Michal Krzysztof

Polak Szarkowicz

118304271

CS3205

Lab Report 2

Introduction

The purpose of this report is to evaluate and compare different ways of visualising text documents. We will use Pipeline and vis-kt to run key-term based clustering on text dataset from different reporting agencies and scientific papers.

Vis-kt Is a tool that lets you perform key term based clustering – clustering of documents based on key terms, uses words and word relevance to guide the clustering and lets you tailor the clustering based on the actual views you gain. Works with ltc – finds out the relevance of words and then assigns the documents in those sets of words.

Everything is available at my GitHub repository

Exercise 1

(a)

I’m going to use an NJ tree and t-SNE for projecting the document sets. Both contain articles from four news outlets (AP, BBC, CNN and Reuters) collected over two days in April 2006. Labelled news contains 381 objects and news contains 2684 objects.

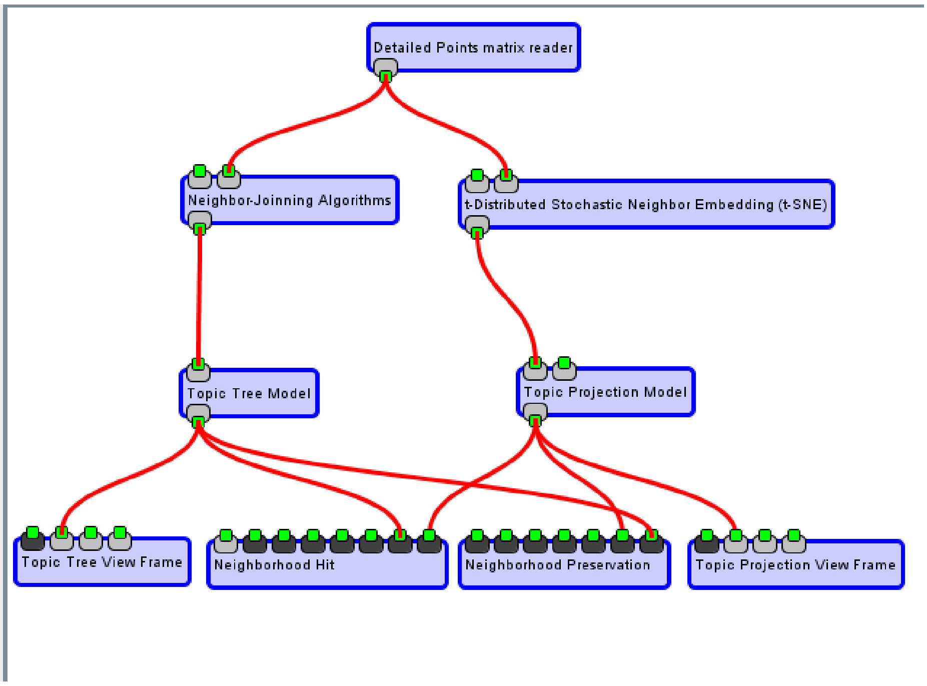

Detailed Points matrix reader
Neighbor-Joinning Algorithms
t-Distributed Stochastic Neighbor Embedding (t-SNE)
Topic Tree Model
Topic Projection Model
Topic Tree View Frame
Neighborhood Hit
Neighborhood Preservation
Topic Projection View Frame

