Tutorial 1

Q1. Consider these documents:

Doc1 breakthrough drug for schizophrenia
Doc2 new schizophrenia drug
Doc3 new approach for treatment of schizophrenia
Doc4 new hopes for schizophrenia patients

a) Draw the term-document incidence matrix for this document collection.
b) Draw the inverted index representation for the collection.

Q2. For the document collection shown in Q1, what are the returned results for these queries:

a) schizophrenia AND drug
b) for AND NOT (drug OR approach)

Q3. The table below gives the sizes of postings lists for tokens a, b, c, d, e and f.

Term	a	b	c	d	e	f
Postings size	174	350	637	9066	950	252

a) Recommend a query processing order for Boolean query:
(a OR d) AND (b OR e) AND (c OR f)

b) Estimate the minimum and maximum possible number of results for query:
(c OR e) AND (NOT a)

Q4. For a conjunctive query (e.g., s1 AND s2 AND s3), is processing postings lists in order of size guaranteed to be optimal? Explain why it is, or give an example where it isn't.

Q5. For the following Porter stemmer rule group:

sses → ss (e.g. caresses → caress)
ies → i (e.g. ponies → poni)
ss → ss (e.g. caress → caress)
s → (e.g. cats → cat)

a) What is the purpose of including an identity rule such as SS => SS?
b) Applying just this rule group, what will the following words be stemmed to?
   i. circus
   ii. canaries
   iii. boss
c) What rule should be added to correctly stem pony, considering the stemming of ponies in the rule group above?

Tutorial 2

Q1. Consider the three words 'fly', 'flier' and 'lier'.
a) List the 3-grams for each word
b) Compute the Jaccard coefficient between the word 'lie' and each of the three words. Which word could be the suggested spell-corrected word for the query 'lie'?
c) Compute the edit distance between 'lie' and each of the three words by using Levenshtein distance algorithm. Which word could be the suggested spell-corrected word for the query 'lie'?

Q2. Given the word "cat":
a) Compute all its possible right rotations for a permuterm query.
b) Generate 5 wild-card queries that can retrieve the word.
c) Can you think of an English term that matches the permuterm query **er$fi***, but does not satisfy the Boolean query fi*mo*er??

Q3. Consider the biword index:

Term	Docs1	Docs2	Docs3
angels fools	1	0	0
angels rush	1	0	1
angels fear	0	0	1
fools rush	1	0	0
fear fools	0	1	0
fear to	0	1	1
where angels	1	0	1
to tread	1	0	0
fear in	0	1	1
rush in	1	0	1

a) Which are the biword boolean queries generated by the following phrase query?
   1. fools rush in
   2. where angels rush in
   3. angels fear to tread

b) Which are (if any) the documents retrieved?

Q4. Which documents (if any) meet each of the three phrase queries of Q3, based on the available positional index?

term	doc1	doc2	doc3
angels	#36, 174, 252, 651$		#15, 123, 412$
fools	#1, 17, 74, 222$	#8, 78, 108, 458$	
fear		#13, 43, 113, 433$	#18, 328, 528$
in	#3, 37, 76, 444, 851$	#10, 20, 110, 470, 500$	#5, 17, 25, 195$
rush	#2, 66, 194, 321, 702$		#4, 16, 404$
to	#47, 86, 234, 999$	#14, 24, 774, 944$	#19, 319, 599, 709$
tread	#57, 94, 333$		
where	#67, 124, 393, 1001$	#11, 41, 101, 421, 431$;	#14, 36, 736$

Q5. We have a three-word text query 'Enjoy a beer'. Below is a table showing the term counts and document term size of each 4 documents. With the information provided, recommend the top 3 documents should be returned to the normalized text query.

	enjoy	a	beer	size
Doc 1	2	10	5	400
Doc 2	3	35	8	500
Doc 3	5	40	3	600
Doc 4	10	10	6	750

Tutorial 3

Q1. Consider the table of term frequencies for 3 documents denoted Doc1, Doc2, Doc3 below. Compute the tf-idf weights for the terms *car, auto, insurance, best*, for each document, using the idf values from the table below.

tf	Doc1	Doc2	Doc3	idf
car	22	4	24	1.65
auto	3	33	0	2.08
insurance	0	33	29	1.62
best	14	0	17	1.5

Q2. Refer to the tf and idf values for four terms and three documents from Q1. Compute the two top scoring documents on the query "best car insurance" for each of the following weighing schemes:

a) nnn.atc
b) ntc.atc

Q3. Consider the following term-document count matrix

	Antony and Cleopatra	Julius Caesar	The Tempest
Antony	157	73	0
Brutus	28	157	0
Caesar	232	227	0
Calpurnia	0	10	0
Cleopatra	23	0	37
Mercy	0	10	15
Worser	2	0	1

a) Compute the cosine similarity and the Euclidean distance between the documents and the query: **"caesar mercy brutus"** based on the above matrix.

b) How does the Euclidean distance change if we normalize the vectors?

