Scores:

1)	5)
2)	6)
3)	7)
4)	8)

CSIT 6000I Search Engines and Applications
Mid term Examination, Fall 2018
Nov 3, 2018
Time Allowed: 80 min

Name: ______________________________ Student ID: ________________________

Note: Answer all questions in the space provided. Answers must be precise and to the point.

1. **[5]** Consider the query: hong kong universities, given to a search engine implementing the plain vector space model, explain the impact on result ranking if the search engine treats hong kong as a phrase representing a single concept or as two individual words.

   If hong kong are not treated as a phrase, the query has three terms, and documents matching hong kong or hong kong university will add up to higher scores. If hong kong is a phrase, documents matching hong kong is considered having one hit instead of two hits, resulting in lower scores.

2. **[5]** In the vector space model, (a) when stemming has been applied to document terms, explain if it is necessary to apply stemming to the query terms. (b) When stopword removal has been applied to document terms, is stopword removal required on the query terms.

   (a) Yes, it is necessary to stem the query terms; otherwise, for the same word, the stemmed and un-stemmed versions will not match.
   (b) No, it is desirable but not a requirement. The same documents will be retrieved, albert with different scores. For example, if the query is science and technology, and the documents have all “and” removed, documents containing “science” and “technology” will still match the query in the vector space model. The tricky part is the word “required”.

3. **[5]** Circle all of the factors below that increase the PageRank of a page when:

   [✓ (a) The number of the page’s parents increases]
   [✓ (b) The number of the page’s grandparents increases]
   [✓ (c) The PageRank of the parents increases]
   (d) The number of the page’s children increases
   (e) The number of children of the page’s parents increases

4. **[5]** Give one reason for why using links to infer the similarity between two linked pages cannot improve web search quality.

   Any one is fine:

   For web search, there are too many truly relevant documents for a query, making the similarity more accurate does not help to distinguish between the truly relevant documents. We need a new dimension (in the case of PageRank, it is authority).

   Links, at best, only indicate the fact that two pages are related but not necessarily similar.

5. **[5]** Give one reason why Clever is inferior to Google PageRank.

   Any one is fine:

Clever computes authority and hub weights in query processing time, whereas Google precomputes the global, static PageRank for all pages and thus have fast query processing.

Clever computes authority and hub weights only on a small subgraph, so the weights computed are not as representative as if all pages related to the queries are used.

6. **[15]** Suppose there are only 5 unique terms (numbered 1 to 5) in the collection, which contains a total of 10 documents. These five term's term frequencies in a document *D* and their document frequencies are given below:

$tf_{D,1} = 2$	$df_1 = 1$	idf=log$_2$(10/1)=3.32
$tf_{D,2} = 0$	$df_2 = 2$	idf=log2(10/2)=2.32
$tf_{D,3} = 1$	$df_3 = 3$	idf=log2(10/3)=1.74
$tf_{D,4} = 5$	$df_4 = 2$	idf=log2(10/2)=2.32
$tf_{D,5} = 2$	$df_5 = 10$	idf=log2(10/10)=0

(a) [5] Write down the document vector when tf/tf$_{max}$ * idf weighting is used.

t_1 = 2/5 * 3.32 = 1.33
t_2 = 0/5 * 2.32 = 0
t_3 = 1/5 * 1.74 = 0.35
t_4 = 5/5 * 2.32 = 2.32
t_5 = 2/5 * 0 = 0

(b) [5] Given the query vector, $Q = \langle 1, 0, 0, 1, 0 \rangle$, compute the cosine similarity values between *Q* and *D*.

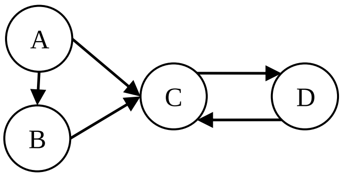
inner product = 1.33 + 2.32 = 3.65
|Q| = sqrt(2) = 1.414
|D|= sqrt(1.33^2 + 0.35^2 + 2.32^2) = sqrt(7.27) = 2.70
Cosine(Q,D) = 3.65 / (1.414 * 2.70) = 0.956

(c) [5] Explain why the normalization factor in cosine similarity is expensive to compute.

