Step 1 Construct features using NLTK

In [1]:
```
import nltk
from nltk.corpus import stopwords
import string
import json

%matplotlib inline
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
from __future__ import division

randomState = np.random.seed(2)


# Import works.json: title, words, sentences of 38 pieces
with open('works_v1028.json', 'r') as f:
    try:
        works = json.load(f)
    # if the file is empty the ValueError will be thrown
    except ValueError:
        works = {}
```

In [2]:
```
from nltk import FreqDist

def getFreqDist(words):
    nltk_text = nltk.Text(words)

    return FreqDist(nltk_text)
```

In [3]:
```
def getCountFrequency(words, freq):
    nltk_text = nltk.Text(words)
    freq_dist = FreqDist(nltk_text)

    count = 0
    for key,value in freq_dist.items():
        if value == freq:
            count += 1

    return count

def getHapaxLegomena(words):
    return getCountFrequency(words, 1)
```

In [4]:
```
def getDisLegomena(words):
    return getCountFrequency(words, 2)
```

In [5]:
```
def getNumberUniqueWords(words):
    return len(getFreqDist(words))
```

In [6]:
```
def getWordLengthDistribution(words):
    result = FreqDist()
    for w in words:
        result[len(w)] += 1

    return result

def printFreqDist(freq_dist):
    for key,value in freq_dist.items():
        print (key,value)
```

In [7]:
```
def getSentenceLengthDistribution(sents):
    result = FreqDist()
    for s in sents:
        result[len(s)] += 1

    return result
```

In [8]:
```
def getAverageSentenceLength(words_all, sents):

    return len(words_all)/len(sents)
```

In [9]:
```
def getPronounDistribution(sents):
    result = FreqDist()
    pronouns = ['I', 'you', 'he', 'she', 'it', 'we', 'you', 'they', 'thou', 'ye']
    for s in sents:
        count = 0
        for p in pronouns:
            count += s.count(p)
        result[count] += 1

    return result
```

In [10]:
```
def getNumberOfPronoun(words):
    pronouns = ['I', 'you', 'he', 'she', 'it', 'we', 'you', 'they', 'thou', 'ye']
    count = 0

    for p in pronouns:
        count += words.count(p)

    return count
```

In [11]:
```
def getConjunctionDistribution(sents):
    result = FreqDist()
    conjunctions = ['for', 'and', 'nor', 'but', 'or', 'yet', 'so', 'after', 'although', 'as',
'because', 'before', 'if',
                'once', 'since', 'though', 'unless', 'until', 'when', 'where', 'while', 'wheth
er', 'either', 'neither',
                'as', 'than', 'even', 'then', 'lest', 'rather', 'which', 'who', 'why', 'that',
'till']
    for s in sents:
        count = 0
        for c in conjunctions:
            count += s.count(c)
        result[count] += 1

    return result
```

In [12]:
```
def getNumberOfConjunction(words):
    count = 0
    conjunctions = ['for', 'and', 'nor', 'but', 'or', 'yet', 'so', 'after', 'although', 'as',
'because', 'before', 'if',
                'once', 'since', 'though', 'unless', 'until', 'when', 'where', 'while', 'wheth
er', 'either', 'neither',
                'as', 'than', 'even', 'then', 'lest', 'rather', 'which', 'who', 'why', 'that',
'till']

    for c in conjunctions:
        count += words.count(c)

    return count
```

In [13]:
```
def getNumberOfComma(text):
    return text.count(",")

def getNumberOfDots(text):
    return text.count(".")

def getNumberOfStringList(words, string_list):
    count = 0

    for c in string_list:
        count += words.count(c)

    return count

def getNumberOfThe(words):
    string_list = ['a', 'an', 'the']

    return getNumberOfStringList(words, string_list)

def getNumberOfIn(words):
    string_list = ['in', 'on', 'to', 'of']

    return getNumberOfStringList(words, string_list)

def getNumberOfIs(words):
    string_list = ['is', 'are', 'was', 'were']

    return getNumberOfStringList(words, string_list)
```

In [14]:
```
# Create a feature table for all acts
# Currently only has some simple features

featureTable = {}

featureTable = {'act': [], 'filename': [], 'HapaxLegomena': [], 'DisLegomena': [],
                'numberUniqueWords': [], 'averageSentenceLength': [], 'numberOfSentences': [],
                'numberOfPronoun': [], 'numberOfConjunction': [], 'numberOfComma': [], 'number
OfDots': [],
                'numberOfThe': [], 'numberOfIn': [], 'numberOfIs': []}


for act, content in works.items():
    words_nonStemmed = content['words_nonStemmed']
    words_all = content['words_all']
    sentences = content['sentences']
    # Normalize by number of all words
    normalizer = len(words_all)
    featureTable['act'].append(act)
    featureTable['filename'].append(content['filename'])
    featureTable['HapaxLegomena'].append(getHapaxLegomena(words_nonStemmed)/normalizer)
    featureTable['DisLegomena'].append(getDisLegomena(words_nonStemmed)/normalizer)
    featureTable['numberUniqueWords'].append(getNumberUniqueWords(words_nonStemmed)/normalize
r)
    featureTable['averageSentenceLength'].append(getAverageSentenceLength(words_all, sentence
s))
    featureTable['numberOfSentences'].append(len(sentences))
    featureTable['numberOfPronoun'].append(getNumberOfPronoun(words_all)/normalizer)
    featureTable['numberOfConjunction'].append(getNumberOfConjunction(words_all)/normalizer)
    featureTable['numberOfComma'].append(getNumberOfComma(content['text'])/normalizer)
    featureTable['numberOfDots'].append(getNumberOfDots(content['text'])/normalizer)
    featureTable['numberOfThe'].append(getNumberOfThe(words_all)/normalizer)
    featureTable['numberOfIn'].append(getNumberOfIn(words_all)/normalizer)
    featureTable['numberOfIs'].append(getNumberOfIs(words_all)/normalizer)

print(len(featureTable['act']))


df = pd.DataFrame(featureTable)
df.set_index(['act'], inplace=True)
```

```
df.head()
```

```
218
```

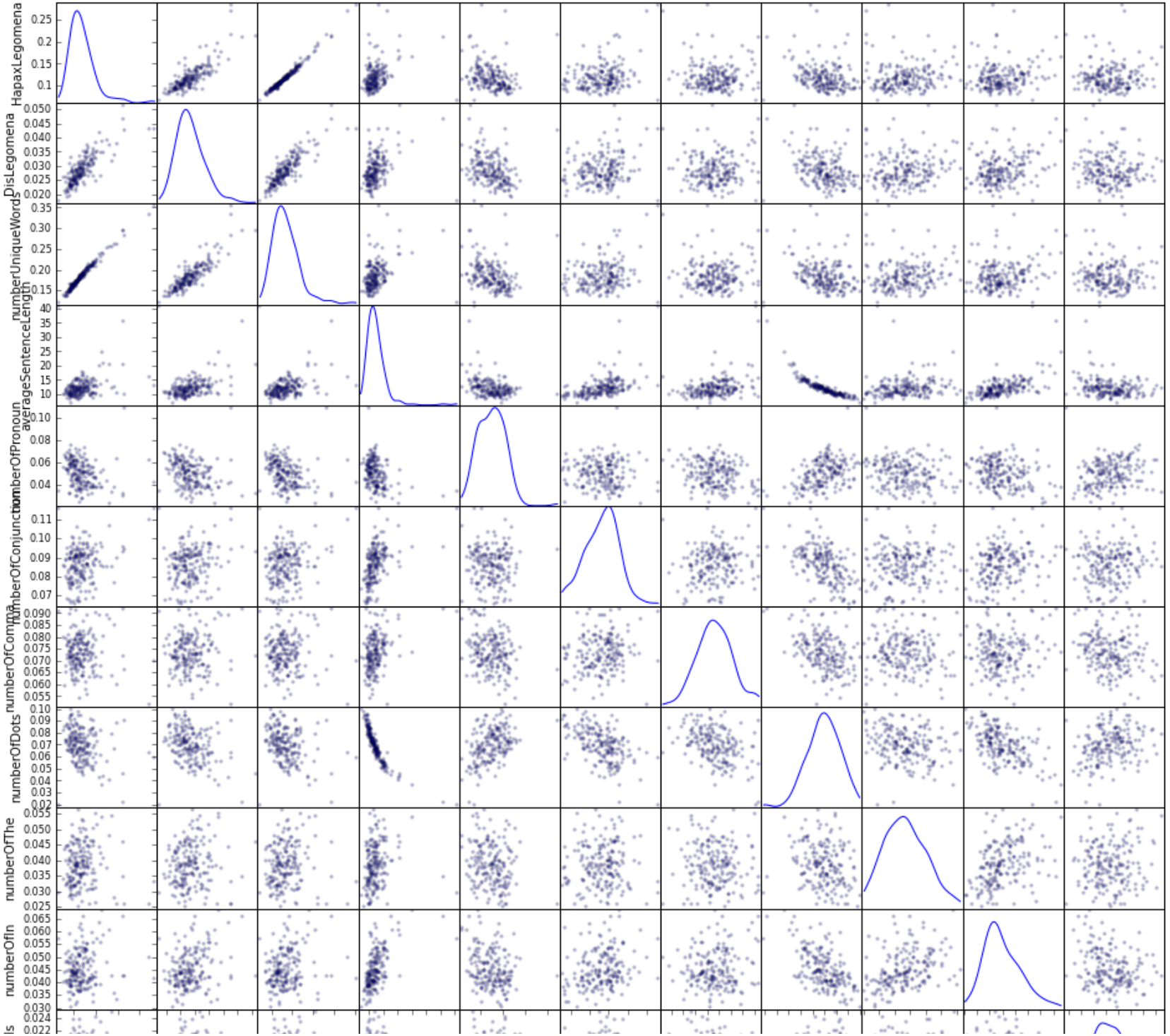
Out[14]:

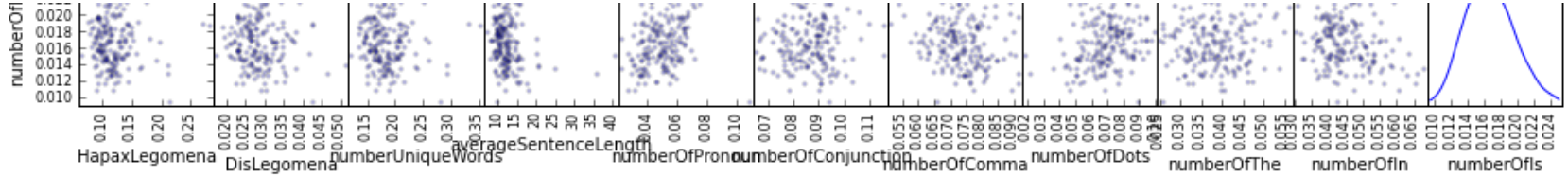
	DisLegomena	HapaxLegomena	averageSentenceLength	filename	numberOfComma	numberOfConju
act						
MEASURE FOR MEASURE ACT II.	0.019071	0.084440	10.955331	21.txt	0.076286	0.094699
THE TEMPEST ACT1612	0.066667	0.384615	65.000000	31.txt	0.076923	0.071795
THE TRAGEDY OF OTHELLO, MOOR OF VENICE ACT III	0.023032	0.089297	9.396991	26.txt	0.071191	0.089789
THE TRAGEDY OF OTHELLO, MOOR OF VENICE ACT V.	0.021995	0.093855	7.056872	26.txt	0.071357	0.073539
KING HENRY THE EIGHTH ACT V.	0.051304	0.213043	20.535714	15.txt	0.072174	0.093043

In [15]:
```
# Filter out wierd acts, e.g. the beginning of a play

# df = df[df['averageSentenceLength'] < 30]

df.drop(["A MIDSUMMER NIGHT'S DREAM ACT1596","ALLS WELL THAT ENDS WELL ACT1603", "AS YOU LIKE
IT ACT1601",
              "CYMBELINE ACT1609","KING HENRY THE EIGHTH ACT1611","KING JOHN ACT1597","KING RI
CHARD III ACT1593",
             "KING RICHARD THE SECOND ACT1596", "LOVE'S LABOUR'S LOST ACT1595", "MEASURE FOR M
EASURE ACT1605",
             "MUCH ADO ABOUT NOTHING ACT1599", "SECOND PART OF KING HENRY IV ACT1598", "THE CO
MEDY OF ERRORS ACT1593",
             "THE FIRST PART OF HENRY THE SIXTH ACT1592", "THE FIRST PART OF KING HENRY THE FO
URTH ACT1598",
             "THE HISTORY OF TROILUS AND CRESSIDA ACT1602", "THE LIFE OF KING HENRY THE FIFTH
ACT1599",
             "THE LIFE OF TIMON OF ATHENS ACT1608", "THE MERCHANT OF VENICE ACT1597", "THE MER
RY WIVES OF WINDSOR ACT1601",
             "THE SECOND PART OF KING HENRY THE SIXTH ACT1591", "THE TAMING OF THE SHREW ACT15
94", "THE TEMPEST ACT1612",
             "THE TRAGEDY OF ANTONY AND CLEOPATRA ACT1607", "THE TRAGEDY OF CORIOLANUS ACT160
8",
              "THE TRAGEDY OF HAMLET, PRINCE OF DENMARK ACT1604","THE TRAGEDY OF JULIUS CAESAR
ACT1599",
             "THE TRAGEDY OF KING LEAR ACT1606", "THE TRAGEDY OF MACBETH ACT1606",
              "THE TRAGEDY OF OTHELLO, MOOR OF VENICE ACT1605", "THE TRAGEDY OF ROMEO AND JULI
ET ACT1595",
             "THE TRAGEDY OF TITUS ANDRONICUS ACT1594", "THE TWO GENTLEMEN OF VERONA ACT1595",
"THE WINTER'S TALE ACT1611",
             "TWELFTH NIGHT; OR, WHAT YOU WILL ACT1602", "THE THIRD PART OF KING HENRY THE SIX
TH ACT1591"], inplace = True)

df_2 = df[['numberUniqueWords', 'averageSentenceLength']]
df_3 = df[['numberUniqueWords', 'averageSentenceLength', 'numberOfPronoun']]
df = df[['HapaxLegomena', 'DisLegomena', 'numberUniqueWords',
        'averageSentenceLength', 'numberOfPronoun', 'numberOfConjunction', 'numberOfComma', 'n
umberOfDots',
               'numberOfThe', 'numberOfIn', 'numberOfIs']]
```

