IR&TM Review II

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 / 2	3 / 2	3 / 3	4 / 3	4 / 4	5 / 4
a	2 / 2	2 / 3	2 / 2	1 / 3	3 / 1	3 / 2	4 / 2	4 / 3	5 / 3
t	3 / 3	3 / 4	3 / 3	3 / 4	2 / 2	2 / 3	3 / 2	2 / 3	4 / 2
s	4 / 4	4 / 5	4 / 4	4 / 5	3 / 3	2 / 4	3 / 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, copy or by copy, copy, copy, insert, insert, insert. In contrast, the minimum number Levenshtein operations with cost 1 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 σ_1 and σ_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 lnc.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?

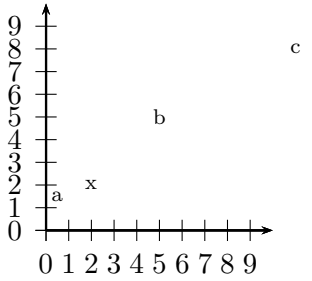
	query					document				
word	tf-raw	tf-wght	df	idf	weight	tf-raw	tf-wght	weight	n'lized	product
digital			10,000							
video			100,000							
phones			50,000							

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 $d_1, d_2, \ldots$, 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 $\vec{a} = (0.5 \quad 1.5)^T$, $\vec{x} = (2 \quad 2)^T$, $\vec{b} = (5 \quad 5)^T$, and $\vec{c} = (11 \quad 8)^T$. 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.

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 = 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 Jelinek-Mercer smoothing with 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:

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

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?

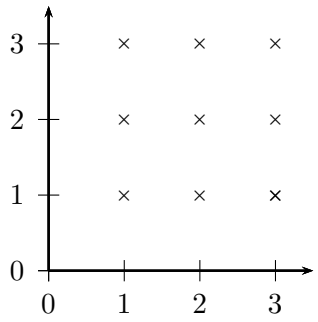
$\vec{a} = (1.01 \quad 2)^T$, $\vec{b} = (1 \quad 1.99)^T$, $\vec{c} = (2 \quad 2)^T$, $\vec{d} = (0 \quad 0)^T$

(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 $\vec{d}$?

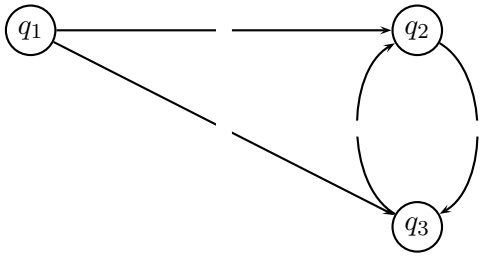
Chapter 16

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 21



Compute PageRank for the web graph in the above 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.

Chapter 19

Exercise

advertiser	bid	CTR
A	\$5.00	0.04
B	\$1.00	0.1
C	\$0.50	0.06
D	\$1.00	0.03

Compute how much advertisers A, B, C, D have to pay for each click in a second price auction as described in class. The minimum amount per click is 0.005.

Exercise

The shingle representations of three documents are as follows: $d_3 = (0,0,1,0,0,0,1)^T$, $d_4 = (0,0,1,0,0,0,0)^T$, $d_5 = (1,1,1,0,1,1,1)^T$

We will use sketches of size 2. The two elements of a sketch are defined by the permutations. $(3 \times n + 2) \mod 7$ and $(5 \times n + 1) \mod 7$. Based on this setup what are the estimates of the three Jaccard coefficients $J(d_3, d_4)$, $J(d_3, d_5)$, and $J(d_4, d_5)$?