoding
s, **kwargs)
     226         self.reset()
     227         try:
--> 228             self._feed()
     229             break
     230         except ParserRejectedMarkup:

/srv/app/venv/lib/python3.6/site-packages/bs4/__init__.py in _feed(self)
     287         self.builder.reset()
     288
--> 289         self.builder.feed(self.markup)
     290         # Close out any unfinished strings and close all the open
tags.
     291         self.endData()

/srv/app/venv/lib/python3.6/site-packages/bs4/builder/_html5lib.py in fee
d(self, markup)
      70         else:
      71             extra_kwargs['encoding'] = self.user_specified_en
coding
----> 72             doc = parser.parse(markup, **extra_kwargs)
      73
      74             # Set the character encoding detected by the tokenizer.

/srv/app/venv/lib/python3.6/site-packages/html5lib/html5parser.py in pars
e(self, stream, *args, **kwargs)
     233         scripting - treat noscript elements as if javascript was
turned on
     234         """
--> 235         self._parse(stream, False, None, *args, **kwargs)
     236         return self.tree.getDocument()
     237

/srv/app/venv/lib/python3.6/site-packages/html5lib/html5parser.py in _par
se(self, stream, innerHTML, container, scripting, **kwargs)
      87
      88         try:
----> 89             self.mainLoop()
      90         except ReparseException:
      91             self.reset()

/srv/app/venv/lib/python3.6/site-packages/html5lib/html5parser.py in main
Loop(self)
     193             new_token = phase.processCharacters(new_t

```

```

oken)
194             elif type == SpaceCharactersToken:
--> 195                 new_token = phase.processSpaceCharacters(
new_token)
196             elif type == StartTagToken:
197                 new_token = phase.processStartTag(new_tok
en)

/srv/app/venv/lib/python3.6/site-packages/html5lib/html5parser.py in proc
essSpaceCharacters(self, token)
406
407     def processSpaceCharacters(self, token):
--> 408         self.tree.insertText(token["data"])
409
410     def processStartTag(self, token):

/srv/app/venv/lib/python3.6/site-packages/html5lib/treebuilders/base.py i
n insertText(self, data, parent)
324                                     self.openElements[-1].na
me
325                                     not in tableInsertModeEl
ements)):
--> 326         parent.insertText(data)
327     else:
328         # We should be in the InTable mode. This means we wan
t to do

/srv/app/venv/lib/python3.6/site-packages/bs4/builder/_html5lib.py in ins
ertText(self, data, insertBefore)
312
313     def insertText(self, data, insertBefore=None):
--> 314         text = TextNode(self.soup.new_string(data), self.soup)
315         if insertBefore:
316             self.insertBefore(text, insertBefore)

/srv/app/venv/lib/python3.6/site-packages/bs4/__init__.py in new_string(s
elf, s, subclass)
309     def new_string(self, s, subclass=NavigableString):
310         """Create a new NavigableString associated with this sou
p."""
--> 311         return subclass(s)
312
313     def insert_before(self, successor):

/srv/app/venv/lib/python3.6/site-packages/bs4/element.py in __new__(cls,
value)
711         """
712         if isinstance(value, str):
--> 713             u = str.__new__(cls, value)
714         else:
715             u = str.__new__(cls, value, DEFAULT_OUTPUT_ENCODING)

```

KeyboardInterrupt:

```
In [131]: all_scripts.keys()
```

```
Out[131]: dict_keys(['10 Things I Hate About You', '12', '12 and Holding', '12 Monk  
eys', '12 Years a Slave', '127 Hours', '1492: Conquest of Paradise', '15  
Minutes', '17 Again', '187', '2001: A Space Odyssey', '2012', '25th Hou  
r'])
```