Figure 1.1 - News Data Set-Up

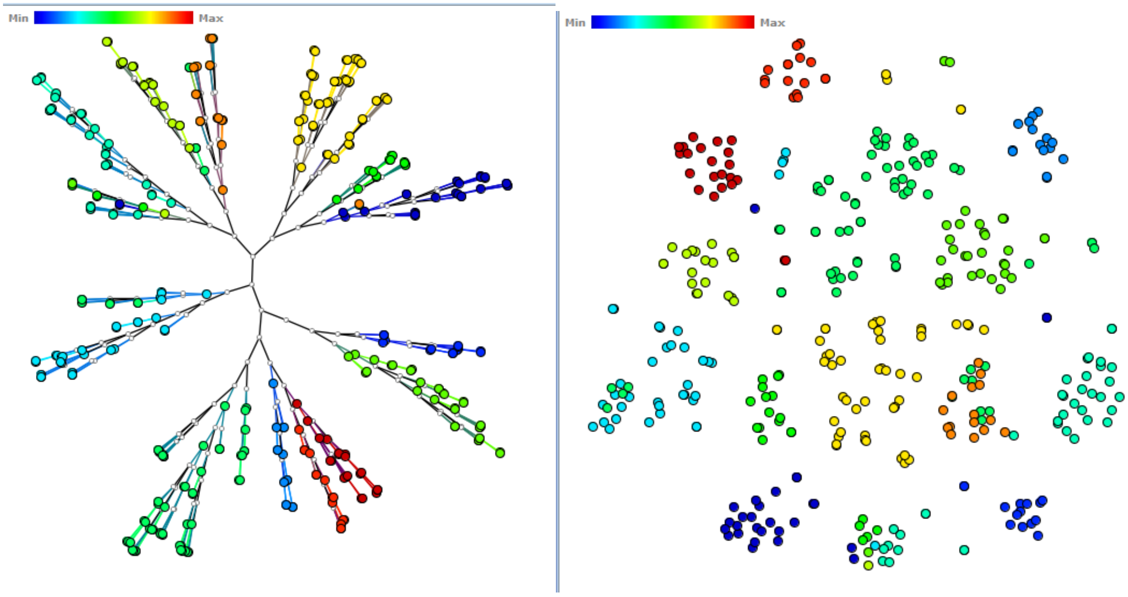


Figure 1.2 – NJ Tree (left) and t-SNE (right) Projection For Labelled News Data

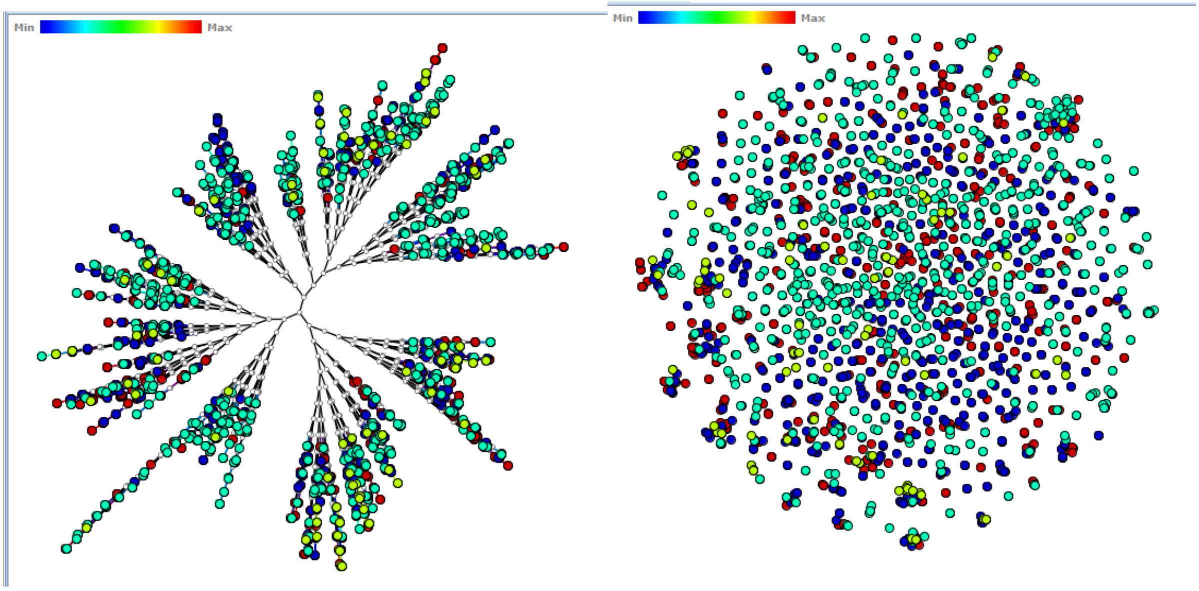


Figure 1.3 – NJ Tree (left) and t-SNE (right) Projection For Unlabelled News Data

(b)

Some of the topics include:

- The Immigration Bill in the Senate
- Terrorist Attacks
- Meredith Vieira Show
- Protests in Kathmandu
- A Jury Suffering a Heart Attack
- Bird Flu spreading in England and Scotland

