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

DEPARTMENT OF COMPUTER SCIENCE - THIRD YEAR LICENCE (ISIL)

# TD 04

---

## Exercise 01

You have the following values of *tf* of three (3) documents from a collection of 806791 documents, and the values of *tf* and *Idf* of four (4) terms:

| | **Doc 1** | **Doc 2** | **Doc 3** | | **term** | **$df_t$** | **$idf_t$** |
|---|---|---|---|---|---|---|---|
| car | 12 | 25 | 39 | | car | 18165 | |
| auto | 2 | 33 | 1 | | auto | 6723 | |
| insurance | 0 | 31 | 20 | | insurance | 19241 | |
| best | 10 | 0 | 12 | | best | 25235 | |

- First, calculate the values of **idf**?

- Compute the two top scoring documents on the query "car car insurance" for the following SMART notation scheme:

    - nnc.ntc
    - ltc.lnc

## Exercise 02

Considering the following table of count vector ($Tf_{raw}$) of a 3 documents and query:

| | Doc 1 | Doc 2 | Doc 3 | Query |
|---|---|---|---|---|
| Two | 2 | 0 | 0 | 0 |
| Tea | 2 | 2 | 0 | 1 |
| Me | 0 | 1 | 2 | 1 |
| You | 0 | 1 | 2 | 0 |

- Calculate the idf (inverse document frequency) for each word represented in the three documents.
- What is the result of the query (Tea me) using NTC.NNN smart notation?