**Assignment 5: 5243: Intro to Data Mining**

Your objective in this lab is to evaluate the efficacy and efficiency of minwise hashing for document similarity. You will use the Reuters dataset. You will begin with the feature vector you created in Assignment 1 (feel free to adapt or develop alternatives if you wish to start from scratch). For example some of you used n-grams or 3-word shingles in your feature vector (results on these may yield better qualitative results as we discussed in the lecture).

**Creating a Baseline:** For the baseline you use the raw feature vectors and for each pair (of documents) report the exact Jaccard similarity. We will call this the true similarity baseline.

**Creating a k-minhash sketch:** You will use any publicly available tool or create your own minhash sketch for each document from its corresponding feature vector following the procedure described in the lecture (Shingling is optional and will net bonus points but does add complexity to the approach). You will then use this k-minhash sketch (varying k between 16, 32, 64, 128, and 256) and report the estimated Jaccard similarity between every pair of documents. We will call this k-minhash estimate of similarity.

Your report should compare both strategies along the axes of efficiency (time to generate estimate from sketch) and efficacy for different values of k. For efficacy or quality you may choose to report mean-squared error (between the estimate and the true similarity) or relative mean error (normalized by the true similarity value). You are also expected to plot or graph these ideas in your report to facilitate comparisons across different sketch sizes. Your report will carry 25% of the grade for this assignment as well as the next assignment so spend time on articulating the main observations.

You can again choose to implement your version of algorithms or download and use free software from kdnuggets.com. We can also provide the source code for BayesLSH (an algorithm from which one can extract the relevant codebase for MinHash, Approximate Linear Permutations etc). Please contact one of the TAs for this source base (I recommend you use this option but you are welcome to explore others. Please note you do not have to run BayesLSH but rather you can extract the relevant codebase for your assignment.) As always you may learn more if you try to implement by yourself. Additionally if you use software from somewhere else please specify and provide credit in your report. Your report should describe any further data transformations you may have had to make to work with these software packages. The report should not exceed 5 pages but at the same time should list all underlying assumptions, and explanations for performance obtained is expected. Use the submit command (**lab 5**) to submit all source files (README, source, makefile, test files, data, report etc.) as before.

**Due date: November 20th 2015 11:59PM. [Gives you ample time to sleep in and cheer the Buckeyes on the 21rst as they go up against the Spartans!]**

**As always possible solutions will be discussed in class.**