```
In [132]: import re

scene_index_list = []
for scene in set(all_scripts['10 Things I Hate About You']['scenes']):
    print(scene)
    indices = [m.start() for m in re.finditer(scene, all_scripts['10 Things
    scene_index_list.extend(indices)
```

HALLWAY - DAY  
INT. BOOK STORE - DAY  
INT. PROM - NIGHT - LATER  
INT. BIANCA'S ROOM - NIGHT  
EXT. CLUB SKUNK - NIGHT  
EXT. DOWNTOWN STREET - NIGHT  
EXT. SCHOOL COURTYARD - DAY  
BOGEY'S KITCHEN - NIGHT  
INT. STRATFORD HOUSE/BATHROOM - NIGHT  
CLASSROOM - DAY  
INT. STRATFORD HOUSE/DEN - DAY  
INT. CLUB - NIGHT  
INT. BOY'S ROOM - DAY  
EXT. PARKING LOT - DAY  
EXT. STRATFORD HOUSE - NIGHT  
INT. MATH CLASS - DAY  
STRATFORD HOUSE/BACKYARD - SUNSET  
INT. CAFETERIA - DAY - CONTINUOUS  
CAFETERIA - DAY  
INSERT - "JOEY DORSEY SAID HI TO ME IN THE HALL! OH! MY  
HOTEL - NIGHT  
INT. KENNY'S THAI FOOD DINER - DAY  
INT. CLUB FOYER - NIGHT  
EXT. SCHOOL PARKING LOT - DAY  
INT. BOGEY'S BATHROOM - NIGHT  
EXT. STRATFORD HOUSE - DAY  
INT. SHOWERS - DAY  
INT. BOGEY LOWENSTEIN'S HOUSE - NIGHT  
CAMERON'S CAR - NIGHT  
INT. HALLWAY - DAY  
INT. BIOLOGY CLASS  
STRATFORD HOUSE - SUNSET  
INT. SOPHOMORE ENGLISH CLASS - DAY  
INT. GYM CLASS - DAY  
EXT. PARKING LOT - MOMENTS LATER  
INT. TUTORING ROOM - DAY  
INT. CLASSROOM - DAY  
EXT. MISS PERKY'S OFFICE - DAY  
KAT'S CAR - NIGHT  
HALLWAY - DAY- CONTINUOUS  
EXT. STREET - NIGHT  
INT. GIRLS' ROOM - DAY  
PADUA HIGH PARKING LOT - DAY  
INT. DETENTION HALL - DAY  
EXT. ARCHERY FIELD - DAY  
EXT HOTEL PARKING LOT - NIGHT  
BIANCA'S ROOM - DAY  
COURTYARD - DAY  
INT. WOODSHOP - DAY

INT. TUTORING ROOM  
 BOGEY LOWENSTEIN'S HOUSE - NIGHT  
 LIVING ROOM - NIGHT  
 INT. STRATFORD HOUSE - NIGHT  
 INT. BOGEY'S LIVING ROOM - NIGHT  
 INT. MISS PERKY'S OFFICE - DAY  
 INT. LIVING ROOM - NIGHT  
 INT. STUDY HALL - DAY  
 EXT. BOGEY LOWENSTEIN'S HOUSE - NIGHT  
 INT. BOGEY'S DINING ROOM - NIGHT  
 INT. KAT'S CAR - NIGHT  
 ENGLISH CLASS - DAY  
 INT. PROM - NIGHT  
 TRACK - DAY  
 STRATFORD HOUSE/BATHROOM - NIGHT  
 INT. STRATFORD HOUSE - DAY  
 INT. SCHOOL COURTYARD - DAY  
 EXT. FIELD HOCKEY FIELD - DAY  
 INT. HALLWAY - DAY- CONTINUOUS  
 PADUA HIGH SCHOOL - DAY  
 INSERT - "O FAIR ONE. JOIN ME AT THE PROM. I WILL BE  
 EXT. SCHOOL CAMPUS LAWN  
 INT. MISS PERKY'S OFFICE - DAY  
 INT. KAT'S ROOM - DAY  
 INT. BOGEY'S KITCHEN - NIGHT  
 INT. DIVE BAR - NIGHT  
 INT. GYM CLASS - DAY  
 INT. ENGLISH CLASS - DAY  
 GUIDANCE COUNSELOR'S OFFICE - DAY  
 INT. STRATFORD HOUSE/UPSTAIRS HALLWAY - NIGHT  
 INT. GUIDANCE COUNSELOR'S OFFICE - DAY  
 INT. CAFETERIA - DAY  
 INT. STRATFORD HOUSE - DAY  
 INT. KAT'S ROOM - NIGHT  
 INT. LADIES ROOM - NIGHT  
 EXT. OUTDOOR ARCADE - DAY  
 INT. BOGEY'S KITCHEN - NIGHT - LATER

In [133]: `len(scene_index_list )`

Out[133]: 154

In [134]: `from nltk.util import ngrams`

```

scene_texts = []
for n in ngrams(sorted(scene_index_list), 2):
    scene_texts.append(all_scripts['10 Things I Hate About You']['text'][n[0]

```

In [135]: `first_scene = scene_texts[0]`

```
In [136]: all_scripts['10 Things I Hate About You']['cast']
```

```
Out[136]: ['KAT',  
           'PATRICK',  
           'BIANCA',  
           'CAMERON',  
           'MICHAEL',  
           'JOEY',  
           'WALTER',  
           'MANDELLA',  
           'MISS PERKY',  
           'MRS. BLAISE',  
           'CHASTITY',  
           'SHARON',  
           'BRUCE']
```

```
In [137]: cast_dict = {}  
  
for c in all_scripts['10 Things I Hate About You']['cast']:  
    cast_dict[c] = []  
    for i, scene in enumerate(scene_texts):  
        if scene.count(c) > 0:  
            cast_dict[c].append(i)
```

