Due: 2018/3/30

Homework 4

Problem 1. (20 points) From the following sequence of γ-coded gaps, reconstruct first the gap sequence and then the postings sequence: 1110001110101011111101101111011.

Solution:

postings sequence	Gap	Length	Offset	γ-code
9	1001=9	1110	001	1110001
15	110=6	110	10	11010
18	11=3	10	1	101
77	111011=59	111110	11011	11111011011
84	111=7	110	11	11011

Problem 2. (30 points)

Table 1: Problem2

	query					document			
word	tf	wf	df	idf	$q_i = wf - idf$	tf	wf	$d_i = normalized\_wf$	$q_i \cdot d_i$
digital			10,000						
video			100,000						
cameras			50,000						

Compute the vector space similarity between the query "digital cameras" and the document "digital cameras and video cameras" by filling out the empty columns in Table 1. Assume $N = 10,000,000$, logarithmic term weighting (wf columns) for query and document, *idf* weighting for the query only and cosine normalization for the document only. Treat *and* as a stop word. Enter term counts in the tf columns. What is the final similarity score?

solution:

Word	Query					document			qi*di
	tf	wf	df	idf	qi=wf-idf	tf	wf	di=normalized wf	
digital	1	1	10,000	3	3	1	1	0.52	1.56
video	0	0	100,000	2	0	1	1	0.52	0
Cameras	1	1	50,000	2.3	2.3	2	1.3	0.68	1.56

Similarity score = 1.56+1.56 = 3.12

Table 2: Problem 3

(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. (30 points) Consider the table of term frequencies for 3 documents denoted $Doc1$, $Doc2$, $Doc3$ in Table 2(a).

a. Compute the tf-idf weights for the terms *car*, *auto*, *insurance*, *best*, for each document, using the *idf* values from Table 2.

b. Compute the Euclidean normalized document vectors for each of the documents, where each vector has four components, one for each of the four terms.

c. Compute the consine similarity between any two of the documents.

d. Compute the two top scoring documents on the query *best car insurance* for each of the following weighing schemes:

   i nnn.atc

   ii ntc.atc

solution:
Consider the table of term frequencies for 3 documents denoted $Doc1$, $Doc2$, $Doc3$ in Table 2(a).

a. Compute the $tf-idf$ weights for the terms *car*, *auto*, *insurance*, *best*, for each document, using the *idf* values from Table 2.

	Doc1	Doc2	Doc3
car	44.55	6.6	39.6
auto	6.24	68.64	0
insurance	0	53.46	46.98
best	21	0	25.5

b. Compute the Euclidean normalized document vectors for each of the documents, where each vector has four components, one for each of the four terms.

$doc1 = [0.8974, 0.1257, 0, 0.4230]$
$doc2 = [0.0756, 0.7867, 0.6127, 0]$
$oc3 = [0.5953, 0, 0.7062, 0.3833]$

c. Compute the cosine similarity between any two of the documents.

Solution:

$Doc1.Doc2 = 204$
$|V(Doc1)||V(Doc2)| = 1,431.5013098143$
$cosinesimilarity = \frac{\vec{V}(Doc1)\times\vec{V}(Doc2)}{|\vec{V}(Doc1)|\times|\vec{V}(Doc2)|} = \frac{204}{1,431.5013098143} = 0.1425077285$

d. Compute the two top scoring documents on the query *best car insurance* for each of the following weighing schemes:

i nnn.atc

ii ntc.atc

1- nnn.atc

the weights of nnn for the documents

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

	query				product		
term	tf (augmented)	idf	tf-idf	atc weight	Doc1	Doc2	Doc3
car	1	1.65	1.65	0.56	15.12	2.24	13.44
auto	0.5	2.08	1.04	0.353	1.06	11.65	0
insurance	1	1.62	1.62	0.55	0	18.15	15.95
best	1	1.5	1.5	0.51	7.14	0	8.67

$Score(Q, doc1) = 15.12 + 1.06 + 0 + 7.14 = 23.32, score(Q, doc2) = 2.24 + 11.65 + 18.15 + 0 = 32.04, score(Q, doc3) = 13.44 + 0 + 15.95 + 8.67 = 38.06$

Ranking: doc3, doc2, doc1

2- ntc.atc

the weight of ntc for Doc1

Term	tf (augmented)	idf	tf-idf	Normalized weights
car	27	1.65	44.55	0.897
auto	3	2.08	6.24	0.125
insurance	0	1.62	0	0
best	14	1.5	21	0.423

the weight of ntc for Doc2

Term	tf (augmented)	idf	tf-idf	Normalized weights
car	4	1.65	6.6	0.075
auto	33	2.08	68.64	0.786
insurance	33	1.62	53.46	0.613
best	0	1.5	0	0

the weight of ntc for Doc3

Term	tf (augmented)	idf	tf-idf	Normalized weights
car	24	1.65	39.6	0.595
auto	0	2.08	0	0
insurance	29	1.62	46.98	0.706
best	117	1.5	25.5	0.383

	query				product		
term	tf (augmented)	idf	tf-idf	atc weight	Doc1	Doc2	Doc3
car	1	1.65	1.65	0.56	0.502	0.042	0.33
auto	0.5	2.08	1.04	0.353	0.044	0.277	0
insurance	1	1.62	1.62	0.55	0	0.337	0.38
best	1	1.5	1.5	0.51	0.216	0	0.19

$Score(Q, doc1) = 0.762, score(Q, doc2) = 0.657, score(Q, doc3) = 0.916$

Ranking: doc3, doc1, doc2

Problem 4. (20 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 $d_1, d_2, \ldots,$, 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.

solution

$\sum(q_i - w_i)^2 = \sum q_i^2 - 2\sum q_i w_i + \sum w_i^2 = 1 - 2\sum q_i w_i + 1 = 2(1 - \sum q_i w_i)$

(Note that for a normalized vector $\vec{x}$, we have: $\sum x_i^2 = 1$.)

Thus: $|\vec{q} - \vec{v}| < |\vec{q} - \vec{w}| \Leftrightarrow |\vec{q} - \vec{v}|^2 < |\vec{q} - \vec{w}|^2 \Leftrightarrow \sum(q_i - v_i)^2 < \sum(q_i - w_i)^2 \Leftrightarrow 2(1 - \sum q_i v_i) < 2(1 - \sum q_i w_i) \Leftrightarrow \sum q_i v_i > \sum q_i w_i \Leftrightarrow \cos(\vec{q}, \vec{v}) > \cos(\vec{q}, \vec{w})$

This proves that ordering normalized vectors according to increasing distance is the same as ordering them according to decreasing cosine similarity.