In [16]:
```
pd.scatter_matrix(df, alpha=0.2, figsize=(16, 16), diagonal='kde')
plt.show()
```

```
/Users/binyan/Python/anaconda/envs/py34/lib/python3.4/site-packages/matplotlib/collections.p
y:590: FutureWarning: elementwise comparison failed; returning scalar instead, but in the fu
ture will perform elementwise comparison
  if self._edgecolors == str('face'):
```


HapaxLegomena
DisLegomena
numberUniqueWords
averageSentenceLength
numberOfPronoun
numberOfConjunction
numberOfComma
numberOfDots
numberOfThe
numberOfIn



Step 2 Construct n-grams using sklearn

In [17]:
```
title = []
text = []

for act in df.index:
    content = works[act]
    title.append(act)
    text.append(content['text'])
```

In [18]:
```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from __future__ import print_function

from sklearn.decomposition import TruncatedSVD
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import HashingVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import Normalizer
from sklearn import metrics

from sklearn.cluster import KMeans, MiniBatchKMeans, AffinityPropagation, AgglomerativeCluster
ing, MeanShift, SpectralClustering, DBSCAN, Birch
import numpy as np
```

In [19]:
```
# excluding stop words, n-grams is an input
def getTfidf(ngram_range):
    count_vect = CountVectorizer(ngram_range=ngram_range, stop_words='english')
    X_counts = count_vect.fit_transform(text)

    tfidf_transformer = TfidfTransformer(use_idf=True)
    X_tfidf = tfidf_transformer.fit_transform(X_counts)
    return X_tfidf
```

In [20]:
```
def getSVD(X_tfidf, n_components):

    svd = TruncatedSVD(n_components = n_components, random_state = randomState)
    lsa = make_pipeline(svd, Normalizer(copy=False))
    X = lsa.fit_transform(X_tfidf)

    explained_variance = svd.explained_variance_ratio_.sum()
    print("Explained variance of the SVD step: {}%".format(
        int(explained_variance * 100)))

    return X
```

In [21]:
```
# Features from n-grams, use fewer components
X_tfidf = getTfidf(ngram_range = (3,5))
X = getSVD(X_tfidf, n_components = 6)
df_ngrams = pd.DataFrame(data=X, index = title,
                         columns=['n-gram 1st component', 'n-gram 2nd compoent', 'n-gram 3rd',
'n-gram 4th', 'n-gram 5th', 'n-gram 6th',])
df_ngrams.head()
```

```
Explained variance of the SVD step: 3%
```

Out[21]:

	n-gram 1st component	n-gram 2nd compoent	n-gram 3rd	n-gram 4th	n-gram 5th	n-gram 6th
MEASURE FOR MEASURE ACT II.	0.529818	-0.394808	-0.717043	-0.182813	-0.083582	-0.094140
THE TRAGEDY OF OTHELLO, MOOR OF VENICE ACT III	-0.308170	0.215696	-0.212219	0.513910	-0.264073	-0.692554
THE TRAGEDY OF OTHELLO, MOOR OF VENICE ACT V.	-0.073108	-0.738725	0.329050	-0.517464	0.186504	-0.195228
KING HENRY THE EIGHTH ACT V.	0.068210	0.101065	-0.186722	-0.137281	-0.944852	-0.196664
THE TRAGEDY OF ROMEO AND JULIET ACT II.	0.343136	0.491263	0.205879	-0.600223	0.232900	-0.428979

Step 3 Combine n-grams and other features

In [22]:
```
# Merge all features
df_all = df.join(df_ngrams)
df_all.head()
```

Out[22]:

	HapaxLegomena	DisLegomena	numberUniqueWords	averageSentenceLength	numberOfPronoun	nur
act						
MEASURE FOR MEASURE ACT II.	0.084440	0.019071	0.136920	10.955331	0.067999	0.0
THE TRAGEDY OF OTHELLO, MOOR OF VENICE ACT III	0.089297	0.023032	0.146323	9.396991	0.057273	0.0
THE TRAGEDY OF OTHELLO, MOOR OF VENICE ACT V.	0.093855	0.021995	0.148254	7.056872	0.058261	0.0
KING HENRY THE EIGHTH ACT V.	0.213043	0.051304	0.295652	20.535714	0.031304	0.0
THE TRAGEDY OF ROMEO	0.105476	0.026264	0.168203	9.955494	0.044286	0.0

AND JULIET ACT II.						

Step 4 Clustering

In [23]:
```
def getLabels(X, n_clusters = 8, method = 'KMeans'):

    if method == 'KMeans':
        cluster = KMeans(n_clusters=n_clusters, init='k-means++', max_iter=500,
                         n_init=5, random_state = randomState)
    elif method == 'AffinityPropagation':
        cluster = AffinityPropagation(max_iter=1000)
    elif method == 'AgglomerativeClustering':
        cluster = AgglomerativeClustering(n_clusters=n_clusters)
    elif method == 'Birch':
        cluster = Birch()
    elif method == 'DBSCAN':
        cluster = DBSCAN(random_state = randomState, eps=0.1, min_samples=3)
    elif method == 'SpectralClustering':
        cluster = SpectralClustering(n_clusters=n_clusters)
    elif method == 'MeanShift':
        cluster = MeanShift()


    return cluster.fit_predict(X)


def printClusters(labels):
    for i in set(labels):
        index = np.squeeze(np.argwhere(labels == i)).tolist()
        print(index)

        print(np.array(title)[index])
        print('\n')

def print2D(X, x_label, y_label, title, labels):
    plt.figure(figsize = (9,7))
    plt.scatter(X[:,0], X[:,1], c=labels, marker='+', s=100, linewidths=2)
    plt.ylabel(y_label, fontsize=14, fontweight='bold')
    plt.xlabel(x_label, fontsize=14, fontweight='bold')
    plt.title(title, fontsize=14, fontweight='bold')

    ax = plt.gca()
    for i in range(len(labels)):
        if labels.tolist().count(labels[i]) <= 3:
```

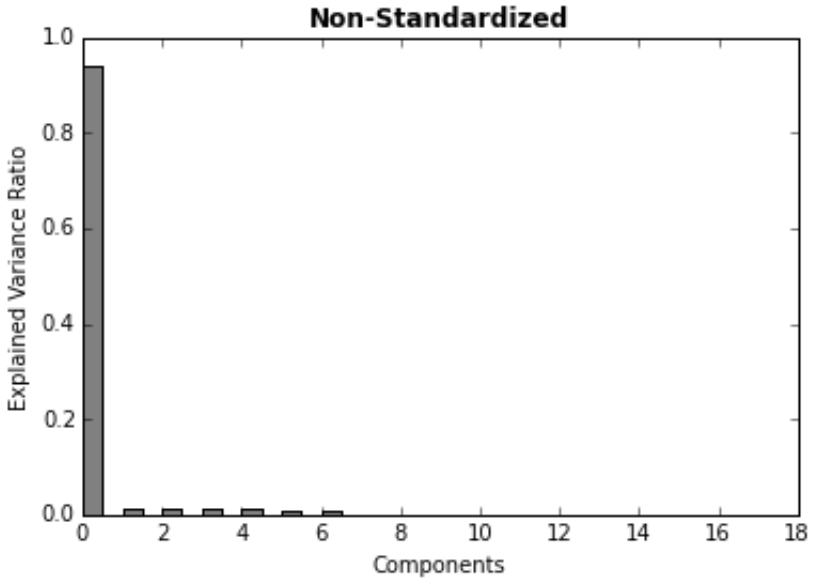
```
            ax.annotate(df_all.index[i], xy=X[i,:], fontweight='bold')

    xmin = min(X[:,0])
    xmax = max(X[:,0])
    ymin = min(X[:,1])
    ymax = max(X[:,1])

    xrange = xmax - xmin
    yrange = ymax - ymin

    xmin = xmin - xrange/10
    xmax = xmax + xrange/10

    ymin = ymin - yrange/10
    ymax = ymax + yrange/10

    plt.xlim([xmin, xmax])
    plt.ylim([ymin, ymax])
    plt.show()
```

In [24]:
```
from sklearn.decomposition import PCA
from sklearn import preprocessing
from sklearn import metrics
from sklearn.manifold import MDS

def clusterDocuments(X, scaled = False, PCA_X = False, n_clusters = 8,
                     method = 'KMeans', visualize = 'PCA', verbose = True):


    if scaled:
        X = preprocessing.scale(X)

    if PCA_X:
        X = PCA().fit_transform(X)

    if PCA_X:
        X_2D_PCA = X[:, 0:2]
    else:
        X_2D_PCA = PCA(n_components=2).fit_transform(X)
        X_2D_MDS = MDS(n_components=2).fit_transform(X)

    labels = getLabels(X, n_clusters, method)

    if visualize == 'PCA':
        print2D(X_2D_PCA, 'PCA 1st component', 'PCA 2nd component', method, labels)
    elif visualize == 'MDS':
        print2D(X_2D_MDS, 'MDS 1st component', 'MDS 2nd component', method, labels)


    print('Number of clusters: ', len(set(labels)))
    print("Silhouette Coefficient: %0.3f"
      % metrics.silhouette_score(X, labels, random_state=randomState))

    if verbose:
        printClusters(labels)

def visualizeCovariance(X, scaled = False):
    if scaled:
        X = preprocessing.scale(X)
        title = 'Standardized'
```

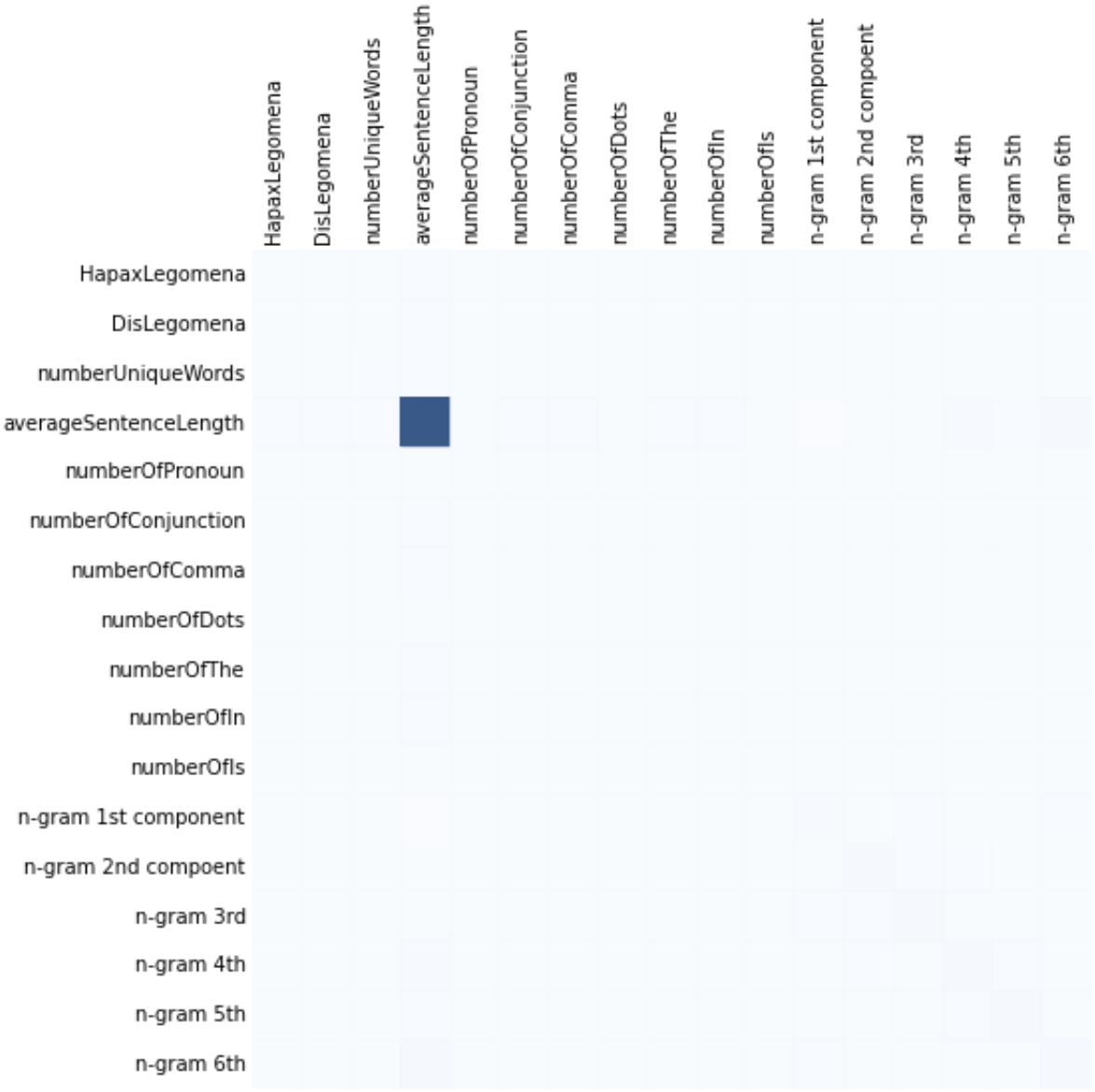
```
else:
    title = 'Non-Standardized'

pca = PCA(n_components=None)
pca.fit(X)

covariance = pca.get_covariance()


width = 0.5
ind = np.arange(len(pca.explained_variance_ratio_))

plt.figure()
ax = plt.gca()
ax.bar(ind, pca.explained_variance_ratio_, width, color='gray')
plt.ylabel('Explained Variance Ratio')
plt.xlabel('Components')
plt.title(title, fontweight='bold')
plt.show()

fig, ax = plt.subplots()
heatmap = ax.pcolor(covariance, cmap=plt.cm.Blues, alpha=0.8)
# Format
fig = plt.gcf()
fig.set_size_inches(8, 8)
# turn off the frame
ax.set_frame_on(False)
# put the major ticks at the middle of each cell
ax.set_yticks(np.arange(covariance.shape[0]) + 0.5, minor=False)
ax.set_xticks(np.arange(covariance.shape[1]) + 0.5, minor=False)
# want a more natural, table-like display
ax.invert_yaxis()
ax.xaxis.tick_top()
# Set the labels
# label source:https://en.wikipedia.org/wiki/Basketball_statistics
labels = df_all.columns
# note I could have used nba_sort.columns but made "labels" instead
ax.set_xticklabels(labels, minor=False)
ax.set_yticklabels(labels, minor=False)
# rotate the
```

