Final - Information Retrieval - Fall 2014

Date: 16 Jan 2015 Marks: Time: 180 mins. Name: Roll Number: *Attempt the examination on the question paper and write concise answers.* ***Do NOT attach any extra sheets, they will not be checked.*** *Use the back of the last page for rough work. Marks would be deducted for superfluous and incorrect answers. You are not allowed to ask any questions. In case of any ambiguity, state an appropriate assumption and solve the question using that assumption. Do not fill the tables titled parts/marks. Good Luck!*

Part	1	2	3	4	5	6	Total
Marks	/20	/9	/15	/15	/10	/11	/80

1. Index Construction

Given the three document corpus and a stop word list below, apply the usual index processing steps and answer the following questions AFTER performing the steps.

d_1: "NU FAST is fast."	**d_3:** "FAST is a university, every university is not NU"
d_2: "NU university is not FAST!"	**Stop words:** a, an, is, not, some, every, all

a. What are the steps usually performed to create an inverted index?
Tokenize, case fold, filter stop word, stem, apply weights, add to index. (You need to mention at least 4 of these) b. How many types are there in the corpus?

7: universe, nu, fast, is, not, a, every
c. How many tokens are there in the corpus?
18 atrix in the table below.
e. Fill the term-document matrix in the table below. Use only term frequencies. (**TF**) f. Calculate the IDF for each term. Fill it in the table below. (**IDF**)
g. Fill the TF*IDF weighted term-document matrix below. (**TF*IDF**)
h. Given the following query q, fill the columns named q for Boolean, IDF and TF*IDF matrices. Use the IDF values calculated from the corpus.

d. Fill the **Boolean** incidence m**q:** "FAST NU University"

i. Given the ad-hoc query **q**, give the results for search using the Boolean retrieval model. {d2,d3} This has to be a set. No ranking.
j. Given the ad-hoc query **q**, give the results for search using the TF weighted inverted index. 1. d2 2. d3 3. d1
k. Given the ad-hoc query **q**, give the results for search using the TF*IDF weighted inverted index. 1. d2 2. d3 3. d1
l. Are the answers for parts i, j and k different? If yes, why? If no, why not?

Part	Marks
a	/1
b	/1
c	/1
d	/1
e	/1
f	/2
g	/1
h	/2
i	/2
j	/2
k	/2
l	/1
m	/1
n	/2

Yes. The results is unranked for Boolean retrieval so it can be in any order. The rest of the results are same despite the difference in length of the documents.

Boolean TF IDF TF*IDF

terms↓ d_1 d_2 d_3 q d_1 d_2 d_3 q d_1 d_2 d_3 q univers 0 1 1 1 0 1 2 1 0.5 0 0.5 1 0.5

nu 1	1	1	1 1	1	1	1 0.33 0.33	0.33	0.33	0.33
fast 1	1	1	1 2	1	1	1 0.33 0.66	0.33	0.33	0.33

Final - Information Retrieval - Fall 2014

Date: 16 Jan 2015 Marks: Time: 180 mins. Name: Roll Number:

m. Skip pointers are used for fast postings list intersection for queries of the form ***foo AND bar***. Why are skip pointers not useful for queries of the form ***foo OR bar***?

In queries of the form "foo OR bar", it is essential to visit every docID in the posting lists of either terms, thus killing the need for skip pointers.

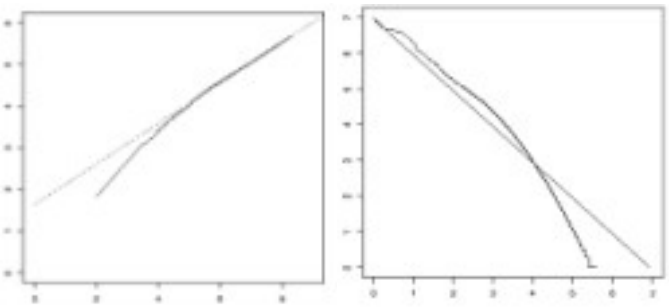
n. Roman Urdu has no normalised form and words such as رتهبنی can be spelled using Roman Urdu in many way e.g. bayhtareen, behtareen, behtarin, bayhtereen, behtereen etc. Given a corpus of messages in Roman Urdu, what extra processing would you suggest to make sure users find results with all variants of the query.

Use soundex modified for Urdu. (Any answer which mentions a normalised representation for alternate spellings would be acceptable here)

2. Index Compression

a. Given below are two graphs depicting Heap's Law and Zipf's law respectively. What do the axes typically represent?

