In [1]:
```
from bs4 import BeautifulSoup, NavigableString, Tag
from datascience import *
from collections import Counter
```

In [2]:
```
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 [3]:

```
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;	: XXyT

... (1165 rows omitted)

In [4]:

```
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;	: XXyT

... (1137 rows omitted)

In [5]:
```
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	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;	: XXyT

... (1136 rows omitted)

In [6]:
```
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	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;	: XXyT

... (1135 rows omitted)

In [ ]:

In [27]:

```
## 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
```

```
---------------------------------------------------------------------------
--
KeyboardInterrupt                         Traceback (most recent call las
t)
<ipython-input-27-2a6a504be4e6> in <module>()
```

```
In [28]: all_scripts.keys()
```

```
Out[28]: 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'])
```

In [29]:
```
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)
```

```
EXT  HOTEL PARKING LOT - NIGHT
INT. HALLWAY - DAY
INT. KENNY'S THAI FOOD DINER - DAY
INT. DIVE BAR - NIGHT
BOGEY'S KITCHEN - NIGHT
INT. BOGEY LOWENSTEIN'S HOUSE - NIGHT
EXT. OUTDOOR ARCADE - DAY
EXT. PARKING LOT - DAY
INT. SCHOOL COURTYARD - DAY
INT. TUTORING ROOM
INT. GIRLS' ROOM - DAY
STRATFORD HOUSE/BATHROOM - NIGHT
HOTEL - NIGHT
INT. BOOK STORE - DAY
INSERT - "JOEY DORSEY SAID HI TO ME IN THE HALL! OH! MY
INT. STRATFORD HOUSE - NIGHT
INT. BOGEY'S BATHROOM - NIGHT
INT. DETENTION HALL - DAY
INT. CLUB - NIGHT
INT.  PROM - NIGHT - LATER
INT. BIOLOGY CLASS
INT. CAFETERIA - DAY
EXT. BOGEY LOWENSTEIN'S HOUSE - NIGHT
INT. BOGEY'S KITCHEN - NIGHT - LATER
GUIDANCE COUNSELOR'S OFFICE - DAY
EXT. FIELD HOCKEY FIELD - DAY
INT. KAT'S ROOM - NIGHT
BOGEY LOWENSTEIN'S HOUSE - NIGHT
INT. KAT'S CAR - NIGHT
INT. HALLWAY - DAY- CONTINUOUS
INT. CLUB FOYER - NIGHT
EXT. STRATFORD HOUSE - NIGHT
EXT. ARCHERY FIELD - DAY
INT. BOY'S ROOM - DAY
INT. LIVING ROOM - NIGHT
HALLWAY - DAY- CONTINUOUS
COURTYARD - DAY
INT. SHOWERS - DAY
PADUA HIGH SCHOOL - DAY
INT.  STRATFORD HOUSE/UPSTAIRS HALLWAY - NIGHT
EXT. PARKING LOT - MOMENTS LATER
EXT. MISS PERKY'S OFFICE - DAY
CAMERON'S CAR - NIGHT
INT.  STRATFORD HOUSE - DAY
INT. SOPHOMORE ENGLISH CLASS - DAY
STRATFORD HOUSE - SUNSET
EXT. SCHOOL PARKING LOT - DAY
CAFETERIA - DAY
ENGLISH CLASS - DAY
```

```
INT. ENGLISH CLASS - DAY
EXT. DOWNTOWN STREET - NIGHT
TRACK - DAY
INT. TUTORING ROOM - DAY
INT.  CAFETERIA - DAY - CONTINUOUS
BIANCA'S ROOM - DAY
LIVING ROOM - NIGHT
INT. STRATFORD HOUSE/DEN - DAY
INT. GUIDANCE COUNSELOR'S OFFICE - DAY
INT. LADIES ROOM - NIGHT
EXT. STRATFORD HOUSE - DAY
INT. WOODSHOP - DAY
INT. BIANCA'S ROOM - NIGHT
EXT. CLUB SKUNK - NIGHT
INT.  MISS PERKY'S OFFICE - DAY
STRATFORD HOUSE/BACKYARD - SUNSET
INT. BOGEY'S LIVING ROOM - NIGHT
INT. MISS PERKY'S OFFICE - DAY
INT. STRATFORD HOUSE/BATHROOM - NIGHT
HALLWAY - DAY
INT. STUDY HALL - DAY
EXT. SCHOOL COURTYARD - DAY
KAT'S CAR - NIGHT
CLASSROOM - DAY
PADUA HIGH PARKING LOT - DAY
EXT. STREET - NIGHT
INT.  MATH CLASS - DAY
INT. GYM CLASS - DAY
INT.  GYM CLASS - DAY
INT.  PROM - NIGHT
INT. STRATFORD HOUSE - DAY
EXT. SCHOOL CAMPUS LAWN
INT. BOGEY'S KITCHEN - NIGHT
INT. BOGEY'S DINING ROOM - NIGHT
INT. KAT'S ROOM - DAY
INSERT - "0 FAIR ONE.  JOIN ME AT THE PROM.  I WILL BE
INT. CLASSROOM - DAY
```