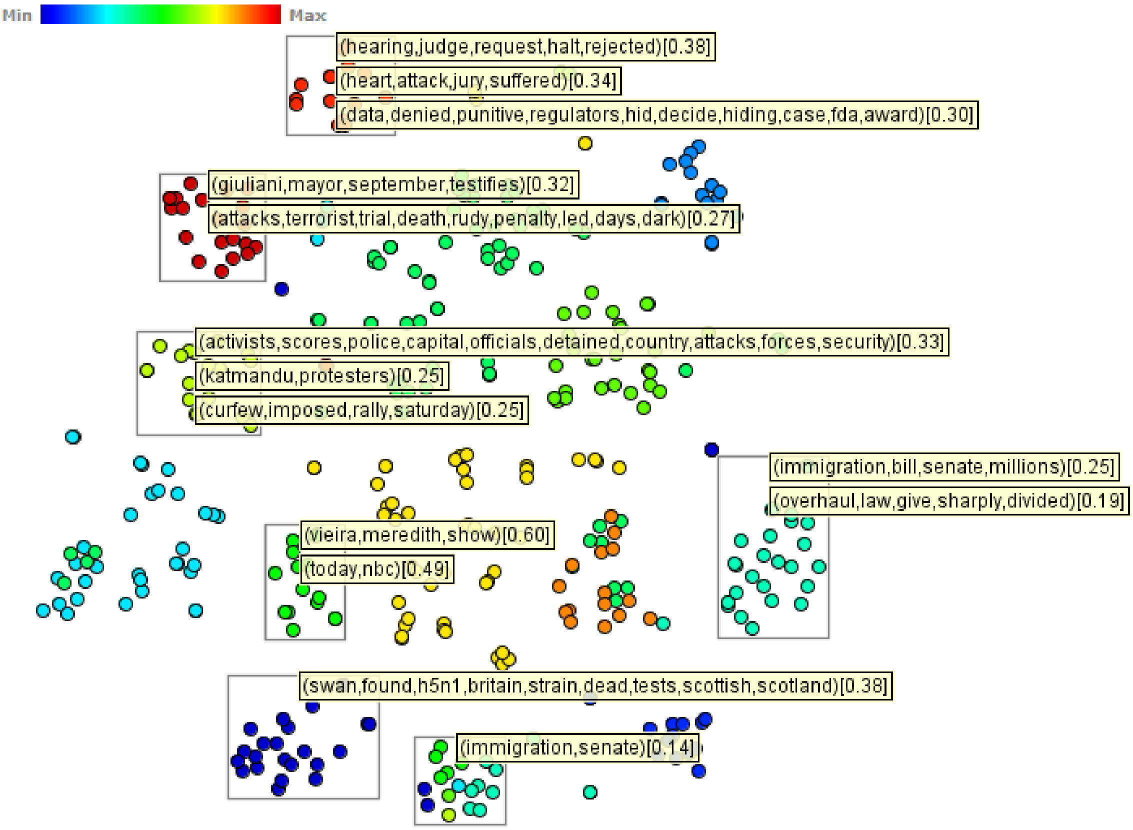


Figure 1.4 - t-SNE Projection of Labelled News Data

(c)

Yes, they're useful, especially in the labelled data set. We can see different topic clusters that would help us in deciding on how we can use or process the data later.

The clusterings resulting from the unlabelled news data are not as useful. They are poorly separated and would require further investigation.

(d)

In for the t-SNE parameters, I've chosen:

- Initial Dimensions as 40 (about 10% of 381)
- Target Dimensions as 2
- Perplexity as 50 (chose the high number to allows more distinct clusters to form)
- Maximum Iterations as 300 (to achieve a good separation)

For the NJ parameters, I've specified the algorithm to:

- Use Leaf Promotion
- Use the Original Neighbour-Joining version

Everywhere I specified to use cosine-based dissimilarity.

For the labelled data:

The silhouette coefficient in the original data is 0.1836292- which is quite good.

The silhouette coefficient of the data in t-SNE is 0.20502892 – better separation than in the original space.

The silhouette coefficient of the data in the NJ Tree is 0.33218148 – which is excellent and mean that the projection has achieved a high degree of separation.