	Heap's Law	Zipf's Law
x-axis	$\log_{10}$(Terms)	$\log_{10}$(Rank)
y-axis	$\log_{10}$(Vocabulary)	$\log_{10}$(collection frequency)



b. Consider the postings list <4, 10, 11, 12, 15, 62, 63, 265, 268, 270, 400>. Find the gap sequence. <4, 6, 1, 1, 3, 47, 1, 202, 3, 2, 130>

c. Assume that the length of the postings list above is stored separately, so the system knows when a postings list is complete. Using variable byte coding, how many bytes will the above postings list require? (Count only space for encoding the sequence of numbers.) 13

Part	Marks
a	/2

b	/1
c	/2
d	/2
e	/2

d. Consider the following sequence of γ-coded gaps: 0111100011101111110101111101011101111. What is the postings list? Gap sequence: 1 19 3 55 6 15
DocID sequence: 1 20 23 78 84 99

e. Given the front-coded string, what should be the lexicon?
n a b ∗ 0 ◊3 ◊b e d 4 ◊i n g 3 ◊i t 3 ◊k 3 ◊o b 2 ◊c a r a t 3 ◊e l l e
Code is invalid. Only decodable strings are nab and nabbed.

3. Evaluation

Final - Information Retrieval - Fall 2014

Date: 16 Jan 2015 Marks: Time: 180 mins. Name: Roll Number: a. To create a gold standard for a sentiment analysis task, two annotators independently annotate 200 documents regarding whether they convey a positive attitude or not. The following table shows how often they agreed. Calculate the Kappa coefficient for the agreement between the two annotators.

P(A) = (120 + 10)/200 = 130/200 = 0.65
P(positive) = (160 + 150)/(200 + 200) = 310/400 = 0.775
P(negative) = (40 + 50)/(200 + 200) = 90/400 = 0.225
P(E) = P(negative)^2 + P(positive)^2 = 0.2252+0.7752 = 0.050625 + 0.600625 =

0.65125

κ = (P(A) − P(E))/(1 − P(E)) = (0.65 − 0.65125)/(1 − 0.65125) = −0.00125/0.34875 = −0.00358422

Judge B

	Positive	*Negative*
Positive	120	30
Negative	40	10

b. Will it be possible to construct a gold standard from the annotated data in part a? If yes, how? If no, why? According to Krippendoff's interpretation, the agreement is too low to be a reliable basis for a gold standard.

c. For a 3-class problem, here's the predicted values for and the truth values are given in the table. Find the Precision, Recall and F score for class *x* only. Round your answer to 3 decimal places.

P= 0.772 R= 0.974 F= 0.861

y	5	25	8
z	31	30	62

Truth

	x	*y*	*z*
x	1323	270	120

d . As an aggregated evaluation metric, we have the option to use either the simple average (macro-F) or the weighted average (micro-F). Which one makes more sense for the problem in part c? What is the value of this metric for this case?
Solution1: Use micro-F. Class distribution is skewed so higher weight(x) should be weighted higher. 0.689

Solution2: Use macro-F. Class distribution is skewed but each class should be equally important. 0.465 (Any answer acceptable as long as the reason is correct)

b	/1
c	/3
d	/2
e	/4
f	/2

e. The table below shows the ranked output of an IR system on two different queries. Crosses represent relevant

Part	Marks
a	/3

documents as marked by human judges. Empty spaces represent irrelevant documents. Assuming that there are no more relevant documents after the 15th rank, calculate 11-point precision for both queries using an appropriate interpolation method.

Rank	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**		**15**	
Q1	X			X	X												
Q2				X		X				X						X	
Q1:R	1/3 =0.33			2/3 = 0.66	3/3 = 1												
Q1:P(R)	1/1 = 1			2/4 = 0.5	3/5 = 0.6												
Q2:R				1/4 = 0.25		2/4 = 0.5				3/4 = 0.75						4/4 = 1	
Q2:P(R)				1/4 = 0.25		2/6 = 0.33				3/1 = 0.3					4/15 = 0.2667		

	P(R=0.0)	P(R=0.1)	P(R=0.2)	P(R=0.3)	P(R=0.4)	P(R=0.5)	P(R=0.6)	P(R = 0.7)	P(R = 0.8)	P(R = 0.9)	P(R = 1.0)	Avg
Q1	1	1	1	1	0.5	0.5	0.5	0.6	0.6	0.6	0.6	
Q2	0.25	0.25	0.25	0.33	0.33	0.33	0.3	0.3	0.2667	0.2667	0.2667	
Avg	0.625	0.625	0.625	0.665	0.415	0.415	0.4	0.45	0.43335	0.43335	0.43335	0.502

Final - Information Retrieval - Fall 2014

Date: 16 Jan 2015 Marks: Time: 180 mins. Name: Roll Number:

f. Recall the TextRank paper by Mihalcea et al. (2004). What sentence similarity measure was used to construct the graph for summarization using sentence extraction?