NB: Compute the vector space using tf-idf formula of Q1

Tutorial 4

Q1. Consider again the data of Tutorial 3 (Q2) with *ntc.atc* for the query-dependent scoring. Suppose that we were given static quality scores of ~~0.2~~ and ~~0.3~~ for Doc1 and Doc2, respectively. Determine what ranges of static quality score for Doc3 result in it being the first, second or third result for the query *best car insurance*.

Q2. Let the static quality scores for Doc1, Doc2 and Doc3 of Tutorial 3 be 0.25, 0.5 and 1, respectively. Sketch the postings for impact ordering (champion list) when each postings list is ordered by the sum of the *ntc* values from Tutorial 3 and the static quality scores (net score).

Q3. The nearest-neighbor problem in the plane is the following: given a set of N data points on the plane, we preprocess them into some data structure such that, given a query point q, we seek the point in N that is closest to q in Euclidean distance. Clearly cluster pruning can be used as an approach to the nearest-neighbor problem in the plane, if we wished to avoid computing the distance from q to every one of the query points. Devise a simple example on the plane so that with 2 leaders, the answer returned by cluster pruning is incorrect (it is not the data point closest to q).

Tutorial 5

Q1. Suppose you travel at speed *P* Km/h for distance *x* Km and speed *R* Km/h for distance *x* Km. The total distance of the whole journey is *2x* Km. Compute the average speed (which is the harmonic mean or F1-measure of the values of *P* and *R*) for the whole journey of *2x*.

Q2. The balanced F measure (a.k.a. F1) is defined as the harmonic mean of precision and recall. What is the advantage of using the harmonic mean rather than "averaging" (using the arithmetic mean)?

Q3. Consider the following retrieval task where there are a total of 20 documents in the corpus and 8 documents are relevant to a query (the remaining 12 are irrelevant). The following left-to-right sequence denotes whether each successively retrieved document is relevant (1) or not (0):

00100 10000 11111 00001

Using manual computation or a spreadsheet software or scripts/code (e.g., matlab),

a) Plot the Precision P versus N (number of documents retrieved) curve, clearly labeling P.
b) Plot the Recall R versus N curve on the same graph as part (a), clearly labeling R.
c) Plot the F1-Measure versus N curve on the same graph as part (a), clearly labeling F.
d) Plot the Arithmetic Mean (M) versus N curve on the same graph as part (a), labeling M.

Q4. Repeat Q3(a) to (d) for the following sequence of documents:

e) 11111 11100 00000 00000
f) 10101 01010 10101 00000
g) 00000 00000 00111 11111
h) 11111 11111 11111 00000

For parts (a)-(c), assume N_R=8 relevant documents, and for part (d), assume N_R=30 relevant documents.

Q5. Based on your plots in Q3 and Q4, state whether the following statements are true or false:

a) The arithmetic mean curve is always sandwiched between the P and R curve.
b) The harmonic mean (F1-measure) curve is always sandwiched between the P and R curve.
c) The arithmetic mean is always larger or equal to the harmonic mean (F-Measure).
d) The P versus N curve always starts from 1.
e) The R versus N curve always starts from 0.
f) The R versus N curve always ends at 1.
g) The P and R curve will always intersect @ N = R_0 (total # of relevant docs in corpus)

Q6. Given that N=# docs retrieved, N_R=# relevant docs retrieved, R_0=# relevant docs in corpus

a) What is the condition for the P and R curve to intersect?
b) How many times will the P and R curve intersect at non-zero values?
c) What is the value of the F1-Measure at the intersection?

Q7. Suppose you have a corpus of Web documents with a vocabulary of 5 words {thaksin, gst, thailand, fine, raise}. Modify the initial query of {gst, raise} based on the 4 returned results below:

- set of relevant documents D_r = {[1 0 1 0 1], [1 1 1 0 1]}, and
- set of irrelevant documents D_n = {[0 0 0 1 0], [0 1 0 1 1]}

using Rocchio's method with the following parameters:

i) α=1, β=1, γ=0.1
j) α=1, β=1, γ=0.5
k) α=1, β=1, γ=1

Q8. In Rocchio's algorithm, what weight setting for α, β, γ does a "Find pages like this one" search correspond to?

