

## **INFS4203/INFS7203 Data Mining (2011)**

### **Assignment 4 – Web Mining & Text Mining**

The goal of this individual written assignment is to explore the area of web mining and text mining. This assignment counts for 5% of the total unit assessment.

The due date of this assignment is Thursday of Week 13 (27 Oct, 2011). You can submit either in hardcopy (during the lecture class), or send it in email to us before the deadline (Please title your email as: INFS4203/INFS7203 Assignment-4 Submission). No late submission will be accepted in normal situation.

#### **Question 1**

Given a query of “Gold Silver Truck” and the following three different documents:

Document 1: < Shipment, Gold, Damaged, Fire >,

Document 2: < Delivery, Silver, Arrived, Silver, Truck >,

Document 3: < Shipment, Gold, Arrived, Truck >.

Use the Vector Space Model, TF/IDF weighting scheme, and Cosine vector similarity measure to find the most relevant document(s) to the query.

**Q1.1:** Calculate DF (document frequency) and IDF (inverse document frequency) for each word.

<b>Word List</b>	<b>DF</b>	<b>IDF</b>

**Q1.2:** Represent each document as a weighted vector by using TF/IDF weight scheme.

**Q1.3:** Represent the query as a weighted vector and find its most relevant document(s) using Cosine Similarity measure.

**Question 2**

**Q2.1** Web mining is one of the major data mining techniques to discover patterns from the Web. Please list three key tasks of web mining.

**Q2.2** Make a brief literature review of two web structural mining methods: HITS and PageRank, and answer the following questions:

1. Briefly describe the fundamental ideas of both HITS and PageRank.
2. Respectively list some applications of HITS and PageRank.