

---

## IR&TM Probeklausur

Note: The real exam on Feb 7 will be somewhat longer. Don't worry if you are not able to complete it. This is not required for getting the top grade.

### 0.0.1 Question – 2 points

Why don't we use a relational database for information retrieval?

**Answer:** Relational databases doesn't support ranking.

### 0.0.2 Question – 2 points

Complex Boolean retrieval systems like Westlaw use many operations that go beyond strictly Boolean operators. Name some of them.

**Answer:** within n words/sentence/paragraph proximity, truncation, phrases

### 0.0.3 Question – 2 points

What is stemming? Give an example that is not also a lemmatization example.

**Answer:** Reducing a word to a "stem" (a distinctive prefix) by heuristic rules such as "increment → incr"

### 0.0.4 Question – 2 points

Name a data structure that supports proximity queries.

**Answer:** positional inverted index

### 0.0.5 Question – 2 points

What is the feast or famine problem?

**Answer:** In Boolean search it's often difficult to write a query that produces a small number of results. Often you get back either nothing or a large number of hits.

### 0.0.6 Question – 2 points

Define the tf-idf vector of a document as the vector whose entry on dimension  $i$  is the tf-idf weight of term/word  $t_i$  in the document.

Why don't we use Euclidean distance of tf-idf vectors to rank documents with respect to a query?

**Answer:** Because document length strongly influences such a ranking and length should not be considered (or only very slightly) in relevance ranking

### 0.0.7 Question – 2 points

What is term-at-a-time processing?

**Answer:** The postings list of one query term is completely processed before processing of the postings list of the next query term is started.

**0.0.8 Question – 2 points**

What is an easy way of maximizing the recall of an IR engine?

**Answer:**

return the entire collection to the user

**0.0.9 Question – 2 points**

What distinguishes a dynamic from a static summary?

**Answer:**

The dynamic summary of document  $d$  is computed for a particular query  $q$ . For two different queries, the dynamic summaries of  $d$  will in general be different. The static summary is always the same.

**0.0.10 Question – 2 points**

What is the time complexity of training a Naive Bayes classifier and why?

**Answer:**

linear in the size of the training set; counts can be computed while scanning the corpus. the actual computation of the parameters is also linear since it is bounded by the number of classes times the size of the training set

**0.0.11 Problem – 10 points**

Given: a collection of 100 documents. The following documents are relevant to query  $q$ :  $d_3$ ,  $d_8$ ,  $d_{12}$ ,  $d_{25}$ ,  $d_{31}$ ,  $d_{72}$ ,  $d_{80}$ . A retrieval system returns the following ranked result:

1.  $d_2$
2.  $d_{80}$
3.  $d_{72}$
4.  $d_{30}$
5.  $d_{24}$
6.  $d_8$
7.  $d_{12}$
8.  $d_{25}$
9.  $d_3$
10.  $d_{31}$

(i) Compute precision, recall and  $F_1$  at rank 5. (ii) Compute precision, recall and  $F_1$  at rank 10.

**Answer:**

(i)  $P = 2/5$ ,  $R = 2/7$ ,  $1/F = 1/2(5/2+7/2) = 3$ ,  $F = 1/3$

(ii)  $P = 7/10$ ,  $R = 1.0$ ,  $1/F = 1/2(10/7+1) = 1/2*17/7 = 17/14$ ,  $F = 14/17$

### 0.0.12 Problem – 10 points

Suppose that a user's initial query is "tunisia hotel". The user examines three documents. She judges  $d_1$  with the content "tunisia hotel tunis hotel" and  $d_2$  with the content "tunisia hotel tunis hotel bizerte beach" relevant and  $d_3$  with content "tunisia unrest march" nonrelevant. Assume that we are using direct term frequency (with no scaling and no document frequency). There is no need to length-normalize vectors. Using Rocchio relevance feedback what would the revised query vector be after relevance feedback? Assume  $\alpha = 1, \beta = 0.5, \gamma = 0.1$ .

