

---

## CS-431 Hands On Information Retrieval

---

### **Question 1:**

**[1 pt]**

When a user submits a query  $Q$  to a vector space Information Retrieval (IR) system operating on a collection of documents, the main goal of the system is:

(Select only one answer)

- to summarize the documents relevant for the query  $Q$
- to filter out all the documents that strictly match the query  $Q$
- to identify the documents that are talking about the topic expressed in the query  $Q$
- to extract the correct answers to the query  $Q$

### **Question 2:**

**[1 pt]**

In the standard tf.idf weighting scheme, what guarantees that the indexing terms that occur most often in a document are given a higher weight?

(Select only one answer)

- the tf component
- the idf component
- both components

**Question 3:****[2 pts]**

In the standard VS-model, is it possible for the document

D = “recycling aluminum can be crucial for the environment”

to be considered as relevant for the query

Q = “ecologic impact of metal”?

(Select only one answer)

yes

no

undecidable

**Question 4:****[2 pts]**

What is the similarity between a document D and the document resulting from the concatenation of 10 copies of D, provided that the cosine similarity measure is used?

(Select only one answer)

0

1/10

1

10

cannot be computed in this general case

**Question 5:****[2 pts]**

For a given query Q, an IR system retrieves 50 documents with a precision  $P = 0.6$ .

If one assumes that the total number of relevant documents for Q is  $N = 100$ , what is the recall R of the system for Q?

(Provide the answer in a form of a fraction)

R =

**Question 6:****[5 pts]**

Two IR systems,  $S_1$  and  $S_2$ , have been evaluated, and, for each of them, the evaluation resulted in the following (Recall, Precision) pairs:

$S_1$	$S_2$
(0.01, 0.80)	(0.01, 0.50)
(0.20, 0.40)	(0.20, 0.40)
(0.60, 0.10)	(0.60, 0.30)

**6.1 [3 pts]** Based on these evaluation results, which system should be selected for an IR application where it is important to find all the documents that are relevant for a given query?

(Select only one answer)

- system  $S_1$
- system  $S_2$
- cannot be decided

**6.2 [2 pts]** Based on these evaluation results, which system should be selected for an IR application performing general purpose IR from the Web?

(Select only one answer)

- system  $S_1$
- system  $S_2$
- cannot be decided