
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: 2019/2020
By Dr. B. LOUNNAS

Exercise 01: Course question (06pt)

1. Cites some differences between exact match retrieval and best match retrieval? (1pt)
2. One of the pre-step in any retrieval model is Stemming. What's the benefit of this task? (0.5pt)
3. CRISP-DM defined and validated a data mining process that could be applicable in any domain application sectors. Shortly what is the objectives of each step? (2pt)
4. What are the steps of building a classifier? (1pt)
5. What is the purpose of calculating *idf* in vector space model? (1pt)
6. In Boolean model we have encountered a problem called "Feast or Famine", what is this problem? (0.5pt)

Exercise 02: Information Retrieval Models (05pt)

You have the following values of *tf* of three (3) documents from a collection of 806791 documents, and the values of *tf* and *idf* of four (4) terms:

	Doc 1	Doc 2	Doc 3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

term	df_t	idf_t
car	18165	1.65
auto	6723	2.08
insurance	19241	1.62
best	25235	1.5

Compute the two top scoring documents on the query "best car insurance" for each of the following SMART notation schemes:

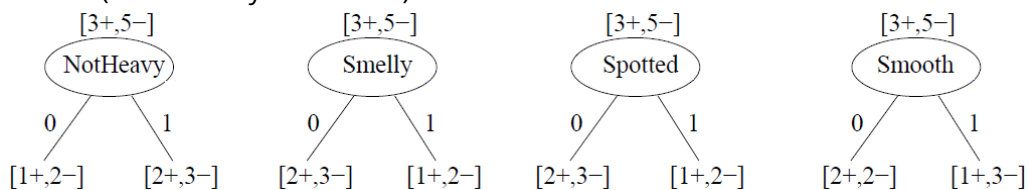
- nnn.ltc
- ntc.nnn

Exercise 03: Decision tree (05pt)

You are stranded on a deserted island. Mushrooms of various types grow widely all over the island, but no other food is anywhere to be found. Some of the mushrooms have been determined as poisonous and others as not (determined by your former companions' trial and error). You are the only one remaining on the island. You have the following data to consider:

Example	NotHeavy	Smelly	Spotted	Smooth	Edible
A	1	0	0	0	1
B	1	0	1	0	1
C	0	1	0	1	1
D	0	0	0	1	0
E	1	1	1	0	0
F	1	0	1	1	0
G	1	0	0	1	0
H	0	1	0	0	0
U	0	1	1	1	?
V	1	1	0	1	?
W	1	1	0	0	?

1. What is the entropy of Edible?
2. Which attribute should you choose as the root of a decision tree? Based only from the picture below (without any calculation)



3. What is the information gain of the attribute you chose in the previous question?
4. Build the decision tree based only on observations like you do in question 2?
5. Classify mushrooms U, V and W using the decision tree?

Exercise 04: Association rules (04pt)

Trace the results of using the Apriori algorithm on the grocery store example with support threshold $s=33.34\%$ and confidence threshold $c=60\%$. Show the candidate and frequent itemsets for each database scan. Enumerate all the final frequent itemsets. Also indicate the association rules that are generated and highlight the strong ones.

Transaction ID	Items
T1	HotDogs, Buns, Ketchup
T2	HotDogs, Buns
T3	HotDogs, Coke, Chips
T4	Chips, Coke
T5	Chips, Ketchup
T6	HotDogs, Coke, Chips

(NB: support threshold $s=33.34\%$ $\Rightarrow$ threshold is at least 2 transactions.)