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.
   a) We means by unstructured data, all information data that are store in papers, documents, book, and text. In other means those data that have no structure such as XML, Data from databases …
   b) No, IR can also work with structured data, which is the simplest task for Information Retrieval.

2. The important of Data Mining is about extracting useful information (known as knowledge)

3. It is preferable to use a dense index instead of a sparse index when the file is not sorted on the indexed field (such as when the index is a secondary index) or when the index file is small compared to the size of memory.

4. The primary index is on the field which specifies the sequential order of the file. There can be only one primary index while there can be many secondary indices.

5. This methodology should make large data mining projects faster, cheaper, more reliable and more manageable.

6. Because the idf will be calculated in the query side.

7. idf values will be useless, when the query contained only one (1) word.

Exercise 02: Indexation (02pt)

1 & 2.

Sparse
123
456
857

SID	Name
123	Milhou
142	Bart
279	Jessica
345	Martin
456	Ralph
512	Nelson
679	Sherri
697	Terri
857	Lisa
912	Windel

Name
Bart
Jessica
Lisa
Martin
Milhouse
Nelson
Ralph
Sherri
Terri
Windel

SID	Name
123	Milhouse
142	Bart
279	Jessica
345	Martin
456	Ralph
512	Nelson
679	Sherri
697	Terri
857	Lisa
912	Windel

3.

It is generally faster to locate a record if we have a dense index rather than a sparse index. However, sparse indices have advantages over dense indices in that they require less space and they impose less maintenance overhead for insertions and deletions…….

Exercise 03: Information Retrieval Models (06pt)

Boolean Retrieval Model

1.

	Doc 1	Doc 2	Doc 3	Doc 4
breakthrough	1	0	0	0
drug	1	1	0	0
for	1	0	1	1
schizophrenia	1	1	1	1
new	0	1	1	1
approach	0	0	1	0
treatment	0	0	1	0
of	0	0	1	0
hopes	0	0	0	1
patients	0	0	0	1

2. A. the query = "schizophrenia AND drug"

	Doc 1	Doc 2	Doc 3	Doc 4
schizophrenia	1	1	1	1
drug	1	1	0	0
AND	**1**	**1**	**0**	**0**
Results = Doc 1 and Doc 2				

B. the query = "for AND NOT (drug OR approach)"

	Doc 1	Doc 2	Doc 3	Doc 4
drug	1	1	0	0
approach	0	0	1	0
OR	**1**	**1**	**1**	**0**
NOT	**0**	**0**	**0**	**1**
for	1	0	1	1
AND	**0**	**0**	**0**	**1**
Results = Doc 4				

Vector Space Model

1 & 2.

Word	breakthrough	drug	for	schizophrenia	new	approach	treatment	hopes	patients	of
df	19000	6000	75000	25	25000	15000	500	7000	1200	500000
idf	1.7212	2.2218	1.1249	4.6020	1.6020	1.8239	3.3010	2.1549	2.9208	0.3010

	Doc 1						Doc 2				
	tf_{raw}	**tf_{weight}**	**weight**	**cosinNor**	**Cosin2**		**tf_{raw}**	**tf_{weight}**	**weight**	**cosinNor**	**Cosin2**
breakthrough	1	1	1.7212	0.3124	0.2448	new	1	1	1.6020	0.2991	0.9723
drug	1	1	2.2218	0.4033	0.0773	schizophrenia	1	1	4.6020	0.8593	0.000972
for	1	1	1.1249	0.2042	0.9664	drug	1	1	2.2218	0.4148	0.2333
schizophrenia	1	1	4.6020	0.8354	0.000322						
Length = 5.5084					=77601.5	Length = 5.3554					=25709.9

	Doc 3						Doc 4				
	tf_{raw}	tf_{weight}	weight	cosinNor	Cosin2		tf_{raw}	tf_{weight}	weight	cosinNor	Cosin2
new	1	1	1.6020	0.2554	0.0493	new	1	1	1.6020	0.2592	0.3149
approach	1	1	1.8239	0.2908	0.0296	hopes	1	1	2.1549	0.3487	0.0881
for	1	1	1.1249	0.1793	0.1480	for	1	1	1.1249	0.1820	0.9448
treatment	1	1	3.3010	0.5264	0.000987	schizophrenia	1	1	4.6020	0.7447	0.000314
of	1	1	0.3010	0.0480	0.9872	patients	1	1	2.9208	0.4726	0.0151
schizophrenia	1	1	4.6020	0.7338	0.0000493						
				Length = 6.2708	=506433.8					Length = 6.1793	=79375.3

3.

	Query					Cosin	Result
	tf_{raw}	tf_{weight}	weight	cosinNor	Query & Doc 1	0.0338	**2**
schizophrenia	1	1	4.6020	0.9005	Query & Doc 2	0.1022	**1**
drug	1	1	2.2218	0.4347	Query & Doc 3	0.000044	**4**
				Length = 5.1102	Query & Doc 4	0.000282	**3**

Exercise 04: Decision tree (06pt)

1. Calculation of the entropy:

$$E(s) = -\sum_{i=1}^{k} p_i log_k(p_i)$$

$$E(s) = -\frac{4}{9}log_2\left(\frac{4}{9}\right) - \frac{5}{9}log_2\left(\frac{5}{9}\right) = 0.9911$$

2. Calculate the information gains of a1, a2 and a3 relative to these training examples.

$$G(s,Q) = E(s) - \sum_{i=1}^{k} p_i * E(s,Q_i)$$

a) $E(s,a1) = \frac{4}{9}\left[-\left(\frac{3}{4}\right)log_2\left(\frac{3}{4}\right) - \left(\frac{1}{4}\right)log_2\left(\frac{1}{4}\right)\right] + \frac{5}{9}\left[-\left(\frac{4}{5}\right)log_2\left(\frac{4}{5}\right) - \left(\frac{1}{5}\right)log_2\left(\frac{1}{5}\right)\right] = 0.7616$

$$G(s,a1) = 0.9911 - 0.7616 = 0.2294$$

b) $E(s,a2) = \frac{5}{9}\left[-\left(\frac{2}{5}\right)log_2\left(\frac{2}{5}\right) - \left(\frac{3}{5}\right)log_2\left(\frac{3}{5}\right)\right] + \frac{4}{9}\left[-\left(\frac{2}{4}\right)log_2\left(\frac{2}{4}\right) - \left(\frac{2}{4}\right)log_2\left(\frac{2}{4}\right)\right] = 0.9839$

$$G(s,a2) = 0.9911 - 0.9839 = 0.0072$$

c) $E(s,a3) = \frac{1}{9}\left[-\left(\frac{1}{1}\right)log_2\left(\frac{1}{1}\right) - \left(\frac{0}{1}\right)log_2\left(\frac{0}{1}\right)\right] + \frac{8}{9}\left[-\left(\frac{3}{8}\right)log_2\left(\frac{3}{8}\right) - \left(\frac{5}{8}\right)log_2\left(\frac{5}{8}\right)\right] = 0.8483$

$$G(s,a3) = 0.9911 - 0.8483 = 0.1428$$

3. According to information gain, a1, produces the best split. Because a1 have the large information gain.