**Creare un sistema di Spam Detection**

**L'azienda ProfessionAI vuole realizzare una libreria capace di fare analisi delle email ricevute.**

**Nello specifico, il CEO ha richiesto di identificare le email di tipo SPAM sulle quali fare analisi contenutistiche.**

**Il CTO nello specifico ti fornisce un dataset e ti chiede di:**

**- Addestrare un classificatore per identificare SPAM**

**- Individuare i Topic principali tra le email SPAM presenti nel dataset**

**- Calcolare la distanza semantica tra i topics ottenuti, per dedurne l'eterogeneità.**

**- Estrarre dalle mail NON SPAM le Organizzazioni presenti.**

**Scarica il dataset!**

**Invia il tuo progetto**

Invia il tuo progetto per la correzione solo quando lo hai ultimato. Per inviarlo:

1. Clicca sul tasto Share in alto a destra.
2. Condividi il notebook con chiunque abbia il link.
3. Assicurati che sia selezionato il ruolo editor.
4. Condividi il progetto, tramite link, nell'ultima lezione di questo corso.

Assicurati di svolgere la procedura correttamente per permetterci di correggere il tuo progetto.

**La correzione richiede tra i 2 e i 7 giorni lavorativi**, dopo tale periodo puoi chiedere informazioni sullo stato della correzione all'indirizzo assignments@profession.ai.

## Anti Spam Software for University Mail

The ProfessionAI company wants to create a library capable of analyzing received emails. Specifically, the CEO requests to identify the SPAM emails on which to carry out content analysis. The CTO specifically provides you with a dataset and asks you to:

* Train a classifier to identify SPAM;
* Identify the main topics among the SPAM emails present in the dataset;
* Calculate the semantic distance between the topics obtained, to deduce their heterogeneity;
* Extract the Organizations quoted the NON-SPAM emails.

The project aims to develop a library capable of analyzing received emails. In particular, the tasks include:

* Training a classifier to identify SPAM emails.
* Identifying the main topics within the SPAM emails in the dataset.
* Calculating the semantic distance between the obtained topics to deduce their heterogeneity.
* Extracting organizations from the non-SPAM emails.

The provided dataset is spam\_dataset.csv.

You can access the project by referring to the Email Analysis Project.ipynb notebook.  
Additionally, the Python library email\_analysis\_library contains all the custom functions utilized throughout the project, which are divided into the modules preprocessing, spam\_detection, lda, and ner.