Script Carga de datos T13.

# Introducción

El siguiente informe detalla la estructura y propósito de cada componente del script de carga de datos. Este proceso es necesario para varios aspectos, primero que nada normalizar y limpiar los datos que vienen desde los archivos entregados por T13, por medio de librerías de limpieza nativas de python (re, datetime) como librerías anexas de python que nos permiten un manejo más profundo de texto y variables (numpy, pandas).

El otro aspecto importante del sistema es el uso de inteligencia artificial para la extracción de entidades. Esto se realiza aplicando el área llamada *procesado de lenguaje natural* o NLP, más específicamente el *reconocimiento de entidades nombradas* o NER. Esto se logra por medio de la librería de python llamada Flair.

T13 nos hizo entrega de varios archivos con información, en formato csv. Esta información tuvo que ser procesada e insertada en la base de datos.

Para esto se elaboró un software escrito en Python que lee los archivos, procesa ciertos campos y luego los ingresa a la database. El programa se divide en 5 archivos: dalc, entities, functions, queries y main.

El archivo dalc.py (del nombre Data Access Layer Component) ofrece una colección de clases y métodos de conexión a la base de datos, por medio de la librería pyodbc, la cual nos permite generar conexiones con base de datos tipo Microsoft SQL Server. A parte del método de conexión existen diferentes clases que comprenden diferentes DAO de las tablas de la base de datos, los cuales hacen las transacciones necesarias. Las queries de la base de datos son guardadas dentro de un diccionario de datos para mantener la simplicidad del código.

El módulo entities.py aplica toda la extracción de entidades a partir de un texto, que en este caso es el contenido de cada publicación.

El archivo queries.py contiene un pequeño DAO que permite generar la información de cada gráfico que es posteriormente implementado en la página web, en este caso sólo 1, pero otorga la estructura necesaria para replicar el formato.

Por último se encuentra el módulo functions.py contiene todas las funciones necesarias para limpieza de texto, procesado de entidades y carga de data a la base de datos.