

---

## IR&TM Review I

### Part 1: Questions

#### Chapter 1

##### Question

Why don't we use grep for information retrieval?

##### Question

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

##### Question

Google does not always interpret the query as a boolean conjunction of its terms. Give examples.

##### Question

What is a term-document incidence matrix?

##### Question

In constructing the index, which step is most expensive/complex?

##### Question

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

#### Chapter 2

##### Question

Define the number of types/tokens in a sentence.

##### Question

An IR system can normalize terms by defining equivalence classes. E.g., "suit" and "suits" could be in an equivalence class. What is the limitation of this model in IR?

##### Question

What is tokenization?

##### Question

Give an example in English where tokenization is nontrivial

**Question**

Give an example in German where tokenization is nontrivial

**Question**

What is a stop list?

**Question**

What is lemmatization? Give an example.

**Question**

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

**Question**

Name a particular stemmer.

**Question**

Give an example of a pair of words that a typical stemmer would put in one equivalence class and we would expect improved performance of the IR system.

**Question**

Give an example of a pair of words that a typical stemmer would put in one equivalence class and we would expect decreased performance of the IR system.

**Question**

Name two data structures that support phrase queries.

**Question**

Name a data structure that supports proximity queries.

**Chapter 3****Question**

Which data structures are typically used for locating the entry for a term in the dictionary?

**Question**

Which data structure is best used for locating the entry for a term in the dictionary if the collection is static?

**Question**

Which data structure is best used for locating the entry for a term in the dictionary if prefix search must be supported?

**Question**

Which special strings are stored in the permuterm index for the word “car”?

**Question**

What sequence of letters is looked up in the permuterm index for the following wildcard queries?  
X, X\*, \*X, \*X\*, X\*Y

**Question**

What is the difference between the regular inverted index used in IR and the k-gram index?

**Question**

Give an example of a query that cannot be corrected using isolated-word spelling correction.

**Question**

Define Levenshtein edit distance.

**Question**

Define Damerau-Levenshtein edit distance.

**Chapter 6****Question**

What is the feast or famine problem?

**Question**

Define the Jaccard coefficient

**Question**

What is the bag of words model?

**Question**

What is the advantage of idf weighting compared to inverse-collection-frequency weighting?

**Question**

What is the tf-idf weight of term  $t$  in document  $d$ ?

**Question**

What is the relationship between term frequency and collection frequency?

**Question**

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

**Question**

Write down the formula for cosine similarity between query  $q$  and document  $d$ .

**Question**

Explain the notation  $ddd.qqq$

**Chapter 7****Question**

What is the advantage of pivot normalization compared to regular cosine normalization?

**Question**

What is document-at-a-time processing?

**Question**

What index organization does document-at-a-time processing require?

**Question**

What is term-at-a-time processing?

**Question**

What data structure does term-at-a-time processing require that document-at-a-time processing does not require?

**Question**

What is a tiered inverted index?

**Question**

Name two criteria that can be used for deciding as to whether to put a document  $d$  in tier 1 of a tiered index.

**Chapter 8****Question**

Name three criteria for evaluating a search engine.

**Question**

What are the components of an information retrieval benchmark?

**Question**

What is the difference between the concepts of query and information need?

**Question**

Define precision

**Question**

Define recall

**Question**

Define  $F_1$

**Question**

What is the harmonic mean of two numbers?

**Question**

Why is  $F_1$  defined as the harmonic mean?

**Question**

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

**Question**

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

**Question**

What is a precision-recall curve?

**Question**

An evaluation benchmark ideally should tell us for any document-query pair whether the document is relevant to the query. Why is Cranfield the only collection that actually satisfies this desideratum?

**Question**

Define the kappa measure

**Question**

What is the minimum and maximum of the kappa measure?

**Question**

What is the significance of kappa being less than / greater than 0?

**Question**

What is A/B testing?

**Question**

What does marginal relevance measure?

**Question**

What distinguishes a dynamic from a static summary?

**Question**

What is a simple heuristic for computing a dynamic summary if you can display  $n$  characters?

**Chapter 9****Question**

What is the difference between adhoc retrieval and relevance feedback?

