4. Analysis of SOU similarities

In [1]:
```
import json
import numpy as np
import matplotlib.pyplot as plt
import string
import re
from math import log, pow
from sklearn.cluster import KMeans
```

In [2]:
```
with open('speeches.json', 'r') as f:
  speeches = json.loads(f.read())
```

(a) Compute the tf-idf vectors for each SOU address. You should lower case all of the text, and remove punctuation. You will have to make choices about the size of the term vocabulary to use—for example throwing out the 20 most common words, and words that appear fewer than, say, 50 times.

In [3]:
```
def clean_and_split(s):
    s = s.lower().replace('-',' ').translate(str.maketrans('', '', string
.punctuation))
    s = re.sub('(\r\n)+',' ',s)
    s = re.sub(" +",' ',s.strip())
    return s.split(' ')

text = np.array([clean_and_split(s['text']) for s in speeches])
```

In [4]:
```
# Count the number of appearances for each word in each document and
# the number of documents with each word.
vcount = {}
dcount = []

for d in text:
    c = {}
    for w in d:
        if w not in c.keys():
            c[w] = 0
        c[w] += 1
    dcount.append(c)
    for k in c.keys():
        if k not in vcount.keys():
            vcount[k] = 0
        vcount[k] +=1
```

In [5]:
```
# Remove all words in less than 50 documents and the 20 most common word
s.
vocab = list(filter(lambda w: vcount[w] >= 50, list(vcount.keys())))
for key, value in sorted(vcount.items(), key=lambda item: item[1], revers
e = True)[:20]:
    vocab.remove(key)
```

In [6]:
```
D = len(dcount)
def getScore(v,d):
    nid = d[v] if v in d.keys() else 0
    return nid * log(D/vcount[v])
scores = np.array([[getScore(v,d) for v in vocab] for d in dcount])
```

In [7]:
```
print("The TF-IDF vectors are:")
print(scores)
print("They are found in the variable \"scores\".")
```

```
The TF-IDF vectors are:
[[ 1.03608355  0.36713734  0.36818607 ...  0.          0.
   0.        ]
 [ 0.34536118  0.22028241  0.55227911 ...  0.          0.
   0.        ]
 [ 0.34536118  0.51399228  0.36818607 ...  0.          0.
   0.        ]
 ...
 [ 0.34536118  0.22028241  0.         ...  1.15785512 31.5412979
   7.15775476]
 [ 0.69072237  0.36713734  0.36818607 ...  3.47356537 42.89616514
   4.29465286]
 [ 0.69072237  0.36713734  0.36818607 ...  3.47356537 35.32625365
   4.29465286]]
They are found in the variable "scores".
```

(b) In terms of this similarity measure, find the

• 50 most similar pairs of SOUs given by different Presidents.
• 50 most similar pairs of SOUs given by the same President.
• 25 most similar pairs of Presidents, averaging the cosine similarity over all pairs of their SOUs.

When you read the above speeches, do they indeed seem similar to you? Comment on what you find, and describe what is needed to construct a better similarity measure between documents.

In [8]:
```
def calcSim(u,v):
    un = np.linalg.norm(u)
    vn = np.linalg.norm(v)
    return u.dot(v)/(un*vn)
sim = [[calcSim(u,v) for v in scores] for u in scores]
```

In [9]:
```
pairs = []
for i in range(len(speeches)):
    for j in range(i+1,len(speeches)):
        pairs.append((i,j, sim[i][j]))
sorted_pairs = sorted(pairs, key=lambda p: p[2], reverse=True)
```

In [10]:
```
sorted_pairs_diff_pres = list(filter(lambda p:
                                speeches[p[0]]['president'] != speeches[p[1
]]['president'],
                                sorted_pairs))[:50]
sorted_pairs_same_pres = list(filter(lambda p:
                                speeches[p[0]]['president'] == speeches[p[1
]]['president'],
                                sorted_pairs))[:50]
```

