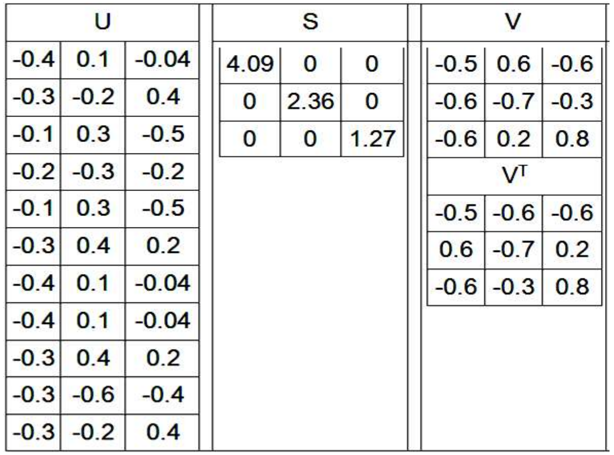
Compute the Adjacency Matrix underlying this web graph.

Compute the Probability Matrix underlying this web graph (Teleport has a probability $\alpha = 0.2$)

Compute an approximation of the PageRank score of the pages of the graph, using three iterations. (Assume surfer starts at page A).

2 + (2 + 2 + 6) = 12

9. A collection consists of the following documents:
D$_1$: Shipment of gold damaged in a fire.
D$_2$: Delivery of silver arrived in a silver truck.
D$_3$: Shipment of gold arrived in a truck.
The following document indexing rules are also used: stop words were not ignored, text was tokenized and lowercased, no stemming was used, terms were sorted alphabetically.
Our goal is to use the Latent Semantic Indexing (LSI) to rank these documents for the query gold silver truck. In order to accomplish the goal,
(i) generate the query vector. Now, assume that the term-document matrix is decomposed using singular value decomposition method and the resultant matrices are shown in table below:



U			S			V		
-0.4	0.1	-0.04	4.09	0	0	-0.5	0.6	-0.6
-0.3	-0.2	0.4	0	2.36	0	-0.6	-0.7	-0.3
-0.1	0.3	-0.5	0	0	1.27	-0.6	0.2	0.8
-0.2	-0.3	-0.2				V^T		
-0.1	0.3	-0.5				-0.5	-0.6	-0.6
-0.3	0.4	0.2				0.6	-0.7	0.2
-0.4	0.1	-0.04				-0.6	-0.3	0.8
-0.4	0.1	-0.04						
-0.3	0.4	0.2						
-0.3	-0.6	-0.4						
-0.3	-0.2	0.4						

A rank-2 approximation by keeping the first two columns of U and V and the first two columns and rows of S.

(ii) Find the new document vector coordinates in the reduced two-dimensional space and the new query vector coordinates. Rank the documents in the decreasing order of query-document cosine similarities.

(2 + 4 + 6) = 12

INFORMATION RETRIEVAL
(CSEN 6148)

Time Allotted : 3 hrs **Full Marks : 70**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and <u>any 5 (five)</u> from Group B to E, taking <u>at least one</u> from each group.

Candidates are required to give answer in their own words as far as practicable.

Group – A
(Multiple Choice Type Questions)

1. Choose the correct alternative for the following: **10 × 1 = 10**

(i) When we have indexed all the text documents in a collection, would we need to do smoothing?
(a) Yes, there will still be unseen words in individual documents
(b) No point in smoothing if we have indexed all terms
(c) No smoothing is required if we have already done stemming
(d) Smoothing would be required only in clustering.

(ii) Consider the following sentence "I left my left food in San Francisco"- how many 2-grams does the sentence have?
(a) 2 (b) 4 (c) 6 (d) 7.

(iii) What would happen if the dimension of latent space in latent semantic analysis is decreased?
(a) It would decrease recall
(b) It has no effect on recall
(c) It would increase recall
(d) Lowers recall and interpolated precision.

(iv) For O'neill, which of the following is the desired tokenization?
(a) O'neill (b) O neill (c) O' neill (d) all of the above.

(v) Stopping and stemming
(a) would reduce the size of an index being constructed
(b) would have no impact on the index being constructed
(c) stopping would decrease, and stemming would increase the size of the index being constructed
(d) stopping would increase, and stemming would decrease the size of the index being constructed.

