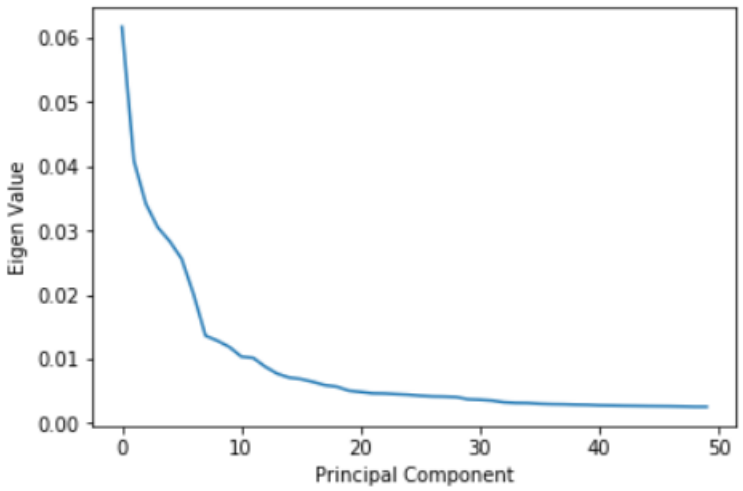
Yogiraj Awati

NUID: 001663431

2 B) Plot the eigenvalues of the principal components. Calculate how many components are needed to explain 50% of the total variance?

```
Drawing plot:
```



('Number of components required for 50% of the total variance: ', 7)

2C) Identify which words are important in each of the principal components. To do so, take the sum of squares of each of the component vectors to check how they are normalized. For each component, then print out the words for which the absolute value of the component is larger than 0.20 of the norm.

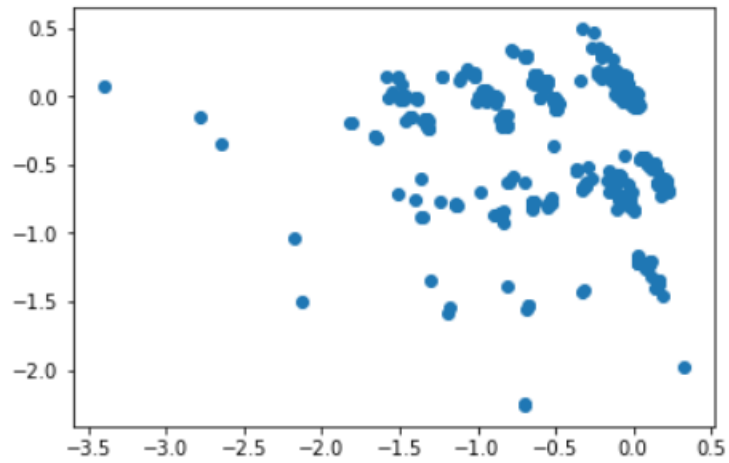
('Important Unique Words: ', 58)

set([u'and', u'control', u'from', u'image', u'detection', u'an', u'high', u'design', u'learning', u'in', u'modeling', u'networks', u'information', u'management', u'based', u'network', u'for', u'application', u'efficient', u'with', u'distributed', u'system', u'by', u'to', u'systems', u'performance', u'adaptive', u'approach', u'method', u'real', u'case', u'web', u'models', u'dynamic', u'fuzzy', u'wireless', u'applications', u'new', u'using', u'data', u'the', u'knowledge', u'a', u'on', u'multi', u'sensor', u'algorithm', u'neural', u'mobile', u'of', u'study', u'analysis', u's', u'algorithms', u'time', u'model', u'parallel', u'software'])

2D) Make a scatter plot of some reasonably sized sample (1k-10k titles). Explain the structure (or lack thereof) you see based on your results from item b-c.