Q9. Why is positive feedback likely to be more useful than negative feedback to an IR system? Why might only using one non-relevant document be more effective than using several?

Tutorial 6

Q1. Consider the 10 class conditioned word probabilities (c_0=non-spam, c_1=spam) in Table 1:

Word w_i	brand	huge	hottest	incredible	million	new	offers	pay	save	family
$p(w_i\|c_0)$	0.10	0.20	0.30	0.05	0.05	0.10	0.20	0.02	0.03	0.40
$p(w_i\|c_1)$	0.98	0.92	0.91	0.99	0.98	0.99	0.93	0.99	0.99	0.02

Table 1

For each of the following 3 email snippets:

d_1: OEM software - throw packing case, leave CD, use electronic manuals. **Pay** for software only and **save** 75-90%! Find **incredible** discounts! See our special **offers**!

d_2: Our **Hottest** pick this year! **Brand new** issue Cana Petroleum! VERY tightly held, in a booming business sector, with a **huge** publicity campaign starting up, Cana Petroleum (CNPM) is set to bring all our readers **huge** gains. We advise you to get in on this one and ride it to the top!

d_3: Dear friend, How is your **family**? hope all of you are fine, if so splendid. Yaw Osafo-Maafo is my name and former Ghanaian minister of finance. Although I was sacked by President John Kufuor on 28 April 2006 for the fact I signed 29 **million** book publication contract with Macmillan Education without reference to the Public Procurement Board and without Parliamentary approval.

Ignoring case, punctuations, and words beyond the 10 known vocabulary words, compute the class conditioned document probabilities for each of the 3 documents, namely $P(d_1|c_0)$, $P(d_2|c_0)$, $P(d_3|c_0)$, $P(d_1|c_1)$, $P(d_2|c_1)$, and $P(d_3|c_1)$, using the Naïve Bayes model:

$$p(d_j \mid c_k) \propto \prod_{i=1}^{t} p(w_i \mid c_k)^{f(w_i,d_j)}$$, where $f(w_i,d_j)$ = frequency of word w_i in document d_j.

Q2. Compute the posterior probabilities of each document in Q1, namely $P(c_0|d_1)$, $P(c_1|d_1)$, $P(c_0|d_2)$, $P(c_1|d_2)$, $P(c_0|d_3)$, and $P(c_1|d_3)$, assuming that 80% of all emails received are spam, and finally decide whether each document is spam or non-spam.

$$p(c_k \mid d_j) \propto p(d_j \mid c_k)p(c_k)$$

Q3. Build a Naïve Bayes classifier using words as features for the training set in Table 2 and use the classifier to classify the test set in the table.

	docID	words in document	in c = *China*?
training set	1	Taipei Taiwan	yes
	2	Macao Taiwan Shanghai	yes
	3	Japan Sapporo	no
	4	Sapporo Osaka Taiwan	no
test set	5	Taiwan Taiwan Sapporo	?

Table 2

Q4. 48% of the 71 PSC (Public Service Commission) scholars lives in Housing Development Board (HDB) flats. 77% of Singapore's population (estimated to be 6 million) lives in HDB. Use the Chi-Square test at significance level p=0.001 (table below) to test whether the distribution of PSC scholars is a reflection of the underlying population distribution.

	PSC Scholars	Non PSC Scholars	
HDB	34	4,619,966	4,620,000
Condo/landed	37	1,379,963	1,380,000
	71	5,999,929	6,000,000

Table 3

p	χ^2 critical value
0.1	2.71
0.05	3.84
0.01	6.63
0.005	7.88
0.001	10.83

Table 4

Q5. Table 5 shows the vector representation of five documents, where the training set includes d_1, d_2, d_3, and d_4, and d_5 belongs to the test set. Among the training documents, only the first three have been labeled as belonging to the target class u_c.

a) Classify document d_5 using vector space classification
   i. Use Euclidean distance
   ii. Use cosine similarity

b) Classify document d_5 using 1NN
   i. Use Euclidean distance
   ii. Use cosine similarity