In [11]:
```
pres_speeches = {}
for (i, s) in enumerate(speeches):
    if s['president'] not in pres_speeches.keys():
        pres_speeches[s['president']] = []
    pres_speeches[s['president']].append(i)

def getPresSim(p1,p2):
    sum = 0
    count = 0
    for s1 in pres_speeches[p1]:
        for s2 in pres_speeches[p2]:
            sum += sim[s1][s2]
            count += 1
    return sum/count

pres_sim = []

for (i,p1) in enumerate(pres_speeches.keys()):
    for p2 in list(pres_speeches.keys())[i+1:]:
        if p1 != p2:
            pres_sim.append((p1, p2, getPresSim(p1,p2)))
pres_sim = sorted(pres_sim, key=lambda p: p[2], reverse=True)[:25]
```

In [12]:
```
print("The 50 most similar pairs of SOUs given by different Presidents ar
e:")
for p in sorted_pairs_diff_pres:
    print(speeches[p[0]]['president'] + " in " + speeches[p[0]]['year']
        + " and " +
          speeches[p[1]]['president'] + " in " + speeches[p[1]]['year']
         + " score:", p[2])

print(" ")
print("The 50 most similar pairs of SOUs given by the same Presidents ar
e:")
for p in sorted_pairs_same_pres:
    print(speeches[p[0]]['president'] + " in " + speeches[p[0]]['year']
            + " and " + " in " + speeches[p[1]]['year'] + " score: ", p[2
])
print(" ")
print("The 25 most similar pairs of Presidents, averaging the cosine simi
larity over all pairs of their SOUs are:")
for p in pres_sim:
    print(p[0] + " and " + p[1] + " score:", p[2])
```

```
The 50 most similar pairs of SOUs given by different Presidents are:
Dwight D. Eisenhower in 1961 and Jimmy Carter in 1981 score: 0.694959763
0564998
Grover Cleveland in 1885 and Benjamin Harrison in 1889 score: 0.67651903
87882957
John Tyler in 1844 and James K. Polk in 1846 score: 0.6737679650974868
Dwight D. Eisenhower in 1956 and Jimmy Carter in 1981 score: 0.664239830
3049182
William J. Clinton in 1994 and Barack Obama in 2010 score: 0.66174267350
99406
Rutherford B. Hayes in 1877 and Grover Cleveland in 1885 score: 0.653579
0336141586
Dwight D. Eisenhower in 1955 and Jimmy Carter in 1981 score: 0.653077583
5598227
John Tyler in 1844 and James K. Polk in 1845 score: 0.6504169118490623
Andrew Jackson in 1836 and Martin Van Buren in 1839 score: 0.64930253200
93996
Theodore Roosevelt in 1907 and William Howard Taft in 1912 score: 0.6446
578474869404
William J. Clinton in 1998 and George W. Bush in 2004 score: 0.642952171
0912769
George Bush in 1992 and William J. Clinton in 1994 score: 0.642130050710
5712
Grover Cleveland in 1885 and Benjamin Harrison in 1891 score: 0.64109913
77422067
William J. Clinton in 1995 and Barack Obama in 2010 score: 0.64086833832
74351
George Bush in 1992 and William J. Clinton in 1995 score: 0.640203828236
4412
William J. Clinton in 1993 and Barack Obama in 2010 score: 0.63535305102
16429
Rutherford B. Hayes in 1880 and Benjamin Harrison in 1889 score: 0.63204
29285297589
Rutherford B. Hayes in 1880 and Grover Cleveland in 1885 score: 0.631960
1421251311
William J. Clinton in 1994 and Barack Obama in 2011 score: 0.63053205950
49154
William J. Clinton in 1993 and Barack Obama in 2011 score: 0.62937143365
23359
Gerald R. Ford in 1976 and Ronald Reagan in 1982 score: 0.62891668378934
38
William J. Clinton in 1993 and Barack Obama in 2009 score: 0.62745351816
22762
Theodore Roosevelt in 1907 and William Howard Taft in 1910 score: 0.6269
854299485835
William J. Clinton in 1994 and Barack Obama in 2012 score: 0.62558650620
98257
Chester A. Arthur in 1881 and Grover Cleveland in 1894 score: 0.62515097
93322017