```
In [138]: cast_dict
```

```
Out[138]: {'BIANCA': [2,  
                     13,  
                     19,  
                     22,  
                     23,  
                     25,  
                     34,  
                     36,  
                     39,  
                     49,  
                     60,  
                     61,  
                     63,  
                     74,  
                     76,  
                     80,  
                     82,  
                     85,  
                     86,  
                     88,  
                     89,  
                     90,  
                     91,  
                     92,  
                     93,  
                     94,  
                     95,  
                     96,  
                     97,  
                     98,  
                     99,  
                     100,  
                     101,  
                     102,  
                     103,  
                     104,  
                     105,  
                     106,  
                     107,  
                     108,  
                     109,  
                     110,  
                     111,  
                     112,  
                     113,  
                     114,  
                     115,  
                     116,  
                     117,  
                     118,  
                     119,  
                     120,  
                     121,  
                     122,  
                     123,  
                     124,  
                     125,  
                     126,  
                     127,  
                     128,  
                     129,  
                     130,  
                     131,  
                     132,  
                     133,  
                     134,  
                     135,  
                     136,  
                     137,  
                     138,  
                     139,  
                     140,  
                     141,  
                     142,  
                     143,  
                     144,  
                     145,  
                     146,  
                     147,  
                     148,  
                     149,  
                     150,  
                     151,  
                     152,  
                     153,  
                     154,  
                     155,  
                     156,  
                     157,  
                     158,  
                     159,  
                     160,  
                     161,  
                     162,  
                     163,  
                     164,  
                     165,  
                     166,  
                     167,  
                     168,  
                     169,  
                     170,  
                     171,  
                     172,  
                     173,  
                     174,  
                     175,  
                     176,  
                     177,  
                     178,  
                     179,  
                     180,  
                     181,  
                     182,  
                     183,  
                     184,  
                     185,  
                     186,  
                     187,  
                     188,  
                     189,  
                     190,  
                     191,  
                     192,  
                     193,  
                     194,  
                     195,  
                     196,  
                     197,  
                     198,  
                     199,  
                     200,  
                     201,  
                     202,  
                     203,  
                     204,  
                     205,  
                     206,  
                     207,  
                     208,  
                     209,  
                     210,  
                     211,  
                     212,  
                     213,  
                     214,  
                     215,  
                     216,  
                     217,  
                     218,  
                     219,  
                     220,  
                     221,  
                     222,  
                     223,  
                     224,  
                     225,  
                     226,  
                     227,  
                     228,  
                     229,  
                     230,  
                     231,  
                     232,  
                     233,  
                     234,  
                     235,  
                     236,  
                     237,  
                     238,  
                     239,  
                     240,  
                     241,  
                     242,  
                     243,  
                     244,  
                     245,  
                     246,  
                     247,  
                     248,  
                     249,  
                     250,  
                     251,  
                     252,  
                     253,  
                     254,  
                     255,  
                     256,  
                     257,  
                     258,  
                     259,  
                     260,  
                     261,  
                     262,  
                     263,  
                     264,  
                     265,  
                     266,  
                     267,  
                     268,  
                     269,  
                     270,  
                     271,  
                     272,  
                     273,  
                     274,  
                     275,  
                     276,  
                     277,  
                     278,  
                     279,  
                     280,  
                     281,  
                     282,  
                     283,  
                     284,  
                     285,  
                     286,  
                     287,  
                     288,  
                     289,  
                     290,  
                     291,  
                     292,  
                     293,  
                     294,  
                     295,  
                     296,  
                     297,  
                     298,  
                     299,  
                     300,  
                     301,  
                     302,  
                     303,  
                     304,  
                     305,  
                     306,  
                     307,  
                     308,  
                     309,  
                     310,  
                     311,  
                     312,  
                     313,  
                     314,  
                     315,  
                     316,  
                     317,  
                     318,  
                     319,  
                     320,  
                     321,  
                     322,  
                     323,  
                     324,  
                     325,  
                     326,  
                     327,  
                     328,  
                     329,  
                     330,  
                     331,  
                     332,  
                     333,  
                     334,  
                     335,  
                     336,  
                     337,  
                     338,  
                     339,  
                     340,  
                     341,  
                     342,  
                     343,  
                     344,  
                     345,  
                     346,  
                     347,  
                     348,  
                     349,  
                     350,  
                     351,  
                     352,  
                     353,  
                     354,  
                     355,  
                     356,  
                     357,  
                     358,  
                     359,  
                     360,  
                     361,  
                     362,  
                     363,  
                     364,  
                     365,  
                     366,  
                     367,  
                     368,  
                     369,  
                     370,  
                     371,  
                     372,  
                     373,  
                     374,  
                     375,  
                     376,  
                     377,  
                     378,  
                     379,  
                     380,  
                     381,  
                     382,  
                     383,  
                     384,  
                     385,  
                     386,  
                     387,  
                     388,  
                     389,  
                     390,  
                     391,  
                     392,  
                     393,  
                     394,  
                     395,  
                     396,  
                     397,  
                     398,  
                     399,  
                     400,  
                     401,  
                     402,  
                     403,  
                     404,  
                     405,  
                     406,  
                     407,  
                     408,  
                     409,  
                     410,  
                     411,  
                     412,  
                     413,  
                     414,  
                     415,  
                     416,  
                     417,  
                     418,  
                     419,  
                     420,  
                     421,  
                     422,  
                     423,  
                     424,  
                     425,  
                     426,  
                     427,  
                     428,  
                     429,  
                     430,  
                     431,  
                     432,  
                     433,  
