

TD02 - Information Retrieval and Algorithms

Third Year ISIL - University of M'Sila

Exercise 01 – Text Processing & Dictionary Construction

Consider the following document 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"
```

- Compute the tokens for each document.
- Normalize the tokens with respect to:
 - Upper/lower case
 - Plurals (e.g., cats → cat, dogs → dog)
- Compute the dictionary (set of distinct normalized terms) relative to the document collection.
- For each of the operations above, write a pseudo-code algorithm for:
 - Tokenization
 - Normalization
 - Dictionary construction

Exercise 02 – Document-Term Incidence Matrix

Starting from the document collection of Exercise 01:

- Build the document-term incidence matrix.
- Use Boolean representation (1 = term appears, 0 = term does not appear).

Exercise 03 – Boolean Retrieval

Using the Boolean model built in Exercise 02, answer the following queries:

- $q_1 = \text{funny AND dog}$
- $q_2 = \text{nice OR dog}$
- $q_3 = \text{big AND dog AND NOT funny}$

Show intermediate Boolean operations.

Exercise 04 – Inverted Index Construction

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
```

Stop words:

is, and, in, the, was, each, other

- Remove stop words.
- Compute the tokens.
- Construct the inverted index in the form:

$\text{term} \rightarrow \{\text{DocIDs}\}$

Exercise 05 – Incidence Matrix & Inverted Index

Consider the following document collection:

```
Doc1 breakthrough drug for schizophrenia
Doc2 new schizophrenia drug
Doc3 new approach for treatment of schizophrenia
Doc4 new hopes for schizophrenia patients
```

- Draw the term-document incidence matrix.
- Construct the inverted index representation.
- Evaluate the following Boolean queries:

- $\text{schizophrenia AND drug}$
- $\text{for AND NOT}(\text{drug OR approach})$

Show the evaluation steps.