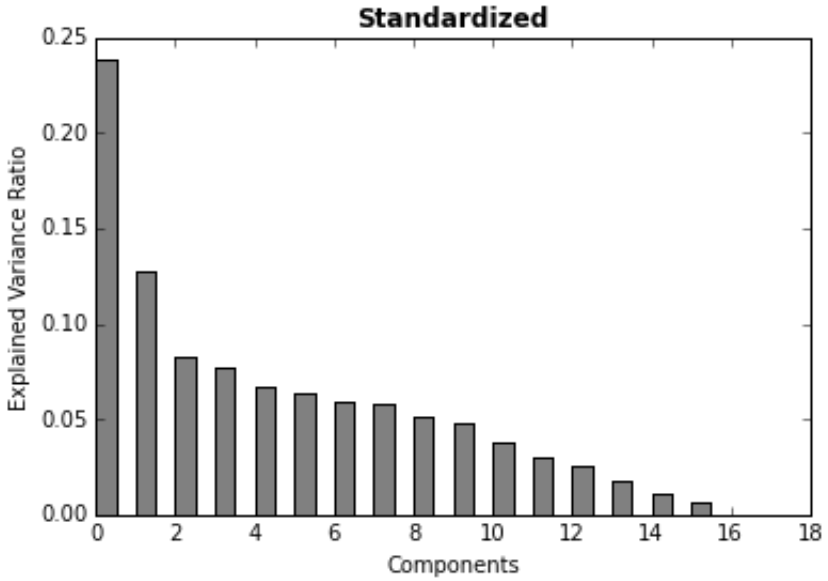
```
    plt.xticks(rotation=90)
    ax.grid(False)
    # Turn off all the ticks
    ax = plt.gca()
    for t in ax.xaxis.get_major_ticks():
        t.tick1On = False
        t.tick2On = False
    for t in ax.yaxis.get_major_ticks():
        t.tick1On = False
        t.tick2On = False
    plt.title(title, fontweight='bold', y=0)
    plt.show()
```

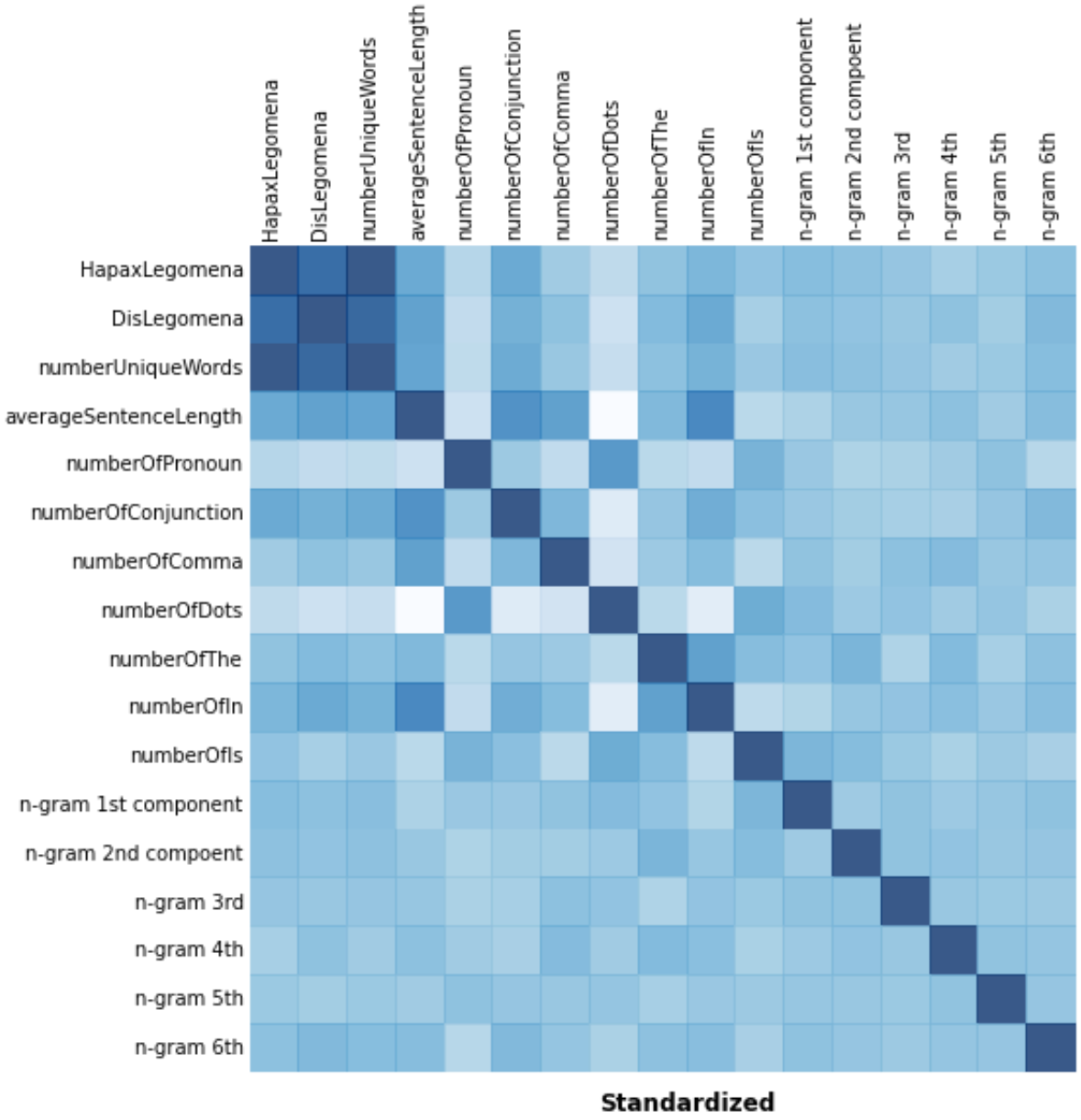
In [25]:
```
X = df_all.values
visualizeCovariance(X, scaled = False)
visualizeCovariance(X, scaled = True)
```




```
/Users/binyan/Python/anaconda/envs/py34/lib/python3.4/site-packages/matplotlib/collections.p
y:590: FutureWarning: elementwise comparison failed; returning scalar instead, but in the fu
ture will perform elementwise comparison
  if self._edgecolors == str('face'):
```


HapaxLegomena
DisLegomena
numberUniqueWords
averageSentenceLength
numberOfPronoun
numberOfConjunction
numberOfComma
numberOfDots
numberOfThe
numberOfIn
numberOfIs
n-gram 1st component
n-gram 2nd compoent
n-gram 3rd
n-gram 4th
n-gram 5th
n-gram 6th
Non-Standardized


Standardized
Explained Variance Ratio
Components
0.25
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8
10
12
14
16
18


HapaxLegomena
DisLegomena
numberUniqueWords
averageSentenceLength
numberOfPronoun
numberOfConjunction
numberOfComma
numberOfDots
numberOfThe
numberOfIn
numberOfIs
n-gram 1st component
n-gram 2nd compoent
n-gram 3rd
n-gram 4th
n-gram 5th
n-gram 6th
Standardized

K-Means

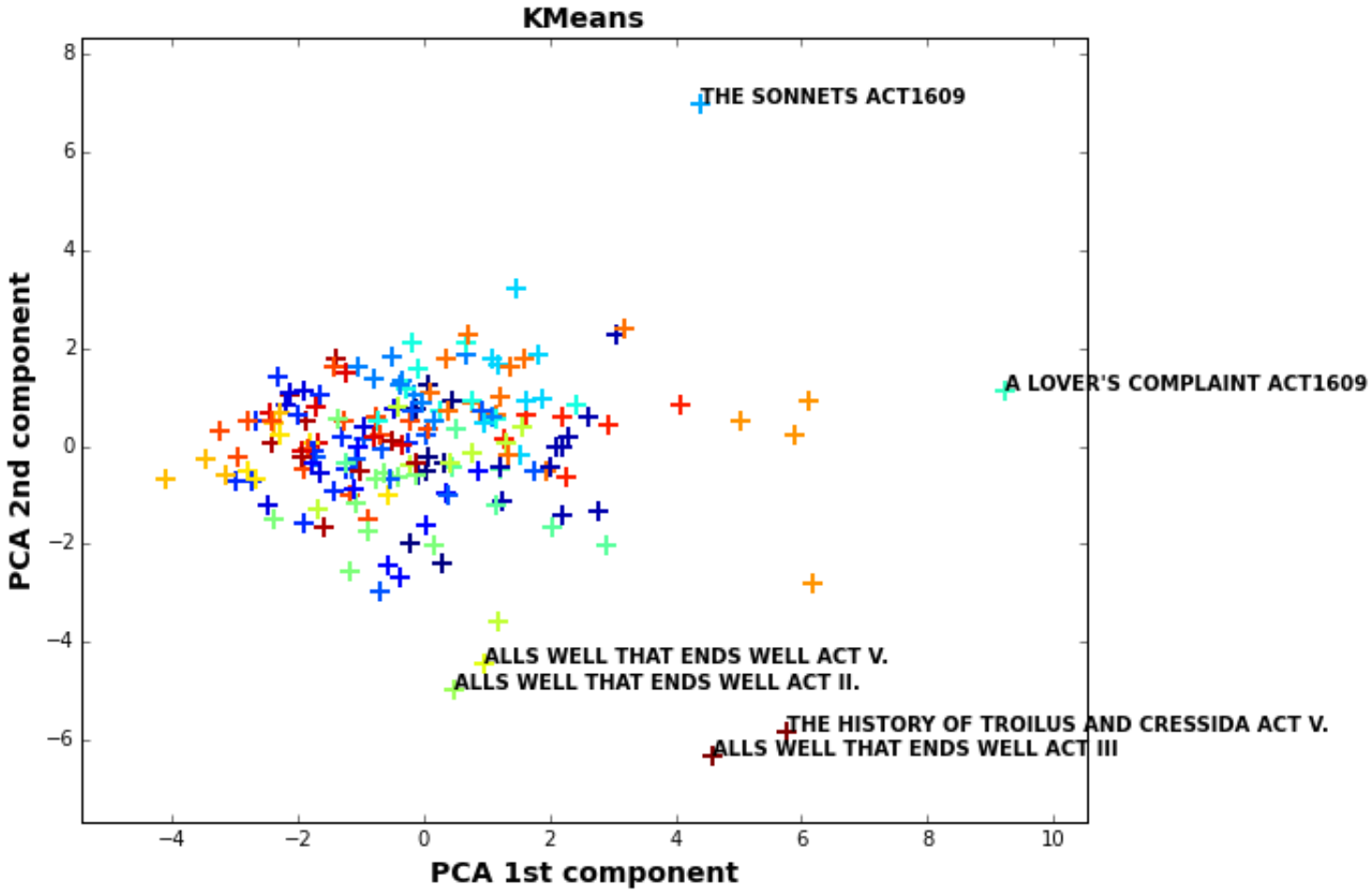
In [41]:
```
# Cluster using all features
X = df_all.values
for n in np.arange(5,30,1):
    clusterDocuments(X, scaled=True, PCA_X = True, n_clusters = n, method = 'KMeans', visualiz
e = 'None', verbose = False)
```

```
Number of clusters:  5
Silhouette Coefficient: 0.078
Number of clusters:  6
Silhouette Coefficient: 0.073
Number of clusters:  7
Silhouette Coefficient: 0.070
Number of clusters:  8
Silhouette Coefficient: 0.069
Number of clusters:  9
Silhouette Coefficient: 0.067
Number of clusters:  10
Silhouette Coefficient: 0.067
Number of clusters:  11
Silhouette Coefficient: 0.078
Number of clusters:  12
Silhouette Coefficient: 0.067
Number of clusters:  13
Silhouette Coefficient: 0.073
Number of clusters:  14
Silhouette Coefficient: 0.078
Number of clusters:  15
Silhouette Coefficient: 0.079
Number of clusters:  16
Silhouette Coefficient: 0.075
Number of clusters:  17
Silhouette Coefficient: 0.082
Number of clusters:  18
Silhouette Coefficient: 0.076
Number of clusters:  19
Silhouette Coefficient: 0.079
Number of clusters:  20
Silhouette Coefficient: 0.077
Number of clusters:  21
Silhouette Coefficient: 0.070
Number of clusters:  22
Silhouette Coefficient: 0.074
Number of clusters:  23
Silhouette Coefficient: 0.071
Number of clusters:  24
Silhouette Coefficient: 0.082
```

```
Number of clusters:  25
Silhouette Coefficient: 0.092
Number of clusters:  26
Silhouette Coefficient: 0.085
Number of clusters:  27
Silhouette Coefficient: 0.076
Number of clusters:  28
Silhouette Coefficient: 0.073
Number of clusters:  29
Silhouette Coefficient: 0.074
```

```
/Users/binyan/Python/anaconda/envs/py34/lib/python3.4/site-packages/numpy/core/_methods.py:5
9: RuntimeWarning: Mean of empty slice.
  warnings.warn("Mean of empty slice.", RuntimeWarning)
```

In [44]:
```
X = df_all.values
clusterDocuments(X, scaled = True, PCA_X = True, n_clusters = 25, method = 'KMeans', visualize
= 'PCA', verbose = True)
```


