Programming Project 2

The aim of the work is to extract knowledge from collections of texts using association rules.

The data you will use corresponds to data from 20 Newsgroups (http://qwone.com/~jason/20Newsgroups/) which correspond to 20 categories (some related and some not related to each other), so that 6 more general categories are also formed (for details on the website).
Specifically, you will use the bydate.tar.gz http://qwone.com/~jason/20Newsgroups/20news- file for your data.

You will develop a program in any language you wish (but python is strongly recommended) that will implement the following phases:

- **Pretreatment:**in this phase you should edit your texts which belong to the training set, to assign to each article a set of important words.
  - Remove headings that do not contribute to the content, remove bullet points punctuation, common words without semantics, and do stemming if you consider it necessary or any other step you want.
  - Weight each word with a weight of your choice (frequency of occurrence,tf-idf, number of impressions, etc.)
  - Save as a documenttop-k of its important words according to the weighting you did
- **Association Rules:**Detect association rules in documents (sets of words) then the preprocessing given support and confidence threshold. Apply a priori considering each document as a transaction and each word as an item. That is, you will generate rules of type, eg: word1, word2→ word3.
- **Text Categorization:**Expand itdataset using also the class of the documents so that the class also belongs to the words in the document and find rules that associate the occurrence of words with the class they belong to. In the right part of these rules we are always interested in the category of the document.
- **Valuation:**Use the rules from the previous phase in the control set and measure system performance. Measure precision/recall per category and macro-average.

In the last two phases you can work with the 6 general categories if you want. You should experiment with different support and confidence values to get a result that satisfies you. It is also your choice how many words you keep per document. If you export multiple categorization rules you can select the best ones to use in valuation. It is your choice how you combine them if you wish. If you have an issue with the amount of data use sampling to reduce.

What you will deliver:

- your code
- a brief report (pdf) in which you will describe the parameters and choices you made for the implementation and you will present the results of the last three phases in any way you think is best (jupyter notebooks are also accepted).