word	$q$	$d_1$	$d_2$	$\mu_R$	$d_3$	$\alpha q$	$\beta \mu_R$	$\gamma d_3$	rocchio
tunisia	1	1	1	1	1	1	0.5	0.1	1.4
hotel	1	2	2	2		1	1		2
tunis		1	1	1			0.5		0.5
bizerte			1	0.5			0.25		0.25
beach			1	0.5			0.25		0.25
unrest					1			0.1	0
march					1			0.1	0

### 0.0.13 Problem – 10 points

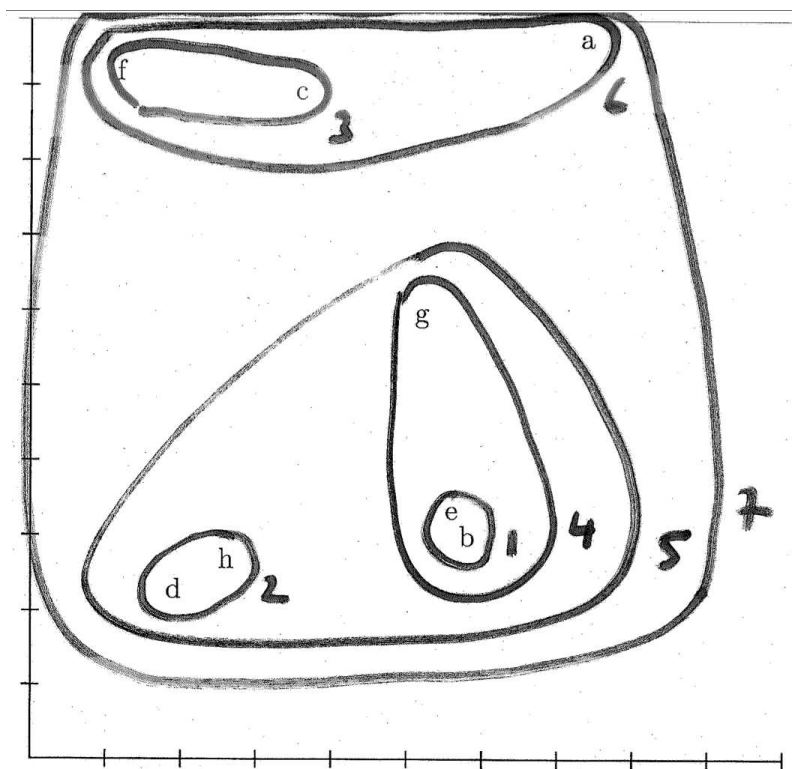
Compute (i) single-link and (ii) complete-link clusterings of this set of points. You can do this visually by drawing circles/ellipses. Mark each circle/ellipsis with a number indicating the temporal sequence of mergers.

- a 7.4 9.6
- b 5.8 3.0
- c 3.6 8.9
- d 1.9 2.3
- e 5.6 3.3
- f 1.2 9.2
- g 5.2 5.9
- h 2.6 2.7