                     434,  
                     435,  
                     436,  
                     437,  
                     438,  
                     439,  
                     440,  
                     441,  
                     442,  
                     443,  
                     444,  
                     445,  
                     446,  
                     447,  
                     448,  
                     449,  
                     450,  
                     451,  
                     452,  
                     453,  
                     454,  
                     455,  
                     456,  
                     457,  
                     458,  
                     459,  
                     460,  
                     461,  
                     462,  
                     463,  
                     464,  
                     465,  
                     466,  
                     467,  
                     468,  
                     469,  
                     470,  
                     471,  
                     472,  
                     473,  
                     474,  
                     475,  
                     476,  
                     477,  
                     478,  
                     479,  
                     480,  
                     481,  
                     482,  
                     483,  
                     484,  
                     485,  
                     486,  
                     487,  
                     488,  
                     489,  
                     490,  
                     491,  
                     492,  
                     493,  
                     494,  
                     495,  
                     496,  
                     497,  
                     498,  
                     499,  
                     500,  
                     501,  
                     502,  
                     503,  
                     504,  
                     505,  
                     506,  
                     507,  
                     508,  
                     509,  
                     510,  
                     511,  
                     512,  
                     513,  
                     514,  
                     515,  
                     516,  
                     517,  
                     518,  
                     519,  
                     520,  
                     521,  
                     522,  
                     523,  
                     524,  
                     525,  
                     526,  
                     527,  
                     528,  
                     529,  
                     530,  
                     531,  
                     532,  
                     533,  
                     534,  
                     535,  
                     536,  
                     537,  
                     538,  
                     539,  
                     540,  
                     541,  
                     542,  
                     543,  
                     544,  
                     545,  
                     546,  
                     547,  
                     548,  
                     549,  
                     550,  
                     551,  
                     552,  
                     553,  
                     554,  
                     555,  
                     556,  
                     557,  
                     558,  
                     559,  
                     560,  
                     561,  
                     562,  
                     563,  
                     564,  
                     565,  
                     566,  
                     567,  
                     568,  
                     569,  
                     570,  
                     571,  
                     572,  
                     573,  
                     574,  
                     575,  
                     576,  
                     577,  
                     578,  
                     579,  
                     580,  
                     581,  
                     582,  
                     583,  
                     584,  
                     585,  
                     586,  
                     587,  
                     588,  
                     589,  
                     590,  
                     591,  
                     592,  
                     593,  
                     594,  
                     595,  
                     596,  
                     597,  
                     598,  
                     599,  
                     600,  
                     601,  
                     602,  
                     603,  
                     604,  
                     605,  
                     606,  
                     607,  
                     608,  
                     609,  
                     610,  
                     611,  
                     612,  
                     613,  
                     614,  
                     615,  
                     616,  
                     617,  
                     618,  
                     619,  
                     620,  
                     621,  
                     622,  
                     623,  
                     624,  
                     625,  
                     626,  
                     627,  
                     628,  
                     629,  
                     630,  
                     631,  
                     632,  
                     633,  
                     634,  
                     635,  
                     636,  
                     637,  
                     638,  
                     639,  
                     640,  
                     641,  
                     642,  
                     643,  
                     644,  
                     645,  
                     646,  
                     647,  
                     648,  
                     649,  
                     650,  
                     651,  
                     652,  
                     653,  
                     654,  
                     655,  
                     656,  
                     657,  
                     658,  
                     659,  
                     660,  
                     661,  
                     662,  
                     663,  
                     664,  
                     665,  
                     666,  
                     667,  
                     668,  
                     669,  
                     670,  
                     671,  
                     672,  
                     673,  
                     674,  
                     675,  
                     676,  
                     677,  
                     678,  
                     679,  
                     680,  
                     681,  
                     682,  
                     683,  
                     684,  
                     685,  
                     686,  
                     687,  
                     688,  
                     689,  
                     690,  
                     691,  
                     692,  
                     693,  
                     694,  
                     695,  
                     696,  
                     697,  
                     698,  
                     699,  
                     700,  
                     701,  
                     702,  
                     703,  
                     704,  
                     705,  
                     706,  
                     707,  
                     708,  
                     709,  
                     710,  
                     711,  
                     712,  
                     713,  
                     714,  
                     715,  
                     716,  
                     717,  
                     718,  
                     719,  
                     720,  
                     721,  
                     722,  
                     723,  
                     724,  
                     725,  
                     726,  
                     727,  
                     728,  
                     729,  
                     730,  
                     731,  
                     732,  
                     733,  
                     734,  
                     735,  
                     736,  
                     737,  
                     738,  
                     739,  
                     740,  
                     741,  
                     742,  
                     743,  
                     744,  
                     745,  
                     746,  
                     747,  
                     748,  
                     749,  
                     750,  
                     751,  
                     752,  
                     753,  
                     754,  
                     755,  
                     756,  
                     757,  
                     758,  
                     759,  
                     760,  
                     761,  
                     762,  
                     763,  
                     764,  
                     765,  
                     766,  
                     767,  
                     768,  
                     769,  
                     770,  
                     771,  
                     772,  
                     773,  
                     774,  
                     775,  
                     776,  
                     777,  
                     778,  
                     779,  
                     780,  
                     781,  
                     782,  
                     783,  
                     784,  
                     785,  
                     786,  
                     787,  
                     788,  
                     789,  
                     790,  
                     791,  
                     792,  
                     793,  
                     794,  
                     795,  
                     796,  
                     797,  
                     798,  