KMeans
THE SONNETS ACT1609
A LOVER'S COMPLAINT ACT1609
ALLS WELL THAT ENDS WELL ACT V.
ALLS WELL THAT ENDS WELL ACT II.
THE HISTORY OF TROILUS AND CRESSIDA ACT V.
ALLS WELL THAT ENDS WELL ACT III
PCA 2nd component
PCA 1st component
8
6
4
2
0
−2
−4
−6
−4
−2
0
2
4
6
8
10

```
Number of clusters:  25
Silhouette Coefficient: 0.081
[32, 34, 44, 85, 106, 114, 127, 174]
['THE COMEDY OF ERRORS ACT III'
 'THE HISTORY OF TROILUS AND CRESSIDA ACT IV.'
 'THE COMEDY OF ERRORS ACT V. ' 'TWELFTH NIGHT; OR, WHAT YOU WILL ACT V. '
 'THE TAMING OF THE SHREW ACT III' 'THE COMEDY OF ERRORS ACT Il.'
 "THE WINTER'S TALE ACT II." 'ALLS WELL THAT ENDS WELL ACT I. ']


[11, 13, 14, 58, 90, 99, 102, 103, 122, 139, 173]
['KING JOHN ACT II.' 'THE LIFE OF TIMON OF ATHENS ACT V. '
 "THE WINTER'S TALE ACT III" 'THE FIRST PART OF HENRY THE SIXTH ACT II.'
 'THE HISTORY OF TROILUS AND CRESSIDA ACT III'
 'THE FIRST PART OF HENRY THE SIXTH ACT III' "LOVE'S LABOUR'S LOST ACT I. "
 'KING RICHARD THE SECOND ACT IV.' 'KING JOHN ACT V. '
 'MEASURE FOR MEASURE ACT I. '
 'THE FIRST PART OF KING HENRY THE FOURTH ACT IV.']


[12, 104, 107, 125, 131, 135, 168]
["A MIDSUMMER NIGHT'S DREAM ACT III" 'THE TRAGEDY OF KING LEAR ACT I. '
 'THE TRAGEDY OF ROMEO AND JULIET ACT IV.'
 'SECOND PART OF KING HENRY IV ACT V. '
 'THE TRAGEDY OF JULIUS CAESAR ACT III' 'THE TAMING OF THE SHREW ACT V. '
 'THE TRAGEDY OF ROMEO AND JULIET ACT III']


[36, 69, 70, 101, 108, 124, 136, 154, 163]
['THE TRAGEDY OF MACBETH ACT IV.' 'THE TEMPEST ACT III'
 'THE TWO GENTLEMEN OF VERONA ACT V. '
 'THE LIFE OF TIMON OF ATHENS ACT III'
 'THE LIFE OF TIMON OF ATHENS ACT IV.' 'THE TRAGEDY OF KING LEAR ACT II.'
 'THE FIRST PART OF HENRY THE SIXTH ACT I. '
 'KING HENRY THE EIGHTH ACT I. ' 'THE TEMPEST ACT II.']


[16, 25, 37, 39, 57, 59, 79, 117, 129, 132, 138, 158, 177]
['MUCH ADO ABOUT NOTHING ACT II.'
 'THE TRAGEDY OF OTHELLO, MOOR OF VENICE ACT II.'
```

```
 'THE MERRY WIVES OF WINDSOR ACT I. ' 'THE TRAGEDY OF KING LEAR ACT V. '
 'THE TRAGEDY OF KING LEAR ACT III'
 'THE FIRST PART OF KING HENRY THE FOURTH ACT II.'
 'THE TRAGEDY OF CORIOLANUS ACT III' 'MUCH ADO ABOUT NOTHING ACT III'
 'THE TRAGEDY OF HAMLET, PRINCE OF DENMARK ACT V. '
 'THE TRAGEDY OF ANTONY AND CLEOPATRA ACT III'
 'THE TRAGEDY OF ROMEO AND JULIET ACT I. ' 'MEASURE FOR MEASURE ACT III'
 "LOVE'S LABOUR'S LOST ACT V. "]


[4, 8, 9, 29, 38, 48, 54, 66, 137, 150, 152, 170]
['THE TRAGEDY OF ROMEO AND JULIET ACT II.' 'AS YOU LIKE IT ACT III'
 'KING RICHARD III ACT II.' 'THE HISTORY OF TROILUS AND CRESSIDA ACT II.'
 'TWELFTH NIGHT; OR, WHAT YOU WILL ACT II.'
 'THE SECOND PART OF KING HENRY THE SIXTH ACT V. ' 'AS YOU LIKE IT ACT IV.'
 'THE TRAGEDY OF JULIUS CAESAR ACT II.' 'CYMBELINE ACT IV.'
 'KING RICHARD THE SECOND ACT V. ' 'KING RICHARD THE SECOND ACT II.'
 "THE WINTER'S TALE ACT I. "]


[18, 28, 35, 86, 87, 100, 120, 123, 147, 149]
['SECOND PART OF KING HENRY IV ACT IV.'
 'THE LIFE OF KING HENRY THE FIFTH ACT IV.'
 'THE SECOND PART OF KING HENRY THE SIXTH ACT IV.'
 'THE THIRD PART OF KING HENRY THE SIXTH ACT I. ' 'KING JOHN ACT IV.'
 'KING RICHARD III ACT I. ' 'CYMBELINE ACT III'
 'THE MERCHANT OF VENICE ACT II.' 'KING RICHARD III ACT IV.'
 'THE TEMPEST ACT I. ']


94
THE SONNETS ACT1609


[53, 60, 89, 115, 130, 133, 157, 162]
['THE TRAGEDY OF TITUS ANDRONICUS ACT 1. '
 'KING RICHARD THE SECOND ACT III'
 'THE FIRST PART OF HENRY THE SIXTH ACT V. '
 'THE LIFE OF KING HENRY THE FIFTH ACT V. ' 'KING RICHARD III ACT V. '
 'THE FIRST PART OF HENRY THE SIXTH ACT IV.'
```

```
 'KING HENRY THE EIGHTH ACT III' 'KING JOHN ACT III']


[24, 40, 47, 63, 73, 81, 118, 121, 143]
['THE TAMING OF THE SHREW ACT I. ' 'THE MERCHANT OF VENICE ACT III'
 'CYMBELINE ACT V. ' 'THE TRAGEDY OF OTHELLO, MOOR OF VENICE ACT I. '
 'AS YOU LIKE IT ACT II.' 'THE FIRST PART OF KING HENRY THE FOURTH ACT I. '
 'THE TRAGEDY OF HAMLET, PRINCE OF DENMARK ACT III'
 'THE FIRST PART OF KING HENRY THE FOURTH ACT III'
 'KING RICHARD III ACT III']


93
A LOVER'S COMPLAINT ACT1609


[22, 26, 33, 41, 68, 75, 88]
["LOVE'S LABOUR'S LOST ACT III" 'THE MERRY WIVES OF WINDSOR ACT V. '
 "A MIDSUMMER NIGHT'S DREAM ACT IV." 'THE TEMPEST ACT IV.'
 'KING HENRY THE EIGHTH ACT IV.' 'SECOND PART OF KING HENRY IV ACT III'
 "A MIDSUMMER NIGHT'S DREAM ACT V. "]


[20, 67, 71, 72, 80, 148, 151, 155, 160, 166]
['THE LIFE OF TIMON OF ATHENS ACT II.' 'THE TRAGEDY OF KING LEAR ACT IV.'
 'THE TRAGEDY OF ANTONY AND CLEOPATRA ACT V. '
 'THE TRAGEDY OF MACBETH ACT II.' 'THE TRAGEDY OF MACBETH ACT III'
 'THE TRAGEDY OF MACBETH ACT V. ' 'THE TRAGEDY OF ROMEO AND JULIET ACT V. '
 "LOVE'S LABOUR'S LOST ACT II."
 'THE TRAGEDY OF ANTONY AND CLEOPATRA ACT I. '
 'THE TRAGEDY OF CORIOLANUS ACT II.']


144
ALLS WELL THAT ENDS WELL ACT II.


[27, 56, 77, 95, 97, 105, 110, 134]
['ALLS WELL THAT ENDS WELL ACT IV ' 'ALLS WELL THAT ENDS WELL ACT IV.'
 'THE MERCHANT OF VENICE ACT I. '
```

```
 'THE LIFE OF KING HENRY THE FIFTH ACT III'
 'MUCH ADO ABOUT NOTHING ACT I. ' 'THE MERCHANT OF VENICE ACT IV.'
 'SECOND PART OF KING HENRY IV ACT I. ' 'THE MERCHANT OF VENICE ACT V. ']


92
ALLS WELL THAT ENDS WELL ACT V.


[46, 74, 113, 126, 161]
['THE LIFE OF TIMON OF ATHENS ACT I. '
 'THE TRAGEDY OF JULIUS CAESAR ACT V. '
 'THE TRAGEDY OF ANTONY AND CLEOPATRA ACT IV.'
 'TWELFTH NIGHT; OR, WHAT YOU WILL ACT IV.'
 'THE TRAGEDY OF JULIUS CAESAR ACT IV.']


[2, 19, 119, 141, 178]
['THE TRAGEDY OF OTHELLO, MOOR OF VENICE ACT V. '
 'THE TRAGEDY OF OTHELLO, MOOR OF VENICE ACT IV.'
 'MUCH ADO ABOUT NOTHING ACT V. ' 'SECOND PART OF KING HENRY IV ACT II.'
 'THE TRAGEDY OF ANTONY AND CLEOPATRA ACT II.']


[3, 6, 21, 116]
['KING HENRY THE EIGHTH ACT V. ' 'THE LIFE OF KING HENRY THE FIFTH ACT I. '
 'THE COMEDY OF ERRORS ACT I. '
 'THE HISTORY OF TROILUS AND CRESSIDA ACT I. ']


[10, 15, 45, 50, 65, 78, 82, 84, 109, 153, 156, 169, 172]
['KING JOHN ACT I. ' 'THE THIRD PART OF KING HENRY THE SIXTH ACT IV.'
 'THE THIRD PART OF KING HENRY THE SIXTH ACT V. '
 'THE SECOND PART OF KING HENRY THE SIXTH ACT III'
 'THE TRAGEDY OF TITUS ANDRONICUS ACT V. '
 'THE TRAGEDY OF TITUS ANDRONICUS ACT IV.'
 'THE THIRD PART OF KING HENRY THE SIXTH ACT III'
 'KING RICHARD THE SECOND ACT I. '
 'THE SECOND PART OF KING HENRY THE SIXTH ACT II.'
 'THE SECOND PART OF KING HENRY THE SIXTH ACT I. '
```

```
 'THE TRAGEDY OF TITUS ANDRONICUS ACT III'
 'THE TRAGEDY OF TITUS ANDRONICUS ACT II.'
 'THE THIRD PART OF KING HENRY THE SIXTH ACT II.']


[17, 42, 51, 55, 62, 64, 83, 128, 140, 142, 159, 164, 176, 180, 181]
['THE TWO GENTLEMEN OF VERONA ACT I. '
 'TWELFTH NIGHT; OR, WHAT YOU WILL ACT I. ' 'CYMBELINE ACT II.'
 'THE TRAGEDY OF CORIOLANUS ACT V. ' 'THE TRAGEDY OF CORIOLANUS ACT IV.'
 'THE FIRST PART OF KING HENRY THE FOURTH ACT V. '
 'THE TRAGEDY OF CORIOLANUS ACT I. ' "LOVE'S LABOUR'S LOST ACT IV."
 'THE COMEDY OF ERRORS ACT IV.' 'CYMBELINE ACT I. '
 'THE MERRY WIVES OF WINDSOR ACT II.' 'AS YOU LIKE IT ACT V. '
 'TWELFTH NIGHT; OR, WHAT YOU WILL ACT III'
 'THE MERRY WIVES OF WINDSOR ACT III' 'THE MERRY WIVES OF WINDSOR ACT IV.']


[5, 91, 96, 98, 111, 112]
['THE TRAGEDY OF MACBETH ACT I. '
 'THE LIFE OF KING HENRY THE FIFTH ACT II.' 'THE TEMPEST ACT V. '
 "A MIDSUMMER NIGHT'S DREAM ACT II." 'KING HENRY THE EIGHTH ACT II.'
 "A MIDSUMMER NIGHT'S DREAM ACT I. "]


[1, 7, 30, 49, 145, 146, 175]
['THE TRAGEDY OF OTHELLO, MOOR OF VENICE ACT III'
 'THE TWO GENTLEMEN OF VERONA ACT IV.'
 'THE TWO GENTLEMEN OF VERONA ACT III' 'MEASURE FOR MEASURE ACT V. '
 'MUCH ADO ABOUT NOTHING ACT IV.' "THE WINTER'S TALE ACT IV."
 'THE TAMING OF THE SHREW ACT Il.']


[0, 23, 43, 61, 76, 165, 167, 171, 179]
['MEASURE FOR MEASURE ACT II.' 'THE TAMING OF THE SHREW ACT IV.'
 'THE TWO GENTLEMEN OF VERONA ACT II.' 'ALLS WELL THAT ENDS WELL ACT V S'
 'THE TRAGEDY OF HAMLET, PRINCE OF DENMARK ACT I. '
 'THE TRAGEDY OF JULIUS CAESAR ACT I. ' 'AS YOU LIKE IT ACT I. '
 'THE TRAGEDY OF HAMLET, PRINCE OF DENMARK ACT IV.'
 "THE WINTER'S TALE ACT V. "]
```

```
[31, 52]
['ALLS WELL THAT ENDS WELL ACT III'
 'THE HISTORY OF TROILUS AND CRESSIDA ACT V. ']
```

```
/Users/binyan/Python/anaconda/envs/py34/lib/python3.4/site-packages/numpy/core/_methods.py:5
9: RuntimeWarning: Mean of empty slice.
  warnings.warn("Mean of empty slice.", RuntimeWarning)
```

By adding the features of average sentence length, number of unique words normalized by total number of words, fequency of conjunctions, and etc., acts are clustered according to the writing style (need to be verified by visualization).

