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 | 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; | : XXyTu |

... (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 | 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; | : XXyT |

... (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 | 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; | : XXyT |

... (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; | : XXyT |

... (1136 rows omitted)

In [129]:
```
data = data.where('title', are.not_equal_to('Back to the Future II & III Scr
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; | : XXyT |

... (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.", "-"] #differenetiate 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[

        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 - "0 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[(
```

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,
  93
```

In [139]:
```
def make_graph(c_dict):
    '''
    This function accepts a dictionary with number of lines and scenes to cr
    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(dat

        # 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 cr
    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(dat

        # 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)
```

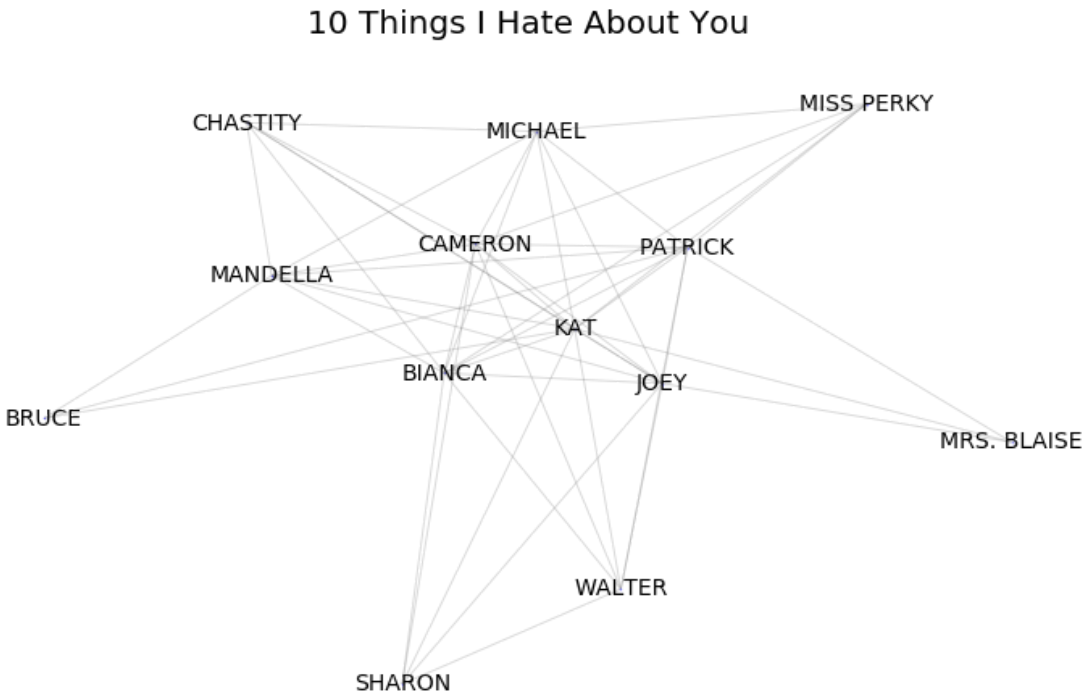
In [153]:
```
import numpy as np
import networkx as nx
from lxml import etree
import itertools
from datascience import *
import matplotlib.pyplot as plt

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,
)
```

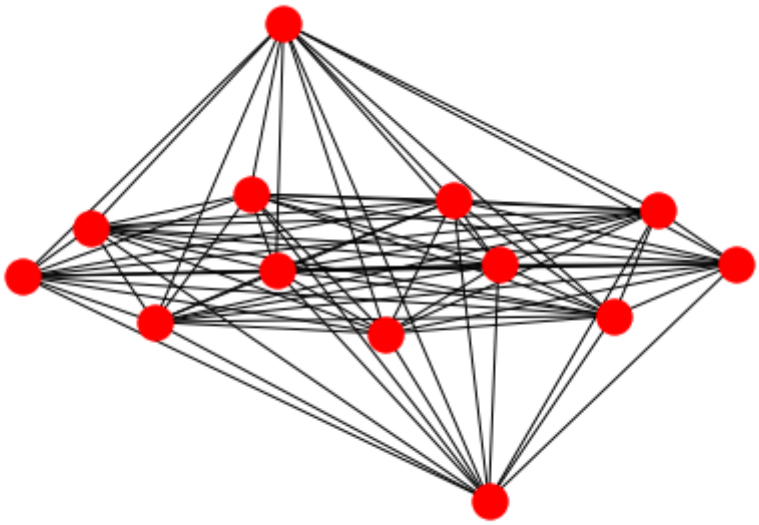


In [143]:
```
import numpy as np
import networkx as nx
from lxml import etree
import itertools
from datascience import *
import matplotlib.pyplot as plt

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 matplotlibr
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 [ ]:

```
```