(vi) What is shingling?
(a) Technique to detect link farming
(b) Technique to detect link spamming
(c) Technique to remove dead links
(d) Technique to detect near-duplicate web pages.

(vii) In a corpus of n documents, one document is randomly picked. The document contains a total of T terms and the term "data" appears K times. What is the correct value for the product of term-frequency and inverse-document-frequency, if the term "data" appears in approximately one-third of the total documents?
(a) KT*Log(3) (b) K*Log(3)/T (c) T*Log(3)/K (d) Log(3)/KT.

(viii) How to calculate Jaccard Index?
(a) ((The number in both sets) / (The number in either sets)) * 100
(b) ((The number in either sets) / (The number in both sets)) * 100
(c) ((The number in both sets) * (The number in either sets)) * 100
(d) none of the above.

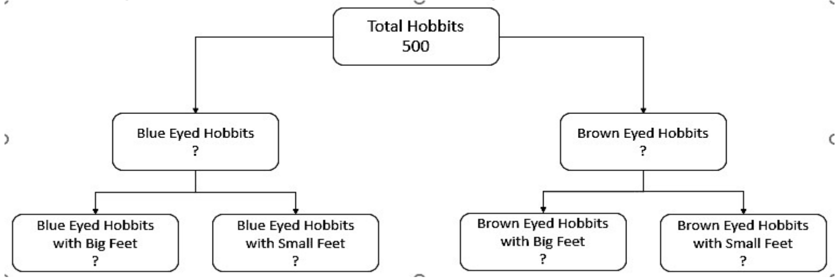
(ix) Can we run through the inverted index intersection in time O(m+n), where m and n are the length of the postings lists for the respective terms? *Information AND NOT retrieval*
(a) Yes, this is the union of the posting lists of information and retrieval
(b) No, this essentially must go over all the documents in the index
(c) No, it is bounded by O(mn)
(d) Yes, this is the difference between the posting lists of information and retrieval.

(x) You have created a document term matrix of the data, treating every tweet as one document. Which of the following statement(s) is/ are correct?
(x) Removal of stop words from the data will affect the dimensionality of data
(y) Normalization of words in the data will reduce the dimensionality of data
(z) Converting all the words in lowercase will not affect the dimensionality of the data.
(a) x and y (b) y and z (c) only x (d) x, y and z.

Group - B

2. (a) Consider you have two Posting Lists as shown:
BRUTUS:
2 → 4→6→8→10→12→14→16→18→20→22→24→26→28→30→32
CAESAR: 28
How many comparisons would be needed for a query *(BRUTUS* AND *CAESAR) u*sing Lists Augmented with Skip Pointers. Show your working, and the sequence of comparisons made to arrive at the result. (Assuming Standard Skip Lengths show the lists augmented with Skip Pointers)

7. (a) This question is about hobbits. Let's say that hobbits come in two different eye colours: Blue and Brown. Let's also say that some hobbits have Small Feet, and some hobbits have Large Feet. Now, suppose we know the following things:
•The probability that a random hobbit has brown eyes is 70%.
• 65% of blue-eyed hobbits have large feet.
• 70% of brown-eyed hobbits have small feet.
The following hierarchical structure is also given to you:


Total Hobbits
500
Blue Eyed Hobbits
?
Brown Eyed Hobbits
?
Blue Eyed Hobbits with Big Feet
?
Blue Eyed Hobbits with Small Feet
?
Brown Eyed Hobbits with Big Feet
?
Brown Eyed Hobbits with Small Feet
?

How many blue-eyed hobbits would you expect?
How many brown-eyed hobbits would you expect?
How many blue-eyed hobbits with big feet would you expect?
How many blue-eyed hobbits with small feet would you expect?
How many brown-eyed hobbits with big feet would you expect?
How many brown-eyed hobbits with small feet would you expect?

(b) If you saw a hobbit with large feet wearing sunglasses, what is the probability that this hobbit has blue eyes?If you saw a hobbit with large feet wearing sunglasses, what eye colour would you guess this hobbit had?

(c) Perform a 2-means clustering to convergence for the points below.
A: (1,1), B: (1, 2), C: (4, 5), D: (6, 7)
Start with the two seeds A and B. For each iteration, give (i) the coordinates of the centroids and (ii) the assignments of points to centroids.

