1 Solution General Questions (20%)

a Data Retrieval vs Information Retrieval:

- IR retrieval system: explain briefly the main concept.
- **IR: representation, storage, organization of, and access to information items.**
- Explain the main differences between Data Retrieval and Information Retrieval in terms of matching and content.
- **Matching - Data Retrieval: Exact Match, InformationRetrieval: Parial Match or Best Match. Content - Data Retrieval: Data, InformationRetrieval: Information.**

b IR Sytem architecture:

- Explain briefly the input and the output of a query expansion step in an IR system. What are the benefits?
- **Input: Given query (after eventually preprocessing step). Output: Expanded query with eventually reweighted terms. Benefits: expanded query closer to user needs and improve evaluation measures**
- What is the purpose of the indexing process and why is there a need for an index in an IR system?
- **Why Index: fast answer**

c Explain briefly the concepts of "user need", "query" and explain the relation between them in the context of an IR system.

d **user need: objective of the user in a search process. Query - translation of the user need in a set of keywords**

2 Solution Multimedia IR (20%)

a Briefly explain the concept of semantic gap in multimedia retrieval.

b **Distance between high level concept (sites, objects, events) and low-level visual/audio features (colour, texture, shape and structure, layout; motion; audio - pitch, energy, etc.). While we have a small semantic gap between a table and its meaning, there is a larger gap between a document and its overall meaning. Further there is also a larger gap between a video and its semantics**

c What are the main differences between Text-based information retrieval and Content-based information retrieval?

d **TBIR: query is a text and serach performed over text sourrounding images. Index based on text. CBIR: query by example (quesry is an image). Index by raw content of the images (features)**

e Histogram color:

- advantages and disadvantages of using color histogram as feature
- **Advantages: Color histogram independent from image resolution. Drawback: Quantization effect**
- explain the steps involved in using color histogram as feature in an image retrieval system
- **General steps: 1 - selection of a color space, 2 - quantization of the color space, 3 - computation of histograms, 4 - derivation of the histogram distance function, 5 - identification of indexing shortcuts**

3 Solution Evaluation

Rank	Document ID	Score	Relevant	Precision	Recall
1	24	.8		0/1	0/7
2	38	.64	x	1/2	1/7
3	40	.62	x	2/3	2/7
4	28	.4	x	3/4	3/7
5	36	.32		3/5	3/7
6	48	.3		3/6	3/7
7	22	.28	x	4/7	4/7
8	44	.26	x	5/8	5/7
9	32	.1		5/9	5/7
10	60	.05		5/10	5/7

b) in web search you don't know the number of relevant documents out there, so recall is not really possible to compute

c) advantage is that it is a single value measure instead of two measures with precision and recall. For the computation we look at precision at all recall levels. This leads to:

$$(1/2.0 + 2/3.0 + 3/4.0 + 4/7.0 + 5/8.0)/5 = 0.623$$

4 Solution Modelling

All results are for log2 and the given equations.

a) document frequencies

```
doc freq: {'stavanger': 1.0, 'oslo': 1.0, 'university': 5.0,
```

```
'northernmost': 1.0, 'norwegian': 1.0, 'tromso': 1.0,
'science': 1.0, 'trondheim': 1.0, 'technology': 1.0, 'bergen': 1.0,
'norway': 3.0}
```

b) term frequencies:

```
({'university': 2, 'norwegian': 1, 'technology': 1, 'trondheim': 1, 'science': 1})
({'university': 2, 'oslo': 1, 'norway': 1})
({'bergen': 2, 'university': 1, 'norway': 1})
({'university': 2, 'tromso': 1, 'northernmost': 1})
({'stavanger': 2, 'university': 2, 'norway': 1})
```

c) boolean retrieval:
and query: no matching results
or query: doc1, doc2, doc3, doc5
d) vector space retrieval

```
tokens: 11
(['stavanger', 'oslo', 'university', 'northernmost', 'norwegian', 'tromso',
'science', 'trondheim', 'technology', 'bergen', 'norway'])

query vector (science norway):
{'science': 1, 'norway': 1, 'northernmost': 0}
[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 2.321928094887362, 0.0, 0.0, 0.0, 0.7369655941662062]
```

ranking:

```
{'norwegian': 1, 'university': 2, 'technology': 1, 'trondheim': 1, 'science': 1}
[0.0, 0.0, 0.0, 0.0, 2.321928094887362, 0.0, 2.321928094887362, 2.321928094887362, 2.321
score: 0.476571274172
{'oslo': 1, 'university': 2, 'norway': 1}
[0.0, 2.321928094887362, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.7369655941662062]
score: 0.0915192825368
{'university': 1, 'norway': 1, 'bergen': 2}
[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 4.643856189774724, 0.7369655941662062]
score: 0.0474158668557
{'stavanger': 2, 'university': 2, 'norway': 1}
[4.643856189774724, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.7369655941662062]
score: 0.0474158668557
{'university': 2, 'tromso': 1, 'northernmost': 1}
[0.0, 0.0, 0.0, 2.321928094887362, 0.0, 2.321928094887362, 0.0, 0.0, 0.0, 0.0, 0.0]
score: 0.0
```