Because the length of each document vector must be computed in cosine similarity and it cannot be pre-computed due to the constant changing of the IDF of a term.

7. **[20]** Given the following Web graph, initial PR values (iteration 0) for all pages are 1/4, and damping factor d=0.8,

(a) **[10]** compute the PR values of the pages in the following table, normalize the PR values with L1 norm in each iteration



Initial value as 1/4

Iteration	0	1	2	3
PageRank(A)	1/4	0.125	0.125	0.125
PageRank(B)	1/4	0.1875	0.15625	0.15625
PageRank(C)	1/4	0.4375	0.375	0.40625

PageRank(D)	1/4	0.25	0.34375	0.3125

Initial value as 1

Iteration	**0**	**1**	**2**	**3**
PageRank(A)	1	0.05	0.125	0.12
PageRank(B)	1	0.15	0.1375	0.15
PageRank(C)	1	0.55	0.3375	0.45
PageRank(D)	1	0.25	0.4	0.28

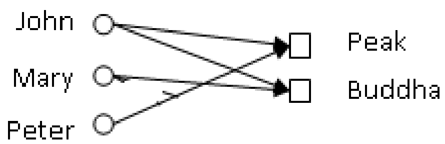
(b) **[5]** Would the PR values converge? If so, comment on the final PR values of A, B, C, and D.

Yes, it will converge. PR(A) and PR(B) will always be 0.2. PR(C) and PR(D) will converge because of the damper factor. PR(C)>PR(D).

(c) **[5**] If d=1, would the PR values converge? If so, comment on the final PR values of A, B, C, and D.
Yes, it will converge. If the damping factor is 1, PR(A) and PR(B) will always be 0. PR(C) will inherit all of PR(D) and PR(D) will inherit all of PR(C). The PR values will converge.
** No, it will not converge. Although PR(A) and PR(B) are zero, but the PR(C) and PR(D) will inherit each other's PR value indefinitely (their PR values swap indefinitely) and hence not converging.

Iteration	0	1	2	3	4	
A	0.25	0	0	0	0	
B	0.25	0.125	0	0	0	
C	0.25	0.625	0.375	0.625	0.375	...
D	0.25	0.25	0.625	0.375	0.625	

8. **[20]** The bipartite graph represents the places visited by a traveler.



A link indicates that the person has visited a place.
(a) [15] Compute the hub and authority weights of the nodes. In the 0^{th} iteration, all weights are 1. Perform two iterations. Normalization is not needed

Iteration	**0**	**1**	**2**
PageRank(John)	1	0, 2	0, 4
PageRank(Mary)	1	0, 1	0, 2
PageRank(Peter)	1	0, 1	0, 2

PageRank(Peak)	1	2, 0	3, 0
PageRank(Buddha)	1	2, 0	3, 0

(b) [5] How would you interpret the meanings of the hub/authority weights of the nodes?

Authority of users and hub of all places are zero. Hub weights of users indicate whether the users have visited many good places, whereas Authority weights of places indicate how popular they are.

