INFORMATION RETRIEVAL – SHORT EXERCISES I – BOOLEAN RETRIEVAL AND NAVIGATIONAL PATTERNS

I. Consider the following documents **D1-D4** using 8 different terms:

D1 = {breakthrough drug schizophrenia}

D2 = {new schizophrenia drug}

D3 = {new approach treatment schizophrenia}

D4 = {new hope schizophrenia patient}

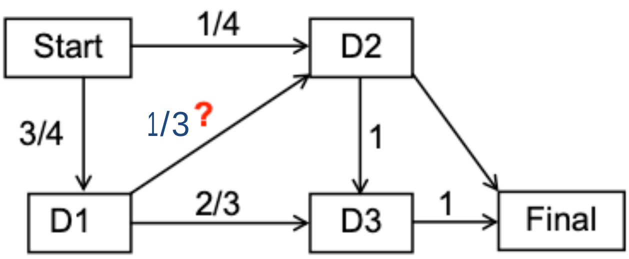
Fill in the term-document incidence matrix for this document collection.

	D1	D2	D3	D4
approach	0	0	1	0
breakthrough	1	0	0	0
drug	1	1	0	0
hope	0	0	0	1
new	0	1	1	1
patient	0	0	0	1
schizophrenia	1	1	1	1
treatment	0	0	1	0

What are the results returned for the below Boolean queries:

- schizophrenia AND drug Answer: 1111 AND 1100 = 1100
- new AND NOT(drug OR approach) Answer: 0111 AND NOT(1100 OR 0010) =
  0111 AND NOT 1110 = 0111 AND 0001 = 0001

II. Given the following four sessions: {D1 D2 D3}, {D1 D3}, {D1 D3}, {D2 D3}, answer the questions related to using the Markov chain for mining navigational patterns.



What is P(D1→D2)? Answer: 1 - 2/3 = 1/3

What is the probability of P(Start→D1→D3)? Answer: 3/4 * 2/3 = 1/2

What is the probability of P(D3|D1)? Answer: 2/3 + 1/3 * 1 = 1

I. Consider a set of terms **T** = {t_1, t_2, t_3, t_4} and the following collection of two documents: **D1** = {t_1 t_2 t_1 t_2 t_3} and **D2** = {t_4 t_2 t_2 t_3}. Consider query **Q** = {t_1 t_4}. Represent D1, D2, and Q using TF (normalized Bag-Of-Words).

TF	t_1	t_2	t_3	t_4	max
D1	2/2	2/2	1/2	0	2
D2	0	2/2	1/2	1/2	2
Q	1	0	0	1	1

Compute IDFs for all four terms (note that only D1 and D2 are included in the collection).

	t_1	t_2	t_3	t_4	N
IDF	log2	log1=0	log1=0	log 2	2

II. Consider the below term-document matrix **C** for the bag-of-words representation of five documents **D1**-**D5** in the space of six terms t_1-t_6. Using the SVD factorization method, matrix **C** has been decomposed into matrices **K**, **S**, and D^T given below. The rank of **C** is 4 (4 ≤ min{6,5}), so 4 concepts (semantic dimensions) were discovered.

C =

	D1	D2	D3	D4	D5
t_1	5	5	0	0	1
t_2	4	5	1	1	0
t_3	5	4	1	1	0
t_4	0	0	4	4	4
t_5	0	0	5	5	5
t_6	1	1	4	4	4

terms -> concepts

K =

t1	-0.27	0.55	-0.78	0
t2	-0.29	0.47	0.44	-0.71
t3	-0.29	0.47	0.44	0.71
t4	-0.45	-0.29	-0.01	0
t5	-0.56	-0.36	-0.02	0
t6	-0.50	-0.18	-0.05	0

concept space

S =

13.74	0	0	0
0	10.88	0	0
0	0	1.36	0
0	0	0	1

docs -> concepts

D^T =

D1	D2	D3	D4	D5
-0.32	-0.32	-0.52	-0.52	-0.5
0.63	0.63	-0.25	-0.25	-0.29
-0.02	-0.02	0.41	0.41	-0.82
0.71	-0.71	0	0	0

Answer the following questions:

- What is the informativeness value of the most important concept? Answer: 13.74
- Based on the informativeness values of all concepts, which seems the most obvious value for the reduced number of dimensions k? Answer: k = 2
- ? not sure What is the (numerical value of the) mapping of term t_6 to the most important (informative) concept? Answer: 0 (i.e., the 4th concept) -0.5
- What is the vector representing document D3 in the space of four discovered concepts? Answer: [-0.52, -0.25 , 0.41 , 0]

INFORMATION RETRIEVAL – SHORT EXERCISES III – EVALUATION IN INFORMATION RETRIEVAL AND PAGERANK

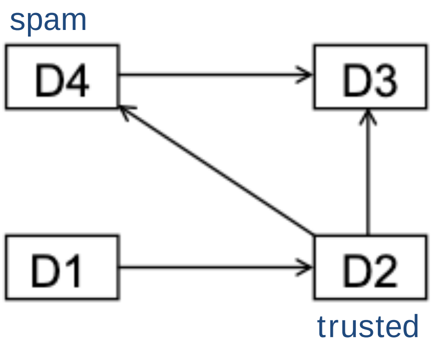
I. Consider an information need for which there are 4 relevant documents in the collection. A system run on this collection returned the top 10 results for which the relevance is judged as follows (R – relevant; N – non-relevant):

R N R N N N N N R R

What is the recall at 6 (R@6)? Answer: (1/4) * (1 + 0 + 1 + 0 + 0 + 0) = 2/4 = 1/2

What is the Mean Average Precision? Answer: (1/4) * (P@1 + P@3 + P@9 + P@10) =
= (1/4) * ((1/1) * 1 + (1/3) * 2 + (1/9) * 3 + (1/10) * 4) = 3/5

II.Consider the web graph presented below to the left. It involves four pages D1-D4 and four links. Fill in the stochastic matrix M given to the right.



