Assignment 5. Document Clustering with ExtMiner

1) **Test ExtMiner user interface**

   Download ExtMiner package from

   http://www.mit.jyu.fi/minurmin/kurssit/dm/extminer.zip and extract it to your working directory (e.g. c:\mytemp\username)

   Start command prompt, chage to same directory as ExtMiner and input

   `extminer -help` (or `java -jar extminer.jar -help`)

   to see command line options.

   Option -c (defaults to `extminer -c default.props`) selects the ”project” to work on. Data and index directories, indexformer, viewer and clustering parameters are defined in props files.

   To try reindexing, use option -R – this takes time on large datasets (c2 wiki & reuters datasets have been pre-indexed).

   To explore clustering parameters, use ”scan” option -s. It clusters the whole dataset multiple times on various dbscan parameters and summarizes results on output file (out.txt by default). For better performance, redirect output stream (eg. `extminer -c html.props -s > dump.txt`). For postprocessing you can also export current index to Matlab-compatible matrix files using -e option.

   Familizarize yourself with ExtMiner user interface with default and Shakespeare (shake.props) datasets. Test zooming, text retrieval, subset clustering and parameter adjusting.

2) **Shakespeare dataset** (36 documents, source: http://www.ibiblio.org/bosak/)

   Reindex shakespeare dataset (`extminer -c shake.props -R`)

   Cluster dataset using hierarchical clustering and dbscan. Combine hierarchical clusters and adjust dbscan parameters (possibly with help of scan option) to get ”natural” clustering.

   Describe the clusters with a few sentences (for both clustering algorithms, if there is a clear difference). ExtMiner shows the most ”descriptive” words from each cluster to help. You can also open some documents from the dataset. Accompany your descriptions with screenshots (ALT+PrintScreen)

3) **c2 wiki dataset** (subset, 328 documents, source: http://c2.com/cgi/wiki)

   use html.props. Don't reindex.

   Explore the clustering structure same way as in Shakespeare dataset. Also try opening the original documents. Does the clustering help navigating the wiki?

4) **Reuters dataset** (small subset, 985 document, source: http://www.daviddlewis.com/resources/testcollections/reuters21578/)

   use reuters.props. Don't reindex.

   Explorer the clustering structure same way as in Shakespeare dataset (you can skip scanning dbscan parameters). Note that all the ”documents” are contained in the same physical file, so you cannot easily browse the original documents.