Natural Language Processing (NLP)

Aim
This lab is about natural language processing. The objective is to learn the term frequency and inverse document frequency measure with the aim to classify documents, extract text summarizations, and perform sentiment analysis. This lab is divided into three parts.

To pass the lab, write down your answers to the questions in Part 1, 2 and 3. During the lab demonstration, explain your answers to the lab assistant. The lab assistant will ask you some general question about the parts of the lab and how you solved them.

Part 1

Read the tutorial Text feature extraction tf-idf available on the course website. This tutorial explains how to calculate the TF-IDF measure and how you can use it to implement text classification and document similarity. Execute the Python scripts contained in the tutorial.

▷ To pass Part 1:

Explain to the lab assistant:

(i) what is the TF-IDF measure

(ii) how to use TF-IDF for:

- document similarity
- classify text

Part 2

In Part 2 we will work on sentiment analysis. Sentiment analysis is the automated process that uses AI to identify positive, negative and neutral opinions from text. Sentiment analysis is widely used for getting insights from social media comments, survey responses, and product reviews, and making data-driven decisions.

Read the tutorial Movie Reviews Sentiment Analysis with Scikit-Learn available on the course website. Download the movie_review dataset from `http://www.nltk.org/nltk_data/` and run the Python scripts contained in the tutorial.

Finally, read the tutorial Parameter Tuning Using Grid Search available on the course website. This tutorial explains how to create a pipeline in Python and how to use grid search to improve the algorithm final score by tuning the pipeline parameters.

▷ To pass Part 2:

> Write a text classification pipeline to classify movie reviews as either positive or negative. Find a good set of parameters for the pipeline you have created by using grid search. Show your result to the lab assistant.

Part 3

The third and last part of this lab is about text summarization.

You can start by reading the tutorial Using TextRank for Summarization.

Then, open a Python shall and install the module summa. Execute the Python scripts of the tutorial with some text you find interesting and see the summarization you get.

▷ To pass Part 3:

(i) Explain the TextRank algorithm and how it works to the lab assistant.

(ii) Show your lab assistant some summaries you created; and discuss the quality of the summaries
(like; does your abstract make any sense? can you create a summary that looks like an abstract from a news article? can you summarize product opinions from customers etc).