1 Assignment 1: Information Retrieval

1. In class, we briefly discussed pre-processing techniques such as stemming, stop-word removal and thesaurus construction. Given a text document, suggest any three additional pre-processing techniques that may be used. Explain the approach and outline the potential benefit of the approach. **(10 marks)**

2. Given the following small sample document collection:

   (a) **D1:** Shipment of gold damaged in a fire

   (b) **D2:** Delivery of silver arrived in a silver truck

   (c) **D3:** Shipment of gold arrived in a truck

   Calculate the term weightings for terms in D1. Show your workings and state any assumptions you make. **(10 marks)**

3. In class we discussed the document collection as term-document matrix, where each cell in the matrix indicates the usefulness of term i in describing document j. We also discussed how we could evaluate the similarity of a query and document.

   Outline a suitable indexing structure to store the information in the matrix (note that matrix is sparse). **(10 marks)**

   Outline at a high level, in pseudo-code, an algorithm to calculate the similarity of a document to a query. **(10 marks)**