|  |  |  |  |
| --- | --- | --- | --- |
| **logo_udlap**    **EDEI** | **Advanced Databases**  **Final Project** | | |
| **DEPARTMENT** | Computing, Electronics and Mechatronics | **COURS** | LIS – 3012 |
| **PROFESSOR** | José Luis Zechinelli Martini | **TERM** | Spring 2022 |

## Document base

## Objective

Define a system to create and manipulate a document base using the indexing and querying techniques covered in the course.

## Definition

From a base of at least 10 documents, build a system that implements LSI technique to represent and consult them. The choice of the collection of documents should look for a topic that satisfies a specific need taking into account public health, safety and welfare, as well as global, cultural, social, environmental and economic factors.

The system should take into account the semantic content of the documents by identifying synonyms and polysemies. For this purpose, it is necessary to define three types of lists: stop list, suffix list, and word stems.

## Expected results

**Frequency matrix**

Construct the frequency matrix FrecT generated from the document base and define a database schema that allows manipulating them for indexing and querying. The idea of representing this matrix is to be able to make comparisons of efficiency and effectiveness with the representation obtained by applying the LSI technique.

**Indexing terms**

Using the SVD (Single Value Decomposition), the system should allow to eliminate the least significant terms to represent the document base. Based on the information provided by the decomposition function, an expert user should be able to choose the terms to be used for indexing.

**Document query using SQL**

The system must allow to execute two types of queries:

* Given two documents D1 and D2, evaluate their degree of similarity.
* Given a query Q, obtain the n most relevant documents to answer Q.

These queries must be able to be executed using two types of functions:

**Query using similarity functions**. Use at least two similarity evaluation techniques (i.e., inner producer, cosines, Dice coefficients, Jaccard coefficients) to compare the documents.

**Query using dissimilarity functions**. Use at least one distance evaluation technique (i.e., Euclidean distance, Manhattan distance) to compare the documents.

## Technique aspects

For implementation, use the results obtained in the previous activities. Regarding persistence, they should make use of a DBMS (Oracle, MySQL, Posgress) for your document base. Design a relational schema that allows you to represent your arrays. Details will be discussed in class or in advising hours.

## Evaluation

Prepare a technical report in which you show your ability to **apply engineering design to produce solutions** that meet specified needs with consideration of public health, safety, and welfare, as well as global, cultural, social, environmental, and economic factors. To organize the above information, the report (10 page maximum) should have sections detailing the following:

* + Title of the project.
  + Brief description.
  + Feasibility analysis.
  + Objectives, goals and hypotheses.
  + Technical specifications.
    - Give the design of your database.
    - Describe the functions of your system and the platform used.
  + Standards and regulations for software development,

see ISO/IEC 12207:2008.

* + Methodology for the development of the project.
  + Results obtained.
    - Give examples that show how the program works.
    - Include copies of windows showing the use of your system.
  + Conclusions.
  + References and bibliography.

Attached you will find the rubric that will be used to evaluate your report:

* “SO2.pdf” (apply engineering design to produce solutions).