```
In [30]: len(scene_index_list )
```

```
Out[30]: 154
```

```
In [31]: 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 [32]: first_scene = scene_texts[0]
```

In [33]:

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

Out[33]:

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

In [34]:

```
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 [35]:

```
cast_dict
```

Out[35]:

```
{'BIANCA': [2,
  13,
  19,
  22,
  23,
  25,
  34,
  36,
  39,
  49,
  60,
  61,
  63,
  74,
  76,
  80,
  82,
  85,
  86,
```

In [36]:

```
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 [78]:

```
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(cast_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']))
        cast_dict[c]

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

    return G
```

In [79]:

```
G = make_graph(cast_dict)
```

In [80]:

```
import numpy as np
import networkx as nx
from lxml import etree
import itertools
from datascience import *
import matplotlib.pyplot as plt

node_size = 0.5
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 [81]:

```
network_tab = Table()
network_tab.append_column(label="Characters", values=[c for c in sorted(cast
network_tab.show()
```

Characters
BIANCA
BRUCE
CAMERON
CHASTITY
JOEY
KAT
MANDELLA
MICHAEL
MISS PERKY
MRS. BLAISE
PATRICK
SHARON
WALTER

In [82]:

```
dc = [x[1] for x in sorted(nx.degree_centrality(G).items(), key=lambda x: x[
network_tab.append_column(label="Degree Centrality", values=dc)
network_tab.show()
```

Characters	Degree Centrality
BIANCA	0.833333
BRUCE	0.25
CAMERON	0.833333
CHASTITY	0.5
JOEY	0.833333
KAT	1
MANDELLA	0.666667
MICHAEL	0.666667
MISS PERKY	0.416667
MRS. BLAISE	0.25
PATRICK	0.833333
SHARON	0.416667
WALTER	0.5

In [83]:

```
bc = [x[1] for x in sorted(nx.betweenness_centrality(G).items(), key=lambda
network_tab.append_column(label="Betweenness Centrality", values=bc)
network_tab.show()
```

Characters	Degree Centrality	Betweenness Centrality
BIANCA	0.833333	0.0454545
BRUCE	0.25	0
CAMERON	0.833333	0.0454545
CHASTITY	0.5	0
JOEY	0.833333	0.0671717
KAT	1	0.159091
MANDELLA	0.666667	0.030303
MICHAEL	0.666667	0.0123737
MISS PERKY	0.416667	0
MRS. BLAISE	0.25	0
PATRICK	0.833333	0.0916667
SHARON	0.416667	0
WALTER	0.5	0.0030303

In [84]:

```
ec = [x[1] for x in sorted(nx.eigenvector_centrality(G).items(), key=lambda
network_tab.append_column(label="Eigenvector Centrality", values=ec)
network_tab.show()
```

Characters	Degree Centrality	Betweenness Centrality	Eigenvector Centrality
BIANCA	0.833333	0.0454545	0.413741
BRUCE	0.25	0	0.0208809
CAMERON	0.833333	0.0454545	0.385503
CHASTITY	0.5	0	0.115439
JOEY	0.833333	0.0671717	0.304199
KAT	1	0.159091	0.49326
MANDELLA	0.666667	0.030303	0.181087
MICHAEL	0.666667	0.0123737	0.309785
MISS PERKY	0.416667	0	0.0908165
MRS. BLAISE	0.25	0	0.0384913
PATRICK	0.833333	0.0916667	0.417197
SHARON	0.416667	0	0.0626333
WALTER	0.5	0.0030303	0.118897

In [85]:

```
def gini(array):
    """Calculate the Gini coefficient of a numpy array."""
    # https://github.com/oliviaguest/gini
    array = np.sort(array) # values must be sorted
    index = np.arange(1, array.shape[0] + 1) # index per array element
    n = array.shape[0] # number of array elements
    return ((np.sum((2 * index - n  - 1) * array)) / (n * np.sum(array))) #
```

In [90]:

```
gini(network_tab.column('Eigenvector Centrality'))
```

Out[90]:

```
0.39558396783323707
```

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 [ ]:
```