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)

Model Answer:

Question (b) is an IR task (1 mark) *because the material is unstructured (text)* (1 mark) *. Question (a) can be answered by a simple query inside the appropriate database (structured data)* (1 mark) *.*

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)

Model Answer:

(a) Success = visitor finds what they were looking for. (1 mark) *Measure = number of return visits* (1 mark) *. (b) Success = user buys what they were looking for.* (1 mark) *Measure = fraction of visitors that are converted to buyers or average time from visit to purchase.* (1 mark)

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)

Model Answer:

Example: Information need = User needs to find out information about the causes of the latest financial crisis. Query = ”causes of the financial crisis”. Result = Article on BBC news entitled: ”Financial crisis causes a 10% increase in cases of clinical depression.” (1 mark)

Result is almost perfect match for user’s query but actually irrelevant to the user’s information need. Hence relevance of results to queries cannot be used to measure user satisfaction. (1 mark)

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

(3 marks)

Model Answer:

In ranked retrieval the system returns an ordering of relevance of the documents in the collection with respect to the query. (1 mark) *Ranked retrieval does not suffer from the feast-or-famine problem of boolean retrieval (i.e. getting either too few or too many results). In ranked retrieval the user is not overwhelmed with too many documents returned since these documents are ordered so we may only show the top N.* (2 marks)

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

Model Answer:

Each document is represented as a vector of dimension equal to the total number of terms in the collection dictionary. Each vector element that corresponds to term t in the dictionary is equal to the number of times t appears in the document. (2 marks)

What is the difference from the **set-of-words** model? (2 marks)

Model Answer:

In the set-of-words model the vector elements are 1 or 0 corresponding to whether the term appears in the document or not. (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)

Model Answer:

$$v\left(docA\right)=\begin{bmatrix}1\\1\\1\\2\\2\\2\\1\end{bmatrix}$$

(1 mark)

$$v\left(docB\right) = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \\ 1 \end{bmatrix}$$

(1 mark)

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

(2 marks)

Model Answer:

The main weakness of the bag-of-words representation is that it ignores word order. In this example, the two documents consist of the same words in different order. Hence they have identical bag-of-words representations despite the fact that their meaning due to word order is very different. (2 marks)

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

(3 marks)

Model Answer:

Document clustering is the process of grouping a document collection into a number of classes so that documents within the same class are similar, while documents in different classes are dissimilar. (1 mark)

Applications include: Better navigation of search results, Improving search recall (when query matches document D, return other documents in same cluster as D), Improving speed of retrieval (searching on cluster centres rather than whole collection) etc. (2 marks)

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

(6 marks)

Model Answer:

Tokenization is the process during which a continuous string of document characters is split into a series of smaller character sequences, each of which (after normalization) is a candidate for index entry. (3 marks) *Problems faced English tokenizers are (1 mark each)*

- *Apostrophes (Finland's capital -¿ Finland? Finlands? Finland's?)*
- *Hyphens Hewlett-Packard, state-of-the-art*
- *One or two tokens (San Francisco)*

- *Numbers (18/10/1979, B-52, 0800 366 277)*

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

(6 marks)

Model Answer:

Informational: user wants to learn about something. (1 mark) *E.g. "quantum mechanics"* (1 mark) *Navigational: user wants to go to a particular page.* (1 mark) *E.g "general motors"* (1 mark) *Transactional: user wants to do something (web-mediated).* (1 mark) *E.g. "rent a car in rome"*

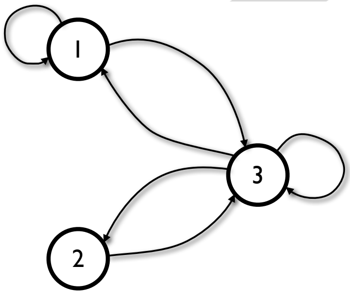
10. Give THREE characteristics of the **world wide web** that make it differ from ordinary document collections (e.g. a collection of legal case documents)

(6 marks)

Model Answer:

Two marks for each. Possibilities are: (1) The web has no design/coordination (2) Distributed content creation, linking, democratization of publishing (3) Content includes truth, lies, obsolete information, contradictions (4) Unstructured (text, html, etc.), semi-structured (XML, annotated photos), structured (Databases) (5) Scale much larger than previous text collections (6) Phenomenal growth

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)

Model Answer:

$$P = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

b) If the surfer is originally in page 1, calculate the probability vector for the state of the surfer after TWO transitions.

