

### Problem 1

Let vocabulary size = 10. Assuming a Zipfian distribution, what is the frequency of the most frequent word in the vocabulary?

Ans:

Let vocabulary size = 10. Assuming a Zipfian distribution, what is the frequency of the most frequent word in the vocabulary?

$$1 / (1 + 1/2 + 1/3 + 1/4 + 1/5 + 1/6 + 1/7 + 1/8 + 1/9 + 1/10)$$
$$= 34\%$$

### Problem 2

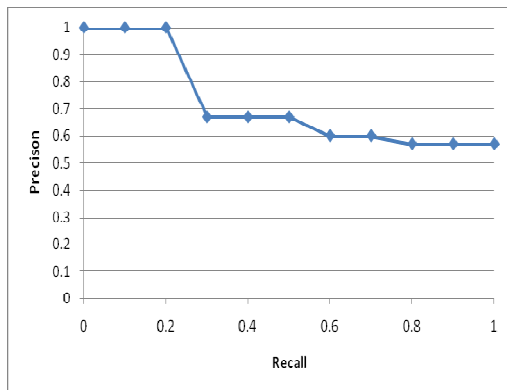
The following table shows the output of a search engine on two queries. Calculate 11-pt avg. interpolated precision for these two queries, and sketch it in a precision-recall graph. Crosses correspond to relevant documents, dashes to irrelevant documents.

Rank	Q1	Q2
1	X	-
2	-	X
3	X	-
4	-	X
5	X	X
6	-	-
7	X	X
8	-	-
9	-	X
10	-	-

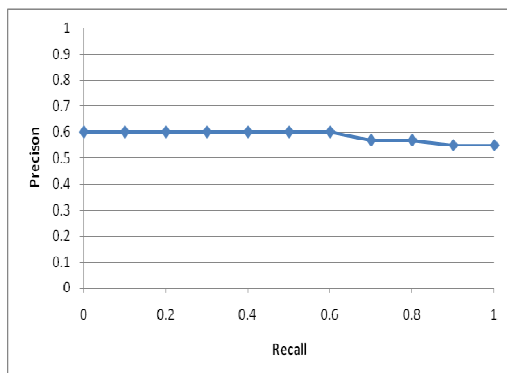
Ans:

11-pt avg. interpolated precision: 0.72

11-pt avg. interpolated precision: 0.59



11-pt avg. interpolated precision: 0.72



11-pt avg. interpolated precision: 0.59

### Problem 3

Consider the document below:

*“A federal judge is evaluating the age of a suspected pirate before deciding whether the suspect's court proceedings should be made public. The suspect, who has not been identified, was arrested in the hijacking of the Maersk Alabama. He arrived in New York late Monday.”*

Using a linear classifier with the following weight vector for class C.

$$W(\text{judge}) = 0.9$$

$$W(\text{pirate}) = 0.8$$

$$W(\text{suspect}) = 0.6$$

$$W(\text{court}) = 0.5$$

$$W(\text{arrested}) = 0.5$$

$$W(\text{hijacking}) = 0.4$$

$$W(\text{New York}) = -1.0$$

$$W(\text{Monday}) = -0.5$$

$$W(\text{public}) = -0.9$$

Determine if this document belongs to the class C given the data above and assuming the prior belief that the document is in class C has a weight of -1. Show all your work.

**Ans:**

$$0.9 + 0.8 + 0.6 + 0.5 + 0.5 + 0.4 - 1.0 - 0.5 - 0.9 - 1.0 = 1.3$$

Or we can count the number of occurrences and multiply that by the weight

We can also use stemming

In all cases  $\rightarrow$  document belongs to the class C

#### Problem 4

Margaret is a new user of WebMovies (a web site that streams movies on demand). So far, she has seen and rated two movies (on a scale from 1 to 5).

*Score (Margaret, Braveheart) = 4*

*Score (Margaret, Halloween) = 2*

What is the next movie that WebMovies will recommend to her given the following scores in its database:

*Score (Andy, Braveheart) = 5*

*Score (Belinda, Braveheart) = 2*

*Score (Charlie, Braveheart) = 1*

*Score (Dana, Braveheart) = 4*

*Score (Andy, Halloween) = 3*

*Score (Belinda, Halloween) = 5*

*Score (Charlie, Halloween) = 1*

*Score (Ellen, Halloween) = 5*

*Score (Andy, Gladiator) = 4*

*Score (Belinda, Gladiator) = 1*

*Score (Ellen, Gladiator) = 5*

*Score (Belinda, Scream) = 4*

*Score (Charlie, Scream) = 1*

*Score (Dana, Scream) = 2*

*Score (Andy, Rainman) = 5*

*Score (Belinda, Rainman) = 4*

*Score (Charlie, Rainman) = 3*

*Score (Dana, Rainman) = 5*

*Score (Ellen, Rainman) = 2*

Make sure that you describe your algorithm very carefully: first in general and then, using the specific data set above.