For the Unlabelled data:

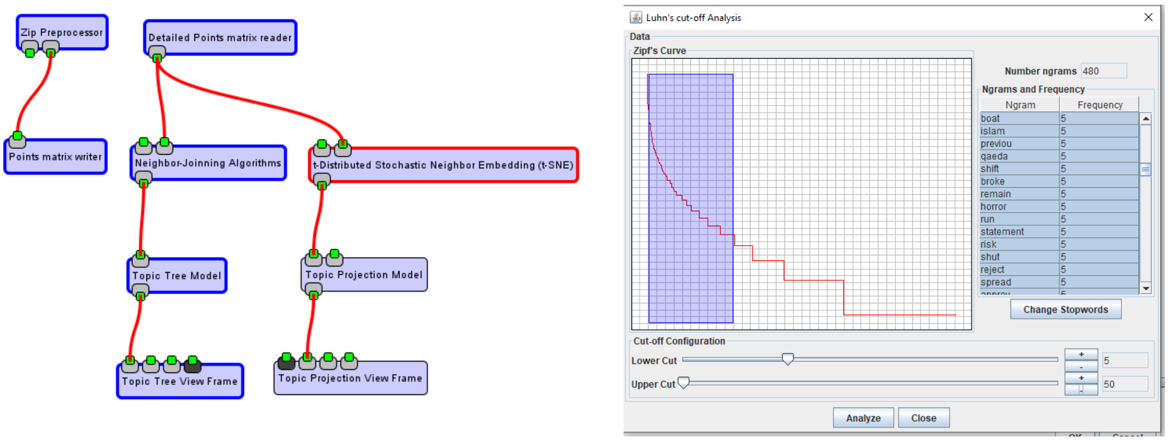
The silhouette coefficient in the original data is 0.002628842- which is quite poor.

The silhouette coefficient of the data in t-SNE is - 0.04919868 – worse separation than in the original space.

The silhouette coefficient of the data in the NJ Tree -0.006279052 – which is worse than t-SNE. It means that the projection has achieved a very small degree of separation. The data is more clustered than in the original space.

Exercise 2

(a)



Number ngrams 480

Ngrams and Frequency

Ngram	Frequency
boat	5
islam	5
previou	5
qaeda	5
shift	5
broke	5
remain	5
horror	5
run	5
statement	5
risk	5
shut	5
reject	5
spread	5

Lower Cut 5

Upper Cut 50

Figure 2.1 - News Data Set-Up (Left) and Choosing Cut-Offs (Right)

(b)

Some of the topics include:

- The Immigration Bill in the Senate
- Terrorist Attacks
- Meredith Vieira Show
- Protests in Kathmandu
- A Jury Suffering a Heart Attack
- Bird Flu spreading in England and Scotland

(c) My new SVM is also useful for separating topics into distinct clusters.

(d) Original silhouette is 0.19261593. Quite good.

The projected silhouette in t-SNE is 0.18591425.

The projected silhouette in the NJ tree is 0.26579228.