**Question**

Give the mathematical definition of the centroid

**Question**

In Rocchio's algorithm, what weight setting for  $\alpha/\beta/\gamma$  does a 'Find pages like this one' search correspond to?

**Question**

Why is relevance feedback not used by most search engines?

**Question**

What is the difference between relevance feedback and manual query expansion?

**Question**

Give an example of ineffective automatic query expansion

**Question**

Search engines log the sequence of queries that a user issues during a session. How can this be exploited for query expansion?

**Question**

Search engines log the documents users click on in response to a query. How can this be exploited for query expansion?

**Chapter 13****Question**

What is the machine learning approach to text classification?

**Question**

Give one advantage and one disadvantage of rule-based classifiers compared to machine-learned classifiers.

**Question**

What is bad about maximum likelihood estimates of the parameters  $P(t|c)$  in Naive Bayes.

**Question**

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

**Question**

What is the main independence assumption of Naive Bayes.

**Question**

What is feature selection?

**Question**

What is feature selection used? Give the two main reasons?

**Question**

In words: what is the meaning of mutual information when used for features selection in text classification?

**Chapter 12****Question**

What is the basic idea of the language model approach to adhoc IR?

**Question**

How is length normalization performed in the vector space model vs the language model approach to IR?

**Question**

What is the number of classes in the Naive Bayes approach to classification? What is the number of classes in the language model approach to IR?

**Chapter 15-1****Question**

What is the definition of a linear classifier?

**Question**

For linearly separable problem, how many different linear decision boundaries are there that separate the two classes of the training set perfectly?

**Question**

Which decision boundary does the linear SVM choose?

**Question**

What is a support vector?

**Question**

How does an SVM classify a test set point in the margin

## Chapter 16

### Question

What is the difference between classification and clustering?

### Question

Why is result set clustering useful?

### Question

What is hard/soft clustering?

### Question

Does K-means always converge and why?

### Question

Does K-means find the global optimum and why?

## Chapter 17

### Question

Name three properties of a dendrogram

### Question

What two different types of hierarchical clustering were introduced in class?

### Question

What is the difference single-link and complete-link HAC?

### Question

What is chaining?

### Question

In general, is single-link or complete-link better to use and why?

## Chapter 21

### Question

When using PageRank for ranking what assumptions are we making about the meaning of hyperlinks?

**Question**

What is a Google bomb? Give an example

**Question**

Why is PageRank a better measure of quality than a simple count of inlinks?

**Question**

What is the meaning of the PageRank  $q$  of a page  $d$  in the random surfer model?

**Question**

What is ergodicity and why is it important for PageRank?

**Part 2: Exercises****Chapter 2****Exercise**

Shown below is a portion of a positional index in the format: term: doc1: ⟨position1, position2, ...⟩; doc2: ⟨position1, position2, ...⟩; etc.

*angels*: 2: ⟨36,174,252,651⟩; 4: ⟨12,22,102,432⟩; 7: ⟨17⟩;  
*fools*: 2: ⟨1,17,74,222⟩; 4: ⟨8,78,108,458⟩; 7: ⟨3,13,23,193⟩;  
*fear*: 2: ⟨87,704,722,901⟩; 4: ⟨13,43,113,433⟩; 7: ⟨18,328,528⟩;  
*in*: 2: ⟨3,37,76,444,851⟩; 4: ⟨10,20,110,470,500⟩; 7: ⟨5,15,25,195⟩;  
*rush*: 2: ⟨2,66,194,321,702⟩; 4: ⟨9,69,149,429,569⟩; 7: ⟨4,14,404⟩;  
*to*: 2: ⟨47,86,234,999⟩; 4: ⟨14,24,774,944⟩; 7: ⟨199,319,599,709⟩;  
*tread*: 2: ⟨57,94,333⟩; 4: ⟨15,35,155⟩; 7: ⟨20,320⟩;  
*where*: 2: ⟨67,124,393,1001⟩; 4: ⟨11,41,101,421,431⟩; 7: ⟨16,36,736⟩;

Which document(s) (if any) match each of the following queries at which positions, where each expression within quotes is a phrase query? (i) “fools rush in” (ii) “fools rush in” AND “angels fear to tread”.

