CS909 Week 10: Text classification, clustering and topic models

Samuel McDermott u1466355

April 18, 2015

1 Introduction

This report demonstrates the use of text classification, clustering and topic models for the Reuters-21578 dataset [1]. This dataset consists of 21578 documents, extracted from the Reuters newswire in 1987, each with multiple or no labels. The aim of this work is to test a variety of features (topic models, n-grams), as well as text classifiers (???) and clustering (???).

The work in this project was done using R and several packages (cited as used). The code associated with this report can be found at `http://git.io/vvNci`.

2 Preprocessing and Data Cleaning

Associated R code: `TestPreprocessing.R`

The Reuters-21578[1] dataset is a `.csv` which consists of the label for document, the title of the article and the text in the article.

Some articles have several labels, and some have none. The first step is to take this information apart, so that in the final dataset, each document has only one label. This means that the same document may appear several times in the corpus, once for each label. This was done to ensure that each label accurately contained each document associated with it.

The next stage is to select the 10 most popular labels in the dataset. These were provided to us and are: *earn, acquisitions, money-fx, grain, crude, trade, interest, ship, wheat, corn.* This reduces the dataset size and provides a more concentrated selection of documents to label. The documents are then randomly ordered, so that k-fold evaluation can be carried out later.

3 Topic models

References

[1] D. Lewis. (2013) Reuters-21578, distribution 1.0. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection