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

_____

Correction of exam of Information Retrieval & Data Mining **University year: 2019/2020**

Exercise 01: Course question (06pt)

1. Cites some differences between exact match retrieval and best match retrieval? (1pt)
   - Exact match does not rank results best does.
   - Problem of "Feast or Famine" could be found in exact match.
   - Exact match simple in implementation.

2. One of the pre-step in any retrieval model is Stemming. What's the benefit of this task? (0.5pt)
   - It will help the IR system to deal only with the base term (root) of a set of driven terms.

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)
   - Business understanding: Focuses on understanding the project objectives and requirements from a business perspective.
   - Data understanding: get familiar with the data.
   - Data preparation: Covers all activities to construct the final dataset from the initial raw data.
   - Modeling: to choose a model to be working with, in order to provide answer of the first step.
   - Evaluation: evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objectives.
   - Deployment: The knowledge gained will need to be organized and presented in a way that the customer can use it.

4. What are the steps of building a classifier? (1pt)
   - Training the model using training data.
   - Testing the model to ensure it is results.

5. What is the purpose of calculating *idf* in vector space model? (1pt)
   - Discrimination between rare terms and frequents terms.

6. In Boolean model we have encountered a problem called "Feast or Famine", what is this problem? (0.5pt)
   - Due to the use of Boolean operators, the system sometimes give a lot of results and other time the system give a few results.

Exercise 02: Information Retrieval Models (05pt)

1- NNN.LTC

NNN weights for documents

Term	***Doc 01***	***Doc 02***	***Doc 03***
Car	27	4	24
Auto	3	33	0
Insurance	0	33	29
Best	4	0	17

LTC weights for query

	Query				***Product***		
Term	***Tf w***	***idf***	***Tf-idf***	***Norm V***	***Doc 01***	***Doc 02***	***Doc 03***
Car	1	1.65	1.65	0.6	16.2	2.4	14.4
Auto	0	2.08	0	0	0	0	0
Insurance	1	1.62	1.62	0.59	0	19.24	17.1
Best	1	1.5	1.5	0.54	2.16	0	9.18

- Score (q, Doc 01) = 16.2+2.16 = 18.36.
- Score (q, Doc 02) = 2.4+19.24 = 21.64.
- Score (q, Doc 03) = 14.4+17.1+9.18 = 40.68.

So the result of this schema are: ***Doc 03***, ***Doc 02***.

2- NTC.NNN

LTC weights for Doc 01

Term	***Tf w***	***idf***	***Tf-idf***	***Norm V***
Car	27	1.65	44.55	0.98
Auto	3	2.08	6.24	0.13
Insurance	0	1.62	0	0
Best	4	1.5	6	0.13

LTC weights for Doc 02

Term	***Tf w***	***idf***	***Tf-idf***	***Norm V***
Car	4	1.65	6.6	0.07
Auto	33	2.08	68.64	0.78
Insurance	33	1.62	53.46	0.61
Best	0	1.5	0	0

LTC weights for Doc 03

Term	***Tf w***	***idf***	***Tf-idf***	***Norm V***
Car	24	1.65	39.6	0.59
Auto	0	2.08	0	0
Insurance	29	1.62	46.98	0.70
Best	17	1.5	25.5	0.38

NNN weights for query

	Query	***Product***		
Term	***Tf raw***	***Doc 01***	***Doc 02***	***Doc 03***
Car	1	0.98	0.07	0.59
Auto	0	0	0	0
Insurance	1	0	0.61	0.70
Best	1	0.13	0	0.38

- Score (q, Doc 01) = 0.98+0.13 = 1.11
- Score (q, Doc 02) = 0.07+0.61 = 0.68
- Score (q, Doc 03) = 0.59+0.70+0.38 = 1.67

So the results in this schema are: ***Doc 03***, ***Doc 01***.

Exercise 03: Decision Tree (05pt)

1. Entropy of Edible:

$$-\frac{3}{8}\cdot\log_2\frac{3}{8}-\frac{5}{8}\cdot\log_2\frac{5}{8}\approx\ 0.9544$$

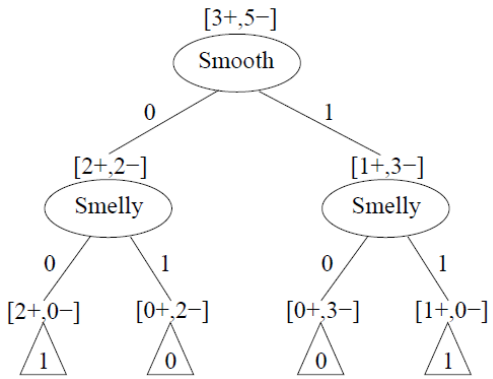
2. ***Smooth*** is the attribute who will be chosen as root.
3. Information gain of Smoot:

$$E_{/Smooth} = \frac{1}{2}\cdot 1+\frac{1}{2}\left(\frac{1}{4}\log_2\frac{1}{4}+\frac{3}{4}\log_2\frac{3}{4}\right)$$

$$\approx 0.4056$$

$$IG_{0/Smooth} = 0.9544-0.4056 = 0.5488$$

4. Building the decision tree



5. Classifying mushrooms U,V and W:

U	$Smooth = 1,\ Smelly = 1 \Rightarrow Edible = 1$
V	$Smooth = 1,\ Smelly = 1 \Rightarrow Edible = 1$
W	$Smooth = 0,\ Smelly = 1 \Rightarrow Edible = 0$

Exercise 04: Association rules (04pt)

Support threshold =33.34% => threshold is at least 2 transactions.
Applying Apriori

Pass (k)	Candidate k-itemsets and their support	Frequent k-itemsets
k=1	HotDogs(4), Buns(2), Ketchup(2), Coke(3), Chips(4)	HotDogs, Buns, Ketchup, Coke, Chips
k=2	{HotDogs, Buns}(2), ~~{HotDogs, Ketchup}~~(1), {HotDogs, Coke}(2), {HotDogs, Chips}(2), ~~{Buns, Ketchup}~~(1), ~~{Buns, Coke}~~(0), ~~{Buns, Chips}~~(0), ~~{Ketchup, Coke}~~(0), ~~{Ketchup, Chips}~~(1), {Coke, Chips}(3)	{HotDogs, Buns}, {HotDogs, Coke}, {HotDogs, Chips}, {Coke, Chips}
k=3	{HotDogs, Coke, Chips}(2)	{HotDogs, Coke, Chips}
k=4	{}	

Note that {HotDogs, Buns, Coke} *and* {HotDogs, Buns, Chips} *are not candidates when* k=3 *because their subsets* {Buns, Coke} *and* {Buns, Chips} *are not frequent.*
Note also that normally, there is no need to go to k=4 *since the longest transaction has only 3 items.*

All Frequent Itemsets: {HotDogs}, {Buns}, {Ketchup}, {Coke}, {Chips}, {HotDogs, Buns}, {HotDogs, Coke}, {HotDogs, Chips}, {Coke, Chips}, {HotDogs, Coke, Chips}.
Association rules:
{HotDogs, Buns} would generate: HotDogs → Buns (2/6=0.33, 2/4=0.5) and
Buns → HotDogs (2/6=0.33, 2/2=1);
{HotDogs, Coke} would generate: HotDogs → Coke (0.33, 0.5) and
Coke → HotDogs (2/6=0.33, 2/3=0.66);
{HotDogs, Chips} would generate: HotDogs → Chips (0.33, 0.5) and
Chips → HotDogs (2/6=0.33, 2/4=0.5);
{Coke, Chips} would generate: **Coke → Chips (3/6=0.5, 3/3=1)** and
Chips → Coke (3/6=0.5, 3/4=0.75);
{HotDogs, Coke, Chips} would generate: HotDogs → Coke ^ Chips (2/6=0.33, 2/4=0.5),
Coke → Chips ^ HotDogs (2/6=0.33, 2/3=0.66),
Chips → Coke ^ HotDogs (2/6=0.33, 2/4=0.5),
HotDogs ^ Coke → Chips(2/6=0.33, 2/2=1),
HotDogs ^ Chips → Coke(2/6=0.33, 2/2=1) and
Coke ^ Chips → HotDogs(2/6=0.33, 2/3=0.66).

With the confidence threshold set to 60%, the Strong Association Rules are (sorted by confidence):

1. Coke → Chips (0.5, 1)
2. Buns → HotDogs (0.33, 1);
3. HotDogs ^ Coke → Chips(0.33, 1)
4. HotDogs ^ Chips → Coke(0.33, 1)
5. Chips → Coke (0.5, 0.75);
6. Coke → HotDogs (0.33, 0.66);
7. Coke → Chips ^ HotDogs (0.33, 0.66)
8. Coke ^ Chips → HotDogs(0.33, 0.66).