

UNIVERSITY OF MOHAMED BOUDIAF – M'SILA
FACULTY OF MATHEMATICS AND INFORMATICS
DEPARTMENT OF COMPUTER SCIENCE
3rd year Bachelor (ISIL) - University year: 2024/2025
Biannual Exam of Information Retrieval

By Dr. B. LOUNNAS

Exercise 01: Multiple Choice Questions (MCQs) (08pts)

Choose the correct answer for each question.

1. In the vector space model, cosine similarity is used to:
A) Count term frequency
☒ B) Measure the angle between document and query vectors
C) Sort terms alphabetically
D) Normalize document lengths to zero
2. If a search engine retrieves 10 documents, 6 of which are relevant, and there are 12 relevant documents in total, what is the precision?
A) 0.5
☒ B) 0.6
C) 0.8
D) 1.0
3. What is the time complexity of merging two sorted posting lists of length m and n using the two-pointer technique?
☒ A) $O(m + n) = O(n)$
B) $O(m * n) = O(n^2)$
C) $O(\log m + \log n)$
D) $O(\max(m, n))$
4. Which of the following is NOT an advantage of a positional index?
A) Allows proximity and phrase queries
B) Increases precision for certain queries
☒ C) Reduces storage space compared to an inverted index
D) Enables advanced query processing
5. Which of the following statements is TRUE about precision and recall?
☒ A) There is often a trade-off between precision and recall
B) Increasing recall always increases precision
C) Increasing the number of retrieved documents always increases precision
D) Precision and recall always have the same value
6. PageRank is mainly based on:
A) Number of terms in a document
B) Length of the document
C) Cosine similarity
☒ D) Number and quality of incoming links
7. A positional index is required to support which type of query?
A) Boolean queries
☒ B) Phrase queries
C) Fuzzy queries
D) Term frequency queries
8. What is the purpose of the robots.txt file?
A) Stores page rank
B) Tracks broken links
☒ C) Tells crawlers which pages they can or cannot visit
D) Ranks search results

Exercise 02: Vector Space Model (06pts)

You are given the following small collection of three documents: D1= "science is fun", D2= "machine learning is part of data", D3= "science and data are important" and Query = "data science"

- Compute the cosine similarity between the query and document 3 and 1 using LTC.LNC smart notation

	Science	is	Fun	machine	learning	part	of	Data	and	are	important
df	2	2	1	1	1	1	1	2	1	1	1
idf	0.176	0.176	0.477	0.477	0.477	0.477	0.477	0.176	0.477	0.477	0.477

doc1	TF _{raw}	idf	FV	FV _{norm}	doc3	TF _{raw}	idf	FV	FV _{norm}
Science	1	0.176	0.176	0.33	Science	1	0.176	0.176	0.20
is	1	0.176	0.176	0.33	and	1	0.477	0.477	0.54
Fun	1	0.477	0.477	0.89	data	1	0.176	0.176	0.20
					are	1	0.477	0.477	0.54
					important	1	0.477	0.477	0.54

$$\text{length} = \sqrt{(0.176)^2 + (0.176)^2 + (0.477)^2} = 0.533$$

$$L_{2\text{norm}} = \sqrt{(\dots)^2 + \dots} = 0.1871$$

query	TF _{raw}	idf	FV	FV _{norm}
data	1	1	1	0.70
Science	1	1	1	0.70

$$\text{Sim}(Q, d_1) = 0.33 * 0.70 = 0.231$$

$$\text{Sim}(Q, d_3) = (0.20 * 0.70) + (0.20 * 0.70) = 0.280$$

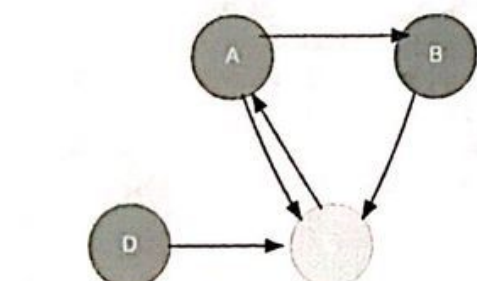
$$L_2 = \sqrt{1^2 + 1^2} = 1.41$$

Exercise 03: PageRank (06pts)

Consider a small web of 4 pages with the following link structure

1. Construct the adjacency matrix and transition matrix for this web with teleportation rate equal to 0.1

2. Give page rank for each pages, using power iteration method (Perform only two iterations)



$$A_{adj} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

$$A_{\text{norm}} = \begin{pmatrix} 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

$$d = 0.1$$

$$T = \begin{pmatrix} 0.1025 & 0.1475 & 0.1475 & 0.1025 \\ 0.1025 & 0.1025 & 0.1925 & 0.1025 \\ 0.1925 & 0.1025 & 0.1025 & 0.1025 \\ 0.1025 & 0.1025 & 0.1925 & 0.1025 \end{pmatrix}$$

choosing initial vector

$$n_0 = (0.125 \quad 0.125 \quad 0.125 \quad 0.125)$$

$$n_0 t = n_1 = (0.125 \quad 0.1375 \quad 0.1875 \quad 0.1025)$$

$$n_1 t = n_2 = (0.1553 \quad 0.137 \quad 0.283 \quad 0.025)$$

PR_A

PR_B

PR_C

PR_D