Lab 7: Data Mining.

Text mining

Text mining refers to extracting information from documents (interpreted as sequence of words). The main text mining tasks are classification and clustering of documents based on their content. The simplest approach for classification/clustering documents is based on the following steps:

- Pre-process the text by:
  - Removing the *stop words* (words which do not provide specific information being rather syntactic components used to link various parts of speech). Lists of stopwords corresponding to different languages can be found at http://www.ranks.nl/stopwords
  - Transform the words by *stemming* (i.e. reduces the inflected variants of words to their root form). The most popular stemming algorithm is that proposed by Porter (see http://tartarus.org/martin/PorterStemmer/). A web service for stemming in various languages is available at http://text-processing.com/demo/stem/

- Construct for each document a *frequency vector* containing quantitative measures of the presence of words belonging to a dictionary in each of the documents. If the dictionary contains N words then to each document in the collection of documents to be processed one have to associate a vector of N elements specifying the number of occurrences of the corresponding word in the document. Since words which are specific to only some documents have a higher discriminative power, instead of using frequencies of terms it is used the so-called *TF-IDF* (term frequency – inverse document frequency) encoding characterized by the fact that the frequency of a term in a given document is divided by the number of documents in the collection which contain that term. Once these numerical vectors are constructed then one can apply any classification/clustering technique.

Exercise 1.

a) Open the file movieReviews.arff (it contains reviews on movies grouped in two categories: positive and negative)

b) Construct the dataset with the occurrence of terms in the collection of reviews by using Filters->Unsupervised->Attribute->StringToWordVector

c) Apply a classifier (e.g. SMO – Support Vector Machine) to the dataset. Remark: it requires to set first the attribute called @@class@@ as class attribute (using Edit, right click on @@class@@ and selecting Attribute as class)

d) Analyze the impact of using a stemming step on the quality of the classification.