

INFS4203/INFS7203 Data Mining (2009)
Assignment 3 – Text Mining

The goal of this individual written assignment is to understand VSM and its related concepts. This assignment counts for 5% of the total unit assessment. Submit either a hardcopy or a soft copy of your assignment through email to shenht@itee.uq.edu.au before or during the Tutorial time 2pm on Friday 9 Oct. 09. Show details of your answers and calculations step-by-step, otherwise your marks will be deducted.

Problem

Given a query of “multimedia retrieval” and a collection of the following five documents:

- Document 1: <Dimension Specific Search for Multimedia Retrieval>
- Document 2: <Bounded Coordinate System Indexing for Video Search>
- Document 3: <Speed Up Interactive Multimedia Retrieval>
- Document 4: <Similarity Search in Large Multimedia Database>
- Document 5: <Dual Dimensionality Reduction for Efficient Video Similarity Search>

Use the Vector Space Model, TF/IDF weighting scheme, and cosine vector similarity to find the top-2 most relevant documents to the query.