Section A - Answer ALL questions

1. Explain briefly why SQL queries are not an example of Information Retrieval.

   (2 marks)

2. Briefly describe how you would evaluate a search engine in terms of relevance of search results to queries.

   (4 marks)

3. A user's query text can be a poor proxy for the actual information need he/she is trying to satisfy. Explain why illustrating with a suitable example.

   (4 marks)

4. Given a search engine and no information about the retrieval model it is based on, how would you determine if it is performing ranked retrieval or boolean retrieval? (Assume multiple query terms are combined by conjunction)

   (2 marks)

5. You are designing a boolean retrieval search engine for a collection of 8 million documents, each of which is about 1000 words long. Assuming there are about 500,000 unique terms in the collection, calculate the size in bytes of the term-document incidence matrix if we store each element by a single bit.

   (1 mark)

   Using the above example as illustration, explain why the inverted index form is a much better representation for the term-document incidence matrix.
   [Hint: estimate the largest possible size of the inverted index representation]

   (3 marks)

6. What is the bag-of-words representation of the documents

   docA: ”Alice is quicker than Bob who is quicker than Charlie”

   docB: ”Charlie is quicker than Alice who is quicker than Bob”

   Assume the dictionary is [Alice, Bob, Charlie, is, quicker, than, who].

   (1 mark)

Using your answer as illustration, state the main weakness of the bag-of-words representation

(1 mark)

7. What do we mean by **document classification**? Briefly describe TWO applications of document classification.

(3 marks)

Briefly describe how K-Nearest Neighbour classification works. (2 marks)

If you were given a reliable and efficient ranked-retrieval search engine, how would you turn that into a KNN classifier?

(2 marks)

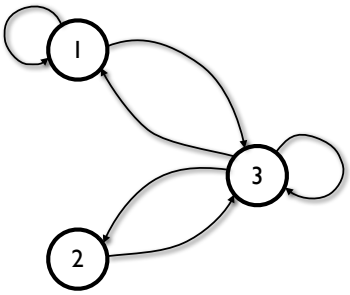
8. What are the THREE main types of user need that a search engine satisfies? Give an example for each.

(6 marks)

9. Give THREE characteristics of the **world wide web** that make it differ from ordinary document collections (e.g. a collection of legal case documents)

(6 marks)

10. This question is about **Pagerank**. The following diagram depicts a small web of three documents (nodes) with links between them (edges).



a) Assuming a random surfer model WITHOUT teleporting, write down the transition probability matrix P.

(3 marks)

b) If a_1, a_2 and a_3 are the long term visit rates for pages 1, 2 and 3 respectively, write down the system of equations that they need to satisfy.

(8 marks)

c) How would you use the solution of this system of equations for performing ranked retrieval?

(2 marks)

END OF SECTION A

Section B - Answer any TWO of the following three questions

11. This question is about **tf-idf**. A search engine uses the following term weighting scheme for a term t and a document d:

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \times \log_{10} \frac{N}{\text{df}_t}.$$

What is signified by the symbols $\text{tf}_{t,d}$, df_t and N? (3 marks)

Identify which part of the formula corresponds to **idf** (1 mark)

.

What is the weight given to a term that appears in ALL documents in the collection?

(1 mark)

Using your answer as illustration explain briefly what is the purpose of idf scores.

(4 marks)

The search engine's collection contains 1,000,000 documents, two of which are the following:

Doc1: "digital cameras and video players"

Doc2: "video cameras and digital video"

Based on this information complete the missing values in the following table:

	Doc1				Doc2			
Dictionary	tf	df	idf	tf-idf	tf	df	idf	tf-idf
and		1,000,000						
cameras						10,000		
digital		100						
video						10,000		
players		10						

(9 marks)

Using the tf-idf weights you found, calculate the **cosine similarity** between the two documents. Show all steps of your calculation.