Analysis for 1k titles:

Scatter plot:



Analysis of steps b,c,d :

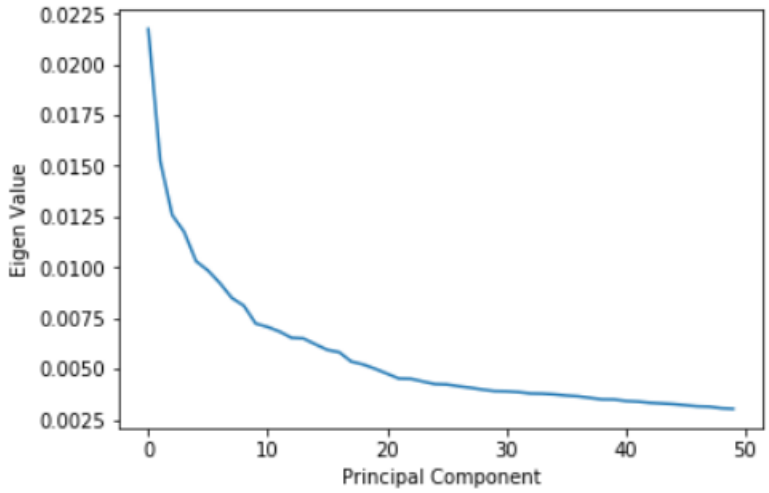
From step we conclude that first 7 components impact 50% of the total variance. After 7th component, the variance of the data set tends to decrease.

There are 58 unique words which impact this variance.

Scatter plot is between first two components of PCA. PC1 shows greater variation compared to PC2. PC1 shows large randomness. From PC2 we understand that the data is being categorized into 3 major categories. If we look around PC2 axis, we can draw 3 horizontal axis, which covers major chunk of the data.

2E) Run a preprocessing step to remove stop words (a list of stop words is provided which is identical to the list used in Spark). Rerun steps b-d and evaluate whether this has improved your representation.

Drawing plot:

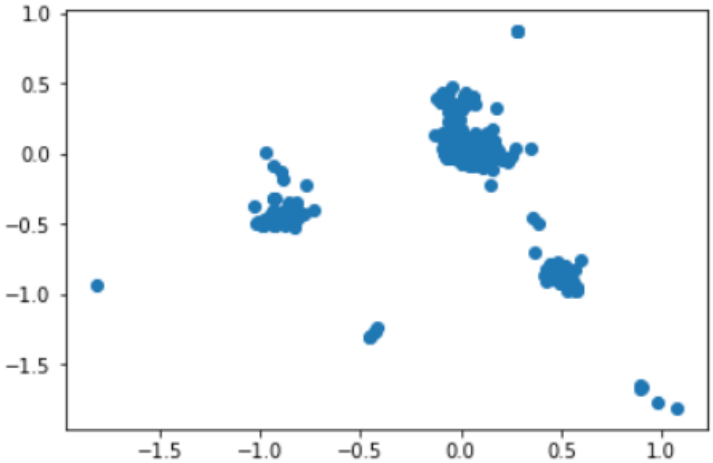


('Number of components required for 50% of the total variance: ', 14)

('Important Unique Words: ', 63)

set([u'control', u'kernel', u'via', u'inference', u'classification', u'search', u'image', u'dynamic', u'gaussian', u'machines', u'high', u'design', u'learning', u'query', u'networks', u'recognition', u'clustering', u'information', u'scale', u'based', u'network', u'decision', u'efficient', u'support', u'distributed', u'system', u'application', u'vector', u'systems', u'time', u'relational', u'adaptive', u'approach', u'xml', u'non', u'object', u'linear', u'models', u'processing', u'processes', u'visual', u'management', u'bayesian', u'using', u'queries', u'data', u'markov', u'dimensional', u'multi', u'algorithm', u'language', u'database', u'neural', u'programming', u'analysis', u'large', u'optimization', u'algorithms', u'sparse', u'databases', u'model', u'estimation', u'temporal'])

Scatter plot:

