1 Assignment 1: Information Retrieval

1. Question 1 (15 marks)

   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. **(5 marks)**

2. Question 2 (10 marks)

   With respect to D1 = {Shipment of gold damage in a fire} and a query = {gold silver truck}, consider how the similarity sim(Q, D1), *should* change for each of the following augmentations to D1.

   (a) D1 = Shipment of gold damaged in a fire. Fire.

   (b) D1 = Shipment of gold damaged in a fire. Fire. Fire.

   (c) D1 = Shipment of gold damaged in a fire. Gold.

   (d) D1 = Shipment of gold damaged in a fire. Gold. Gold.

   Note, there is no need to show any calculations; the question pertains to how the similarity should change.

3. Question 3 (10 marks)

   In the term weighting schemes covered in class thus far, we have considered the tf factor, the idf factor and normalisation approaches.

   Assuming that your document collection consists of all the scientific articles published in the Communications of the ACM (www.acm.org/dl), identify two other sources of evidence (features or sets of features) one could consider and suggest a weighting scheme that incorporates these features.

   Your answer should define the evidence/feature, your reason for including it, and a means to include it in the weighting scheme.