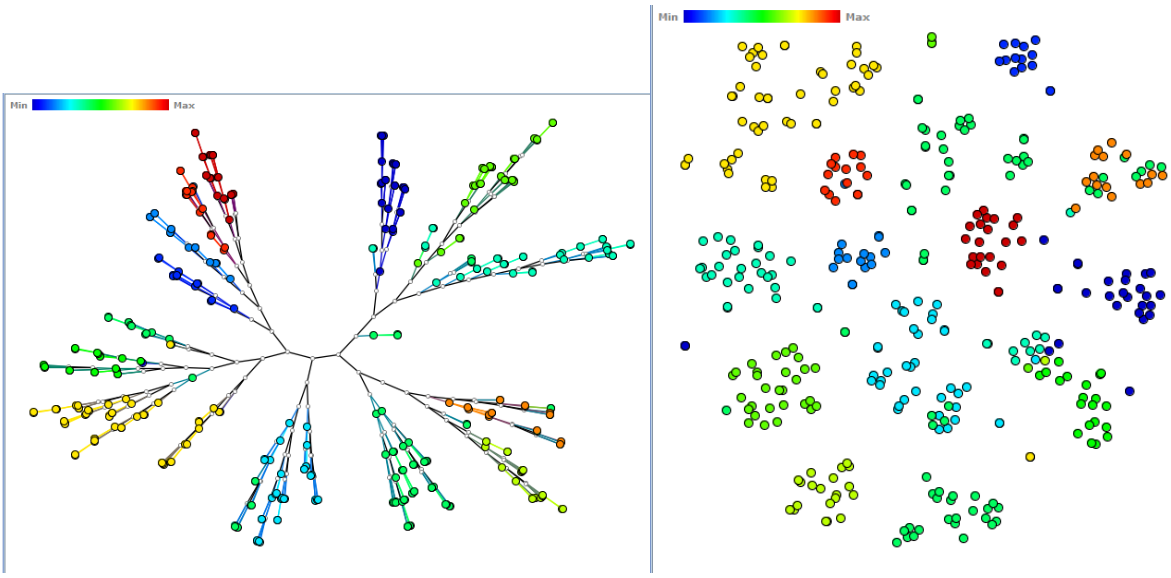


Figure 2.2 – NJ Tree (left) and t-SNE (right) Projection for My SVM

Exercise 3

a)

Cluster0 – gaza, hamas, strip, Palestine, Israeli

Cluster1 - night, session, los angeles, game

Cluster2 – prices, stocks, sales, data

Cluster3 – Iraq, Baghdad, Shiite, suicide, killing

Cluster4 – Texas, police, gas, protesters, kathmandu

Cluster5 – trail, chief, judge, execution

Cluster6 – wash, London, drug, federal, flu

Cluster7 – Washington, bush, house, immigrant, illegal

Exercise 4

(a) There are 179 papers, and they are clustered into 6 groups,

Cluster 0 and 1 mostly contained the papers about data visualisation, keywords included consistency, clumpy, wdbc, minghim and scagnostics.

Cluster 5 contains papers that write about projections and plots. It's key-terms include merge, subspace, pipeline and scatterplot.

Cluster 2 covers bioacoustic.

Clusters 3 covers genomics and Euler diagrams.

Clusters 4 covers Twitter and tweets.

Evaluate the quality of multi-dimensional projections, they try to map multidimensional data, but it has to lose data, techniques are focused on preserving certain data types of interest from the regional (vector) space. And have different priorities, eg preserve distances or preserve neighbourhoods

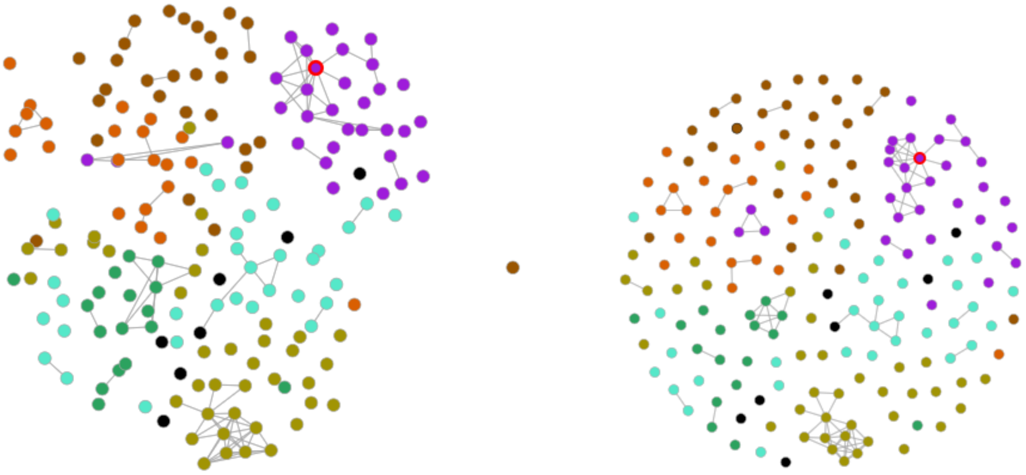


Figure 4.1 – Initial t-SNE (left) and Force (Right) Projections for Papers Dataset

(b)

The t-SNE silhouette coefficient is 0.149. This is alright and means that the algorithm was separated the clusters, but too well.

There exists quite a bit of overlap between clusters 0, 1 and 5. We can observe that on the t-SNE projection. This makes sense since they cover similar topics – projections and visualisations. This means the topics share many key terms and cover similar topics.