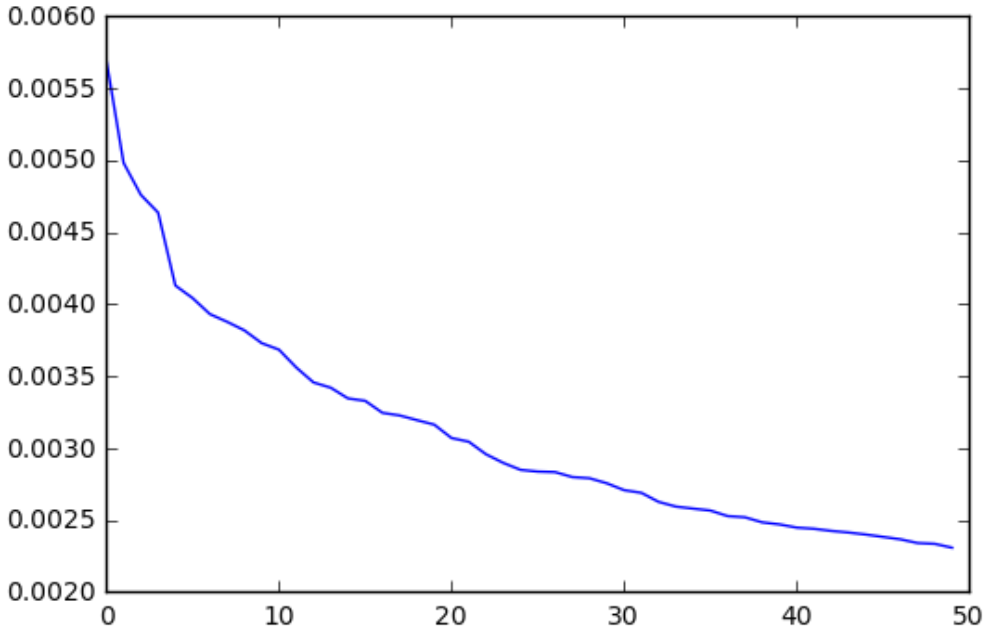


Evaluation:

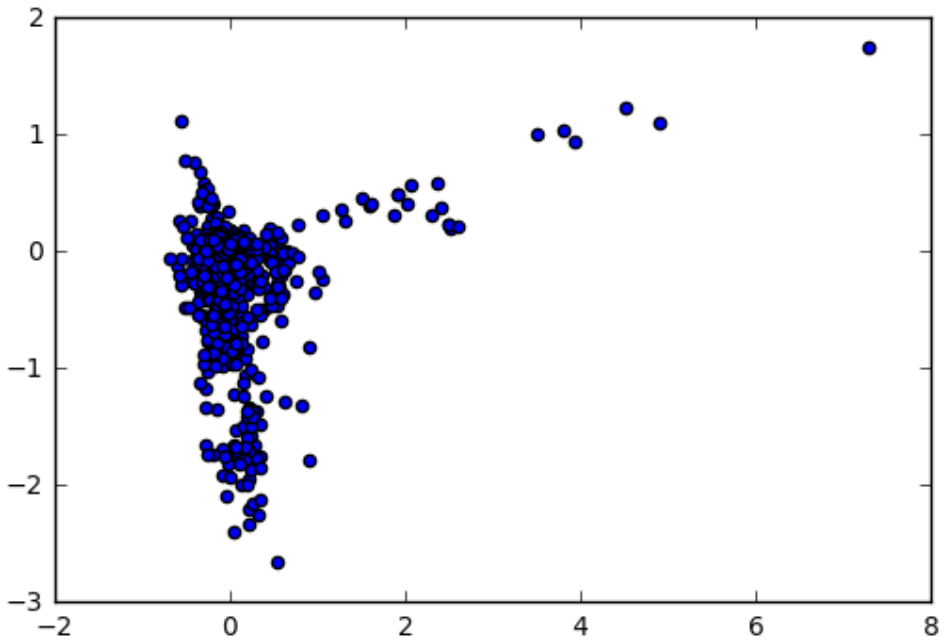
Yes, this has improved our representation. Now, there are 14 major components which depict 50 % variance which was 7 in the previous case. This means that the data has been distributed more along the components compared to previous. This increases the credibility of the results since the results are more meaningful because we are removing the stop words. Moreover, the scatter plot has become more narrow, i.e it depicts more concrete information. The data is better clustered compared to previous results.

2F) Calculate TF-IDF features for all titles and rerun the operations in parts b-d of this exercise. How have your results changed?

