Assignment 4

11/01/2017 - Due 20/01/2017 BIL 614 - Text Mining

Text Categorization using Label Propagation

In this assignment you will work on Text Categorization. The dataset you will be using is based on the Reuters dataset, where the corpora consists of news articles selected from 8 different categories.

Two files are provided. The r8-train-stemmed.txt file contains the text documents in each line, where the first word in the line is the assigned category to the document. There are 8 categories namely; acq, crude, earn, grain, interest, money-fx, ship, trade. The category is separated by a tab character and is located at the beginning of each line. Also using the same ordering with this file, sims.mat file contains the similarity matrix for the documents.

Building the Graph

A document pairwise similarity matrix is provided in sims.mat file. The cosine similarities between document pairs can be found in this file. For example, the 3rd number in first line is the similarity between document 3 and document 1 in the r8-train-stemmed.txt file.

Using this similarity matrix you should construct a graph, as we discussed in the lectures (you can use k most similar documents are neighbours).

Propagate Labels

Build a matrix for label probabilities. Choose a subset of documents from this file, for which you will provide the algorithm with their true categories. So, for example choose a 100 document and initialize their label probabilities by assigning probability 1 to its true category, and 0 assigned to other categories. Use the iterative label propagation algorithm until convergence. The final probabilities you learned can be used to assign categories to all the documents. Try to experiment with different number of known labels. Also note that if there is an imbalance between category labels, the results may be misleading. Think of this problem as if we start with more documents from category earn, more earn probability will be propagated to the rest of the graph. You can calculate accuracy of your label assignments.

Submission

Your assignment is due for 20/01 Friday. As usual you have 3 late submission days. You can upload all your work with a short Readme file explaining your program and how to run it. Please try to include short and clear instructions. Please report your experiments with different number of seed documents (the documents for which you know what their category is). Do not include the data files in your submission. You can upload your assignment to the link below. Upload your solution in a single zip file.

https://script.google.com/macros/s/AKfycbyP-qNzlBtVYTFlrysdXkhnMmrYikJVBRE5yyEOJMT4QqT03Snh/exec