Due: 2022/3/18

Homework 4

Student Number:
Name:

Problem 1. (20 points) Use variable byte codes to encode the posting list of the term $COMPUTER$ in page 9 of the slides.

Problem 2. (30 points) Consider the table of term frequencies for 3 documents denoted $Doc1$, $Doc2$, $Doc3$ in Table 1(a). Compute the tf-idf weights for the terms *car*, *auto*, *insurance*, *best*, for each document, using the idf values from Table 1.

Table 1: Problem 1

(a) Term Frequency

	Doc1	Doc2	Doc3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

(b) IDF

term	df_t	idf_t
car	18165	1.65
auto	6723	2.08
insurance	19241	1.62
best	25235	1.5

Problem 3. (20 points) Refer to the tf-idf weights computed in Problem 1. Compute the Euclidean normalized document vectors for each of the documents, where each vector has four components, one for each of the four terms.

Problem 4. (30 points) One measure of the similarity of two vectors is the *Euclidean distance* (or L_2 distance) between them:

$$|\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^{M}(x_i - y_i)^2} \tag{1}$$

Given a query q and documents $d_1, d_2, ...,$, we may rank the documents d_i in order of increasing Euclidean distance from q. Show that if q and the d_i are all normalized to unit vectors, then the rank ordering produced by Euclidean distance is identical to that produced by cosine similarities.