Lab 4

# 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:

1. **What is the TF-IDF measure?**

TF-IDF (term frequency – inverse document frequency) is a measure of how important a word is in a document. The algorithm also considers how important the word is compared to all words in a set of documents. It is used by search engines to rank documents based on a user query. The word is weighted in logarithmic scale.

1. **How to use TF-IDF for:**
   1. **Document similarity**

Document similarity, in this case cosine similarity, measures how related two documents are based on an angle computed in a normalized space. The previously computed TF-IDF score is used to compare two sentences.



The normalized space and angle mean that no magnitude is considered while computing. For example, the Euclidian distance between a sentence/document that appears 50 times versus one who appears 10 times is large. The angle between the documents can still be small, which means that they are related.

* 1. **Classify text.**

dwgsga

# 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. 1/2 Machine Learning for Social Media (TNM108) Pierangelo Dell’Acqua HT 2019 Lab 4.

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.**

rdasgsdhgr

# 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:

1. Explain the TextRank algorithm and how it works to the lab assistant.
2. 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)