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>