
UNIVERSITY OF MOHAMED BOUDIAF – M'SILA
FACULTY OF MATHEMATICS AND INFORMATICS
DEPARTMENT OF COMPUTER SCIENCE
2nd year Master (IDO)

Time duration: 1h:30m - Biannual Exam of Information Retrieval & Data Mining - University year: 2018/2019

By Dr. B. LOUNNAS

Exercise 01: Course question (06pt)

1. Does the indexation of data has any role in the process of information retrieval, and why? (1pt)
2. In what case we use multilevel indices? (1pt)
3. What is the difference between Information retrieval and Data mining? (0.5pt)
4. One of the differences of exact and best matching is: (0.5pt)
 - a. **Exact matching:** Query specifies precise retrieval criteria.
 - b. **Best matching:** Query describes retrieval criteria for desired documents
What does that means?
5. Why we moved from term-document incidence to inverted index? (1pt)
6. What is the difference between the second and the third phases of CRISP-DM (Data understanding, and Data preparation)? (1pt)
7. We have three attributes: Age, Salary, and Position. After calculation of information gain we found that Salary attribute is the best choose to be a root.
The question is, if you did not choose Salary as a root, instead you used Age as a root. Would your decision tree gives false result or not? (1pt)

Exercise 02: Information Retrieval Models (08pt)

1. We have the following term-document incidence: (3pt)

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia						
Cleopatra	1	0	0	0	0	0

- What is the result of the following query: **(Brutus OR Caesar) AND NOT (Antony OR Cleopatra)**
 - Complete the values of Calpurnia based on the following:
 - Document Julius Caesar mentioned the word Calpurnia 156 times.
 - Documents The Tempest, Antony and Cleopatra, Hamlet, Othello, and Macbeth never mentioned the word Calpurnia.
 - After completing the values of Calpurnia, and assuming that those terms are the only ones, what is the irrelevant document of this retrieval system?
2. Describe by graphical how the Merge Algorithm works on the following inverted indices: (2pt)

Brutus → 1 → 2 → 4 → 11 → 31 → 45 → 173 → 174

Calpurnia → 2 → 31 → 54 → 101

(Obs: Merge algorithm in the above example return 2, and 31 in linear time complexity $O(n)$)

- Write the algorithm?

3. Considering the following table of count vector (Tf_{raw}) of a 3 documents and query: (3pt)

	Doc 1	Doc 2	Doc 3	Query
Two	2	0	0	0
Tea	2	2	0	1
Me	0	1	2	1
You	0	1	2	0

- Calculate the idf (inverse document frequency) for each word represented in the three documents.
- What is the result of the query (Tea me) using NTC.NNN smart notation?

Exercise 04: Decision tree (06pt)

Imagine you only ever do four things at the weekend: go shopping, watch a movie, play tennis or just stay in. What you do depends on three things: the weather (windy, rainy or sunny); how much money you have (rich or poor) and whether your parents are visiting.

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay in
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema

- Calculate the entropy of this collection of training examples. (2.5pt)
- Calculate the information gains of Weather, Parents and Money relative to these training examples. (2.5pt)
- What is the best split (among Weather, Parents, and Money) according to the information gain? Explain? (1pt)