Plot:



Scatter plot:

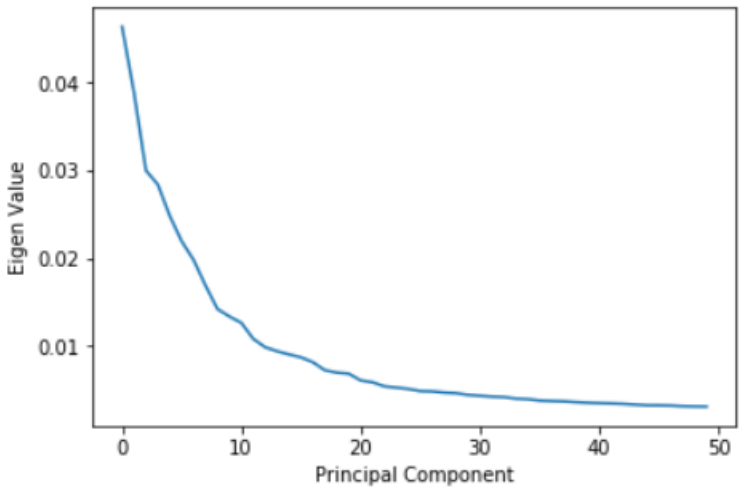


Evaluation:

The data is more distributed in TF-IDF since the words are normalized according to document length. The plot is more wide, covering more PCA components. Scatter plot also become clustered to the core, reducing the variance along first and second component. There might also be the case that noise has been reduced in the data set.

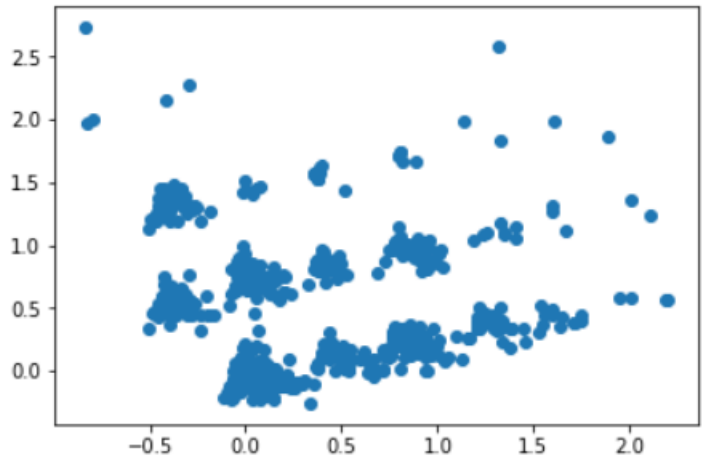
2l) Merge the two sets of titles. Construct both word count vectors and TF-IDF features. Repeat steps b-d and compare word count results to TF-IDF results.

Plot:



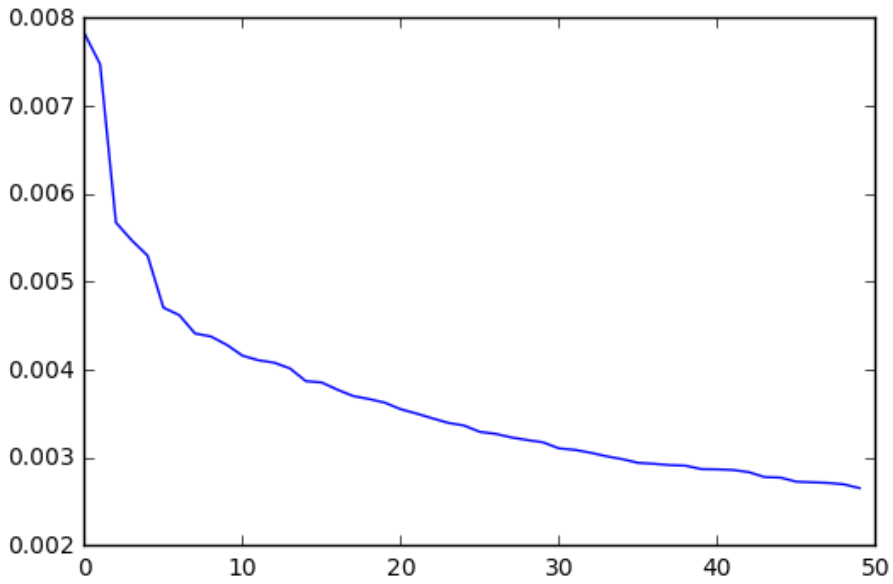
('Number of components required for 50% of the total variance: ', 9)