                     799,  
                     800,  
                     801,  
                     802,  
                     803,  
                     804,  
                     805,  
                     806,  
                     807,  
                     808,  
                     809,  
                     810,  
                     811,  
                     812,  
                     813,  
                     814,  
                     815,  
                     816,  
                     817,  
                     818,  
                     819,  
                     820,  
                     821,  
                     822,  
                     823,  
                     824,  
                     825,  
                     826,  
                     827,  
                     828,  
                     829,  
                     830,  
                     831,  
                     832,  
                     833,  
                     834,  
                     835,  
                     836,  
                     837,  
                     838,  
                     839,  
                     840,  
                     841,  
                     842,  
                     843,  
                     844,  
                     845,  
                     846,  
                     847,  
                     848,  
                     849,  
                     850,  
                     851,  
                     852,  
                     853,  
                     854,  
                     855,  
                     856,  
                     857,  
                     858,  
                     859,  
                     860,  
                     861,  
                     862,  
                     863,  
                     864,  
                     865,  
                     866,  
                     867,  
                     868,  
                     869,  
                     870,  
                     871,  
                     872,  
                     873,  
                     874,  
                     875,  
                     876,  
                     877,  
                     878,  
                     879,  
                     880,  
                     881,  
                     882,  
                     883,  
                     884,  
                     885,  
                     886,  
                     887,  
                     888,  
                     889,  
                     890,  
                     891,  
                     892,  
                     893,  
                     894,  
                     895,  
                     896,  
                     897,  
                     898,  
                     899,  
                     900,  
                     901,  
                     902,  
                     903,  
                     904,  
                     905,  
                     906,  
                     907,  
                     908,  
                     909,  
                     910,  
                     911,  
                     912,  
                     913,  
                     914,  
                     915,  
                     916,  
                     917,  
                     918,  
                     919,  
                     920,  
                     921,  
                     922,  
                     923,  
                     924,  
                     925,  
                     926,  
                     927,  
                     928,  
                     929,  
                     930,  
                     931,  
                     932,  
                     933,  
                     934,  
                     935,  
                     936,  
                     937,  
                     938,  
                     939,  
                     940,  
                     941,  
                     942,  
                     943,  
                     944,  
                     945,  
                     946,  
                     947,  
                     948,  
                     949,  
                     950,  
                     951,  
                     952,  
                     953,  
                     954,  
                     955,  
                     956,  
                     957,  
                     958,  
                     959,  
                     960,  
                     961,  
                     962,  
                     963,  
                     964,  
                     965,  
                     966,  
                     967,  
                     968,  
                     969,  
                     970,  
                     971,  
                     972,  
                     973,  
                     974,  
                     975,  
                     976,  
                     977,  
                     978,  
                     979,  
                     980,  
                     981,  
                     982,  
                     983,  
                     984,  
                     985,  
                     986,  
                     987,  
                     988,  
                     989,  
                     990,  
                     991,  
                     992,  
                     993,  
                     994,  
                     995,  
                     996,  
                     997,  
                     998,  
                     999,  
                     1000,  
                     1001,  
                     1002,  
                     1003,  
                     1004,  
                     1005,  
                     1006,  
                     1007,  
                     1008,  
                     1009,  
                     1010,  
                     1011,  
                     1012,  
                     1013,  
                     1014,  
                     1015,  
                     1016,  
                     1017,  
                     1018,  
                     1019,  
                     1020,  
                     1021,  
                     1022,  
                     1023,  
                     1024,  
                     1025,  
                     1026,  
                     1027,  
                     1028,  
                     1029,  
                     1030,  
                     1031,  
                     1032,  
                     1033,  
                     1034,  
                     1035,  
                     1036,  
                     1037,  
                     1038,  
                     1039,  
                     1040,  
                     1041,  
                     1042,  
                     1043,  
                     1044,  
                     1045,  
                     1046,  
                     1047,  
                     1048,  
                     1049,  
                     1050,  
                     1051,  
                     1052,  
                     1053,  
                     1054,  
                     1055,  
                     1056,  
                     1057,  
                     1058,  
                     1059,  
                     1060,  
                     1061,  
                     1062,  
                     1063,  
                     1064,  
                     1065,  
                     1066,  
                     1067,  
                     1068,  
                     1069,  
                     1070,  
                     1071,  
                     1072,  
                     1073,  
                     1074,  
                     1075,  
                     1076,  
                     1077,  
                     1078,  
                     1079,  
                     1080,  
                     1081,  
                     1082,  
                     1083,  
                     1084,  
                     1085,  
                     1086,  
                     1087,  
                     1088,  
                     1089,  
                     1090,  
                     1091,  
                     1092,  
                     1093,  
                     1094,  
                     1095,  
                     1096,  
                     1097,  
                     1098,  
                     1099,  
                     1100,  
                     1101,  
                     1102,  
                     1103,  
                     1104,  
                     1105,  
                     1106,  
                     1107,  
                     1108,  
                     1109,  
                     1110,  
                     1111,  
                     1112,  
                     1113,  
                     1114,  
                     1115,  
                     1116,  
                     1117,  
                     1118,  
                     1119,  
                     1120,  
                     1121,  
                     1122,  
                     1123,  
                     1124,  
                     1125,  
                     1126,  
                     1127,  
                     1128,  
                     1129,  
                     1130,  
                     1131,  
                     1132,  
                     1133,  
                     1134,  
                     1135,  
                     1136,  
                     1137,  
                     1138,  
                     1139,  
                     1140,  
                     1141,  
                     1142,  
                     1143,  
                     1144,  
                     1145,  
                     1146,  
                     1147,  
                     1148,  
                     1149,  
                     1150,  
                     1151,  
                     1152,  
                     1153,  
                     1154,  
                     1155,  
                     1156,  
                     1157,  