(6 marks)

Model Answer:

[5/12, 1/6, 5/12]

END OF SECTION A

Total: 50 marks

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)

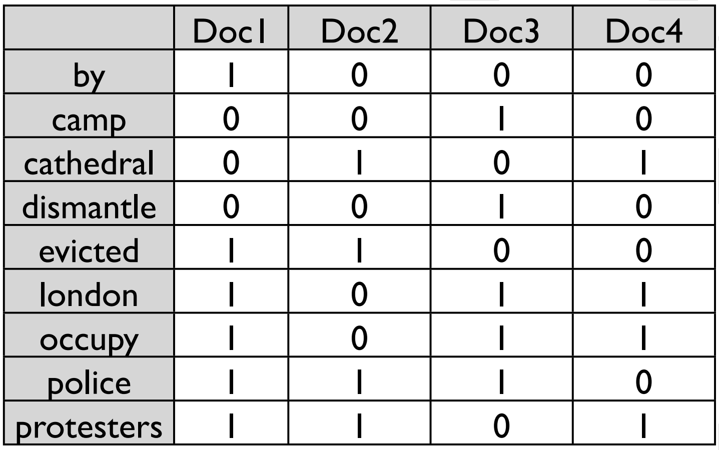
Model Answer:

Under the boolean model a query is given as a set of terms that are joined with boolean operators. To evaluate a document against the query, we substitute each term in the query with True or False depending on whether the term appears in the document or not. We then compute the truth-value of the resulting boolean expression. If it is True, the document is deemed to be relevant. Otherwise it is discarded. (4 marks) *Examples of boolean retrieval are email searching, library catalogues, Mac OS X Spotlight, Patent/legal case searching etc.* (1 mark)

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

(10 marks)

Model Answer:

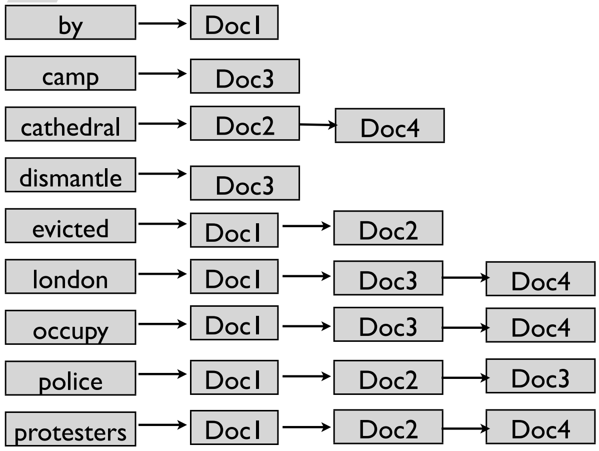


	Doc1	Doc2	Doc3	Doc4
by	1	0	0	0
camp	0	0	1	0
cathedral	0	1	0	1
dismantle	0	0	1	0
evicted	1	1	0	0
london	1	0	1	1
occupy	1	0	1	1
police	1	1	1	0
protesters	1	1	0	1

(10 marks)

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

Model Answer:


by → Doc1
camp → Doc3
cathedral → Doc2 → Doc4
dismantle → Doc3
evicted → Doc1 → Doc2
london → Doc1 → Doc3 → Doc4
occupy → Doc1 → Doc3 → Doc4
police → Doc1 → Doc2 → Doc3
protesters → Doc1 → Doc2 → Doc4

(6 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)

Model Answer:

The full term-document incidence matrix is sparse in general (consists mostly of zeros). The inverted index representation stores only the indices of the non-zero positions and hence is much more memory-efficient. In the example above, we need 40 numbers for the full matrix representation versus 18 numbers for the inverted index representation (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)

Model Answer:

(i) doc2. (1 mark) *(ii) doc4.* (1 mark)

13. This question is about **content-based image retrieval**.

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

Model Answer:

A human manually labels each image in the collection with a set of textual tags. At query time the user enters a text query that is then matched against the tags in the collection using traditional text retrieval techniques (tf.idf etc).