Formally, given two sentences S_i and S_j, with a sentence being represented by the set of N_i words that appear in the sentence: $S_i = w_1^i, w_2^i, ..., w_{N_i}^i$, the similarity of S_i and S_j is defined as:

$$Similarity(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{log(|S_i|) + log(|S_j|)}$$

4. Classification

a. Based on the data below, estimate a multinomial Naive Bayes classifier using add-one smoothing and apply

the classifier to the test document. Calculate the probability that the classifier assigns the test document to j = Japan or n = not Japan.

docID words in document class
Training set 1 Kyoto Tokyo j
2 Japan Kyoto j
3 Taipei Taiwan n
4 Macao Taiwan n
Test set 5 Taiwan Taiwan Kyoto ?

P(j) = P(n) = ½ = 0.5
P(Taiwan|j) = (0 + 1)/(4 + 6) = 1/10 = 0.1
P(Kyoto|j) = (2 + 1)/(4 + 6) = 3/10 = 0.3
P(Taiwan|n) = (2 + 1)/(4 + 6) = 3/10 = 0.3
P(Kyoto|n) = (0 + 1)/(4 + 6) = 1/10 =0.1
P(j|d) ∝ P(j).P(Taiwan|j).P(Taiwan|j).P(Kyoto|j) = (0.5)(0.1)(0.1)(0.3) = 0.0015 P(n|d) ∝ P(n).P(Taiwan|n).P(Taiwan|n).P(Kyoto|n) = (0.5)(0.3)(0.3)(0.1) = 0.0045 P(n|d) > P(j|d), therefore class = n

Part	Marks
a	/4
b	/2
c	/3
d	/3
e	/2
f	/1

b. Given the number of features and the length of the document, a multinomial model should work better than a Bernoulli model. Comment.
Incorrect for the document above. The Bernoulli model works best for short docs and fewer features

c. In binary Rocchio classification, a vector $\vec{x}$ is on the decision boundary if it has equal distance to the two class centroids, say $\vec{\mu}$ and $\vec{\nu}$. Show, mathematically, that the decision boundary $|\vec{x}-\vec{\mu}| = |\vec{x}-\vec{\nu}|$ forms a linear classifier.

$$|\vec{x}-\vec{\mu}| = |\vec{x}-\vec{\nu}| \Rightarrow |\vec{x}-\vec{\mu}|^2 = |\vec{x}-\vec{\nu}|^2 \Rightarrow |\vec{x}|^2 + |\vec{\mu}|^2 + 2\vec{x}\vec{\mu} = |\vec{x}|^2 + |\vec{\nu}|^2 + 2\vec{x}\vec{\nu}$$

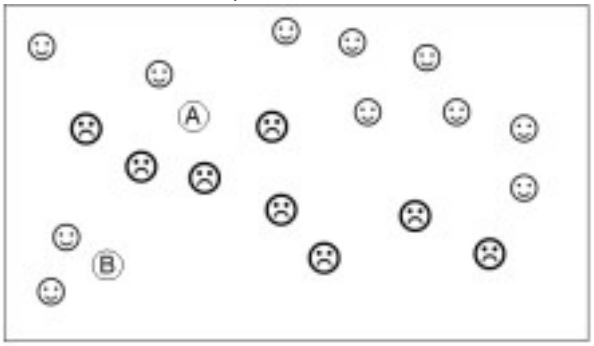
$$\Rightarrow 2(\vec{\mu}-\vec{\nu})\vec{x} + (|\vec{\mu}|^2 - |\vec{\nu}|^2) = 0$$

Which is the equation of a line of the form $\vec{w}\vec{x} + b = 0$

Final - Information Retrieval - Fall 2014

Date: 16 Jan 2015 Marks: Time: 180 mins. Name: Roll Number: d. The figure below shows the sitting arrangement of a room. Some people are happy in the room while the others are sad. Two new persons, A and B, walk in and sit as shown in the figure. What moods should be assigned to these people if their mood is determined by k-nearest neighbours where k = 1, k = 3 and k = 5?