                     1158,  
                     1159,  
                     1160,  
                     1161,  
                     1162,  
                     1163,  
                     1164,  
                     1165,  
                     1166,  
                     1167,  
                     1168,  
                     1169,  
                     1170,  
                     1171,  
                     1172,  
                     1173,  
                     1174,  
                     1175,  
                     1176,  
                     1177,  
                     1178,  
                     1179,  
                     1180,  
                     1181,  
                     1182,  
                     1183,  
                     1184,  
                     1185,  
                     1186,  
                     1187,  
                     1188,  
                     1189,  
                     1190,  
                     1191,  
                     1192,  
                     1193,  
                     1194,  
                     1195,  
                     1196,  
                     1197,  
                     1198,  
                     1199,  
                     1200,  
                     1201,  
                     1202,  
                     1203,  
                     1204,  
                     1205,  
                     1206,  
                     1207,  
                     1208,  
                     1209,  
                     1210,  
                     1211,  
                     1212,  
                     1213,  
                     1214,  
                     1215,  
                     1216,  
                     1217,  
                     1218,  
                     1219,  
                     1220,  
                     1221,  
                     1222,  
                     1223,  
                     1224,  
                     1225,  
                     1226,  
                     1227,  
                     1228,  
                     1229,  
                     1230,  
                     1231,  
                     1232,  
                     1233,  
                     1234,  
                     1235,  
                     1236,  
                     1237,  
                     1238,  
                     1239,  
                     1240,  
                     1241,  
                     1242,  
                     1243,  
                     1244,  
                     1245,  
                     1246,  
                     1247,  
                     1248,  
                     1249,  
                     1250,  
                     1251,  
                     1252,  
                     1253,  
                     1254,  
                     1255,  
                     1256,  
                     1257,  
                     1258,  
                     1259,  
                     1260,  
                     1261,  
                     1262,  
                     1263,  
                     1264,  
                     1265,  
                     1266,  
                     1267,  
                     1268,  
                     1269,  
                     1270,  
                     1271,  
                     1272,  
                     1273,  
                     1274,  
                     1275,  
                     1276,  
                     1277,  
                     1278,  
                     1279,  
                     1280,  
                     1281,  
                     1282,  
                     1283,  
                     1284,  
                     1285,  
                     1286,  
                     1287,  
                     1288,  
                     1289,  
                     1290,  
                     1291,  
                     1292,  
                     1293,  
                     1294,  
                     1295,  
                     1296,  
                     1297,  
                     1298,  
                     1299,  
                     1300,  
                     1301,  
                     1302,  
                     1303,  
                     1304,  
                     1305,  
                     1306,  
                     1307,  
                     1308,  
                     1309,  
                     1310,  
                     1311,  
                     1312,  
                     1313,  
                     1314,  
                     1315,  
                     1316,  
                     1317,  
                     1318,  
                     1319,  
                     1320,  
                     1321,  
                     1322,  
                     1323,  
                     1324,  
                     1325,  
                     1326,  
                     1327,  
                     1328,  
                     1329,  
                     1330,  
                     1331,  
                     1332,  
                     1333,  
                     1334,  
                     1335,  
                     1336,  
                     1337,  
                     1338,  
                     1339,  
                     1340,  
                     1341,  
                     1342,  
                     1343,  
                     1344,  
                     1345,  
                     1346,  
                     1347,  
                     1348,  
                     1349,  
                     1350,  
                     1351,  
                     1352,  
                     1353,  
                     1354,  
                     1355,  
                     1356,  
                     1357,  
                     1358,  
                     1359,  
                     1360,  
                     1361,  
                     1362,  
                     1363,  
                     1364,  
                     1365,  
                     1366,  
                     1367,  
                     1368,  
                     1369,  
                     1370,  
                     1371,  
                     1372,  
                     1373,  
                     1374,  
                     1375,  
                     1376,  
                     1377,  
                     1378,  
                     1379,  
                     1380,  
                     1381,  
                     1382,  
                     1383,  
                     1384,  
                     1385,  
                     1386,  
                     1387,  
                     1388,  
                     1389,  
                     1390,  
                     1391,  
                     1392,  
                     1393,  
                     1394,  
                     1395,  
                     1396,  
                     1397,  
                     1398,  
                     1399,  
                     1400,  
                     1401,  
                     1402,  
                     1403,  
                     1404,  
                     1405,  
                     1406,  
                     1407,  
                     1408,  
                     1409,  
                     1410,  
                     1411,  
                     1412,  
                     1413,  
                     1414,  
                     1415,  
                     1416,  
                     1417,  
                     1418,  
                     1419,  
                     1420,  
                     1421,  
                     1422,  
                     1423,  
                     1424,  
                     1425,  
                     1426,  
                     1427,  
                     1428,  
                     1429,  
                     1430,  
                     1431,  
                     1432,  
                     1433,  
                     1434,  
                     1435,  
                     1436,  
                     1437,  
                     1438,  
                     1439,  
                     1440,  
                     1441,  
                     1442,  
                     1443,  
                     1444,  
                     1445,  
                     1446,  
                     1447,  
                     1448,  
                     1449,  
                     1450,  
                     1451,  
                     1452,  
                     1453,  
                     1454,  
                     1455,  
                     1456,  
                     1457,  
                     1458,  
                     1459,  
                     1460,  
                     1461,  
                     1462,  
                     1463,  
                     1464,  
                     1465,  
                     1466,  
                     1467,  
                     1468,  
                     1469,  
                     1470,  
                     1471,  
                     1472,  
                     1473,  
                     1474,  
                     1475,  
                     1476,  
                     1477,  
                     1478,  
                     1479,  
                     1480,  
                     1481,  
                     1482,  
                     1483,  
                     1484,  
                     1485,  
                     1486,  
                     1487,  
                     1488,  
                     1489,  
                     1490,  
                     1491,  
                     1492,  
                     1493,  
                     1494,  
                     1495,  
                     1496,  
                     1497,  
                     1498,  
                     1499,  
                     1500,  
                     1501,  
                     1502,  
                     1503,  
                     1504,  
                     1505,  
                     1506,  
                     1507,  
                     1508,  
                     1
```