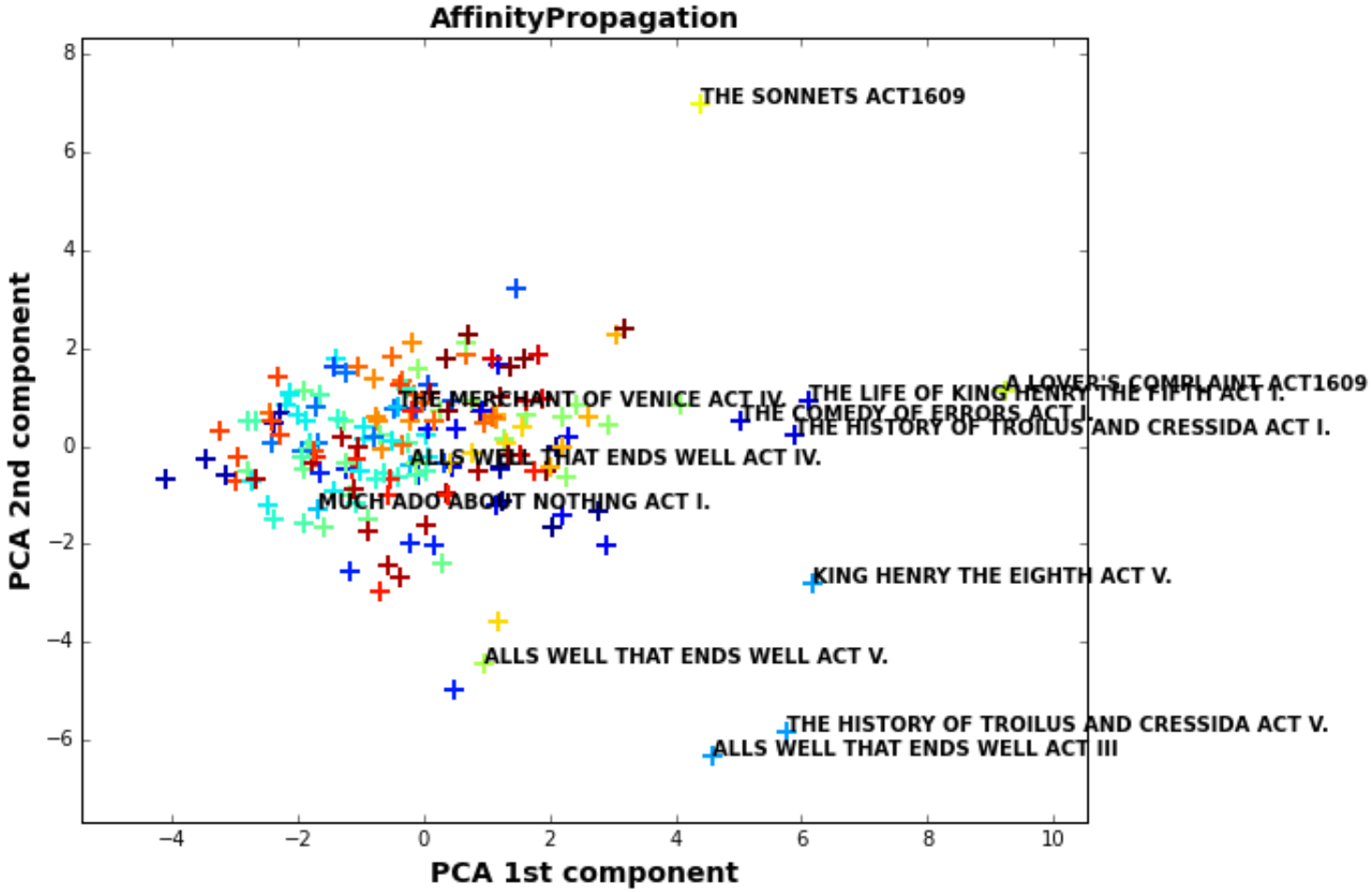
Affinity Propagation

In [28]:
```
print('No mormalization')
clusterDocuments(X, scaled = False, PCA_X=False, method = 'AffinityPropagation', visualize =
'None', verbose = False)
print('Normalize to N(0,1)')
clusterDocuments(X, scaled = True, method = 'AffinityPropagation', visualize = 'None', verbose
= False)
```

```
No mormalization
Number of clusters:  18
Silhouette Coefficient: 0.130
Normalize to N(0,1)
Number of clusters:  27
Silhouette Coefficient: 0.070
```

```
/Users/binyan/Python/anaconda/envs/py34/lib/python3.4/site-packages/numpy/core/_methods.py:5
9: RuntimeWarning: Mean of empty slice.
  warnings.warn("Mean of empty slice.", RuntimeWarning)
```

In [29]:
```
clusterDocuments(X, scaled = True, PCA_X = True, method = 'AffinityPropagation')
```


AffinityPropagation
PCA 2nd component
PCA 1st component
THE SONNETS ACT1609
A LOVER'S COMPLAINT ACT1609
THE LIFE OF KING HENRY THE FIFTH ACT I.
THE MERCHANT OF VENICE ACT IV.
THE COMEDY OF ERRORS ACT I.
THE HISTORY OF TROILUS AND CRESSIDA ACT I.
ALLS WELL THAT ENDS WELL ACT IV.
MUCH ADO ABOUT NOTHING ACT I.
KING HENRY THE EIGHTH ACT V.
ALLS WELL THAT ENDS WELL ACT V.
THE HISTORY OF TROILUS AND CRESSIDA ACT V.
ALLS WELL THAT ENDS WELL ACT III

```
Number of clusters:  26
Silhouette Coefficient: 0.070
[13, 41, 90, 99, 139]
['THE LIFE OF TIMON OF ATHENS ACT V. ' 'THE TEMPEST ACT IV.'
 'THE HISTORY OF TROILUS AND CRESSIDA ACT III'
 'THE FIRST PART OF HENRY THE SIXTH ACT III' 'MEASURE FOR MEASURE ACT I. ']


[2, 19, 119, 141, 178]
['THE TRAGEDY OF OTHELLO, MOOR OF VENICE ACT V. '
 'THE TRAGEDY OF OTHELLO, MOOR OF VENICE ACT IV.'
 'MUCH ADO ABOUT NOTHING ACT V. ' 'SECOND PART OF KING HENRY IV ACT II.'
 'THE TRAGEDY OF ANTONY AND CLEOPATRA ACT II.']


[6, 21, 116]
['THE LIFE OF KING HENRY THE FIFTH ACT I. ' 'THE COMEDY OF ERRORS ACT I. '
 'THE HISTORY OF TROILUS AND CRESSIDA ACT I. ']


[14, 22, 26, 33, 68, 88, 102, 128, 130, 152, 173]
["THE WINTER'S TALE ACT III" "LOVE'S LABOUR'S LOST ACT III"
 'THE MERRY WIVES OF WINDSOR ACT V. ' "A MIDSUMMER NIGHT'S DREAM ACT IV."
 'KING HENRY THE EIGHTH ACT IV.' "A MIDSUMMER NIGHT'S DREAM ACT V. "
 "LOVE'S LABOUR'S LOST ACT I. " "LOVE'S LABOUR'S LOST ACT IV."
 'KING RICHARD III ACT V. ' 'KING RICHARD THE SECOND ACT II.'
 'THE FIRST PART OF KING HENRY THE FOURTH ACT IV.']


[20, 34, 57, 107, 144, 155]
['THE LIFE OF TIMON OF ATHENS ACT II.'
 'THE HISTORY OF TROILUS AND CRESSIDA ACT IV.'
 'THE TRAGEDY OF KING LEAR ACT III'
 'THE TRAGEDY OF ROMEO AND JULIET ACT IV.'
 'ALLS WELL THAT ENDS WELL ACT II.' "LOVE'S LABOUR'S LOST ACT II."]


[32, 44, 53, 106, 114, 140]
['THE COMEDY OF ERRORS ACT III' 'THE COMEDY OF ERRORS ACT V. '
 'THE TRAGEDY OF TITUS ANDRONICUS ACT 1. '
```

```
 'THE TAMING OF THE SHREW ACT III' 'THE COMEDY OF ERRORS ACT Il.'
 'THE COMEDY OF ERRORS ACT IV.']


[7, 12, 43, 49, 145, 146, 175]
['THE TWO GENTLEMEN OF VERONA ACT IV.' "A MIDSUMMER NIGHT'S DREAM ACT III"
 'THE TWO GENTLEMEN OF VERONA ACT II.' 'MEASURE FOR MEASURE ACT V. '
 'MUCH ADO ABOUT NOTHING ACT IV.' "THE WINTER'S TALE ACT IV."
 'THE TAMING OF THE SHREW ACT Il.']


[3, 31, 52]
['KING HENRY THE EIGHTH ACT V. ' 'ALLS WELL THAT ENDS WELL ACT III'
 'THE HISTORY OF TROILUS AND CRESSIDA ACT V. ']


[56, 97, 105]
['ALLS WELL THAT ENDS WELL ACT IV.' 'MUCH ADO ABOUT NOTHING ACT I. '
 'THE MERCHANT OF VENICE ACT IV.']


[0, 9, 23, 66, 74, 76, 85, 117, 125, 129, 131, 135, 158, 168, 171, 179]
['MEASURE FOR MEASURE ACT II.' 'KING RICHARD III ACT II.'
 'THE TAMING OF THE SHREW ACT IV.' 'THE TRAGEDY OF JULIUS CAESAR ACT II.'
 'THE TRAGEDY OF JULIUS CAESAR ACT V. '
 'THE TRAGEDY OF HAMLET, PRINCE OF DENMARK ACT I. '
 'TWELFTH NIGHT; OR, WHAT YOU WILL ACT V. '
 'MUCH ADO ABOUT NOTHING ACT III' 'SECOND PART OF KING HENRY IV ACT V. '
 'THE TRAGEDY OF HAMLET, PRINCE OF DENMARK ACT V. '
 'THE TRAGEDY OF JULIUS CAESAR ACT III' 'THE TAMING OF THE SHREW ACT V. '
 'MEASURE FOR MEASURE ACT III' 'THE TRAGEDY OF ROMEO AND JULIET ACT III'
 'THE TRAGEDY OF HAMLET, PRINCE OF DENMARK ACT IV.'
 "THE WINTER'S TALE ACT V. "]


[47, 67, 71, 80, 148, 160, 166]
['CYMBELINE ACT V. ' 'THE TRAGEDY OF KING LEAR ACT IV.'
 'THE TRAGEDY OF ANTONY AND CLEOPATRA ACT V. '
 'THE TRAGEDY OF MACBETH ACT III' 'THE TRAGEDY OF MACBETH ACT V. '
 'THE TRAGEDY OF ANTONY AND CLEOPATRA ACT I. '
```

```
 'THE TRAGEDY OF CORIOLANUS ACT II.']


[16, 25, 39, 79, 127, 151, 159, 167, 177]
['MUCH ADO ABOUT NOTHING ACT II.'
 'THE TRAGEDY OF OTHELLO, MOOR OF VENICE ACT II.'
 'THE TRAGEDY OF KING LEAR ACT V. ' 'THE TRAGEDY OF CORIOLANUS ACT III'
 "THE WINTER'S TALE ACT II." 'THE TRAGEDY OF ROMEO AND JULIET ACT V. '
 'THE MERRY WIVES OF WINDSOR ACT II.' 'AS YOU LIKE IT ACT I. '
 "LOVE'S LABOUR'S LOST ACT V. "]


[17, 42, 46, 51, 61, 62, 75, 83, 104, 142, 164, 174]
['THE TWO GENTLEMEN OF VERONA ACT I. '
 'TWELFTH NIGHT; OR, WHAT YOU WILL ACT I. '
 'THE LIFE OF TIMON OF ATHENS ACT I. ' 'CYMBELINE ACT II.'
 'ALLS WELL THAT ENDS WELL ACT V S' 'THE TRAGEDY OF CORIOLANUS ACT IV.'
 'SECOND PART OF KING HENRY IV ACT III' 'THE TRAGEDY OF CORIOLANUS ACT I. '
 'THE TRAGEDY OF KING LEAR ACT I. ' 'CYMBELINE ACT I. '
 'AS YOU LIKE IT ACT V. ' 'ALLS WELL THAT ENDS WELL ACT I. ']


[5, 24, 40, 63, 81, 91, 96, 98, 111, 112, 121, 165]
['THE TRAGEDY OF MACBETH ACT I. ' 'THE TAMING OF THE SHREW ACT I. '
 'THE MERCHANT OF VENICE ACT III'
 'THE TRAGEDY OF OTHELLO, MOOR OF VENICE ACT I. '
 'THE FIRST PART OF KING HENRY THE FOURTH ACT I. '
 'THE LIFE OF KING HENRY THE FIFTH ACT II.' 'THE TEMPEST ACT V. '
 "A MIDSUMMER NIGHT'S DREAM ACT II." 'KING HENRY THE EIGHTH ACT II.'
 "A MIDSUMMER NIGHT'S DREAM ACT I. "
 'THE FIRST PART OF KING HENRY THE FOURTH ACT III'
 'THE TRAGEDY OF JULIUS CAESAR ACT I. ']


92
ALLS WELL THAT ENDS WELL ACT V.


93
A LOVER'S COMPLAINT ACT1609
```