k→	1	3	5
A	☺		
B	☺	☺	



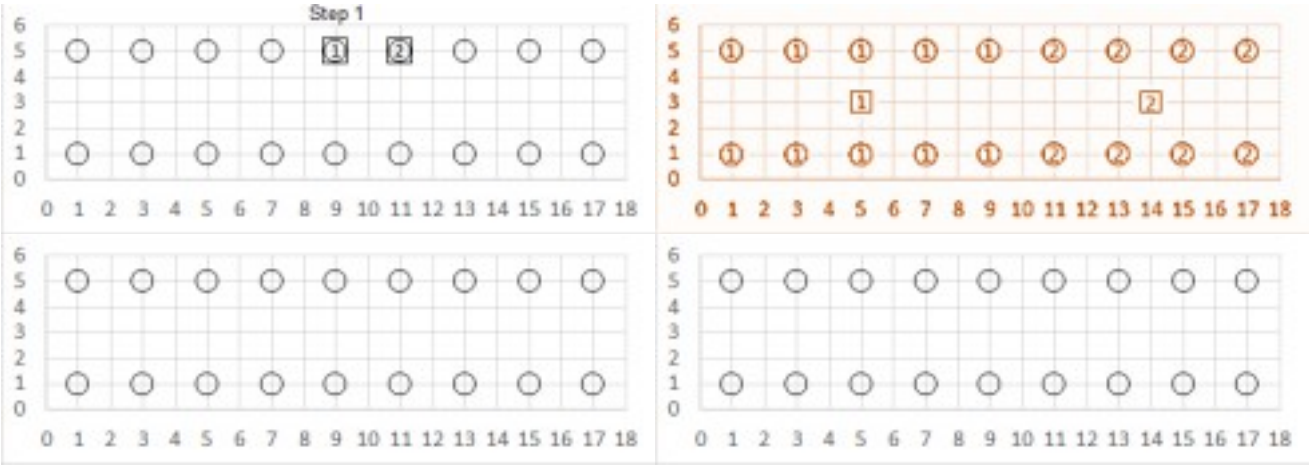
e. Assuming the one dimensional non-linear binary classification problem given below, write a kernel function $\phi: \mathbb{R}^1 \rightarrow \mathbb{R}^2$ which makes the black and white data points linearly separable.

$$x \rightarrow (x, x^2)$$

f. Under what conditions should we prefer using cross-validation for evaluation of a classifier instead of a percentage split? When there's not enough data to have a representative sample while splitting.

5. Clustering

a. Apply k-means clustering (k = 2) to the graph below with initial centers (9, 5) and (11, 5). Show the clusters obtained at each step by marking each point with ① or ② and marking the centers with a box. Number the steps.



b. Give one advantage of hierarchical clustering over k-means clustering, and one advantage of k-means clustering over hierarchical clustering.
HAC: Don't need to know how many clusters you're after / Can cut hierarchy at any level to get any number of clusters K-Means: Can be much faster than hierarchical clustering, depending on data / Can incorporate new data and reform clusters easily

c. The data below shows records of 5 bowlers in a cricket match.

Player	A	B	C	D	E	F
Overs	0	1	2	4	5	5
Wickets	0	2	1	1		

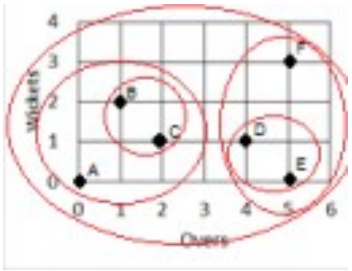
Final - Information Retrieval - Fall 2014

Date: 16 Jan 2015 Marks: Time: 180 mins. Name: Roll Number:

$\vec{\mu} = \frac{1}{N}\sum \vec{x} \quad x \in \omega$

Part	Marks
a	/3
b	/2
c	/4
d	/1

Plot the dataset in the graph. Then apply Hierarchical Agglomerative clustering to the data where the distance between two clusters is the Euclidean distance between their centers. The center $\vec{\mu}$ of a cluster ω with N elements is defined as above. Work out the distances below and create appropriate clusters on the figure by drawing ellipses.



Solution: The distance matrices have been calculated using squared Euclidean distance.

	A	B	C	D	E
B	5				
C	5	2			
D	17	10	4		
E	25	20	10	2	
F	34	17	13	5	9

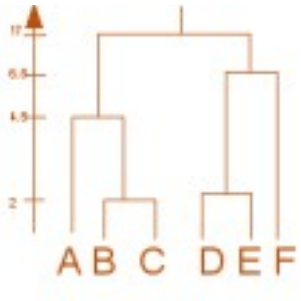
	A	BC	D	E
BC	4.5			
D	17	6.5		
E	25	14.5	2	
F	34	14.5	5	9

	A	BC	DE
BC	4.5		
DE	20.5	10	
F	34	14.5	6.5

	ABC	DE
DE	14.125	
F	23.125	6.5

	ABC
DEF	17

BC (1.5, 1.5)
DE (4.5, 0.5)
ABC (0.75, 0.75)
DEF (4.75, 1.75)
ABCDEF (2.833, 1.1667)



d. Draw a dendrogram showing the merging sequence and labelling the axes.

