Assignment 1

Colm O'Riordan

October 2020

1 Assignment 1: Information Retrieval

1. Question 1 (10 marks)

   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

   and the following query:

   Q: gold silver truck

   Using a suitable weighting scheme (e.g. the basic tf-idf one covered in class will suffice), rank the documents D1, D2, D3 with respect to the query.

   Showing how the calculations will proceed is sufficient. If you wish you may write code (or use Excel). The emphasis is showing an understanding of the process.

   State any assumptions you make with respect to term weighting and normalisation.

2. Question 2 (10 marks)

   With respect to D1 above, 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.

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