Benjamin Harrison in 1892 and Grover Cleveland in 1894 score: 0.62396770
7341668
Benjamin Harrison in 1891 and Grover Cleveland in 1894 score: 0.62392362
08153234
William J. Clinton in 1994 and Barack Obama in 2013 score: 0.62144490875
98246
Gerald R. Ford in 1975 and Jimmy Carter in 1981 score: 0.621202435902216
6
```

```
William McKinley in 1899 and William Howard Taft in 1912 score: 0.620666
1808006034
George Bush in 1992 and William J. Clinton in 1993 score: 0.617294054968
1873
William J. Clinton in 1998 and Barack Obama in 2011 score: 0.61659998226
60431
Benjamin Harrison in 1891 and Grover Cleveland in 1893 score: 0.61277470
16867883
William J. Clinton in 1993 and Barack Obama in 2012 score: 0.60853402146
29221
Chester A. Arthur in 1884 and Grover Cleveland in 1885 score: 0.60814473
41559101
William J. Clinton in 2000 and George W. Bush in 2004 score: 0.606464729
4737429
William J. Clinton in 1995 and Barack Obama in 2011 score: 0.60517198821
87322
Theodore Roosevelt in 1903 and William Howard Taft in 1912 score: 0.6040
066425507227
Chester A. Arthur in 1881 and Grover Cleveland in 1888 score: 0.60362711
79978956
Rutherford B. Hayes in 1880 and Benjamin Harrison in 1891 score: 0.60261
34999371929
Gerald R. Ford in 1975 and Ronald Reagan in 1981 score: 0.60139010274149
92
Gerald R. Ford in 1977 and Jimmy Carter in 1981 score: 0.601344872369140
1
John F. Kennedy in 1962 and Jimmy Carter in 1981 score: 0.60113462268166
22
Chester A. Arthur in 1881 and Grover Cleveland in 1886 score: 0.60049166
11280416
Zachary Taylor in 1849 and Millard Fillmore in 1851 score: 0.59993677096
1276
William J. Clinton in 1993 and Barack Obama in 2013 score: 0.59942811851
79814
Rutherford B. Hayes in 1877 and Benjamin Harrison in 1889 score: 0.59932
11078255823
Theodore Roosevelt in 1903 and William Howard Taft in 1910 score: 0.5986
494750576417
William J. Clinton in 1994 and Barack Obama in 2009 score: 0.59840257248
65817
William J. Clinton in 1995 and Barack Obama in 2012 score: 0.59776484195
23952

The 50 most similar pairs of SOUs given by the same Presidents are:
Barack Obama in 2012 and  in 2013 score:  0.9856400032688407
William McKinley in 1899 and  in 1900 score:  0.7548644048666687
William J. Clinton in 1997 and  in 1998 score:  0.7511945730215503
William J. Clinton in 1998 and  in 2000 score:  0.7508684193461371
William J. Clinton in 1998 and  in 1999 score:  0.7486873791988522
William J. Clinton in 1994 and  in 1995 score:  0.742226414590187
William Howard Taft in 1910 and  in 1912 score:  0.7412891546930293
Barack Obama in 2010 and  in 2012 score:  0.7391573205300171
Barack Obama in 2011 and  in 2012 score:  0.7379054772636774
William J. Clinton in 1999 and  in 2000 score:  0.7364126590107125
Barack Obama in 2010 and  in 2013 score:  0.73063282597191
Dwight D. Eisenhower in 1955 and  in 1956 score:  0.7266763213165804
Barack Obama in 2011 and  in 2013 score:  0.7221488933429239
Barack Obama in 2009 and  in 2010 score:  0.7200392750755801
```

```
Theodore Roosevelt in 1905 and  in 1907 score:  0.7174739185077383
James K. Polk in 1845 and  in 1846 score:  0.7165289361146102
William McKinley in 1898 and  in 1899 score:  0.7160897923374993
Grover Cleveland in 1894 and  in 1896 score:  0.70837795142263
Theodore Roosevelt in 1906 and  in 1907 score:  0.7030744952258874
William Howard Taft in 1909 and  in 1910 score:  0.7025380386658665
William J. Clinton in 1997 and  in 1999 score:  0.7006614001138864
Ronald Reagan in 1981 and  in 1982 score:  0.6959087910710089
Theodore Roosevelt in 1907 and  in 1908 score:  0.6901496293999095
