CS 378 **Introduction to Data Mining** **Spring 2009**

Homework 1

Instructor: Inderjit Dhillon Date Due: Feb 5th, 2009

Keywords: *Linear Algebra, Matlab, Vector Space Model*

Here is a collection of documents ($d = 10$), where the terms used in the analysis are underlined ($w = 11$):

c_1**:** Indian government goes for open-source software

c_2**:** Debian 3.0 Woody released

c_3**:** Wine 2.0 released with fixes for Gentoo 1.4 and Debian 3.0

c_4**:** gnuPOD released iPod on Linux... with GPLed software

c_5**:** Gentoo servers running an open-source mySQL database

m_6**:** Dolly the sheep not totally identical clone

m_7**:** DNA news: introduced low-cost human genome DNA chip

m_8**:** Malaria-parasite genome database on the Web

m_9**:** UK sets up genome bank to protect rare sheep breeds

m_{10}**:** Dolly's DNA Damaged

Answer the following questions based on the above data:

1. Transform the data into a term-document matrix A (an 11×10 matrix in this case) in the *Vector Space Model*, where each document vector is normalized to have unit L2-norm.
2. Compute the cosine similarity between each pair of documents, i.e., compute $A^T A$.
3. Compute the *Singular Value Decomposition* of matrix A. (Use the Matlab command svd.)
4. Plot the first two left and right singular vectors respectively. (Use the Matlab command plot.)
5. Plot the projected document vectors in the space spanned by the first two left singular vectors.
6. Plot the projected term vectors in the space spanned by the first two right singular vectors.