6. Web

a. Given the following two documents, what is the Jaccard similarity if we use bigram shingles?
D1: a rose is a rose is a rose D2: is a rose a rose
Unique bigrams: B1: a rose, rose is, is a B2: is a, a rose, rose a
Jaccard's similarity = $|B1 \cap B2| / |B1 \cup B2| = 2/4 = 0.5$

Partial Marks for nonUnique bigrams: B1: a rose, rose is, is a, a rose, rose is, is a, a rose B2: is a, a rose, rose a, a rose Jaccard's similarity = $|B1 \cap B2| / |B1 \cup B2| = 3/8 = 0.375$

b. List the properties of an ergodic Markov chain and succinctly describe how PageRank ensures each of them are satisfied. A MC is ergodic if (1) you have a path from any state to any other (2) For any start state, after a finite transient time T0, the probability of being in any state at a fixed time T>T0 is nonzero. PageRank ensure both by adding the teleportation capability.

c. Figure below represents the questions and submitted answers for a Question Answering task. Correct answers are underlined. Calculate the mean reciprocal rank.

A. How far is Lahore from Karachi?
1. Distances Summary and More Information. Your Travel Starts at Lahore, Punjab, Pakistan. It Ends... 2. 13 h 50 min (1,231.7 km) via N-5
3. What is the distance between Karachi and Lahore - Find out the distance between cities
4. From Lahore to Karachi: Calculate distance between Lahore and Karachi in miles and kilomenter... 5. Lahore, Pakistan to Karachi, Pakistan distance • How many miles / kilometer from Karachi to Lahore?

Final - Information Retrieval - Fall 2014

Date: 16 Jan 2015 Marks: Time: 180 mins. Name: Roll Number:

B. How many software engineers does it take to change a light bulb?
1. None. That's a hardware problem.
2. Only one, but fixing the lightbulb will cause 4 more to go out.
3. The light bulb works fine on my machine
4. Have you tried it turning it on and off again?
5. 0.99999999999999

C. Where was Allama Iqbal born?
1. Sir Muhammad Iqbal, also known as Allama Iqbal, was a philosopher, poet and politician in India... 2. DATE OF IQBAL'S BIRTH. S. A. VAHID. The daily "Inqilab" of Lahore published the following note ... 3. Allama Iqbal's Biography. Birth. Iqbal was born in the Punjab on February 22, 1873.
4. Allama Iqbal was born 9 November 1877 in Sialkot.
5. Iqbal was born on 9th November, 1877.

D. When did Justin Bieber die?
1. NIL
2. Jan 9, 2014 - R.I.P. Justin Bieber, died crashing his Ferrari going 125 mph.
3. Has Justin Bieber died in a car crash?
4. Justin Bieber died in car accident 15 mins ago!? Watch (VIDEO)
5. Justin Bieber died in a single vehicle crash on Route 80 between Morristown and Roswell...

MMR = (1/2 +1+ 1/3 +0)/4 = 0 .458

What do the entries of C^TC represent? The (i,j) entry is the number of terms that documents i and j have in common.

d. Another method to evaluate Question Answering is to use confidence-weighted score (CWS). Systems submit one answer per question, but they place the questions for which they are most confident at the top of the return file.

$$\text{CWS} = \frac{1}{n}\sum_{i=1}^{n}\frac{c_i}{i}$$

where c_i is the number of correct answers up to rank i and n is the number of questions. Find the CWS for the above example if only the first answer is submitted and the confidence scores for the questions are given below.
conf(A) = 0.73, conf(B) = 0.53, conf(C) = 0.32, conf(D) = 0.88
Order of submission is D A B C. Note that the user submits only one answer so we consider only the top answer. CWS = ¼*(0/1+ 0/2 +1/3 +1/4) = 0 .146

Part	Marks
a	/2
b	/2
c	/1
d	/2
e	/2
f	/2

e. Suppose that C is a term-document incidence matrix.

f. Assume that machines in MapReduce have 100 GB of disk space each. Assume further that the postings list of the term *the* has a size of 200 GB. Then the MapReduce algorithm as described cannot be run to construct the index. How would you modify MapReduce so that it can handle this case?

partition by docid as well as term for very frequent terms

Final - Information Retrieval - Fall 2014

Date: 16 Jan 2015 Marks: Time: 180 mins. Name: Roll Number:

This space intentionally left blank. Trees have given their life for it. Use it wisely. Best of luck.