Andrew Jackson in 1834 and  in 1835 score:  0.6898242946399905
Theodore Roosevelt in 1901 and  in 1905 score:  0.6891896368239985
Barack Obama in 2010 and  in 2011 score:  0.6891414493057106
James Buchanan in 1858 and  in 1859 score:  0.6856170014658136
Theodore Roosevelt in 1904 and  in 1905 score:  0.6846761581109158
Grover Cleveland in 1893 and  in 1894 score:  0.6845681359718145
William Howard Taft in 1909 and  in 1912 score:  0.6844344114860198
Rutherford B. Hayes in 1879 and  in 1880 score:  0.681629890256377
Grover Cleveland in 1885 and  in 1886 score:  0.6811369086667697
James K. Polk in 1846 and  in 1847 score:  0.6806590084648411
Grover Cleveland in 1886 and  in 1888 score:  0.6775992069570572
William J. Clinton in 1997 and  in 2000 score:  0.6766369731762988
Dwight D. Eisenhower in 1954 and  in 1956 score:  0.6759880641937046
William Howard Taft in 1911 and  in 1912 score:  0.6750267147239583
Dwight D. Eisenhower in 1954 and  in 1955 score:  0.6748159834094205
William Howard Taft in 1910 and  in 1911 score:  0.6723935238695319
Theodore Roosevelt in 1905 and  in 1906 score:  0.6690293719017278
Theodore Roosevelt in 1904 and  in 1907 score:  0.6684389058358007
Chester A. Arthur in 1881 and  in 1882 score:  0.6664907499756225
William J. Clinton in 1996 and  in 1997 score:  0.6662687494308936
William McKinley in 1897 and  in 1898 score:  0.66493092313906
Theodore Roosevelt in 1901 and  in 1907 score:  0.6601660706680892
William J. Clinton in 1995 and  in 1998 score:  0.6586219763946931
Benjamin Harrison in 1891 and  in 1892 score:  0.6564162683845197
Theodore Roosevelt in 1901 and  in 1902 score:  0.6560298441513013
Ronald Reagan in 1982 and  in 1983 score:  0.6543189157748855
William J. Clinton in 1993 and  in 1994 score:  0.6518506577221261

The 25 most similar pairs of Presidents, averaging the cosine similarity
over all pairs of their SOUs are:
Zachary Taylor and Millard Fillmore score: 0.5600710385007703
William J. Clinton and Barack Obama score: 0.55836300063597
Rutherford B. Hayes and Benjamin Harrison score: 0.5445367775464199
Rutherford B. Hayes and Chester A. Arthur score: 0.5406694833022884
Chester A. Arthur and Benjamin Harrison score: 0.5397896491511647
Theodore Roosevelt and William Howard Taft score: 0.5263365638594164
Grover Cleveland and Benjamin Harrison score: 0.5184061441047045
Benjamin Harrison and William Howard Taft score: 0.5168837148621588
Chester A. Arthur and William Howard Taft score: 0.5163181317227092
George Bush and William J. Clinton score: 0.5157968856178523
Rutherford B. Hayes and Grover Cleveland score: 0.5106146376348016
William McKinley and William Howard Taft score: 0.5043554659892407
Andrew Jackson and Martin Van Buren score: 0.5029530779370834
Gerald R. Ford and Jimmy Carter score: 0.5009770902469973
William J. Clinton and George W. Bush score: 0.5006018418570832
Chester A. Arthur and Grover Cleveland score: 0.4979612721482962
Zachary Taylor and James Buchanan score: 0.4944091153988176
James K. Polk and Millard Fillmore score: 0.49270689066883916
Millard Fillmore and James Buchanan score: 0.48875146666834896
```

```
Franklin Pierce and James Buchanan score: 0.48533253516601715
Ulysses S. Grant and William McKinley score: 0.4846084406903425
Ronald Reagan and William J. Clinton score: 0.48396446324270737
Millard Fillmore and Franklin Pierce score: 0.48362887020154827
Rutherford B. Hayes and William Howard Taft score: 0.4833911462849727
Chester A. Arthur and William McKinley score: 0.48216680047333366
```