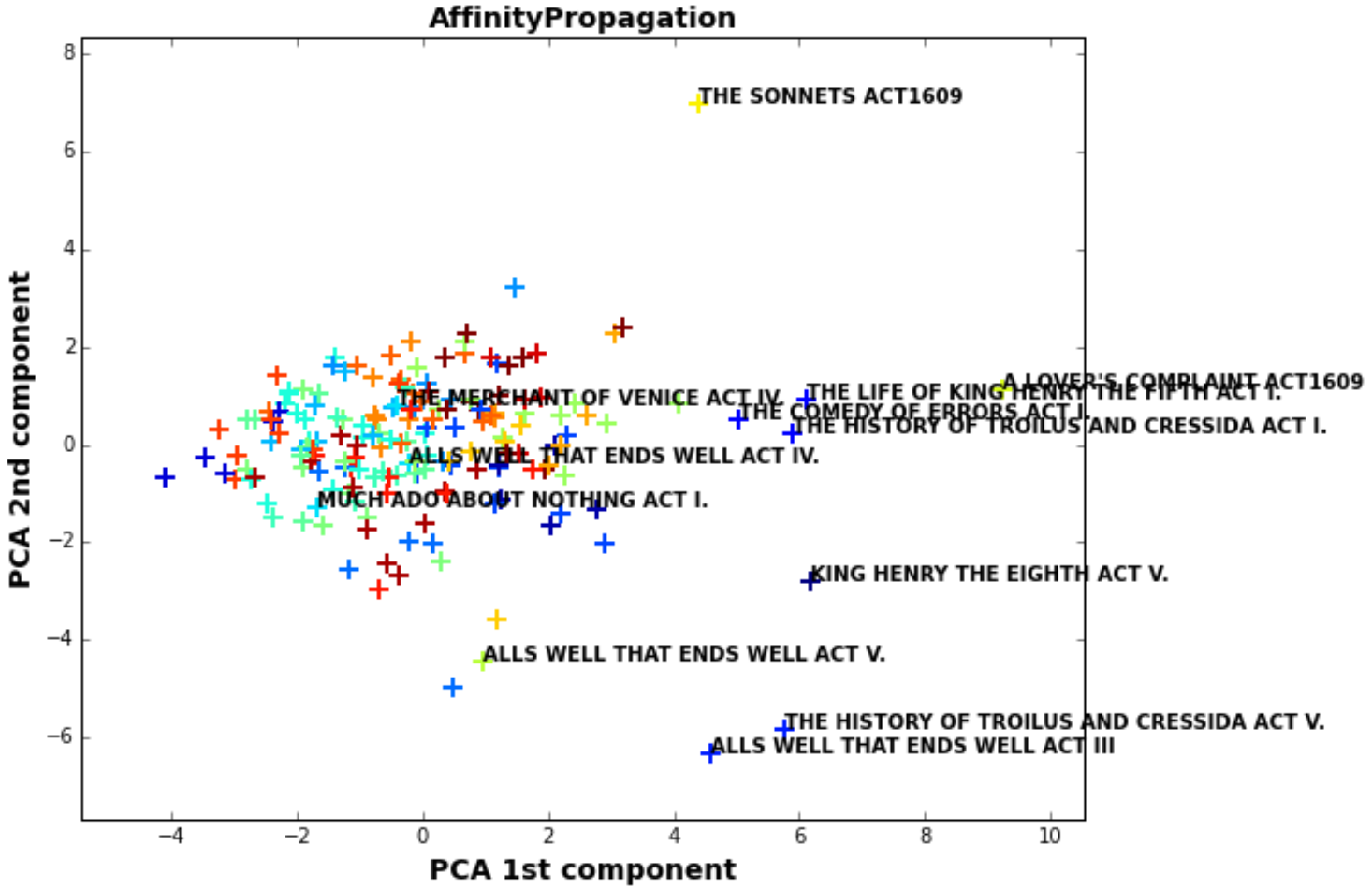
```
94
THE SONNETS ACT1609


[27, 77, 95, 110, 134]
['ALLS WELL THAT ENDS WELL ACT IV ' 'THE MERCHANT OF VENICE ACT I. '
 'THE LIFE OF KING HENRY THE FIFTH ACT III'
 'SECOND PART OF KING HENRY IV ACT I. ' 'THE MERCHANT OF VENICE ACT V. ']


[11, 50, 58, 103, 122]
['KING JOHN ACT II.' 'THE SECOND PART OF KING HENRY THE SIXTH ACT III'
 'THE FIRST PART OF HENRY THE SIXTH ACT II.'
 'KING RICHARD THE SECOND ACT IV.' 'KING JOHN ACT V. ']


[54, 55, 64, 73, 118, 143, 147]
['AS YOU LIKE IT ACT IV.' 'THE TRAGEDY OF CORIOLANUS ACT V. '
 'THE FIRST PART OF KING HENRY THE FOURTH ACT V. ' 'AS YOU LIKE IT ACT II.'
 'THE TRAGEDY OF HAMLET, PRINCE OF DENMARK ACT III'
 'KING RICHARD III ACT III' 'KING RICHARD III ACT IV.']


[1, 18, 28, 30, 35, 87, 100, 120, 123, 149, 157]
['THE TRAGEDY OF OTHELLO, MOOR OF VENICE ACT III'
 'SECOND PART OF KING HENRY IV ACT IV.'
 'THE LIFE OF KING HENRY THE FIFTH ACT IV.'
 'THE TWO GENTLEMEN OF VERONA ACT III'
 'THE SECOND PART OF KING HENRY THE SIXTH ACT IV.' 'KING JOHN ACT IV.'
 'KING RICHARD III ACT I. ' 'CYMBELINE ACT III'
 'THE MERCHANT OF VENICE ACT II.' 'THE TEMPEST ACT I. '
 'KING HENRY THE EIGHTH ACT III']


[37, 59, 86, 113, 132, 176, 180, 181]
['THE MERRY WIVES OF WINDSOR ACT I. '
 'THE FIRST PART OF KING HENRY THE FOURTH ACT II.'
 'THE THIRD PART OF KING HENRY THE SIXTH ACT I. '
```

```
 'THE TRAGEDY OF ANTONY AND CLEOPATRA ACT IV.'
 'THE TRAGEDY OF ANTONY AND CLEOPATRA ACT III'
 'TWELFTH NIGHT; OR, WHAT YOU WILL ACT III'
 'THE MERRY WIVES OF WINDSOR ACT III' 'THE MERRY WIVES OF WINDSOR ACT IV.']


[4, 8, 29, 38, 48, 126, 137, 150, 170]
['THE TRAGEDY OF ROMEO AND JULIET ACT II.' 'AS YOU LIKE IT ACT III'
 'THE HISTORY OF TROILUS AND CRESSIDA ACT II.'
 'TWELFTH NIGHT; OR, WHAT YOU WILL ACT II.'
 'THE SECOND PART OF KING HENRY THE SIXTH ACT V. '
 'TWELFTH NIGHT; OR, WHAT YOU WILL ACT IV.' 'CYMBELINE ACT IV.'
 'KING RICHARD THE SECOND ACT V. ' "THE WINTER'S TALE ACT I. "]


[60, 89, 115, 133, 162]
['KING RICHARD THE SECOND ACT III'
 'THE FIRST PART OF HENRY THE SIXTH ACT V. '
 'THE LIFE OF KING HENRY THE FIFTH ACT V. '
 'THE FIRST PART OF HENRY THE SIXTH ACT IV.' 'KING JOHN ACT III']


[36, 69, 70, 72, 78, 101, 108, 124, 136, 138, 154, 161, 163]
['THE TRAGEDY OF MACBETH ACT IV.' 'THE TEMPEST ACT III'
 'THE TWO GENTLEMEN OF VERONA ACT V. ' 'THE TRAGEDY OF MACBETH ACT II.'
 'THE TRAGEDY OF TITUS ANDRONICUS ACT IV.'
 'THE LIFE OF TIMON OF ATHENS ACT III'
 'THE LIFE OF TIMON OF ATHENS ACT IV.' 'THE TRAGEDY OF KING LEAR ACT II.'
 'THE FIRST PART OF HENRY THE SIXTH ACT I. '
 'THE TRAGEDY OF ROMEO AND JULIET ACT I. ' 'KING HENRY THE EIGHTH ACT I. '
 'THE TRAGEDY OF JULIUS CAESAR ACT IV.' 'THE TEMPEST ACT II.']


[10, 15, 45, 65, 82, 84, 109, 153, 156, 169, 172]
['KING JOHN ACT I. ' 'THE THIRD PART OF KING HENRY THE SIXTH ACT IV.'
 'THE THIRD PART OF KING HENRY THE SIXTH ACT V. '
 'THE TRAGEDY OF TITUS ANDRONICUS ACT V. '
 'THE THIRD PART OF KING HENRY THE SIXTH ACT III'
 'KING RICHARD THE SECOND ACT I. '
 'THE SECOND PART OF KING HENRY THE SIXTH ACT II.'
```

```
 'THE SECOND PART OF KING HENRY THE SIXTH ACT I. '
 'THE TRAGEDY OF TITUS ANDRONICUS ACT III'
 'THE TRAGEDY OF TITUS ANDRONICUS ACT II.'
 'THE THIRD PART OF KING HENRY THE SIXTH ACT II.']
```

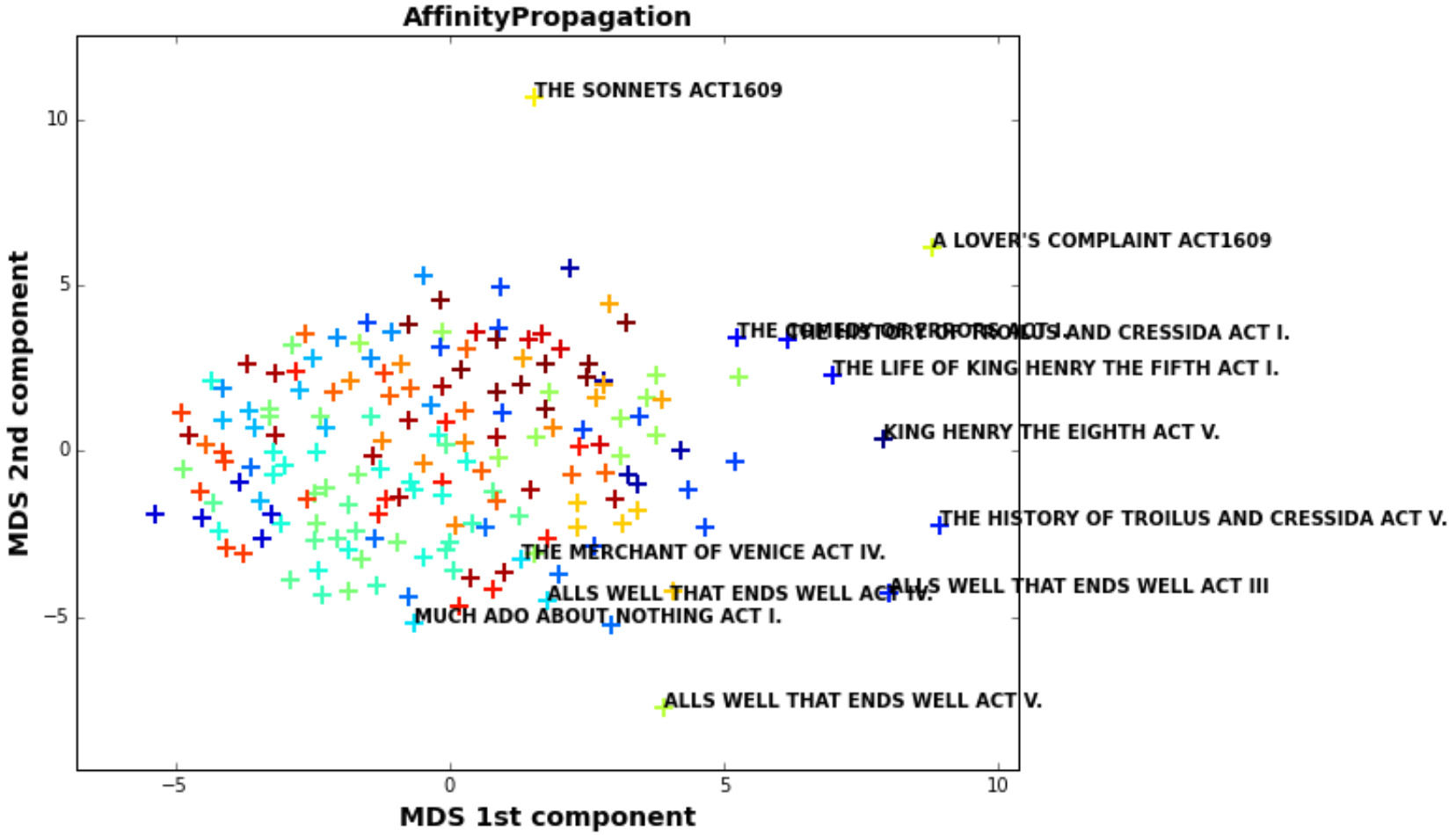
```
/Users/binyan/Python/anaconda/envs/py34/lib/python3.4/site-packages/numpy/core/_methods.py:5
9: RuntimeWarning: Mean of empty slice.
  warnings.warn("Mean of empty slice.", RuntimeWarning)
```

In [35]:
```
clusterDocuments(X, scaled = True, method = 'AffinityPropagation', visualize = 'PCA', verbose
= False)
clusterDocuments(X, scaled = True, method = 'AffinityPropagation', visualize = 'MDS', verbose
= False)
```




```
Number of clusters:  27
Silhouette Coefficient: 0.070
```

```
/Users/binyan/Python/anaconda/envs/py34/lib/python3.4/site-packages/numpy/core/_methods.py:5
9: RuntimeWarning: Mean of empty slice.
  warnings.warn("Mean of empty slice.", RuntimeWarning)
```




```
Number of clusters:  27
Silhouette Coefficient: 0.070
```

