

Homework 2

Lecturer: Prof. Pradeep Ravikumar

Date Due: Mar 03, 2014

Keywords: *Linear Algebra, Document Analysis*

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

c_1 : Apple has released its new software suite for its major computers
 c_2 : An operating system is an example of computer software
 c_3 : Computer virus nearly knocks out business
 c_4 : Windows 7 system released: an operating system for PC
 c_5 : The newly released McAfee can scan virus in Windows 7 system
 m_6 : IMF: The recovery of US economy could take a while
 m_7 : UK economy exited recession in fourth quarter
 m_8 : Growth in British economy suggests end to recession
 m_9 : US economy experiences fastest growth in six years.
 m_{10} : Yahoo: the fastest growth IT company in Internet business

Answer the following questions using **Matlab** based on the above data and report your results.

1. Data transformation and normalization:

- (1 point) Transform the data into a term-document matrix A (an 11×10 matrix in this case) in the Vector Space Model, where A_{ij} is the number of occurrences of term i in document j .
- (1 point) Normalize matrix A to get matrix B , where each column (document) vector of B has unit ℓ_2 -norm.

2. (1 point) Compute the cosine similarity between each pair of documents, i.e., compute $B^T B$.

3. (1 point) Compute the Singular Value Decomposition of matrix B , and report the left/right singular vectors and singular values. (Use the **Matlab** command `svd`.)

4. (2 points) Plot the first two left and right singular vectors respectively. (Use the **Matlab** command `plot`.)

5. (2 points) Plot the projected document vectors in the space spanned by the first two left singular vectors.

6. (2 points) Plot the projected term vectors in the space spanned by the first two right singular vectors.