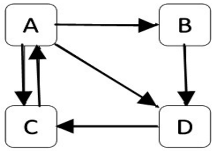
(d) Why are documents that do not use the same term for the concept ***car*** likely to end-up in the same cluster in k-means clustering?

(0.5 × 6) + (2 + 2) + 4 + 1 = 12

Group - E

8. (a) Explain with examples the terms: Polysemy and Synonymy

(b) The following toy Web Graph consists of four webpages A, B, C, and D showing the in links and out links.



(b) Consider a Bigram Index for Wildcard Queries. Give an example of a sentence that falsely matches the wildcard query *mon*h* if the search were to simply use a conjunction of bigrams. (Assume the actual indexed term is *month).*

(c) Find the Levenshtein Edit Distance between the terms *WINNING* and *WHINING*. Show the Backtrack and corresponding Alignment. [Insertion/Deletion Cost = 1, Replacement Cost = 2]
For the technique you used, what is the Time Complexity of finding Minimum Edit Distance for two terms, one of size *m* and one of size *n*?

2 + 2 + (5 + 2 + 1) = 12

3. (a) Tha table below contains the four documents along with the words stored in them.

Doc$_1$: whale, sea, sea, whale, boat, boat, boat, boat, boat
Doc$_2$: whales, sea, sea, water
Doc$_3$: whale, water, water, whale, whale
Doc$_4$: whales, whales, whales

(i) Construct the term-document incidence matrix under the assumption that the terms are not stemmed.
(ii) Using boolean model, find the documents that contain the terms water, whale, but NOT whales.

(b) What is the disadvantage of using a term-document incidence matrix? How can it be overcome?

(c) Draw the inverted index representation for the collection given in Table above.

(3 + 2) + 4 + 3 = 12

Group - C

4. (a) Define implicit and explicit feedback. Which one is more reliable and why?

(b) 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 "CD software cheap CDs" relevant and D$_2$ with content "cheap thrills DVDs" non-relevant. Assume that we are using direct term frequency (with no scaling and no document frequency). Do not length-normalize vectors. Using Rocchio relevance feedback, what would the revised query vector be after relevance feedback? Assume that α=1, β=0.8, γ=0.2.

(c) Heap's law belongs to a class of law called Power Law. Write Heap's law in the form of power law.

(d) We have 100 million documents containing a total of 9 million terms. How many posting entries are there using the simple Zipf approximation? You may assume that the natural log of 9 million is 16.

4 + 4 + 2 + 2 = 12

5. (a) How does blocked-sort based indexing differ from single pass-in-memory indexing?

(b) An unranked document retrieval approach is tested on a test set that consists of 300 documents. In response to a query, 200 documents are retrieved of which 170 documents are relevant to the query and 30 not relevant. For the entire test corpus, 190 documents are considered to be relevant for the mentioned query. Calculate precision, recall, accuracy and f-measure of the presented classifier.

(c) Why do we usually have to face a trade-off between precision and recall.

2 + 8 + 2 = 12

Group - D

6. (a) Use Query Likelihood Model to predict how the two documents will be ranked for the given query:
D1: Has 50 Terms, with apple occurring 2 times and iPad occurring 3 times
D2: Has 50 Terms, with apple occurring 3 times and iPad occurring 2 times
Entire collection has 1,000,000 terms; apple occurring 200 times and iPad occurring 100 times.
Query q: apple iPad (Use λ =0.5)

(b) Look at the following Table, where documents are represented as vectors, and their tf-idf scores are listed. For the first four documents their classification is also given. You need to Classify the Query Document D5.

Document	China	Japan	Tokyo	Macao	Beijing	Shanghai	Classification
D$_1$	0	0	0	0	1	0	C
D$_2$	0	0	0	0	0	1	C
D$_3$	0	0	0	1	0	0	C
D$_4$	0	0.71	0.71	0	0	0	C'
D$_5$	*0*	*0.71*	*0.71*	*0*	*0*	*0*	*?*

i. Use Rocchio Classification Method to classify the document D$_5$. Show your working clearly.
ii. For the same problem, this time use K-Nearest Neighbours method to classify the document D$_5$. Again, clearly show your working. (Hint: Choose your K judiciously).
iii. "KNN is called a Lazy Learner". Would you consider Rocchio to be lazy learner too? Justify your answer.

5 + (3 + 3 + 1) = 12