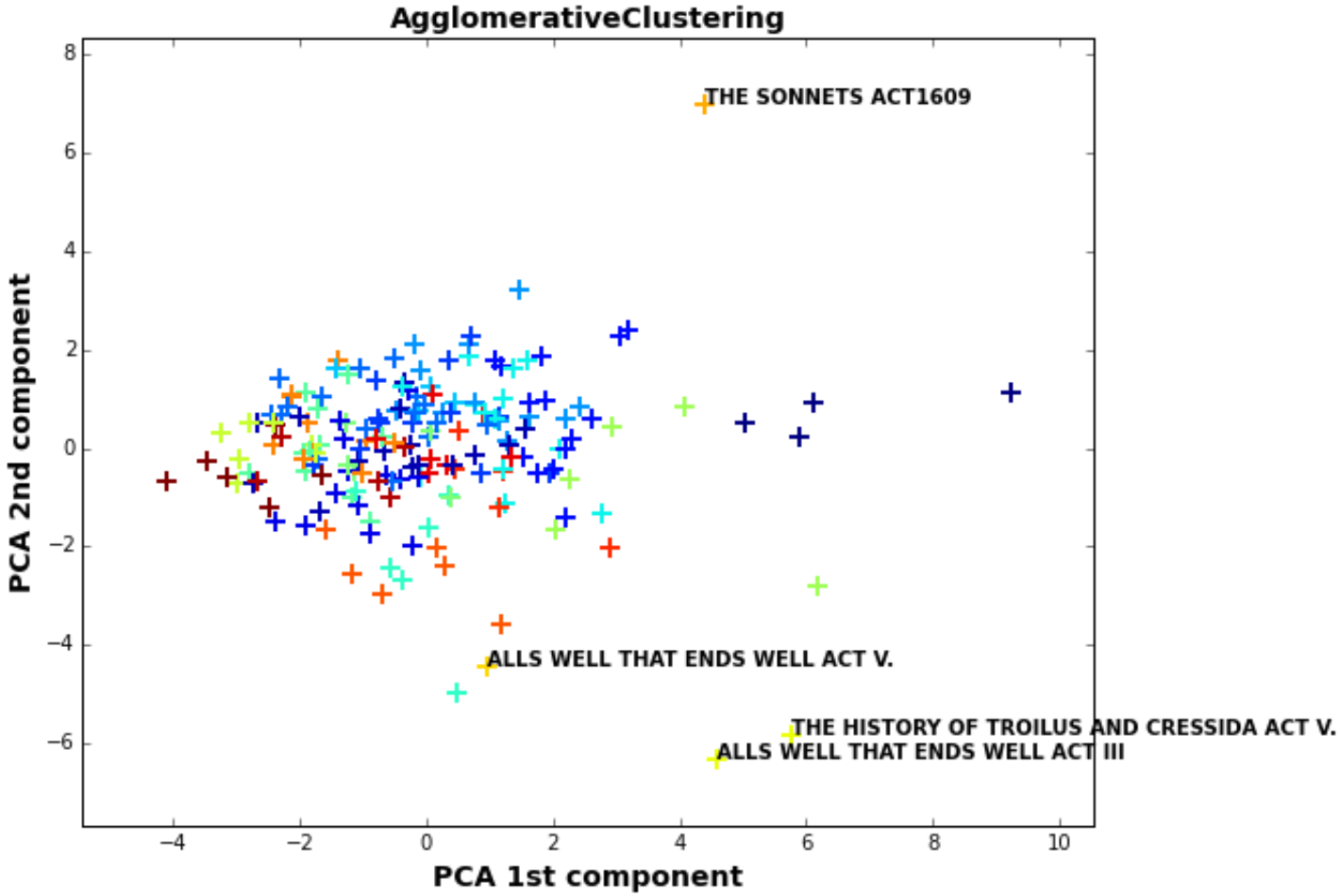
In [31]:
```
X = df_all.values
for n in np.arange(2,30,1):
    clusterDocuments(X, scaled=True, PCA_X = True, n_clusters = n, method = 'AgglomerativeClus
tering', visualize = False, verbose = False)
```

```
Number of clusters:  2
Silhouette Coefficient: 0.112
Number of clusters:  3
Silhouette Coefficient: 0.085
Number of clusters:  4
Silhouette Coefficient: 0.057
Number of clusters:  5
Silhouette Coefficient: 0.063
Number of clusters:  6
Silhouette Coefficient: 0.062
Number of clusters:  7
Silhouette Coefficient: 0.059
Number of clusters:  8
Silhouette Coefficient: 0.062
Number of clusters:  9
Silhouette Coefficient: 0.061
Number of clusters:  10
Silhouette Coefficient: 0.051
Number of clusters:  11
Silhouette Coefficient: 0.054
Number of clusters:  12
Silhouette Coefficient: 0.054
Number of clusters:  13
Silhouette Coefficient: 0.057
Number of clusters:  14
Silhouette Coefficient: 0.061
Number of clusters:  15
Silhouette Coefficient: 0.065
Number of clusters:  16
Silhouette Coefficient: 0.067
Number of clusters:  17
Silhouette Coefficient: 0.071
Number of clusters:  18
Silhouette Coefficient: 0.067
Number of clusters:  19
Silhouette Coefficient: 0.070
Number of clusters:  20
Silhouette Coefficient: 0.073
Number of clusters:  21
Silhouette Coefficient: 0.072
```

```
Number of clusters:  22
Silhouette Coefficient: 0.075
Number of clusters:  23
Silhouette Coefficient: 0.081
Number of clusters:  24
Silhouette Coefficient: 0.081
Number of clusters:  25
Silhouette Coefficient: 0.083
Number of clusters:  26
Silhouette Coefficient: 0.085
Number of clusters:  27
Silhouette Coefficient: 0.085
Number of clusters:  28
Silhouette Coefficient: 0.088
Number of clusters:  29
Silhouette Coefficient: 0.090
```

```
/Users/binyan/Python/anaconda/envs/py34/lib/python3.4/site-packages/numpy/core/_methods.py:5
9: RuntimeWarning: Mean of empty slice.
  warnings.warn("Mean of empty slice.", RuntimeWarning)
```

In [32]:
```
clusterDocuments(X, n_clusters = 23, scaled = True, PCA_X = True, method = 'AgglomerativeClustering')
```


AgglomerativeClustering
THE SONNETS ACT1609
ALLS WELL THAT ENDS WELL ACT V.
THE HISTORY OF TROILUS AND CRESSIDA ACT V.
ALLS WELL THAT ENDS WELL ACT III
PCA 2nd component
PCA 1st component
8
6
4
2
0
−2
−4
−6
−4
−2
0
2
4
6
8
10

```
Number of clusters:  23
Silhouette Coefficient: 0.081
[6, 21, 93, 116]
['THE LIFE OF KING HENRY THE FIFTH ACT I. ' 'THE COMEDY OF ERRORS ACT I. '
 "A LOVER'S COMPLAINT ACT1609"
 'THE HISTORY OF TROILUS AND CRESSIDA ACT I. ']


[8, 16, 25, 28, 54, 56, 77, 95, 97, 105, 110, 117, 134, 158, 165]
['AS YOU LIKE IT ACT III' 'MUCH ADO ABOUT NOTHING ACT II.'
 'THE TRAGEDY OF OTHELLO, MOOR OF VENICE ACT II.'
 'THE LIFE OF KING HENRY THE FIFTH ACT IV.' 'AS YOU LIKE IT ACT IV.'
 'ALLS WELL THAT ENDS WELL ACT IV.' 'THE MERCHANT OF VENICE ACT I. '
 'THE LIFE OF KING HENRY THE FIFTH ACT III'
 'MUCH ADO ABOUT NOTHING ACT I. ' 'THE MERCHANT OF VENICE ACT IV.'
 'SECOND PART OF KING HENRY IV ACT I. ' 'MUCH ADO ABOUT NOTHING ACT III'
 'THE MERCHANT OF VENICE ACT V. ' 'MEASURE FOR MEASURE ACT III'
 'THE TRAGEDY OF JULIUS CAESAR ACT I. ']


[34, 39, 47, 57, 67, 71, 72, 80, 129, 138, 151, 160, 166]
['THE HISTORY OF TROILUS AND CRESSIDA ACT IV.'
 'THE TRAGEDY OF KING LEAR ACT V. ' 'CYMBELINE ACT V. '
 'THE TRAGEDY OF KING LEAR ACT III' 'THE TRAGEDY OF KING LEAR ACT IV.'
 'THE TRAGEDY OF ANTONY AND CLEOPATRA ACT V. '
 'THE TRAGEDY OF MACBETH ACT II.' 'THE TRAGEDY OF MACBETH ACT III'
 'THE TRAGEDY OF HAMLET, PRINCE OF DENMARK ACT V. '
 'THE TRAGEDY OF ROMEO AND JULIET ACT I. '
 'THE TRAGEDY OF ROMEO AND JULIET ACT V. '
 'THE TRAGEDY OF ANTONY AND CLEOPATRA ACT I. '
 'THE TRAGEDY OF CORIOLANUS ACT II.']


[11, 14, 48, 58, 60, 84, 89, 103, 115, 122, 130, 133, 162, 173]
['KING JOHN ACT II.' "THE WINTER'S TALE ACT III"
 'THE SECOND PART OF KING HENRY THE SIXTH ACT V. '
 'THE FIRST PART OF HENRY THE SIXTH ACT II.'
 'KING RICHARD THE SECOND ACT III' 'KING RICHARD THE SECOND ACT I. '
 'THE FIRST PART OF HENRY THE SIXTH ACT V. '
 'KING RICHARD THE SECOND ACT IV.'
```

```
 'THE LIFE OF KING HENRY THE FIFTH ACT V. ' 'KING JOHN ACT V. '
 'KING RICHARD III ACT V. ' 'THE FIRST PART OF HENRY THE SIXTH ACT IV.'
 'KING JOHN ACT III' 'THE FIRST PART OF KING HENRY THE FOURTH ACT IV.']


[10, 15, 50, 55, 64, 82, 101, 109, 118, 124, 136, 147]
['KING JOHN ACT I. ' 'THE THIRD PART OF KING HENRY THE SIXTH ACT IV.'
 'THE SECOND PART OF KING HENRY THE SIXTH ACT III'
 'THE TRAGEDY OF CORIOLANUS ACT V. '
 'THE FIRST PART OF KING HENRY THE FOURTH ACT V. '
 'THE THIRD PART OF KING HENRY THE SIXTH ACT III'
 'THE LIFE OF TIMON OF ATHENS ACT III'
 'THE SECOND PART OF KING HENRY THE SIXTH ACT II.'
 'THE TRAGEDY OF HAMLET, PRINCE OF DENMARK ACT III'
 'THE TRAGEDY OF KING LEAR ACT II.'
 'THE FIRST PART OF HENRY THE SIXTH ACT I. ' 'KING RICHARD III ACT IV.']


[1, 9, 35, 38, 59, 100, 120, 123, 125, 131, 137, 141, 149, 150, 157, 177]
['THE TRAGEDY OF OTHELLO, MOOR OF VENICE ACT III'
 'KING RICHARD III ACT II.'
 'THE SECOND PART OF KING HENRY THE SIXTH ACT IV.'
 'TWELFTH NIGHT; OR, WHAT YOU WILL ACT II.'
 'THE FIRST PART OF KING HENRY THE FOURTH ACT II.'
 'KING RICHARD III ACT I. ' 'CYMBELINE ACT III'
 'THE MERCHANT OF VENICE ACT II.' 'SECOND PART OF KING HENRY IV ACT V. '
 'THE TRAGEDY OF JULIUS CAESAR ACT III' 'CYMBELINE ACT IV.'
 'SECOND PART OF KING HENRY IV ACT II.' 'THE TEMPEST ACT I. '
 'KING RICHARD THE SECOND ACT V. ' 'KING HENRY THE EIGHTH ACT III'
 "LOVE'S LABOUR'S LOST ACT V. "]


[5, 12, 24, 40, 53, 63, 73, 81, 91, 112, 121, 143]
['THE TRAGEDY OF MACBETH ACT I. ' "A MIDSUMMER NIGHT'S DREAM ACT III"
 'THE TAMING OF THE SHREW ACT I. ' 'THE MERCHANT OF VENICE ACT III'
 'THE TRAGEDY OF TITUS ANDRONICUS ACT 1. '
 'THE TRAGEDY OF OTHELLO, MOOR OF VENICE ACT I. ' 'AS YOU LIKE IT ACT II.'
 'THE FIRST PART OF KING HENRY THE FOURTH ACT I. '
 'THE LIFE OF KING HENRY THE FIFTH ACT II.'
 "A MIDSUMMER NIGHT'S DREAM ACT I. "
```

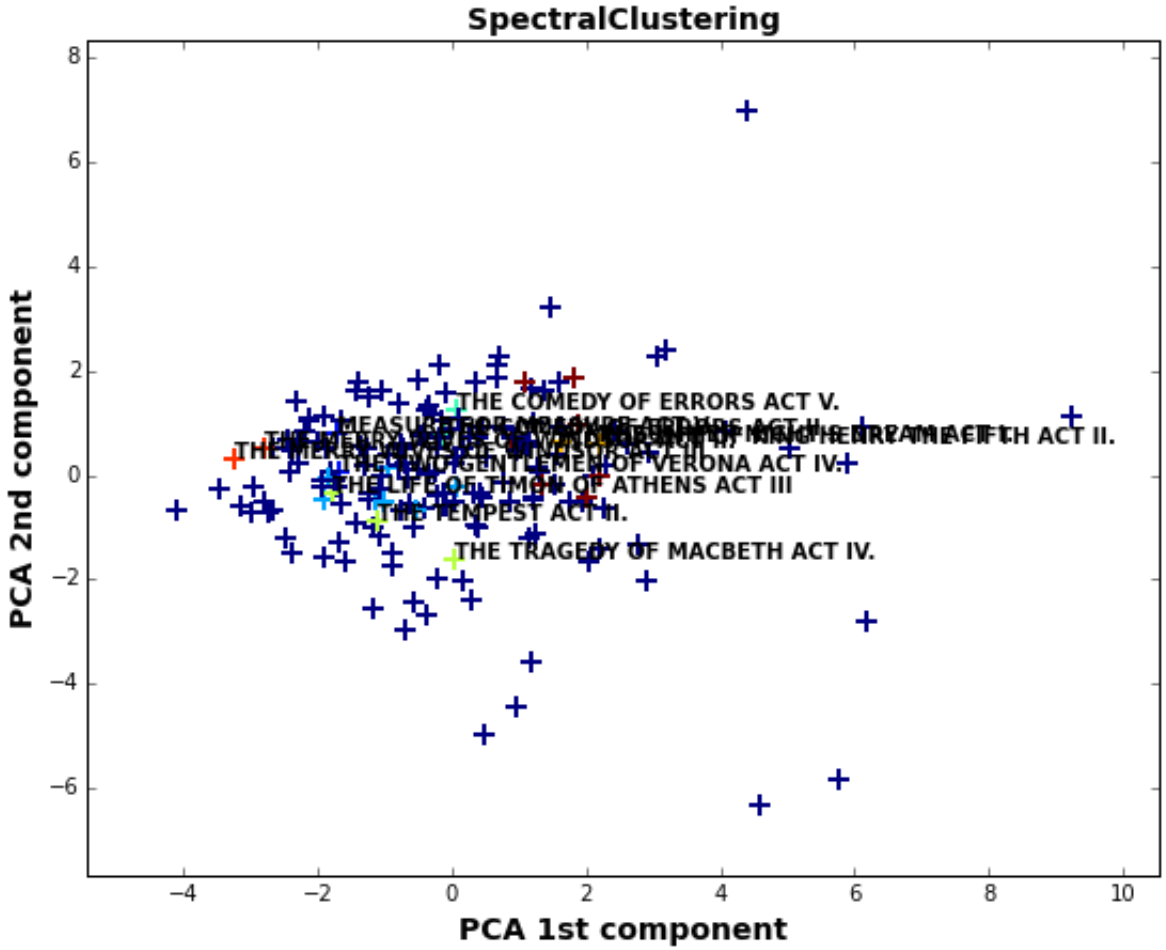
```
 'THE FIRST PART OF KING HENRY THE FOURTH ACT III'
 'KING RICHARD III ACT III']


[32, 44, 114, 140]
['THE COMEDY OF ERRORS ACT III' 'THE COMEDY OF ERRORS ACT V. '
 'THE COMEDY OF ERRORS ACT Il.' 'THE COMEDY OF ERRORS ACT IV.']


[13, 18, 65, 86, 87, 90, 99, 102, 139, 152, 153, 156, 172]
['THE LIFE OF TIMON OF ATHENS ACT V. '
 'SECOND PART OF KING HENRY IV ACT IV.'
 'THE TRAGEDY OF TITUS ANDRONICUS ACT V. '
 'THE THIRD PART OF KING HENRY THE SIXTH ACT I. ' 'KING JOHN ACT IV.'
 'THE HISTORY OF TROILUS AND CRESSIDA ACT III'
 'THE FIRST PART OF HENRY THE SIXTH ACT III' "LOVE'S LABOUR'S LOST ACT I. "
 'MEASURE FOR MEASURE ACT I. ' 'KING RICHARD THE SECOND ACT II.'
 'THE SECOND PART OF KING HENRY THE SIXTH ACT I. '
 'THE TRAGEDY OF TITUS ANDRONICUS ACT III'
 'THE THIRD PART OF KING HENRY THE SIXTH ACT II.']


[36, 69, 70, 144, 154, 163]
['THE TRAGEDY OF MACBETH ACT IV.' 'THE TEMPEST ACT III'
 'THE TWO GENTLEMEN OF VERONA ACT V. ' 'ALLS WELL THAT ENDS WELL ACT II.'
 'KING HENRY THE EIGHTH ACT I. ' 'THE TEMPEST ACT II.']


[7, 17, 42, 46, 49, 51, 83, 104, 108, 142, 145, 146]
['THE TWO GENTLEMEN OF VERONA ACT IV.'
 'THE TWO GENTLEMEN OF VERONA ACT I. '
 'TWELFTH NIGHT; OR, WHAT YOU WILL ACT I. '
 'THE LIFE OF TIMON OF ATHENS ACT I. ' 'MEASURE FOR MEASURE ACT V. '
 'CYMBELINE ACT II.' 'THE TRAGEDY OF CORIOLANUS ACT I. '
 'THE TRAGEDY OF KING LEAR ACT I. ' 'THE LIFE OF TIMON OF ATHENS ACT IV.'
 'CYMBELINE ACT I. ' 'MUCH ADO ABOUT NOTHING ACT IV.'
 "THE WINTER'S TALE ACT IV."]


[4, 62, 75, 128, 164, 170]
```

```
['THE TRAGEDY OF ROMEO AND JULIET ACT II.'
 'THE TRAGEDY OF CORIOLANUS ACT IV.' 'SECOND PART OF KING HENRY IV ACT III'
 "LOVE'S LABOUR'S LOST ACT IV." 'AS YOU LIKE IT ACT V. '
 "THE WINTER'S TALE ACT I. "]


[3, 41, 96, 98, 111]
['KING HENRY THE EIGHTH ACT V. ' 'THE TEMPEST ACT IV.'
 'THE TEMPEST ACT V. ' "A MIDSUMMER NIGHT'S DREAM ACT II."
 'KING HENRY THE EIGHTH ACT II.']


[37, 74, 132, 159, 176, 180, 181]
['THE MERRY WIVES OF WINDSOR ACT I. '
 'THE TRAGEDY OF JULIUS CAESAR ACT V. '
 'THE TRAGEDY OF ANTONY AND CLEOPATRA ACT III'
 'THE MERRY WIVES OF WINDSOR ACT II.'
 'TWELFTH NIGHT; OR, WHAT YOU WILL ACT III'
 'THE MERRY WIVES OF WINDSOR ACT III' 'THE MERRY WIVES OF WINDSOR ACT IV.']


[31, 52]
['ALLS WELL THAT ENDS WELL ACT III'
 'THE HISTORY OF TROILUS AND CRESSIDA ACT V. ']


92
ALLS WELL THAT ENDS WELL ACT V.


94
THE SONNETS ACT1609


[0, 23, 43, 66, 76, 79, 167, 168, 171, 179]
['MEASURE FOR MEASURE ACT II.' 'THE TAMING OF THE SHREW ACT IV.'
 'THE TWO GENTLEMEN OF VERONA ACT II.'
 'THE TRAGEDY OF JULIUS CAESAR ACT II.'
 'THE TRAGEDY OF HAMLET, PRINCE OF DENMARK ACT I. '
 'THE TRAGEDY OF CORIOLANUS ACT III' 'AS YOU LIKE IT ACT I. '
```