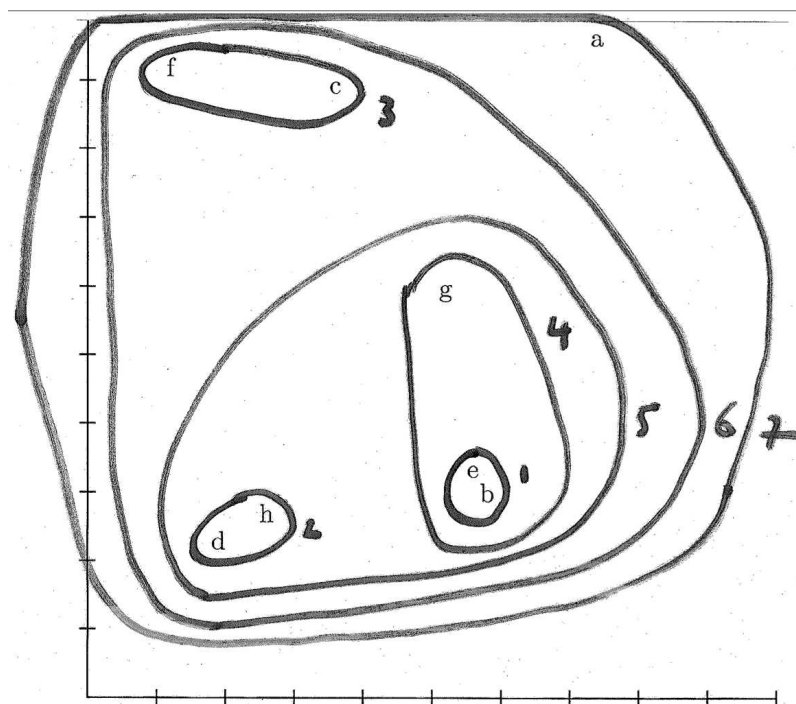
**Answer:**

0.36-e-b 0.81-h-d 2.42-f-c 2.63-g-e 2.96-g-b 3.06-h-e 3.21-h-b 3.40-g-c 3.83-d-e 3.86-c-a 3.96-d-b 4.12-h-g  
4.30-g-a 4.88-g-d 5.19-f-g 5.95-e-c 6.21-f-a 6.28-h-c 6.30-b-c 6.55-e-a 6.65-h-f 6.79-b-a 6.82-d-c 6.94-f-d  
7.36-f-e 7.72-f-b 8.41-h-a 9.14-d-a

6.79-a-b 3.86-a-c 9.14-a-d 6.55-a-e 6.21-a-f 4.30-a-g 8.41-a-h 6.79-b-a 6.30-b-c 3.96-b-d 0.36-b-e 7.72-b-f  
2.96-b-g 3.21-b-h 3.86-c-a 6.30-c-b 6.82-c-d 5.95-c-e 2.42-c-f 3.40-c-g 6.28-c-h 9.14-d-a 3.96-d-b 6.82-d-c  
3.83-d-e 6.94-d-f 4.88-d-g 0.81-d-h 6.55-e-a 0.36-e-b 5.95-e-c 3.83-e-d 7.36-e-f 2.63-e-g 3.06-e-h 6.21-f-a  
7.72-f-b 2.42-f-c 6.94-f-d 7.36-f-e 5.19-f-g 6.65-f-h 4.30-g-a 2.96-g-b 3.40-g-c 4.88-g-d 2.63-g-e 5.19-g-f  
4.12-g-h 8.41-h-a 3.21-h-b 6.28-h-c 0.81-h-d 3.06-h-e 6.65-h-f 4.12-h-g



single-link clustering:



complete-link clustering:

#### 0.0.14 Problem – 15 points

Compute PageRank for the web graph in Figure 1 for each of the four pages. Also give the relative ordering of the 4 nodes indicating any ties.

Assume that at each step of the PageRank random walk, we teleport to a random page with probability 0.1, with a uniform distribution over which particular page we teleport to.

Using symmetries to simplify and solving with linear equations might be easier than using iterative methods.

**Answer:**

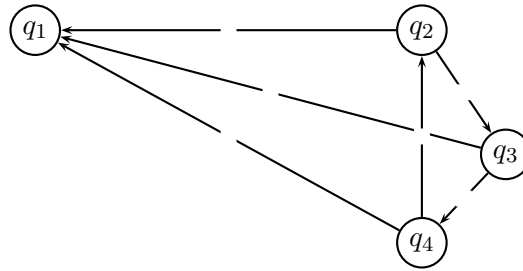


Abbildung 1: A web graph

transition probability from q2 to q1/q3:  $0.45 + 1/40 = 18/40 + 1/40 = 19/40$

	q1	q2	q3	q4
q1	1/4	1/4	1/4	1/4
q2	19/40	1/40	19/40	1/40
q3	19/40	1/40	1/40	19/40
q4	19/40	19/40	1/40	1/40