**Ans:**

One method: represent each user by a vector holding the rating they gave to each movie and find the user most similar (i.e. closest by some measure, cosine similarity, Euclidean distance, etc.) to Margaret. Return the highest-rated movie by the most similar user that is unseen by the customer.

<Braveheart Halloween Gladiator Scream Rainman>

Andy is the closest user to Margaret and the movie that he recommends is Rainman.

### Problem 5

Describe three techniques for building a collection of stop words.

Ans:

### Problem 6

Consider the following output of the Porter stemmer.

*a british consortium pledg tuesdai to spend up to \$14.5 million in research grant to find out what i caus  
a seriou declin in bee and other pollin insect.*

*those insect includ honei bee bumbl bee butterfli and moth plai an essenti role in pollin mani vital crop  
but their number have been declin steadili in recent year scientist sai.*

*in the unit kingdom alon the number of pollin ha fallen between and percent in the past two year accord  
to the biotechnologi and biolog scienc research council a government-sponsored research group.*

Indicate which words were improperly stemmed. Why? How would you have stemmed them?

**Ans:** Examples of words that were improperly stemmed:

*tuesdai* → *Tuesday*

*is* → *is*

*caus* → *cause*

*seriou* → *serious*

*declin* → *decline*

*honei* → *honey*

...

*And several others*

**Problem 7**

What is the Levenshtein edit distance between these two words:

GOOGLE  $\leftrightarrow$  YAHOO

Show all your work. What is the edit sequence that transforms one of the words to the other?

Ans:

		Y A H O O				
	0	1	2	3	4	5
G	1	1	2	3	4	5
O	2	2	2	3	3	4
O	3	3	3	3	3	3
G	4	4	4	4	4	4
L	5	5	5	5	5	5
E	6	6	6	6	6	6

		G O O G L E					
	0	1	2	3	4	5	6
Y	1	1	2	3	4	5	6
A	2	2	2	3	4	5	6
H	3	3	3	3	4	5	6
O	4	4	3	3	4	5	6
O	5	5	4	3	4	5	6

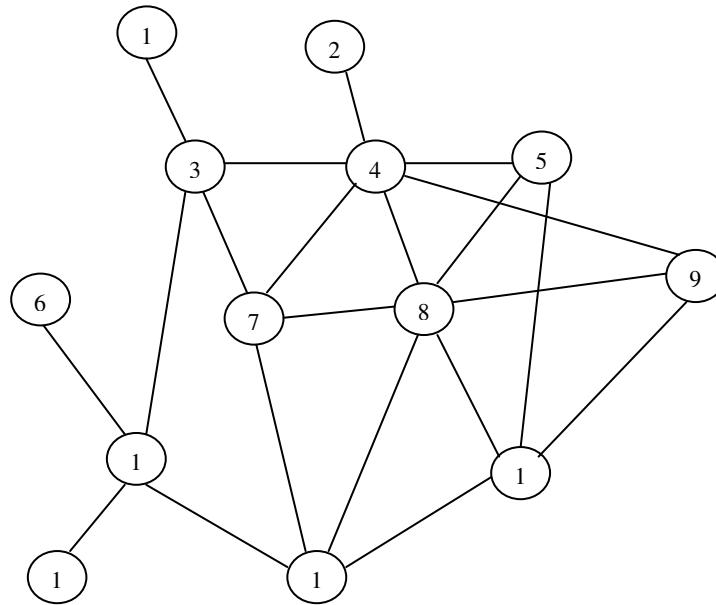
**Problem 8**

Give 8 criteria that can be combined to produce a ranking formula for documents given a query Q. Assume a web-based, hyperlinked collection of documents.

Ans:

**Problem 9**

What is the clustering coefficient in the graph in the following figure? Use any of the formulas used in class.

**Ans:**

Watts -Strogatz

Node	C
1	0
2	0
3	0.166667
4	0.266667
5	0.666667
6	0
7	0.5
8	0.466667
9	0.666667
10	0
11	0.5
12	0
13	0.333333
avg	0.274359

Newman clustering coefficient:  $8 * 3 / 66 = 0.3636$

Triangles (8 triangles, 66 triples):

13 5 8    13 8 9    4 5 8    12 13 8    12 7 8    3 4 7

### Problem 10

Find all paths of length 2 in the previous graph using matrix multiplication. Show the two input matrices and the output matrix.