**Chapter 3****Exercise**

Compute the Levenshtein matrix for the distance between the strings “apfel” (input) and “poems” (output). Use this format (as introduced in class):

		f	a	s	t
	0	1 1	2 2	3 3	4 4
c	1 1	1 2 2 1	2 3 2 2	3 4 3 3	4 5 4 4
a	2 2	2 2 3 2	1 3 3 1	3 4 2 2	4 5 3 3
t	3 3	3 3 4 3	3 2 4 2	2 3 3 2	2 4 3 2
s	4 4	4 4 5 4	4 3 5 3	2 3 4 2	3 3 3 3

**Exercise**

We saw in class that the Levenshtein sequence of operations for converting strings into each other is not unique. For example, “cat” can be transformed into “catcat” either by insert, insert, insert, copy, copy or by copy, copy, copy, insert, insert, insert. In contrast, the minimum number of cost-1 Levenshtein operations for converting one string to another is fixed since the minimum is unique. Let  $n_i, n_d, n_r$  be the number of inserts, deletes and replaces in a sequence of operations. Give an example of a pair of strings and two different sequences of operations  $\sigma_1$  and  $\sigma_2$  that convert the first string into the second such that  $n_i(\sigma_1) \neq n_i(\sigma_2)$  or  $n_d(\sigma_1) \neq n_d(\sigma_2)$  or  $n_r(\sigma_1) \neq n_r(\sigma_2)$ . Or prove that this is not possible.

**Chapter 6/7**

**Exercise**

Compute the Inc.ltn similarity between the query “digital phones” and the document “digital phones and video phones and other phones” by filling out the empty columns in the table below. Assume  $N = 10,000,000$ . Treat *and* and *other* as stop words. What is the final similarity score? What is the corresponding Jaccard score?

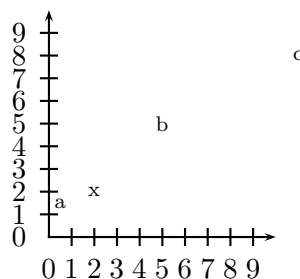
word	query				document				product
	tf-raw	tf-wght	df	idf weight	tf-raw	tf-wght	weight	n'lized	
digital			10,000						
video			100,000						
phones			50,000						

**Exercise**

One measure of the similarity of two vectors is the Euclidean distance between them:  $|\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^M (x_i - y_i)^2}$ . Given a query  $q$  and documents  $d_1, d_2, \dots$ , 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.

**Exercise**

In the figure below, which of the three vectors  $\vec{a}$ ,  $\vec{b}$ , and  $\vec{c}$  is (i) most similar to  $\vec{x}$  according to dot product similarity ( $\sum_i x_i \cdot y_i$ ), (ii) most similar to  $\vec{x}$  according to cosine similarity ( $\sum_i x_i \cdot y_i / (|x||y|)$ ), (iii) closest to  $\vec{x}$  according to Euclidean distance? The vectors are  $\vec{a} = (0.5 \ 1.5)^T$ ,  $\vec{x} = (2 \ 2)^T$ ,  $\vec{b} = (5 \ 5)^T$ , and  $\vec{c} = (11 \ 8)^T$ . Compute the relevant dot products, cosines and distances. Assume that higher dot product indicates higher similarity.



## Chapter 8

### Exercise

Below is a table showing how two human judges assigned documents to the class “English” (0 = is not written in English, 1 = is written in English). Let us assume that you’ve written a classifier that assigns the documents {2, 5, 6, 7, 8} to “English”.

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

(i) Calculate precision, recall, and  $F_1$  of your system if a document is considered relevant only if the two judges agree it is relevant. (ii) Calculate precision, recall, and  $F_1$  of your system if a document is considered relevant if either judge thinks it is relevant. (iii) Calculate kappa

## Chapter 9

### Exercise

Suppose that a user’s initial query is “cheap CDs cheap DVDs extremely cheap CDs”. The user examines two documents,  $d_1$  and  $d_2$ . She judges  $d_1$ , with the content “CDs software cheap CDs” relevant and  $d_2$  with content “cheap thrills DVDs” 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 as in Equation 9.3 what would the revised query vector be after relevance feedback? Assume  $\alpha = 1, \beta = 0.8, \gamma = 0.2$ .

