Tipología y ciclo de vida de los datos.
Wiki Práctica 1

Yaiza Santana Santana

1. Memoria.

Este documento contiene los ocho enunciados a responder propuestos en la memoria de la Práctica 1.

2. Carpeta python.

Contiene tres ficheros python:

- *diccionario_total_palabras_empleo.py*: se entrega de manera adicional. Con este script se extrajeron las palabras de los textos formación, experiencia y conocimientos explicados en la memoria (extraídos de BB.DD. Propias).
- *preprocesado_texto_genero.py*: se entrega de manera adicional. Con este script se redujo el conjunto de datos 1, eliminando plurales, palabras repetidas, y términos que a priori no encontrará la RAE (nombre de software, abreviaciones, etc.). Éstos últimos se almacenaron en una tabla de la BB.DD. como terminología de empleo.
- *web_scraping_rae_lematizacion.py*: este script contiene las tareas para llevar a cabo el web (javascript) scraping para obtener el conjunto de datos 2.

3. Carpeta conjunto 1.

Contiene dos archivos:

- *diccionario_reducido_octubre_2.xlsx*: contiene los primeros 1700 términos primeros del diccionario de empleo obtenido tras ejecutar *diccionario_total_palabras_empleo.py* y *preprocesado_texto_genero.py.*
- *diccionario_reducido_octubre_2.csv: contiene los primeros 1700 términos primeros del diccionario de empleo obtenido tras ejecutar diccionario_total_palabras_empleo.py* y *preprocesado_texto_genero.py.*

4. Carpeta conjunto 2.

Contiene dos archivos:

- *lematizacion_rae.xlsx*: contiene el lema y tipo de los primeros 1700 términos primeros del diccionario de empleo obtenido después del *Web Scraping*.
- *lematizacion_rae.csv*: contiene el lema y tipo de los primeros 1700 términos primeros del diccionario de empleo obtenido después del *Web Scraping*.