	D1	D2	D3	D4	
	0	0	0	0	D1
	1	0	0	0	D2
	0	1 1/2	0	1	D3
	0	1 1/2	0	0	D4

Write the equation for PR(D3) without dumping factor *q*? Answer: PR(D3) = 0*PR(D1) + 1/2 1*PR(D2) + 0*PR(D3) + 1*PR(D4)

Which page has the greatest PageRank (without computing the exact PR values)? Answer: D3

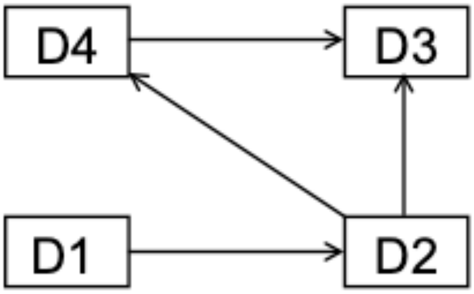
An oracle has evaluated D2 as trusted and D4 as spam. What is the starting vector d for TrustRank?

Answer: d = [0 , 1 , 0 , 0]

(Of course, apart from d = [0, 0, 0, 0])

INFORMATION RETRIEVAL – SHORT EXERCISES IV – HITS, RELEVANCE FEEDBACK AND SPELLING CORRECTION

I. Consider the web graph presented below to the left. It involves four pages D1-D4 and four links. Fill in the adjacency matrix L given to the right.



0	1	0	0
0	0	1	1
0	0	0	0
0	0	1	0

The principal eigenvector of LL^T is [0, 1.618, 0, 1] and the principal eigenvector of L^TL is [0, 0, 1.618, 1].

What is $h(D_4)$? Answer: …1.
(not normalized)

max

The page with the greatest authority score is: D3

II. Compute the Levenshtein distance for ”LEGIA” and ”LECHIA”.

		L	E	C	H	I	A
	0	1	2	3	4	5	6
L	1	0	1	2	3	4	5
E	2	1	0	1	2	3	4
G	3	2	1	1	2	3	4
I	4	3	2	2	2	2	3
A	5	4	3	3	3	3	**2**

Information Retrieval – Short Exercises V – Collaborative Filtering and ADWORDS

I. Given the below user-item rating matrix, predict rating of user U7 for item I4:

	I1	**I2**	I3	**I4**	*sim*(U7,U·)	Average
U1	5	**4**	4	**4**	0.0	(5 + 4 + 4) / 3 = 4.3(3)
U2	5	**3**	7	**3**	1.0	5
U3	4	**3**	2	**3**	-0.5	3
U4	6	**4**	5	**4**	0.5	5
U5	3	**4**	2	**4**	-1.0	3
U6	4	**3**	5	**3**	1.0	4
U7	4	**3**	5	**?**		4

a) Employ user-based CF with k=2 and either simple average or weighted average?

Answer: U7(I4) = $(3 + 3) / 2 = 3$

b) Employ user-based CF with k=2 and modify U7's average rating by the weighted modification of its nearest neighbors averages:

Answer: U7(I4) = $4 + \frac{1.0 * (3 - 4) + 1.0 * (3 - 5)}{1.0 + 1.0} = 4 - 1.5 = 2.5$

c) Which item should be analyzed to predict the rating when using item-based CF with k=1? What would be the predicted rating?

Answer: item - I2 and prediction – 3

II. Four advertisers A, B, C, and D with a daily budget of \$2 bid for the following keywords (\$1 each): A: w, x; B: x, z; C: x, y; D: y, z. Use a simplified version of BALANCE to select the ads for the following query stream (in the case of a tie use the following order for breaking it A > B > C > D):

query stream	x	y	w	z	z	w	y	x
BALANCE	A	C	A	**?** B	**?** D	**?** -	**?** C	**?** C

B

I. Consider the following fragment of a term-based positional index in the format:

term: doc1: <position1,position2,…>; doc2: <position1,…>; etc.

Gates: 1: <3>; 2: <6>; 3: <2,17>; 4: <1>;

IBM: 4: <3>; 7: <14>;

Microsoft: 1: <1>; 2: <1,21>; 3: <3>; 5: <16,22,51>;

The **/k** operator, **word1 /k word2** finds occurrences of word1 within k words of word2 (on either side), where **k** is a positive integer argument. Which document(s) satisfy the query "**Gates /2 Microsoft**"?

Answer: 1: [3]   3: [2]
1: [1]   3: [3]

II. Build a suffix array for "couscous$" using the *qsufsort* algorithm.

	i	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
h	x_i	c	o	u	s	c	o	u	s	$
	A[i]	9	1	5	2	6	4	8	3	7
	V[A[i]]	1	3	3	5	5	7	7	9	9
1	V[A[i]+h]		5	5	9	9	3	1	7	7
	A[i]	9	1	5	2	6	8	4	3	7
	V[A[i]]	1	3	3	5	5	6	7	9	9
2	V[A[i]+h]		9	9	7	6			3	1
	A[i]	9	1	5	6	2	8	4	7	3
	V[A[i]]	1	3	3	4	5	6	7	8	9
4	V[A[i]+h]		3	1						
	A[i]	9	5	1	6	2	8	4	7	3

III. Encode 15 in γ. Answer: 0001111

IV. Decode 00111000001 written in the δ-code. Answer: N+1 = 00111 = 7 => N = 6
Answer: 2^6 + 1 = 65