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: 2017/2018
By Dr. B. LOUNNAS

Exercise 01: Course question (06pt)

1. Information retrieval system is a tool of finding material (usually documents) of unstructured data …….
   a) What do we mean by unstructured data?
   b) And, is it IR only used on unstructured data? Explain?

2. Describe shortly what is the important of Data Mining?

3. When is it preferable to use a dense index rather than a sparse index? Explain?

4. What is the difference between a primary index and a secondary index?

5. CRISP-DM defined and validated a data mining process that could be applicable in any domain application sectors.
   a) Cite some benefits of using such process?

6. As far, the best known combination of SMART notation in Vector Space Model is LNC.LTC
   The calculation of the idf adds a greats values to the search results, however we see that it was deleted from the document side (LNC where N stand for no idf). Yet this remains the best known combination, why?

7. In what situation the calculation of idf values will be useless?

Exercise 02: Indexation (02pt)

Assume we have the following database table:

SID	Name	Level	Point
123	Milhouse	10	3.1
142	Bart	10	2.3
279	Jessica	10	4
345	Martin	8	2.3
456	Ralph	8	2.3
512	Nelson	10	2.1
679	Sherri	10	3.3
697	Terri	10	3.3
857	Lisa	8	4.3
912	Windel	8	3.1

1. Create a Sparse index on SID column, assuming that each entry of the index hold at max 4 rows.
2. Create a Dense index on Name column.
3. Cite some advantages and disadvantages for both indices.

Exercise 03: Information Retrieval Models (06pt)

Boolean Retrieval Model

1. **Draw the term-document incidence matrix for the following documents. This retrieval system take consideration for the stop-word (Do not delete them):**

   - **Doc 1: "breakthrough drug for schizophrenia"**
   - **Doc 2: "new schizophrenia drug"**
   - **Doc 3: "new approach for treatment of schizophrenia"**
   - **Doc 4: "new hopes for schizophrenia patients"**

2. **What are the returned results for these queries using term-document incidence matrix from above. Explain in details how you get the results :**

   a) **"schizophrenia AND drug"**
   b) **"for AND NOT (drug OR approach)"**

Vector Space Model

Considering a collection of documents contains 1 000 000 documents including Doc 1, Doc 2, Doc 3 and Doc 4 from above.

Given the df (document frequency) for each word represented in the four documents:

Word	breakthrough	drug	for	schizophrenia	new	approach	treatment	hopes	patients	of
df	19000	6000	75000	25	25000	15000	500	7000	1200	500000

1. **For each documents calculate tf_{raw}, tf_{weight}, idf, and the final weight (tf-idf = tf_{weight} x idf).**
2. **Normalize the vectors of each document using cosine normalization.**
3. **What are the results of the following query using LNC.LTC SMART notation:**
   Query = "schizophrenia drug"

Exercise 04: Decision tree (06pt)

Consider the training examples shown in table below for a binary classification problem.

Instance	a1	a2	a3	Target Class
01	T	T	<2	+
02	T	T	>2	+
03	T	F	>2	-
04	F	F	>2	+
05	F	T	>2	-
06	F	T	>2	-
07	F	F	>2	-
08	T	F	>2	+
09	F	T	>2	-

1. **Calculate the entropy of this collection of training examples.**
2. **Calculate the information gains of a1, a2 and a3 relative to these training examples.**
3. **What is the best split (among a1, a2, and a3) according to the information gain? Explain?**