**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.

Lab3 vid 1 + 2

Table of attributes = .data

Distance matrix = .dmat

Points > read and write data (.data)

Trees are like hierarchical clustering

Silhouette, the lower the more mixed

IDMAP wants to spread things around

Creating topics (i), s= show

Lsp = 10%

Unlabelled = colours mean nothing

Attributes and distance matrix

Lab4

Evaluate the quality of a multi d projection, 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

Stress and silhouette, > data analysis

(but need labelled projection view NOT topic projection model)

Stress curve(how distance is behaving) if the distance is preserved then the point should fall on the line

Labelled projection model vs topic projection model

There is NJ, LSP, IDMAP, t-SNE, MDS

Disturb distance to preserve neighbourhoods

Neighbour hit = for each point, it checks if its neighbours are of the same label, % of neighbours of the same label, how well it preserves neighbours relative to the labels, eg for 15 neighbours 58% are of the same label

For tree use, tree model

Neighbourhood preservation = % of distances being preserved, >0.30 is very good

Lab5 video 1 – covers vis-kt, V3

<https://ucc.instructure.com/courses/32085/modules/items/926597>

<https://ucc.instructure.com/courses/32085/modules/items/926598>

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

BLACK = more than one cluster at once

**DSA1**

= most clusters are quite spread out in terms of projection

9 or so clusters

Some are kinda connected in some regions

t-sne is not a good projection when the data is sparse

Force-based displays by force by adjusting the distance between points which is proportional to the weight in the edge, the weight is the similarity

Cluster number too large = gets confused and groups unrelated articles together

Too small = few large clusters, poor separation

Add cluster > right click > addterm

Also add words in plural

Tells it to put those words in higher weights than usual

eg if mixed or interesting

Compare silhouette

Lots of heterogeneous data (many topics, many documents, some have only a few words in them). Harder to cluster, hard to improve

Dense groups = more news

**DSA3**

4 Clusters

Visualising documents

**Uploading Documents**

Click upload documents

DSA2

Show the list of documents

Preprocess – lemmatized, remove numbers

Might cluster together because of the lack of information

**Stop words and Stemming**

<https://ucc.instructure.com/courses/32085/modules/items/917459>

**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.

Diagram

Description automatically generated

*Figure 1.1 - News Data Set-Up*

*A picture containing shape

Description automatically generatedChart, scatter chart

Description automatically generated*

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

Chart, shape

Description automatically generatedChart, scatter chart

Description automatically generated

*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

Timeline

Description automatically generated

*Figure 1.3 - 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)**

**Graphical user interface, diagram, application

Description automatically generatedChart

Description automatically generated**

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

**(b)**

**(c)**

**(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.

**A picture containing diagram

Description automatically generatedChart, scatter chart

Description automatically generated**

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

**Exercise 3**

**(a)**

**(b)**

**(c)**

**(d)**

**(e)**

**(f)**

**(g)**

**(h)**

**Exercise 4**

**(a)**

**(b)**

**(c)**

**Conclusion**