CS909: 2013-14

Week 10: Text classification, clustering and topic models

Objective

The objective of this exercise is to evaluate your understanding of representing documents as a set of features and performing classification and clustering on the documents, as discussed in lectures.

To do this exercise you will need to download the Reuters-21578 data set from http://www.daviddlewis.com/resources/testcollections/reuters21578/reuters21578.tar.gz. This data set is one of the standard corpora used by text mining researchers to test their algorithms. The data set consists of 21,578 documents, split into 21 SGML files, each with multiple tags. For more details on the data set read the associated README file at: http://www.daviddlewis.com/resources/testcollections/reuters21578/readme.txt.

Use the TOPICS tag to determine the class of each document. Ignore the DATE, PLACES, PEOPLE, ORGS, EXCHANGES, COMPANIES and UNKNOWN tags. Use the LEWISSPLIT attribute of each document to split the data into a training set and a test set.

Tasks

1. Explore the data and undertake any cleaning/pre-processing that you deem necessary for the data to be analysed.

2. Obtain feature representations of the documents/news articles as discussed in the text mining lectures. Try a version where you use topic models as features. Try topic models on their own as well as in conjunction with other features. Your final report must provide a summary of your assumptions and features and a rationale for why you decided to use them, as well as an explanation of how you obtained them.

3. Build classifiers, using either R or Python libraries to predict the TOPICS tags for documents. Focus your efforts on the 10 most populous classes, namely: (*earn, acquisitions, money-fx, grain, crude, trade, interest, ship, wheat, corn*). Use the training data for building the models and comparing their accuracy, precision and recall. Use the test set only to get an estimate of the accuracy, precision and recall of the "optimal" model based on your analysis using the training data. For the performance on the training data, report micro and macro averaged measures and explain the difference between the two.

4. Now consider all the data and use the best performing features of part (3) to represent the documents and apply three clustering algorithms of your choice, selected from those discussed in lectures. Justify your choice. Provide appropriate measures of cluster quality. Is there a correspondence between clusters and the original TOPICS labels?

5. Produce a report that clearly explains your work and contains appropriate results and table matrices. You will be evaluated on the basis of the report, you will not have the chance to explain your work in person, so this has to be well written and clear. It should include sections on:

    a. A description of the data, any pre-processing performed on it and why.
    b. Explanation of the features used, why these were chosen and how they were obtained. Numbers for different types of features. Details on any feature selection implemented.

c. Details on which classification algorithms were used and why. What parameters were used in each case and why.
d. Full evaluation of the classification algorithms, building on your experience with exercise seven. Compare different algorithms as appropriate and provide a final choice of an algorithm.
e. Details of which clustering algorithms were used and why. Provide an evaluation of the clusters in terms of appropriate quality measures and how they correspond to the original TOPICS tags.
f. Upload your code to github (https://github.com/) and provide the link to it in the report. If the link is password protected, please provide the password as well in the report. Include a short readme.txt file with an example of how to run your code.
g. Where appropriate in the report provide references to your code. You can include some of your code in the appendix.

Marking Scheme
10% Data Preprocessing
20% Feature Engineering
15% Classification process
10% Evaluation of classification
20% Clustering process
10% Evaluation of clustering
15% Clarity of report and tables, code documentation and examples.

Submission deadline: Midday, 8th May 2014 via Tabula.