Section A - Answer ALL questions

1. On a visit to the University Library you ask the librarian the following two questions: (a) ”My library card number is 32515, can you please check if there are any fines I need to pay?” and (b) ”Can you point me towards some good books on the early Bronze age in the Middle East?” Which of these two questions is an **Information Retrieval** task and why isn't the other one such a task?

   (3 marks)

2. Evaluating search engines is a very difficult task. This is partly because the notion of a good search engine can vary widely depending on the user. Illustrate this by considering how we might **define** AND **measure** a successful user experience if the user is (a) a visitor to a web search engine and (b) a potential buyer in an e-commerce website.

   (4 marks)

3. User satisfaction from a search engine is commonly associated with the **relevance** of the search results to the user's **information need**. The measurement of relevance to user queries is problematic. Explain why with an appropriate example different to the one given in the notes.

   (2 marks)

4. What is **ranked retrieval** ? Give one reason why it is preferable to the boolean model.

   (3 marks)

5. What is the **bag-of-words** model for documents? (2 marks)

   What is the difference from the **set-of-words** model? (2 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].

(2 marks)

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

(2 marks)

7. What do we mean by **document clustering**? Briefly describe TWO applications of document clustering to information retrieval.

(3 marks)

8. Briefly describe the process of **tokenization** and describe THREE types of problem (with examples) faced by tokenizers of English.

(6 marks)

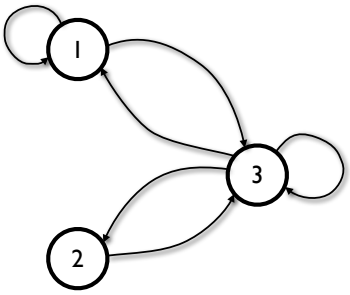
9. What are the THREE main types of user need that a search engine satisfies? Give an example for each.

(6 marks)

10. 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)

11. 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 the surfer is originally in page 1, calculate the probability vector for the state

of the surfer after TWO transitions. (6 marks)

END OF SECTION A

Section B - Answer any TWO of the following three questions

12. This question concerns **boolean retrieval**. We are given the following collection of (very short) documents:

Doc1: ”occupy london protesters evicted by police”

Doc2: ”police evicted cathedral protesters”

Doc3: ”police dismantle london occupy camp”

Doc4: ”london protesters occupy cathedral”

a) Define the **boolean model** for information retrieval. State an example of where this model is used. (5 marks)

b) Write down the term-document incidence matrix for the document collection given above. (10 marks)

c) Draw the inverted index representation for this collection. (5 marks)

d) Using the two representations you computed as illustration, explain why inverted indices are much more memory efficient than the full term-document incidence matrix. (3 marks)

e) For the document collection above, what are the returned results for these queries?

i) cathedral AND evicted (1 mark)

ii) london AND NOT (by OR dismantle) (1 mark)

13. 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. Using the two images of the Notre Dame cathedral shown below, identify THREE factors that make image matching a hard problem.

(6 marks)

d) 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 second row and third column. (2 marks)

If you were using Local Binary Patterns for image retrieval, how would you use the value computed above? (2 marks)

e) In video editing, a very common task is to identify shot cuts in film footage (points in the film where the scene changes abruptly). These shot cuts can then be used to delete entire shots, insert new ones etc. Describe briefly how you would use Colour Histogram Intersection to automate this task.

(10 marks)

14. 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) Design a simple system for image retrieval using the **bag of visual words** model. Make sure to describe both the necessary off-line pre-processing of the image collection as well as the runtime processing of the query image.

(14 marks)

END OF EXAMINATION PAPER