When I read them, they do indeed seem similar. They are not super similar, but definitely have similar styles of rheteric and even protray the same ideas in some cases. However, the sentence structure is not too similar. This is what I would expect given the limitations of the similarity scores as the scores do not reflect the ordering of words at all.

(c) Using this vector representation, cluster the speeches using k-means.

The options here limit the number of iterations of kmeans to 50, the number of clusters to 10, the clusters are initialized randomly.
Experiment with different number of clusters, and display the clusters obtained (in some manner that you choose). Comment on the clustering results, and whether or not the results are interpretable.

In [13]:
```
c=10
model = KMeans(n_clusters=c, max_iter=50)
sou_clust=model.fit(scores)
labels = model.predict(scores)
clen = []
for i in range(c):
    print("Cluster", i)
    cl = 0
    for c in np.where(labels == i)[0]:
        cl += 1
        print(speeches[c]['president'], "in", speeches[c]['year'])
    clen.append(cl)
```

```
Cluster 0
William J. Clinton in 1993
William J. Clinton in 1994
William J. Clinton in 1995
William J. Clinton in 1996
William J. Clinton in 1997
William J. Clinton in 1998
William J. Clinton in 1999
William J. Clinton in 2000
Barack Obama in 2009
Barack Obama in 2010
Barack Obama in 2011
Barack Obama in 2012
Barack Obama in 2013
Cluster 1
Franklin D. Roosevelt in 1945
Harry S Truman in 1947
Harry S Truman in 1948
Harry S Truman in 1949
Harry S Truman in 1950
Harry S Truman in 1951
Harry S Truman in 1952
Harry S Truman in 1953
Dwight D. Eisenhower in 1953
Dwight D. Eisenhower in 1954
Dwight D. Eisenhower in 1955
Dwight D. Eisenhower in 1956
Dwight D. Eisenhower in 1957
Dwight D. Eisenhower in 1958
Dwight D. Eisenhower in 1959
Dwight D. Eisenhower in 1960
Dwight D. Eisenhower in 1961
John F. Kennedy in 1961
John F. Kennedy in 1962
John F. Kennedy in 1963
Lyndon B. Johnson in 1964
Lyndon B. Johnson in 1965
Lyndon B. Johnson in 1966
Lyndon B. Johnson in 1967
Lyndon B. Johnson in 1968
Lyndon B. Johnson in 1969
Richard M. Nixon in 1970
Richard M. Nixon in 1971
Richard M. Nixon in 1972
Richard M. Nixon in 1974
Gerald R. Ford in 1975
Gerald R. Ford in 1976
Gerald R. Ford in 1977
Jimmy Carter in 1978
Jimmy Carter in 1979
Jimmy Carter in 1980
Ronald Reagan in 1981
Ronald Reagan in 1982
Ronald Reagan in 1983
Ronald Reagan in 1984
Ronald Reagan in 1985
Ronald Reagan in 1986
```

```
Ronald Reagan in 1987
Ronald Reagan in 1988
George Bush in 1989
George Bush in 1990
George Bush in 1991
George Bush in 1992
George W. Bush in 2001
George W. Bush in 2002
George W. Bush in 2003
George W. Bush in 2004
George W. Bush in 2005
George W. Bush in 2006
George W. Bush in 2007
George W. Bush in 2008
Cluster 2
James Monroe in 1824
Andrew Jackson in 1829
Andrew Jackson in 1830
Andrew Jackson in 1833
Andrew Jackson in 1834
Andrew Jackson in 1835
Andrew Jackson in 1836
Martin Van Buren in 1837
Martin Van Buren in 1838
Martin Van Buren in 1839
Martin Van Buren in 1840
John Tyler in 1841
John Tyler in 1842
John Tyler in 1843
John Tyler in 1844
James K. Polk in 1845
James K. Polk in 1848
Zachary Taylor in 1849
Millard Fillmore in 1850
Millard Fillmore in 1851
Millard Fillmore in 1852
Franklin Pierce in 1853
Franklin Pierce in 1854
Franklin Pierce in 1855
Franklin Pierce in 1856
James Buchanan in 1857
James Buchanan in 1858
James Buchanan in 1859
James Buchanan in 1860
Abraham Lincoln in 1862
Andrew Johnson in 1866
Andrew Johnson in 1867
Andrew Johnson in 1868
Ulysses S. Grant in 1869
Ulysses S. Grant in 1870
Ulysses S. Grant in 1872
Ulysses S. Grant in 1873
Ulysses S. Grant in 1874