```

In [139]: def make_graph(c_dict):
    '''
    This function accepts a dictionary with number of lines and scenes to create a
    NetworkX graph object
    '''
    # setup graph object
    G = nx.Graph()

    # add nodes with attributes of number of lines and scenes
    for c in c_dict.keys():
        if c_dict[c]["num_lines"] > 0:
            G.add_node(
                c,
                number_of_lines=c_dict[c]["num_lines"],
                scenes=c_dict[c]["scenes"]
            )

    # make edges by iterating over all combinations of nodes
    for (node1, data1), (node2, data2) in itertools.combinations(G.nodes(data=True)):
        # count scenes together by getting union of their sets
        scenes_together = len(set(data1['scenes'] & set(data2['scenes'])))

        if scenes_together:
            # add more weight for more scenes together
            G.add_edge(node1, node2, weight=scenes_together)

    return G

```

```

In [149]: import numpy as np
import networkx as nx
from lxml import etree
import itertools
from datascience import *
import matplotlib.pyplot as plt

def make_graph(c_dict):
    '''
    This function accepts a dictionary with number of lines and scenes to create a
    NetworkX graph object
    '''
    # setup graph object
    G = nx.Graph()

    # add nodes with attributes of number of lines and scenes
    for c in cast_dict.keys():
        G.add_node(
            c,
            scenes = cast_dict[c]
        )

    # make edges by iterating over all combinations of nodes
    for (node1, data1), (node2, data2) in itertools.combinations(G.nodes(data=True)):
        # count scenes together by getting union of their sets
        scenes_together = len(set(data1['scenes']) & set(data2['scenes']))

        if scenes_together:
            # add more weight for more scenes together
            G.add_edge(node1, node2, weight=scenes_together)

    return G

```

```

In [150]: G = make_graph(cast_dict)

```

```

In [153]: import numpy as np
import networkx as nx
from lxml import etree
import itertools
from datascience import *
import matplotlib.pyplot as plt

node_color = 'blue'

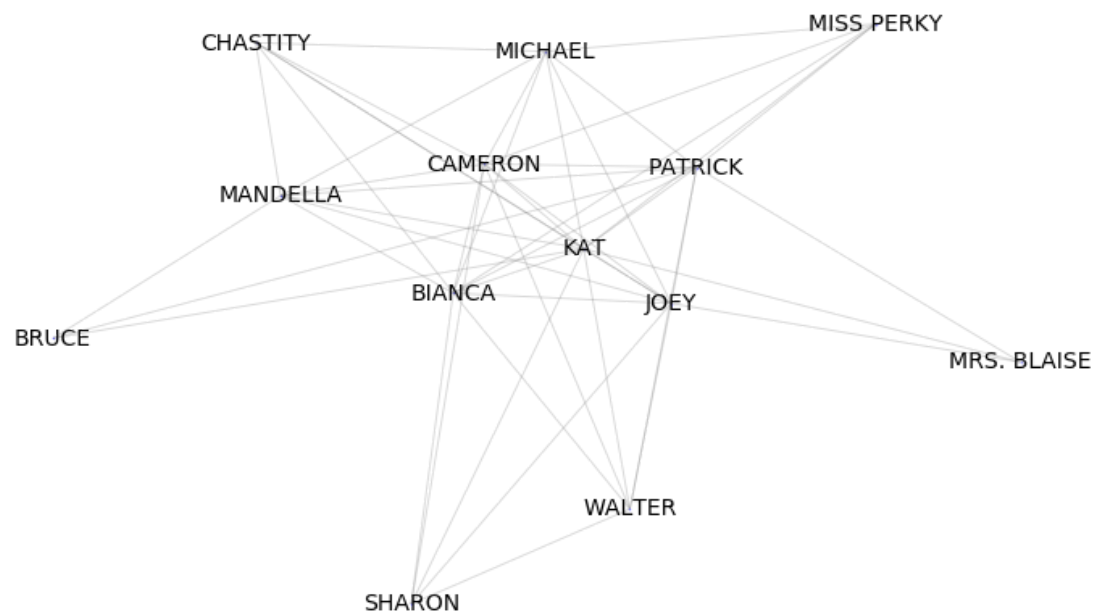
plt.figure(figsize=(13,8)) # make the figure size a little larger
plt.axis('off') # remove the axis, which isn't meaningful in this case
plt.title("10 Things I Hate About You", fontsize=20)

