CS1573, Spring 2011
Project 2
Clustering

The last project focused on features. This project will focus on variations in the method.

The task is to use hierarchical agglomerative clustering to cluster the blog data provided by Segaran.

Use whatever programming language you want to use. You can build from the code in the Segaran text or from the pseudo code in the Manning et al. text. Note that Figure 17.8 has pseudo code for a more efficient version of HAC using a priority queue.

The blogs should be represented as TF.IDF vectors.

In addition to implementing the basis system, You should do one interesting things, chosen from the following list:

- Experiment with different similarity/distance metrics (e.g., cosine similarity, Euclidean distance, Pearson correlation)
- Experiment with different similarity computations, i.e., single-link, complete-link, group-average and centroid similarity
- Experiment with the four methods on page 380 of Manning et al. for cutting the dendrogram.
- Experiment with bottom-up (HAC) versus top-down ("divisive"; see Section 17.6) clustering
- Implement two methods for labeling clusters (See Manning et al. Section 17.7)

1 Data

We will use half the blog data provided by Segaran. Alex has labeled the data using delicious tags. You may treat these as gold standard tags.

2 Your Evaluations

In this assignment, you will need to compare variations of metric, method, etc., as outlined above. In order to do this, you need a way to quantitatively evaluate the quality of induced clusters. A popular metric for cluster evaluation is Rand Index (RI). RI sees the clustering as a set of decisions whether an instance pair is in the same cluster or not and computes the percentage of correct decisions (Please see the chapter 16 for more details). You can accept the provided tags as gold-standard cluster labels, which should give you an opportunity to apply RI.

In the described setting, putting a politics blog in the same cluster as a technology blog is penalized by RI in the same way as putting a science blog in the same cluster as a technology blog. This is probably a rough approach since a science blog is more similar to a technology blog than it is to a politics blog. You should modify the RI metric in order to take into account similarities between tags to accomplish a finer evaluation. A straightforward way of doing that is considering a blog pair with different but similar tags in the same cluster as a partially correct decision, incrementing true positives (TP) by x (0¡x¡1). In addition, if a similar tagged blog pair are placed in different clusters, then they should count as a partially correct decision increasing true negatives (TN) by x (0¡x¡1) instead of 1.

3 Partnering

There are several aspects to the project. Meet with your partner early to split up your tasks and create a schedule. You will both need to stay on schedule so you do not hold each other up.

4 Grading

Grading:

- Fundamentals, the clustering algorithm with documents represented as TF.IDF features and your version of the Rand Index implementated: 45
- The interesting thing you will experiment with, from the above list: 30
- Your report: 25%

5 Report

Your report should describe what you implemented, and show the results of comparitive experiments.

Present your results in readable tables with informative captions, and discuss your results in the text.

Describe your ideas for what you would try next, if you had more time to work on the project.

6 Submission Instructions

To submit the project, please email the TA an archive containing all of your source code and report and README file before the deadline. The README file should specify how to execute your code.

Please send the archive to alexander.p.conrad@gmail.com.