$M = [$

	1	2	3	4	5	6	7	8	9	10	11	12	13
1			1										
2				1									
3	1			1			1			1			
4		1	1		1		1	1	1				
5				1				1			1		
6										1			
7			1	1				1					1
8				1	1		1		1		1		1
9				1				1			1		
10			1			1						1	1
11					1			1	1				1
12										1			
13							1	1		1	1		

$]$

$MM = [$

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1	1	0	0	1	0	0	1	0	0	1	0	0
2	0	0	1	1	0	1	0	1	1	1	0	0	0
3	0	0	1	4	1	1	1	1	2	1	0	0	1
4	1	0	1	6	1	0	2	3	1	1	3	0	2
5	0	0	1	1	1	3	0	2	2	3	0	1	0
6	0	0	0	1	0	0	1	0	0	0	0	0	1
7	1	1	1	1	2	2	0	4	2	2	2	2	0
8	0	0	1	2	3	2	0	2	6	2	1	3	0
9	0	0	1	1	1	3	0	2	2	3	0	1	0
10	1	0	0	1	0	0	2	1	0	4	1	0	0
11	0	0	0	3	1	0	2	3	1	1	4	0	1
12	0	0	0	1	0	0	1	0	0	0	0	0	1
13	0	0	0	2	2	2	1	1	2	2	0	1	1

Any non-negative term indicates a path of length 2.

**Problem 11**

How does pseudo relevance feedback work? Give an example with specific sentences (as documents).

**Ans:**

**Problem 12**

Consider a random walk on a small web graph with  $A \rightarrow B$ ,  $A \rightarrow C$ ,  $A \rightarrow D$ ,  $B \rightarrow C$ ,  $D \rightarrow C$ ,  $D \rightarrow B$ ,  $C \rightarrow A$ . If after  $k$  iterations, the probability distribution over the four nodes is  $(A,B,C,D) = (0.25,0.125,0.125,0.5)$ , what is the probability distribution of the random walk after  $k+1$  iterations?

**Ans:**

[0.25 0.125 0.125 0.5]

\*

0	1/3	1/3	1/3
0	0	1	0
1	0	0	0
0	1/2	1/2	0

=

0.125 0.3333 0.458 0.0833

**Problem 13**

Assuming Heaps' Law with parameters:  $k=50$  and  $b = 0.5$ .

Consider a blog corpus with 10,000 archived postings and 1,000 newly received postings. How many new words are likely to appear in the newly received part compared to the archive?

**Ans** – 244 (Around 250)

**Problem 14**

A query session has resulted in 10 false positives, 4 false negatives, 20 true positives, and 96 true negatives. What is the value of the F-measure for this session, assuming that Precision and Recall are equally important.

Ans:

$$20/27=0.74$$

**Problem 15**

A cat is chasing a mouse in a maze. The maze consists of 5 rooms, connected with doors as follows: 1-2, 2-3, 3-4, 1-4, 4-5. The mouse is in room 1, the cat -- in room 3. A piece of cheese is located in room 5. The mouse walks randomly from room to room. The cat is too lazy to chase the mouse so he stays in his room all the time. If the mouse gets to the cheese, the game is over and the mouse wins. If the mouse gets to the room with the cat, the game is also over and the win is given to the cat. What is the probability that the mouse will win this game?

Ans:

$$2/7$$

**Problem 16**

Consider the following sentences (some of which are manually translated from Chinese, others are automatically translated).

The case for sure will go to the US Supreme Court.

The ruling is likely to be appealed in U.S. Supreme court.

It will almost be turned to the US Supreme Court.

There is no doubt that the ruling will be appealed to US Supreme Court.

For preprocessing, use punctuation removal but **no** stemming and **no** lowercasing. Then, using the Jaccard coefficient, build a dendrogram of these four sentences. For the merge operation, use set union (that is, the merged output of "A B C" and "A C D" is the union of these two sets, or "A B C D").

Show the corresponding Venn diagram to be used in hierarchical clustering.

HINT:

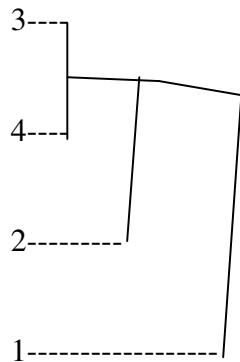
Let's consider the following simpler example:

D N T R  
A B C C S  
A D E F

D E F G

We build a 4x4 matrix of similarities. The last two items are the most similar so we merge them into one using set union: A D E F G. Then we build a new 3x3 matrix, etc.

BONUS QUESTION (+3 points). Can you think of a different similarity metric that would result in a different Venn diagram?



Bonus – If cosine similarity or weighted n-grams is considered and it is shown that the results are actually different, full marks are awarded.