UNIVERSITY OF MOHAMED BOUDIAF – M'SILA
FACULTY OF MATHEMATICS AND INFORMATICS
2023/2024

DEPARTMENT OF COMPUTER SCIENCE - THIRD YEAR LICENCE (ISIL)

TD 02

Exercise 01

Consider the following documents collection

d1 = "Big cats are nice and funny"
d2 = "Small dogs are better than big dogs"
d3 = "Small cats are afraid of small dogs"
d4 = "Big cats are not afraid of small dogs"
d5 = "Funny cats are not afraid of small dogs"

(a) Compute the tokens for each document.
(b) Normalize the tokens with respect to plurals and upper/lower case.
(c) Compute the dictionary relative to the documents collection.

- For each of the questions above write a pseudo code of the algorithm.

Exercise 02

Assume the following documents

Doc1 "italy is world champion 2006"
Doc2 "germany and italy played each other in the semifinal"
Doc3 "germany was in the semifinal 2006"
Doc4 "germany won the semifinal in italy 1990"

Assume that the following terms are stop words: **is, and, in, the, was, each, other**. Construct an inverted index.

Exercise 03

Starting from the document collection of exercise 01, build the documents-terms incidence matrix as required by the Boolean model.

Exercise 04

Starting from the documents collection of Exercise 1 consider a Boolean model.

(a) Answer the query q1 = funny AND dog
(b) Answer the query q2 = nice OR dog
(c) Answer the query q3 = big AND dog AND NOT funny

Exercise 05

Draw the term-document incidence matrix and the inverted index representation for the following document collection:

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

For the document collection above, what are the returned results for these queries:

– schizophrenia AND drug
– for AND NOT(drug OR approach)