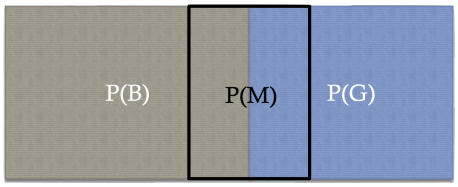
	term weights					
vector	Chinese	Japan	Tokyo	Macao	Beijing	Shanghai
$\vec{d}_1$	0	0	0	0	1.0	0
$\vec{d}_2$	0	0	0	0	0	1.0
$\vec{d}_3$	0	0	0	1.0	0	0
$\vec{d}_4$	0	0.71	0.71	0	0	0
$\vec{d}_5$	0	0.71	0.71	0	0	0
$\vec{\mu}_c$	0	0	0	0.33	0.33	0.33
$\vec{\mu}_{\bar{c}}$	0	0.71	0.71	0	0	0

Table 5

Tutorial 7

Q1. Suppose 100 people attended a party at Zouk, 40 girls (G) and 60 boys (B), among which are 20 married couples (1 married couple M = 1 girl + 1 boy). Compute the following probabilities. (Hint: use the Venn diagram below)

a. What is the probability that you will run into a boy if you attend the party?
b. Suppose you throw a bunch of flowers randomly at a person in the party, what is the probability that it will land on a girl?
c. Suppose you see an attractive girl at the party, what is the probability that she is married?
d. Suppose you see a gorgeous hunk at the party, what is the probability that he is married?



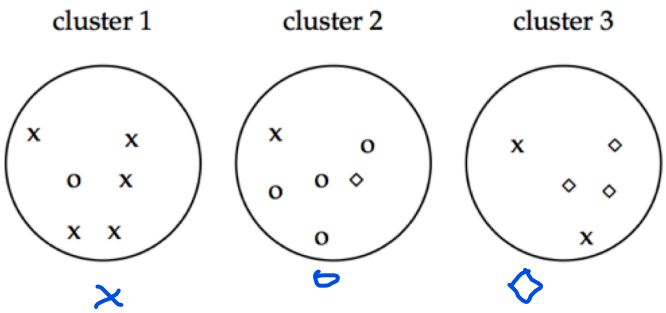
Q2. Briefly explain why kNN handles multimodal classes better than the standard vector space model (VSM) where you classify based on cosine similarity.

Q3. Consider a British news article about Toyota automobiles and an American news article about Toyota cars, where the former uses "automobile" and the latter uses "car". Why are documents that do not use the same term for the concept *car* likely to end up in the same cluster in K-means clustering?

Q4. Two of the possible termination conditions for K-means are (1) assignment does not change, and (2) centroids do not change. Do these two conditions imply each other? Explain.

Q5. Give an example of a set of points and three initial centroids (which need not be members of the set of points) for which 3-means converges to a clustering with an empty cluster.

Q6. Explain how to compute the Rand Index using the example in Figure 1.



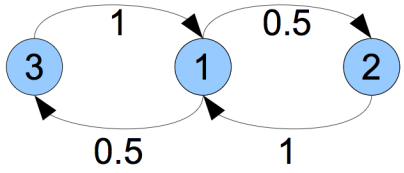
► **Figure 1** Purity as an external evaluation criterion for cluster quality. Majority class and number of members of the majority class for the three clusters are: x, 5 (cluster 1); o, 4 (cluster 2); and ⋄, 3 (cluster 3). Purity is $(1/17) \times (5+4+3) \approx 0.71$.

Tutorial 8

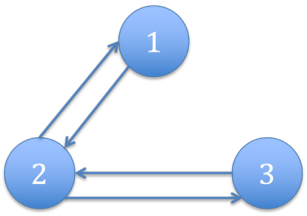
Q1. The Goto method ranked click-through advertisements matching a query by bid: the highest-bidding advertiser got the top position, the second- highest the next, and so on. What can go wrong with this when the highest-bidding advertiser places an advertisement that is irrelevant to the query? Why might an advertiser with an irrelevant advertisement bid high in this manner?

Q2. Each of two Web search engines A and B generates a large number of pages uniformly at random from their indexes. 30% of A's pages are present in B's index, while 50% of B's pages are present in A's index. What is the number of pages in A's index, relatively to B's?

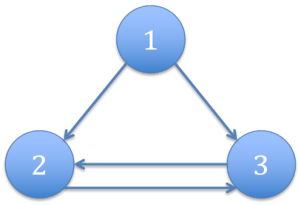
Q3. Write down the transition probability matrix for the following Markov chain. Is this Markov chain ergodic?



Q4. Write down the transition probability matrices for the surfer's walk with teleporting probability $\alpha = 0.3$.



Q5. Compute the Hubs and Authorities scores for each of the 3 nodes in the following graph. Normalize your hubs and authorities values to lie within [0, 1].



Q6. If all the hub and authority scores are initialized to 1, what is the hub/authority score of a node after the first iteration?