Scatter Plot:



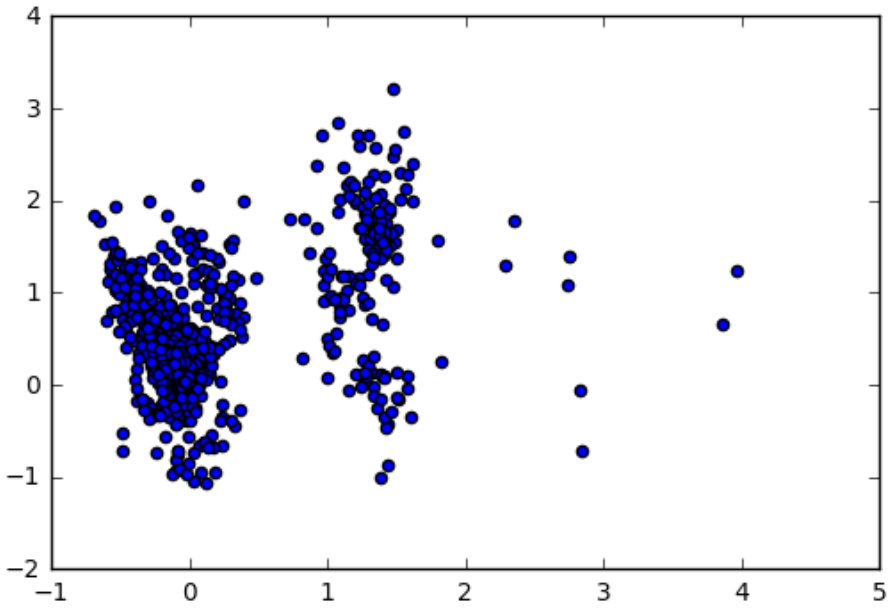
TF-IDF:

Plot:



('Number of components required for 50% of the total variance: ', 20)

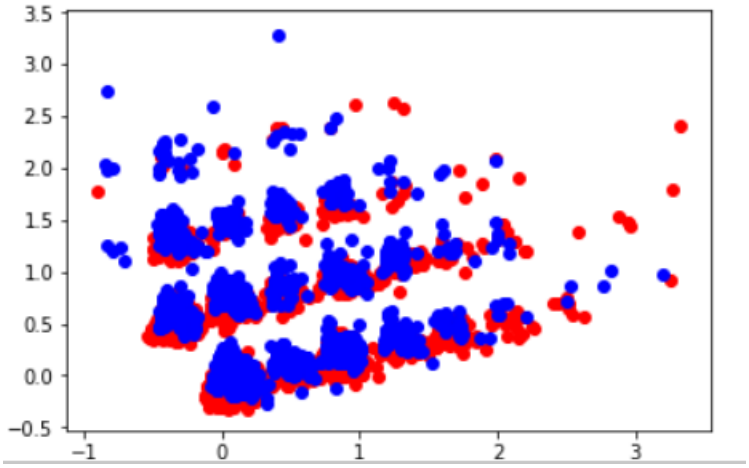
Scatter plot:



Evaluation:

1) PCA components for TF-IDF are more informative i.e 20 leading to better analysis
2) TF-IDF scatter plot is more clustered
3) Compared to normal analysis, I guess TF-IDF has less noise due normalization of the words according to document length

2J) Now make a scatter plot of these two principal components, showing the titles from each subset in different colors. Again compare word counts and TF-IDF.
Did PCA succeed in uncovering the differences between the communities?



From the graph, I hardly believe that PCA was successful in uncovering the differences. Since you cannot clearly distinguish the behavior of 2 components.

Collabration :

Discussed strategy with professor, Rushikesh Badami and Ashish Kalbhor

Note:

Submitting md file since I am not sure whether the notebook is successfully downloaded to local machine