## Chapter 13

### Exercise

Based on the data below, estimate a multinomial Naive Bayes classifier (the type of NB classifier we introduced in class) and apply the classifier to the test document.

	docID	words in document	in $c = \text{Japan?}$
training set	1	Kyoto Osaka Taiwan	yes
	2	Japan Kyoto	yes
	3	Taipei Taiwan	no
	4	Macao Taiwan Shanghai	no
test set	5	Taiwan Taiwan Kyoto	?

## Chapter 12

### Exercise

Rank the documents in collection  $\{d_1, d_2\}$  for query  $q$  using the language model approach to IR introduced in class. Use the mixture coefficient  $\lambda = 0.4$ .

- $d_1$ : Scottish sheep getting smaller due to climate change study says
- $d_2$ : The analysis has shown a dramatic shift in the natural ranges for US Bird species in response to climate change
- Query  $q$ : climate change

## Chapter 15

### Exercise

The decision boundary of the support vector machine  $S$  is defined by:

$$(2 \ 1)^T \vec{x} - 4 = 0$$

In this exercise, use the labels +1 and -1 for the two classes.

(i) Let  $S$  be an SVM that doesn't make a decision for points in the margin. Which of the following points does  $S$  make a decision on and what is the decision?

$$\vec{a} = (1.01 \ 2)^T, \vec{b} = (1 \ 1.99)^T, \vec{c} = (2 \ 2)^T, \vec{d} = (0 \ 0)^T$$

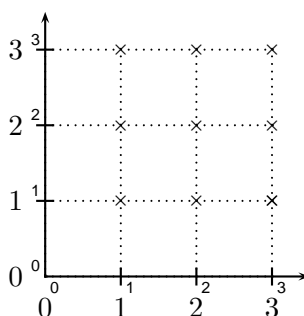
(ii) Let  $S$  be an SVM that always makes a decision, even for points in the margin. In this case, what is the decision for the four points  $\vec{a}$ ,  $\vec{b}$ ,  $\vec{c}$ , and  $\vec{d}$ ?

## Chapter 16

### Exercise

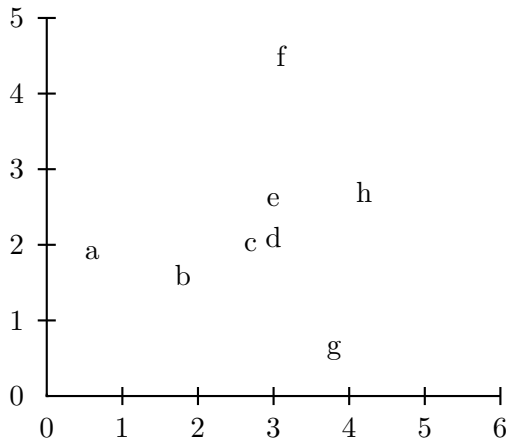
a) Perform a 3-means clustering for the points below. Draw a different diagram for each iteration to show the assignments and the centroids. If a tie occurs during an assignment step, you can freely choose any of the possible assignments.

b) There are several clusterings that 3-means can converge to in this case. Give an example of such a clustering that is different from the one in a.



## Chapter 17

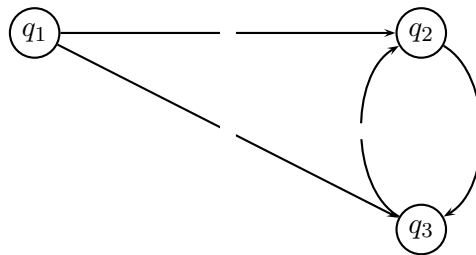
### Exercise



The points have the following coordinates: a: (0.6,1.9), b: (1.8,1.6), c: (2.7,2.0), d: (3.0,2.1), e: (3.0,2.6), f: (3.1,4.5), g: (3.8,0.6), h: (4.2,2.7). Define the similarity of two points as  $-(x_1 - x_2)^2 - (y_1 - y_2)^2$ .

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/ellipse with a number indicating the temporal sequence of mergers.

## Chapter 21



Compute PageRank for the web graph in Figure for each of the three pages. Also give the relative ordering of the 3 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.