Ulysses S. Grant in 1875
Rutherford B. Hayes in 1877
Rutherford B. Hayes in 1878
Rutherford B. Hayes in 1879
Chester A. Arthur in 1882
```

```
Chester A. Arthur in 1883
Chester A. Arthur in 1884
Cluster 3
Rutherford B. Hayes in 1880
Chester A. Arthur in 1881
Grover Cleveland in 1885
Grover Cleveland in 1886
Grover Cleveland in 1888
Benjamin Harrison in 1889
Benjamin Harrison in 1890
Benjamin Harrison in 1891
Benjamin Harrison in 1892
Grover Cleveland in 1893
Grover Cleveland in 1894
Grover Cleveland in 1895
Grover Cleveland in 1896
William McKinley in 1897
William McKinley in 1898
William McKinley in 1899
William McKinley in 1900
Theodore Roosevelt in 1903
William Howard Taft in 1909
Cluster 4
James K. Polk in 1846
James K. Polk in 1847
Cluster 5
Jimmy Carter in 1981
Cluster 6
Harry S Truman in 1946
Cluster 7
Theodore Roosevelt in 1901
Theodore Roosevelt in 1904
Theodore Roosevelt in 1905
Theodore Roosevelt in 1906
Theodore Roosevelt in 1907
Theodore Roosevelt in 1908
William Howard Taft in 1912
Cluster 8
George Washington in 1790
George Washington in 1791
George Washington in 1792
George Washington in 1793
George Washington in 1794
George Washington in 1795
George Washington in 1796
John Adams in 1797
John Adams in 1798
John Adams in 1799
John Adams in 1800
Thomas Jefferson in 1801
Thomas Jefferson in 1802
Thomas Jefferson in 1803
Thomas Jefferson in 1804
Thomas Jefferson in 1805
Thomas Jefferson in 1806
Thomas Jefferson in 1807
Thomas Jefferson in 1808
James Madison in 1809
```

```
James Madison in 1810
James Madison in 1811
James Madison in 1812
James Madison in 1813
James Madison in 1814
James Madison in 1815
James Madison in 1816
James Monroe in 1817
James Monroe in 1818
James Monroe in 1819
James Monroe in 1820
James Monroe in 1821
James Monroe in 1822
James Monroe in 1823
John Quincy Adams in 1825
John Quincy Adams in 1826
John Quincy Adams in 1827
John Quincy Adams in 1828
Andrew Jackson in 1831
Andrew Jackson in 1832
Abraham Lincoln in 1861
Abraham Lincoln in 1863
Abraham Lincoln in 1864
Andrew Johnson in 1865
Ulysses S. Grant in 1871
Ulysses S. Grant in 1876
Grover Cleveland in 1887
Theodore Roosevelt in 1902
Woodrow Wilson in 1913
Woodrow Wilson in 1914
Woodrow Wilson in 1915
Woodrow Wilson in 1916
Woodrow Wilson in 1917
Woodrow Wilson in 1918
Woodrow Wilson in 1919
Woodrow Wilson in 1920
Warren G. Harding in 1921
Warren G. Harding in 1922
Calvin Coolidge in 1923
Calvin Coolidge in 1924
Calvin Coolidge in 1925
Calvin Coolidge in 1926
Calvin Coolidge in 1927
Calvin Coolidge in 1928
Herbert Hoover in 1929
Herbert Hoover in 1930
Herbert Hoover in 1931
Herbert Hoover in 1932
Franklin D. Roosevelt in 1934
Franklin D. Roosevelt in 1935
Franklin D. Roosevelt in 1936
Franklin D. Roosevelt in 1937
Franklin D. Roosevelt in 1938
Franklin D. Roosevelt in 1939
Franklin D. Roosevelt in 1940
Franklin D. Roosevelt in 1941
Franklin D. Roosevelt in 1942
Franklin D. Roosevelt in 1943
```

```
Franklin D. Roosevelt in 1944
Richard M. Nixon in 1973
Cluster 9
William Howard Taft in 1910
William Howard Taft in 1911
```

Its clear that the kmeans algorithm tends to cluster the speeches with similar simularity scores together. After experimenting with different amounts of clusters, the speeches with the higher similarity scores that we determined earlier stay clustered when more clusters are added and the less similar speeches tend to split off. This is what I would expect to be the case, but it is cool to see how this occurs and how we can use kmeans to group the speeches in this sense. I found that 10 clusters was a good median for where the speeches in each cluster have a high enough corresponding similarity score.