9. **[10]** In Latent Semantic Indexing, we decompose a term-document matrix into U∑V^T.
(a) [5] Using the following example, which is reproduced from the lecture notes, explain how to do a rank-2 approximation.

$$A = U\Sigma V^T$$

$$U = \begin{pmatrix} 0.6977 & 0.0931 & -0.0175 & 0.6951 & 0 & -0.0157 & -0.1441 & 0 & 0 \\ 0.2619 & -0.2966 & -0.4681 & -0.1969 & 0 & 0.2468 & 0.1570 & -0.6356 & 0.3099 \\ 0.3527 & 0.4491 & 0.1017 & -0.4013 & -0.7071 & 0.0066 & 0.0493 & 0 & 0 \\ 0.1121 & -0.1410 & 0.1478 & 0.0733 & 0 & -0.4842 & 0.8402 & 0 & 0 \\ 0.2619 & -0.2966 & -0.4681 & -0.1969 & 0 & 0.2468 & 0.1570 & 0.6356 & -0.3099 \\ 0.1874 & -0.3747 & 0.5049 & -0.1270 & 0 & 0.2287 & -0.0338 & -0.3099 & -0.6356 \\ 0.3527 & 0.4491 & 0.1017 & -0.4013 & 0.7071 & 0.0066 & 0.0493 & 0 & 0 \\ 0.2104 & -0.3337 & -0.0954 & -0.2820 & 0 & -0.7340 & -0.4657 & 0 & 0 \\ 0.1874 & -0.3747 & 0.5049 & -0.1270 & 0 & 0.2287 & -0.0338 & 0.3099 & 0.6356 \end{pmatrix}$$

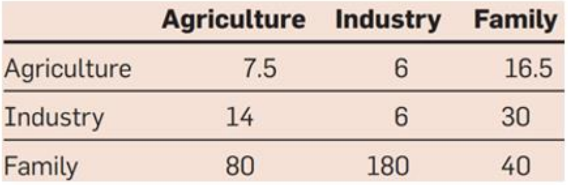
$$\Sigma = \begin{pmatrix} 1.5777 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1.2664 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.1890 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.7962 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.7071 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5664 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.1968 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$V = \begin{pmatrix} 0.1680 & -0.4184 & 0.6005 & -0.2256 & 0 & 0.5710 & -0.2432 \\ 0.4471 & -0.2280 & -0.4631 & 0.2185 & 0 & 0.4872 & 0.4986 \\ 0.2687 & -0.4226 & -0.5009 & -0.4900 & 0 & -0.2451 & -0.4451 \\ 0.3954 & -0.3994 & 0.3929 & 0.1305 & 0 & -0.6132 & 0.3697 \\ 0.4708 & 0.3028 & 0.0501 & 0.2609 & 0.7071 & -0.0113 & -0.3405 \\ 0.3162 & 0.5015 & 0.1210 & -0.7128 & 0 & 0.0166 & 0.3542 \\ 0.4708 & 0.3028 & 0.0501 & 0.2609 & -0.7071 & -0.0113 & -0.3405 \end{pmatrix}$$

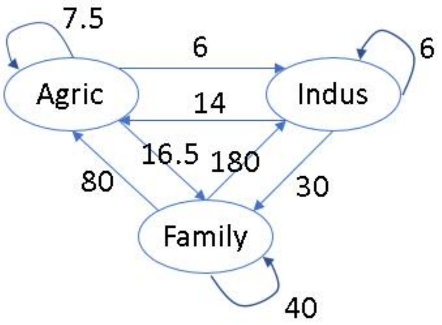
Make all diagonal elements beyond the first two zero, and multiple Usigma V^T

(b) [5] This question is independent of the example in (a). In a rank-2 approximation of the original term-document matrix, which of the following statement(s) are(is) true (circle all of them):
(i) There are two major clusters of documents in the document set
(ii) There are two major clusters of terms in the document set
(iii) There are two latent features in the document set

10. **[10]** In the econometric example discussed in class, which is repeated below. An entry in the table indicates the quantity an economic sector produces for another sector (which could be itself). When PageRank is computed on the graph and converges, what does the PageRank values mean? Justify your answer.



	Agriculture	Industry	Family
Agriculture	7.5	6	16.5
Industry	14	6	30
Family	80	180	40



The PageRank obtained by an economic sector is the price it can charge on its output. It is given in the paper. When converges, the total cost of the inputs (i.e., quantity * price) is the same as the value of the output.