CSI 5387
Fall 2012
Course project

Project reports (max. 10 pages) due by email to me Dec. 14, 2012.

The purpose of this course project is to get you hands-on experience with text mining, and specifically with text classification. You will be working with real-life data taken from a real-life application, i.e. so called Systematic Reviews in Medicine. There was an introduction to the problem and its setting in class. The two relevant papers are also available from the class website. As this is a project, and not an assignment, it may seem to you open-ended, i.e. not everything is given to you up front. You are expected to make some reasoned decisions about the representation, classifier, evaluation, and tools that you will use. As in a real data mining task, you will have to make the decisions concerning the data representation, data preprocessing, the modeling tools, the evaluation criteria, etc. You will be provided with basic tools in R to convert text data into a bag of words (or tf/idf) representation. It is very important that you start on this project as soon as possible, do not wait until close to the deadline. In particular, data preprocessing may require some time to complete, and you need to do this prior to using any classifier.

Your task in this project will be to learn, from the data, a classifier categorizing abstracts of medical papers in one of two classes: E and I. The data is given as in files Estrogens, OralHypoglemics, Triptans and BB.

Each instance consist of five parts: K, T, A, P, M.

K is class (E or I)
T is the title of the abstract
A is the text of the abstract
P is the Publication type
M is the set of MeSH (Medical Subject Headings) categories for this abstract

See papers [1] and [2] for further explanation of the data.

Note that as explained in class, the problem is imbalanced.

Your task will be to:

1. pre-process the data and convert it into the format you will use, either BOW or tf-idf. Work with datasets Estrogens, NSAIDs, and OralHypoglacemics. Please note that there are missing values in the data. Once you have obtained satisfactory results with those, you may try the BB data (BetaBlockers). This is a much larger data set.

You are welcome to apply any type of data preprocessing that may improve the performance. You may find TextDirectoryLoader converter and StringToWordVector filter in Weka useful.

2. using 5x2 cross-validation, apply more than one classifier on the data, e.g. some Bayesian classifier (e.g. MNB), decision tree, SVM, k-NN, etc. Use AUC as the comparison measure. You have to try several and decide which ones you will use for the final comparison and discussion. Then compare the results, using statistical significance.

3. discuss the results. The objective is to get a good AUC value.

4. paper by A. Cohen is posted under Project on the website. While I only give you 3 groups (reviews), try to beat his results on these groups. The best report that does that will be offered a 50% RA to complete the work on all 15 datasets. There is a strong possibility of publication from this work.

Your report will present the results and justify all the decisions you have made while getting these results. Be innovative in your solutions! You will be graded on the correctness of the methods, on your creativity, and on the quality of the results themselves.

[1] Cohen A. Optimizing feature representation for automated systematic review work prioritization. AMIA Annu Symp Proc 2008:121-5

[2] Matwin**,** Kouznetsov, A., **S.,** Inkpen, D., Frunza, O., O’Blenis, P. “Using Factorized Complement Naïve Bays and weight Engineering for Reducing Workload in Evidence-Based Medicine Systematic Reviews”, Journal of the American Medical Informatics Association, (JAMIA), Vol. 17, pp. 446-453, 2010