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

Model Answer:

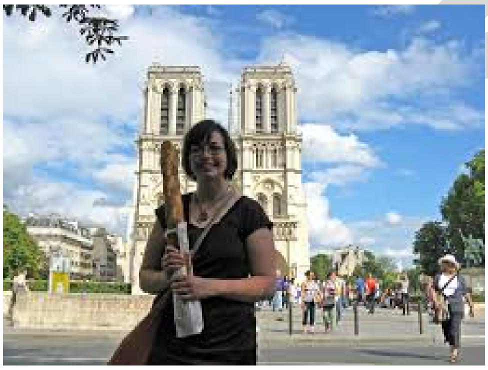
1. Descriptions from different persons will differ

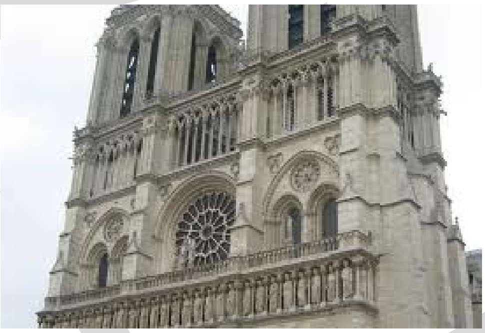
2. Words convey inexact information

3. Not everything can be expressed with words (e.g. mood, visual properties like texture etc)

4. Impractical - Too many images

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)

Model Answer:

Two points each. Possibilities include (1) Colour variation due to different pixel sensors, (2) Occlusion, (3) Perspective distortion, (4) lighting, (5) scale and (6) viewpoint variation.

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)

Model Answer:

192

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

(2 marks)

Model Answer:

When coming across a LBP value of 192 this would increase the 192nd bin of the LBP histogram.

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)

Model Answer:

A video sequence is just a set of images given in a temporal order. We can assume that during a shot the difference in appearance between one frame and the next is very small. However when a cut occurs, the appearance variation will be significantly larger.(3m) One possible algorithm is to compute colour histograms for each frame in the video and then compute Histogram Intersection between frame t and frame t+1. (3m) Whenever the Histogram Intersection score drops below a certain threshold (say 40%) between frames t and t+1, then there is a good chance that a shot cut has occured between those two frames.(4m)

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)

Model Answer:

1. The local image structure around the interest point is rich in terms of local information content

2. It is stable under local and global perturbations in the image domain (due to viewpoint changes, illumination/brightness variations etc) such that the interest points can be reliably computed with high degree of reproducibility.

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

Model Answer:

A keypoint descriptor is a mathematical process for mapping a local image region to a real-valued n-dimensional vector.

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

(3 marks)

Model Answer:

1. Robust: Descriptor is invariant under local and global perturbations in the image domain.

2. Distinctive: Visually different image regions will correspond to dissimilar descriptor vectors.

3. Sparse: The number of elements of the vector is much less than the number of pixels in the image region described.

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)

Model Answer:

This descriptor would be 100% distinctive as every patch gets a completely different descriptor. However it suffers from complete lack of robustness as even the same image region in two different images will be described with entirely different vectors. It is also not sparse as it has the same number of elements as the number of pixels in the image region.

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

(2 marks)

Model Answer:

1. Multiplying all pixel intensities in the region by a constant.

2. Adding a constant to all pixel intensities in the region.

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)

Model Answer:

In our pre-processing step, we run a keypoint detector and descriptor on each image in our collection. We then collect the descriptor vectors from all images and run a clustering algorithm on these vectors (e.g. k-means). Alternatively we use some other type of quantisation algorithm, e.g. k-d trees. The goal is to quantise the vectors so that descriptors corresponding to the same visual object in different images are identified as being the same. (4m)

For each image in the collection we can form a histogram where the bins are the identified descriptor clusters and each bin counts how many keypoint descriptors belong to the corresponding cluster. (2m)

When a query image arrives, this can be converted into the same type of histogram as follows: Run the same keypoint detector and descriptor on the query image. For each descriptor vector, find the nearest cluster (in Euclidean distance). In each bin, count all descriptor vectors that have picked that cluster as the nearest. (4m)

Now that our query image has been reduced to a histogram of quantised descriptors we can use standard text IR techniques like TF.IDF to perform ranked retrieval. The term and document frequencies of TF.IDF will be replaced by the frequencies that the quantised descriptor vectors appear in the collection and query images.(4m)

END OF EXAMINATION PAPER

Total: 125 marks