# The 'k' argument determines how spaced out the nodes will be from
# one another on the graph.
pos = nx.spring_layout(G, k=0.5)

nx.draw_networkx(
    G,
    pos=pos,
    node_size=node_size,
    node_color=node_color,
    edge_color='gray', # change edge color
    alpha=0.3, # make nodes more transparent to make labels clearer
    font_size=14,
)

```

## 10 Things I Hate About You

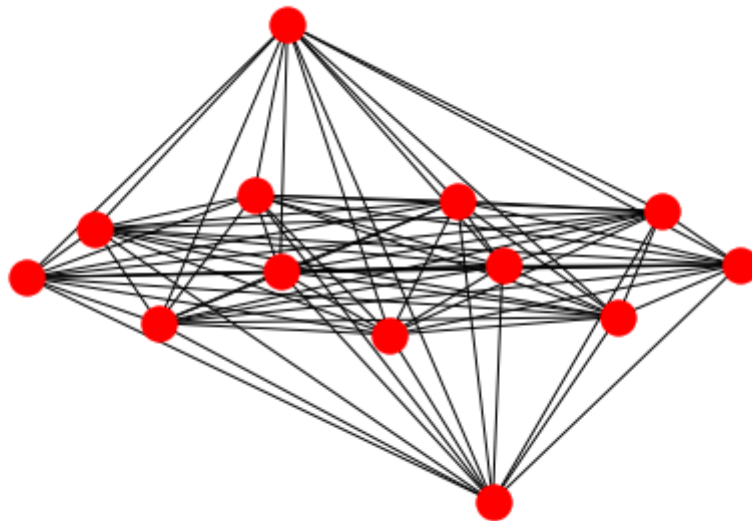


```
In [143]: import numpy as np
import networkx as nx
from lxml import etree
import itertools
from datascience import *
import matplotlib.pyplot as plt

nx.draw(
    G)

plt.show()
```

```
/srv/app/venv/lib/python3.6/site-packages/networkx/drawing/nx_pylab.py:12
6: MatplotlibDeprecationWarning: pyplot.hold is deprecated.
    Future behavior will be consistent with the long-time default:
    plot commands add elements without first clearing the
    Axes and/or Figure.
    b = plt.ishold()
/srv/app/venv/lib/python3.6/site-packages/networkx/drawing/nx_pylab.py:13
8: MatplotlibDeprecationWarning: pyplot.hold is deprecated.
    Future behavior will be consistent with the long-time default:
    plot commands add elements without first clearing the
    Axes and/or Figure.
    plt.hold(b)
/srv/app/venv/lib/python3.6/site-packages/matplotlib/__init__.py:917: Use
rWarning: axes.hold is deprecated. Please remove it from your matplotlib
r c and/or style files.
    warnings.warn(self.msg_depr_set % key)
/srv/app/venv/lib/python3.6/site-packages/matplotlib/rcsetup.py:152: User
Warning: axes.hold is deprecated, will be removed in 3.0
    warnings.warn("axes.hold is deprecated, will be removed in 3.0")
```



In [ ]:

In [ ]:

In [ ]:

In [87]: `'hello'.find('e')`

Out[87]: 1

In [7]: `soup = BeautifulSoup(raw, 'html5lib')`

In [8]: `bolded = soup.find('td', {'class': 'scrtext'} ).find_all('b')`

In [9]: `b_text = [b.text.strip() for b in bolded]`

In [10]: `bolded_text = [b for b in b_text if len(b) > 0]`

In [11]: `sift_out = ['INT.', "EXT.", "-"]`

```
characters = []
for c in bolded_text:
    character = True
    for s in sift_out:
        if s in c:
            character = False

    if character == True:
        characters.append(c)
```

In [12]: `from collections import Counter`

In [13]: `[c[0] for c in Counter(characters).most_common() if c[1] > 5]`

Out[13]: `['KAT',  
'PATRICK',  
'BIANCA',  
'CAMERON',  
'MICHAEL',  
'JOEY',  
'WALTER',  
'MANDELLA',  
'MISS PERKY',  
'MRS. BLAISE',  
'CHASTITY',  
'SHARON',  
'BRUCE']`

In [ ]:

In [ ]:





