CS3243: ASSIGNMENT 3 (2009)
Text Categorization

Title: Text Categorization

Total Marks: 13

Due date: 13th April 2009 (by 1400 Hrs)

Tutor In-charge: Zheng Yan-Tao (yantaozheng@gmail.com)

Descriptions:

- To implement a system to perform text categorization. You are to implement at least 2 categorization methods: kNN and Decision Tree. For Decision Tree, you may down-load package from the Web (e.g C5.0) and adapt it for your purpose.
- You will be given a set of codes to perform feature extraction and compute document to document similarity. You will also be given a set of training documents in 10 categories (number of documents varying from 10 to 40), and a corresponding set of test documents.
- You are to perform feature selection, and classification using decision tree (with different tree pruning parameters) and kNN (with different k), and able to show your results interactively through a simple UI.
- You are to provide detailed comparison of results on the test set provided. You need to compare the performance of different variants of DT, kNN (with different k) and different features size.
- You are required to demonstrate the effectiveness of your system by performing "live" retrieval based on a new set of test documents provided.

Input:

You will be given: (a) simple program to extract feature and perform similarity matching written in Java; and, (b) a set of training and test documents in 10 categories. During evaluation, separate set of documents will be used to test your system.

Report: To submit:

a) Not more than 6-page report on the techniques and design, structure of program, and comparison/analysis of results of employing different variants of classifiers and feature set.
b) PPT file for <= 5 minute presentation
c) Source codes and executable codes.

Submit all online before the deadline, and submit (a) and (b) in hardcopy during lecture.

Remarks:

Flexibility and effectiveness of system is most important.

<u>**Late Submissions**</u>: Usual late submission rules apply.