$$q_2 = 0.45 * q_2 + 1/40 * 3 * q_2 + 1/4 * q_1$$

$$q_1 + 3 * q_2 = 1$$

$$q_2 = 0.45q_2 + 1/40 * 3 * q_2 + 1/4 * (1 - 3 * q_2)$$

$$q_2(1 - 0.45 - 3/40 + 3/4) = 1/4$$

$$q_2(40/40 - 18/40 - 3/40 + 30/40) = 1/4$$

$$q_2(49/40) = 1/4$$

$$q_2 = 40/49 * 1/4 = 10/49$$

$$1 - 3 * 10/49 = 19/49$$

$$q_1 = 19/49, q_2 = q_3 = q_4 = 10/49$$

initial probs 0.4 0.2

0 0.385 0.205

1 0.388375 0.203875

2 0.387615625 0.204128125

3 0.387786484375 0.204071171875

4 0.387748041016 0.204083986328

5 0.387756690771 0.204081103076

6 0.387754744576 0.204081751808

7 0.38775518247 0.204081605843

8 0.387755083944 0.204081638685

9 0.387755106113 0.204081631296

code:

```

q1 = 0.4
q2 = 0.2
for n in range(10):
    if n==0:
        print 'initial probs',q1,q2
  
```

```

tmp = (q2+q2+q2)*(0.45+1/40.)+1/4.*q1
q2 = (0.45+3*1/40.)*q2+1/4.*q1
q1 = tmp
print n,q1,q2

```

### 0.0.15 Problem – 10 points

The shingle representations of two documents are as follows:  $d_1 = (0, 0, 1, 0)^T$ ,  $d_2 = (1, 1, 1, 0)^T$ .

We will use sketches of size 2. The two elements of a sketch are defined by the permutations:  $(3 * n + 1) \bmod 4$  and  $(5 * n + 1) \bmod 4$ . Based on this setup what is the estimate of the Jaccard coefficient  $J(d_1, d_2)$ ?

**Answer:**

(permutation 1)

```

doing [0, 0, 1, 0]
pi(i) at 1 is 0
shingle at dim 1 does not occur
pi(i) at 2 is 3
shingle at dim 2 does not occur
pi(i) at 3 is 2
min at dim 3 is 2
pi(i) at 4 is 1
shingle at dim 4 does not occur
min for doc [0, 0, 1, 0] is 2

```

```

doing [1, 1, 1, 0]
pi(i) at 1 is 0
min at dim 1 is 0
pi(i) at 2 is 3
min at dim 2 is 0
pi(i) at 3 is 2
min at dim 3 is 0
pi(i) at 4 is 1
shingle at dim 4 does not occur
min for doc [1, 1, 1, 0] is 0

```

(permutation 2)

```

doing [0, 0, 1, 0]
pi(i) at 1 is 2
shingle at dim 1 does not occur
pi(i) at 2 is 3
shingle at dim 2 does not occur
pi(i) at 3 is 0
min at dim 3 is 0
pi(i) at 4 is 1
shingle at dim 4 does not occur
min for doc [0, 0, 1, 0] is 0

```

```

doing [1, 1, 1, 0]
pi(i) at 1 is 2

```

min at dim 1 is 2  
 pi(i) at 2 is 3  
 min at dim 2 is 2  
 pi(i) at 3 is 0  
 min at dim 3 is 0  
 pi(i) at 4 is 1  
 shingle at dim 4 does not occur  
 min for doc [1, 1, 1, 0] is 0

[0, 0, 1, 0], permutation 1:

1	-	$\infty$
2	-	$\infty$
3	2	2
4	-	2

[1, 1, 1, 0], permutation 1:

1	0	0
2	...	
3	...	
4	...	

[0, 0, 1, 0], permutation 2: 0

[1, 1, 1, 0], permutation 2:

1	2	2
2	3	2
3	0	0
4	-	0

estimated Jaccard:  $1/2 = 0.5$