Assessment: Text Analytics using Rapid Miner
Value: 30%

Date Due: Wed. March 21st
Demo: 10%; Report: 20%

Objective: Mine non-structured data using Rapid Miner.

To do:

This assessment requires you to source samples of text for mining. **No two people can use the same texts.**

1. Decide on three categories of text documents. Two topics should be relatively similar, and the third quite different.
2. Source 13 texts for each category: 10 for training; and 3 for testing. Try to use Rapidiners Web crawler to find some of the texts. Identify a suitable root for each category, and use regular expression to create a topical crawler for each topic.
3. Download training samples and test samples, and store in folders organsied by class label.
4. Create a word cloud for each category to determine frequently occuring terms, potential stop words and potential synonyms. Comment on each word cloud.
5. Experiment with the texts to answer the following questions:

   a) Preprocessing: which options produced the best list of words and the most accurate models?

   - Compare Lovins stemmer with Porters stemmer.
   - Can these be improved by using your own stemming dictionary?
   - Within what range of occurences are the words that are most predictive?
   - Does the generic stop word list, along with pruning short words, remove all the stop words? Is there a need to add your own list of stop words?
   - Compare the accuracy of models based on a binary document vector with models based on both an occurences vector and a TF-IDF vector.

   b) Clustering

   - Do any of the clustering algorithms accurately identify the three clusters of text documents? If not, analyse the results to determine why not. Give details of the algorithms you tried, and what the results were.
   - Does using a cosine meaures improve performance over using one of the symmetric measures?
   - Is mininum linkage or maximum linkage more suitable for the clusters in your

data set? Why do you think that is the case?

c) Classification

- Which classifiers produced the best model when classifying the 9 test documents (3 per category)?
- Of the terms chosen by the algorithms as being the most predictive, do they concur with the terms you thought would be the best predictors?
- Discuss what you think would be an optimal number of documents and terms to classify your three groups of texts.

Note: If you want to use SVM which is limited to a binary class label, seperate one class from the other two, and build a model that just predicts membership of that class. You can then repeat the exercise for the other two classes. This can be done by giving two folders the same label in the 'Process Document From Files' operator.

What to hand up.

A report following the stages of CRISP-DM, to include:

1. Overview of the project detailing your categories, sources of data, and how useful you found the web crawler. **(1 week)**

2. Include text visualisations such as word clouds for each category of text.

3. Data preparation done: Discuss the stemmers and filters you used as per the points raised above. Evaluate the list of terms produced. **(2 weeks)**

4. Data Mining:

4.1. Clustering: Which algorithms did you choose and why? Again, discuss as per points made above.

4.2. Classification: Which algorithms did you choose and why? Discuss as per points made above.

(2 weeks)

4. Appendix: The RapidMiner processes you developed in XML format.

Notes:

1. You man include screen shots from Rapid Miner to illustrate points made.
2. The preparation and mining done will be specific to your sample of documents. As each person will have different documents, and different categories of documents, I would expect discussions in each report to be varied.
3. Marks will be awarded based on evidence of your understanding of what is involved in mining text data. Your report should illustrate that you understand the options

available for both preparation and mining text, and can apply them effectively to a dataset of texts. Interpret all results.

4. Text mining is an iterative process of trial and error - this should also come accross in your report.

5. Any evidence of plagiarism will result in all those involved getting 0%

How to hand it up:

Demo (10%): Demo work done on **March 21st**, during the scheduled lab. The demo will cover:

- Experimentation done with a number of preprocessing options
- A good understanding of how to create a useful bag of words from your texts
- Ability to discuss the output from a number of modelling algorithms.

Report (20%): Upload your report to Moodle by Wednesday, March 21st, 10pm

Note: Your final report **WILL NOT** be corrected until you have demo'd your work