UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

School of Computer Science and Information Technology

CS3205 Data Visualization for Analytics Applications

Rosane Minghim

March 10, 2021

Lab Report 2

User-centric document exploration and clustering

Deadline: March 22nd

This lab report summarizes the findings of exploring document sets employing multidimensional projections as well as visual keyterm-based clustering.

Multidimensional Projections are mappings from multidimensional spaces (represented by a table of attributes or a similarity matrix) into visual spaces (2D or 3D). In order to employ this techique for documents, a Vector Space Model (VSM) is generated. Additionally, extracting topics can support making sense of general contents for groups of points formed in the projection. The tool Vispipeline is an experimental tool that supports the whole document exploration pipeline based on VSMs, projections, and topic extraction.

Key-term based clustering is a method for grouping documents based on relevance of words. It supports intuitive user interaction with the clustering process, by allowing feedback trough word and cluster addition, removal aor re-ordering. Vis-kt is an open system that employs various visualizations to support the process. It has been developed in cooperation between Dalhousie University - Canada and University of Sao Paulo, Brazil.

The Table below shows the data sets involved in the tasks for this report.

	Data set	File/User	Size	Description
Q1	**News**	AP_BBC_CNN_Reuters*.zip	2,684	Unlabeled NEWS RSS files collected from 4 news outlets for two days in April 2006
	News_vsm	AP_BBC_CNN_Reuters*.data	2,684	2.2K features extracted from the News data set
Q1, Q2	**News_labeled**	NewsSeparate_classified.zip	381	Part of the **News** data set labeled by an informed user
	News_labeled_vsm	NewsSeparate*.data	381	499 attributes extracted from the **News_labeled** data set
Q3	**News1**	DSA1	722	Part of the News Data set corresponding to News from Associated Press
Q3	**News2**	DSA2	677	Part of the News Data set corresponding to News from Reuters
Q4	**Papers**	DSA3	179+	Research articles, initially clustered in 6 clusters

Vis-kt current hosting url is `http://vis-kt.vicg.icmc.usp.br/IC3/`

The tasks you should perform for this report are:

1. Explore the document sets News and News_labeled using Vispipeline, employing at least two different projections or a projection and a NJ tree. Make the observations on the following items:
    (a) The overall document set and their content.
    (b) The topics generated for at least 4 recognizable groups of points on the projection.
    (c) The usefulness (or not) of the visualizations.
    (d) The quality of the projections for each of the data set. Choose one measure in each case.
2. Generate your own VSM for the News_labeled data set, and repeat the tasks in Exercise 1
3. For users DSA1 and DSA2 in vis-kt, perform the exploration of the initial clusters (session Initial) using the visualizations provides, t-sne projection and changing the level of connection for the force-based projection. Make the following observations.
    (a) The contents for each cluster and how coherent they are with the projection.
    (b) The keyterms and contents for at least 2 clusters and the existence of more than one topic within those clusters.
    (c) The usefulness of the clustering process to understand the content of a document collection.
    (d) Change the cluster type and verify improvement or decay in data set partition in relation to content.
    (e) Compare in general observations the difference in news content between DSA1 and DSA2.
    (f) For user DSA1 only: starting from the Session 'Initial', try to split Cluster 6 in coherent content. Report on the results. Save your session using your student number.
    (g) The differences between t-sne and force-based projection to support understanding of document clusterings.
    (h) The difference between DSA1 (Associated Press) and DSA2 (Reuters) on the news for the same two days.
4. for the Papers data set:
    (a) Explore the Session 'Initial' and make observations on cluster content and their coherence with the keyterms of each cluster.
    (b) Make observations on the clustering process, its effectiveness and try to explain the clusters formed.
    (c) Register and report the changes made to the Initial session, what the purpose was, and the results.

Report Format

Identification (Name, Module, Report Title)

1. **Introduction**
   Explain the problem, the data sets, and the goal of the report
2. **Exercise 1**
   Result of exercise 1
3. **Exercise 2**
   Result of exercise 2
4. **Exercise 3**
   Result of exercise 3
5. **Exercise 4**
   Result of exercise 4
6. **Conclusions**
   Explain your conclusions on the data sets, on the exercise itself, and on the use of visual clustering for document analysis.

OBS: Submit a single .zip file containing a pdf file of your report and a .data file of the VSM generated in Exercise 2