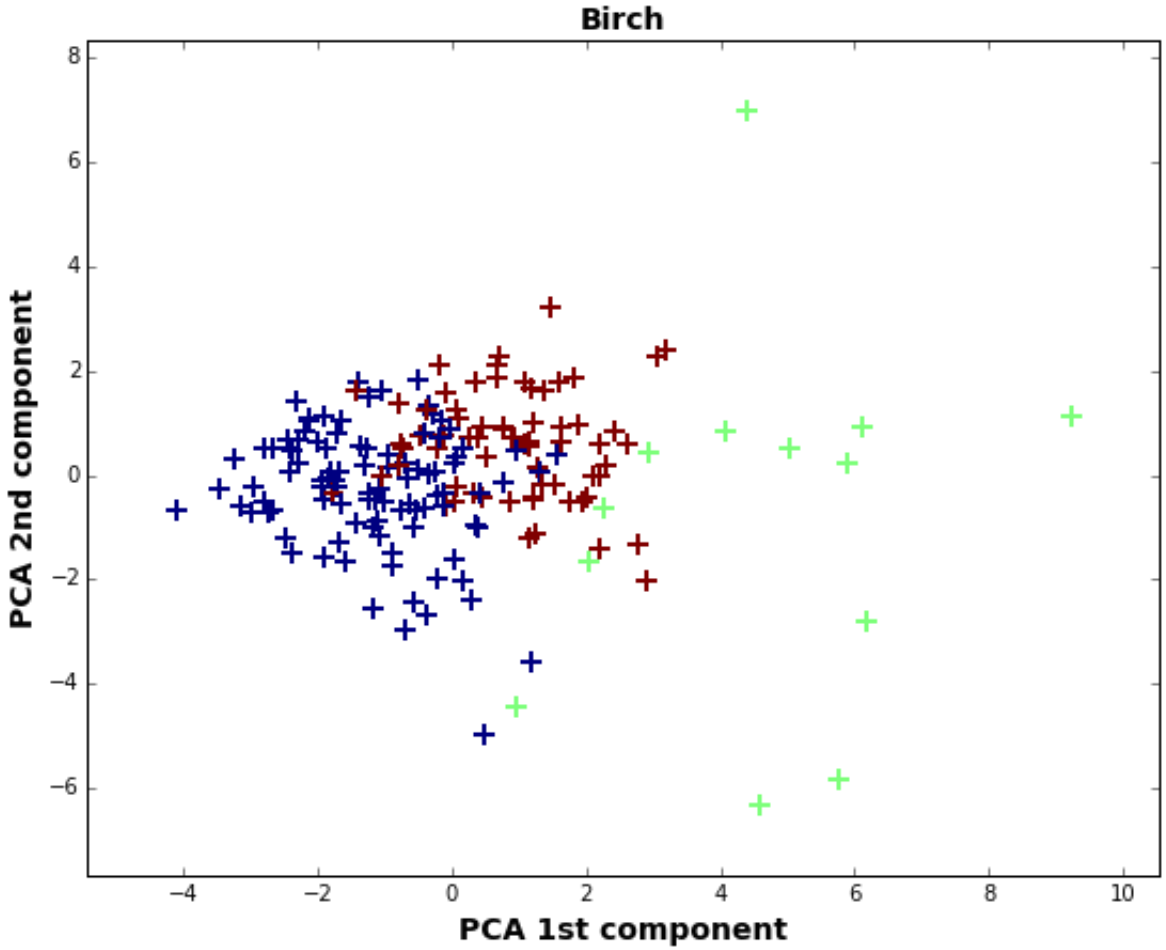
```
 'THE TRAGEDY OF ROMEO AND JULIET ACT III'
 'THE TRAGEDY OF HAMLET, PRINCE OF DENMARK ACT IV.'
 "THE WINTER'S TALE ACT V. "]


[20, 27, 29, 61, 155, 174]
['THE LIFE OF TIMON OF ATHENS ACT II.' 'ALLS WELL THAT ENDS WELL ACT IV '
 'THE HISTORY OF TROILUS AND CRESSIDA ACT II.'
 'ALLS WELL THAT ENDS WELL ACT V S' "LOVE'S LABOUR'S LOST ACT II."
 'ALLS WELL THAT ENDS WELL ACT I. ']


[22, 26, 33, 68, 88]
["LOVE'S LABOUR'S LOST ACT III" 'THE MERRY WIVES OF WINDSOR ACT V. '
 "A MIDSUMMER NIGHT'S DREAM ACT IV." 'KING HENRY THE EIGHTH ACT IV.'
 "A MIDSUMMER NIGHT'S DREAM ACT V. "]


[45, 78, 85, 106, 127, 169, 175]
['THE THIRD PART OF KING HENRY THE SIXTH ACT V. '
 'THE TRAGEDY OF TITUS ANDRONICUS ACT IV.'
 'TWELFTH NIGHT; OR, WHAT YOU WILL ACT V. '
 'THE TAMING OF THE SHREW ACT III' "THE WINTER'S TALE ACT II."
 'THE TRAGEDY OF TITUS ANDRONICUS ACT II.'
 'THE TAMING OF THE SHREW ACT Il.']


[30, 113, 126, 148, 161]
['THE TWO GENTLEMEN OF VERONA ACT III'
 'THE TRAGEDY OF ANTONY AND CLEOPATRA ACT IV.'
 'TWELFTH NIGHT; OR, WHAT YOU WILL ACT IV.'
 'THE TRAGEDY OF MACBETH ACT V. ' 'THE TRAGEDY OF JULIUS CAESAR ACT IV.']


[2, 19, 107, 119, 135, 178]
['THE TRAGEDY OF OTHELLO, MOOR OF VENICE ACT V. '
 'THE TRAGEDY OF OTHELLO, MOOR OF VENICE ACT IV.'
 'THE TRAGEDY OF ROMEO AND JULIET ACT IV.' 'MUCH ADO ABOUT NOTHING ACT V. '
 'THE TAMING OF THE SHREW ACT V. '
 'THE TRAGEDY OF ANTONY AND CLEOPATRA ACT II.']
```

```
/Users/binyan/Python/anaconda/envs/py34/lib/python3.4/site-packages/numpy/core/_methods.py:5
9: RuntimeWarning: Mean of empty slice.
  warnings.warn("Mean of empty slice.", RuntimeWarning)
```

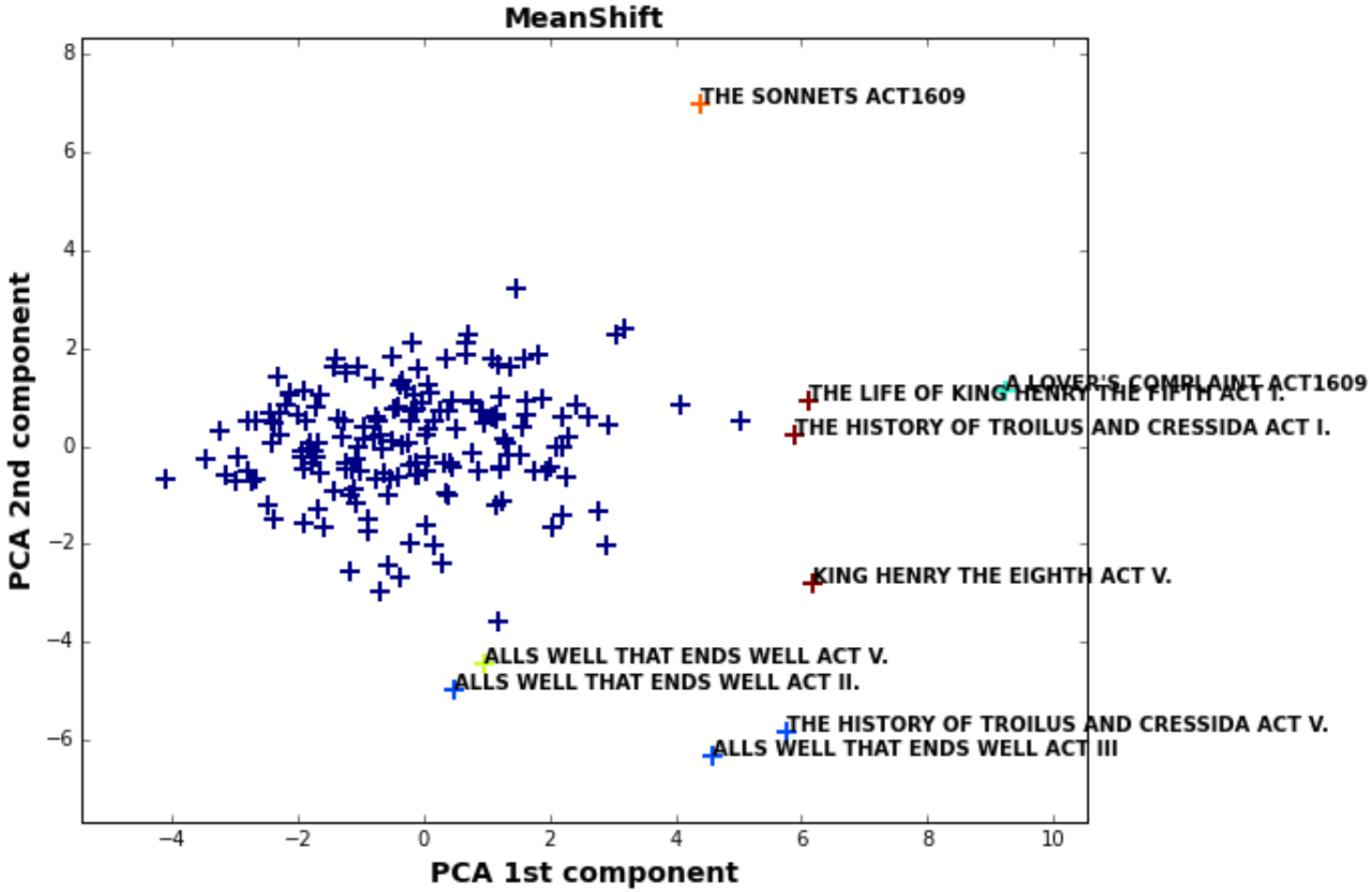
In [33]:
```
clusterDocuments(X, n_clusters = 8, scaled = True, PCA_X = True, method = 'SpectralClusterin
g', verbose = False)
clusterDocuments(X, scaled = True, PCA_X = True, method = 'Birch', verbose = False)
#clusterDocuments(X, scaled = True, PCA_X = True, method = 'DBSCAN')
clusterDocuments(X, scaled = True, PCA_X = True, method = 'MeanShift', verbose = False)
```




```
Number of clusters:  8
Silhouette Coefficient: -0.191
```




```
Number of clusters:  3
Silhouette Coefficient: 0.085
```




```
Number of clusters:  6
Silhouette Coefficient: 0.315
```

```
/Users/binyan/Python/anaconda/envs/py34/lib/python3.4/site-packages/numpy/core/_methods.py:5
9: RuntimeWarning: Mean of empty slice.
  warnings.warn("Mean of empty slice.", RuntimeWarning)
```