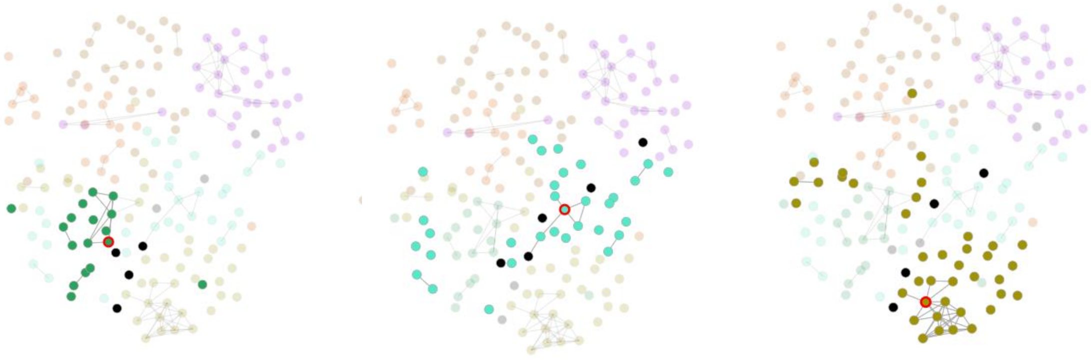


Figure 4.2 – Overlap between clusters 0, 1 and 5

There is a small amount of overlap between cluster 3 and 4

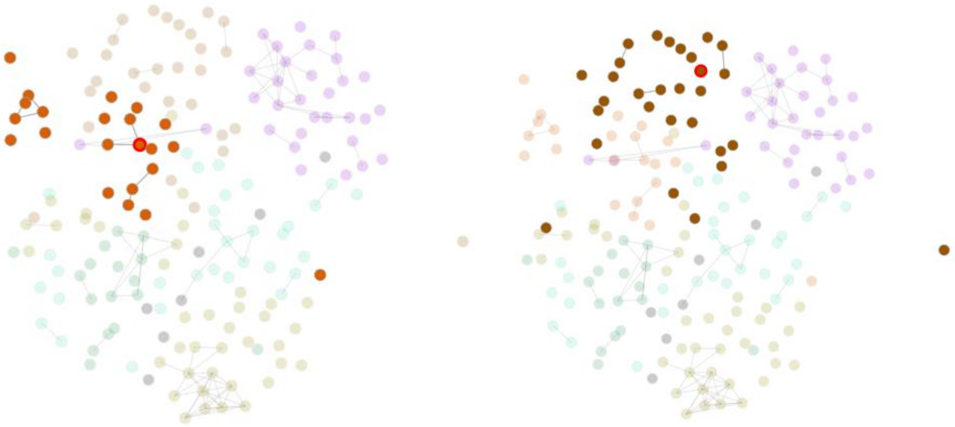


Figure 4.3 – Overlap between clusters 3, 4

Cluster 2 is separated very clearly. Suggesting that it contains paper that key terms that are a distance from the other clusters.

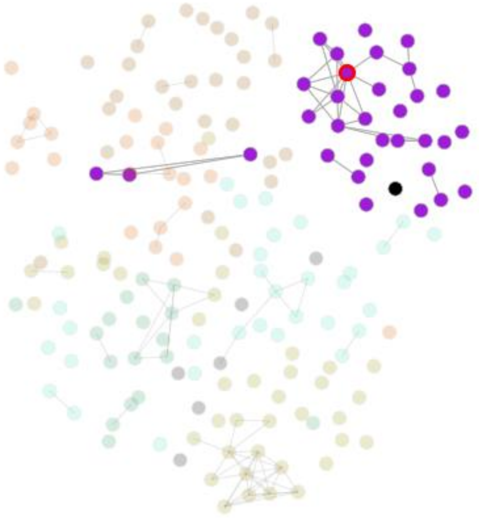


Figure 4.4 – Cluster 2 separated well from the other clusters

(c)

There are now 7 clusters in total. The key terms in the new clusters are

As we can see. The separation for the clusters is better in this choice than when we chose 7 clusters.

The purpose of adding more clusters was to see if it helps to improve the separation of documents.