(7 marks)

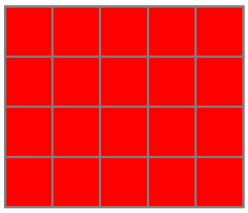
12. This question is about **content based image retrieval**.

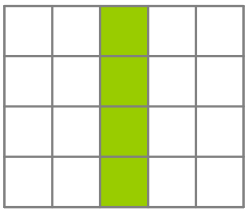
a) Briefly describe traditional image retrieval (non-content-based). (2 marks)

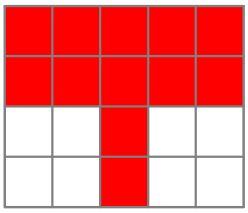
b) Identify THREE problems with this approach. (3 marks)

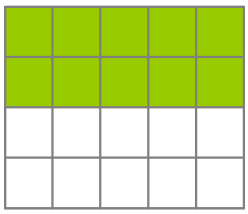
c) Content-based image retrieval is based on **image matching**, i.e. deciding whether or not two images represent the same object, person or scene. Name THREE factors that make image matching a hard problem. (6 marks)

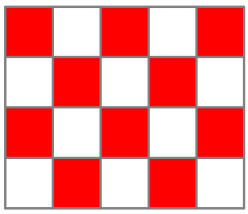
d) Colour Coherence Vectors can be used to perform image retrieval based on colour similarity. Assume we have a collection of the following 7 images shown below.

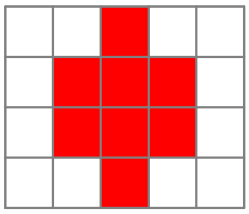

1

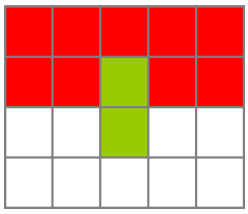

2


3

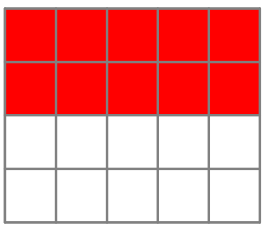

4


5


6


7

Further assume we have another image Q (shown below) that we want to use to retrieve similar images in a ”query by example” fashion.


Q

i) Compute the **Colour Coherence Vectors** of all 8 images and the query in

numerical form.

(4 marks)

ii) Perform image retrieval based on **histogram intersection**. Give all intersection similarity scores and provide the ranking for the collection of images in descending similarity.

(3 marks)

iii) By considering the score for image 5, explain how colour coherence vectors improve upon simple colour histogram intersection.

(3 marks)

e) Local Binary Patterns can be used to perform image retrieval based on texture similarity. Assume we are given a grayscale image, part of which is shown below:

12	50	60	19
10	3	50	30
50	10	3	30
50	50	10	5

Compute the Local Binary Pattern for the pixel in the third row and second column.

(2 marks)

If you were using Local Binary Patterns for image retrieval, how would you use the value computed above?

(2 marks)

13. This question is about image retrieval using the **bag of visual words** model.

a) Identify and explain TWO desirable properties of a good **keypoint detector**.

(2 marks)

b) What is a **keypoint descriptor**? (2 marks)

c) Identify and explain THREE desirable properties of a good keypoint descriptor.

(3 marks)

d) Consider a, somewhat quirky, keypoint descriptor that, for any image region, always outputs a random vector to describe it, that is equal in size to the number of pixels in the image region. Which of the properties you identified above does it satisfy? Justify your answer.

(2 marks)

e) **Zero-normalized patches** are a particularly robust keypoint descriptor. Identify TWO image intensity transformations under which this descriptor is invariant.

(2 marks)

f) We would like to create a mobile application for guiding visitors in a museum. Users of this application should be able to obtain information on exhibits around the museum by taking pictures of those exhibits using a hand-held device. Design a simple system for solving this problem using the **bag-of-visual-words** model.

(14 marks)

END OF EXAMINATION PAPER