This time clusters 0 and 1 are better separated.

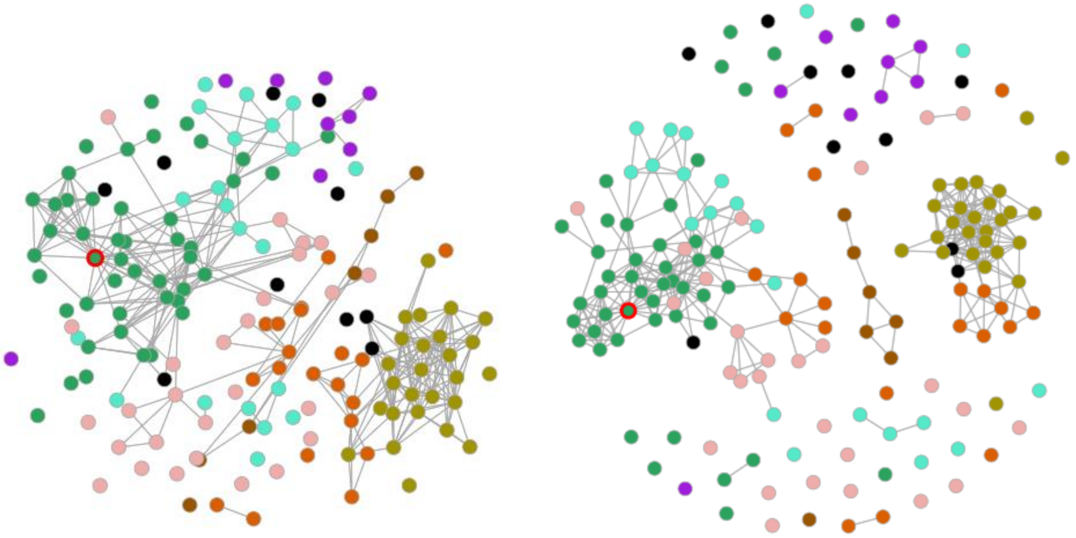
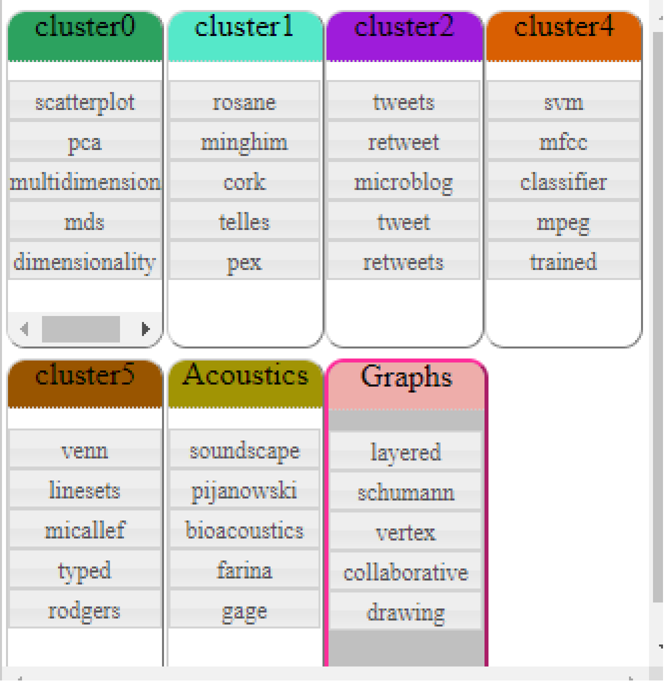


Figure 4.5 – Better Separation of Clusters (t-SNE and Force Based Projections)



cluster0	cluster1	cluster2	cluster4
scatterplot	rosane	tweets	svm
pca	minghim	retweet	mfcc
multidimension	cork	microblog	classifier
mds	telles	tweet	mpeg
dimensionality	pex	retweets	trained

cluster5	Acoustics	Graphs
venn	soundscape	layered
linesets	pijanowski	schumann
micallef	bioacoustics	vertex
typed	farina	collaborative
rodgers	gage	drawing

Figure 4.6 – Clusters and Included Key Terms

Conclusion

The conclusion is that visually clustering and projecting documents using tools like key term based clustering can be useful in separating documents.