Universidad Rey Juan Carlos

TEMA 1 – HOJA DE EJERCICIOS I

Se plantean diferentes ejercicios para familiarizarse con el entorno de trabajo propuesto en la asignatura. En particular, se utiliza el proyecto de ejemplo YouTube Spam Detection, el cual sigue una estructura típica de proyecto reproducible de procesamiento de lenguaje natural y minería de texto. Este repositorio presenta una separación por etapas de datos, scripts de preparación y experimentos, y notebooks de análisis.

Ejercicio 1. Preparación del entorno de desarrollo

Utilizando el proyecto de ejemplo Youtube Spam Detection proporcionado en data-science-project-example, se pide:

a. Instalar Visual Studio Code y uv siguiendo el método indicado para tu sistema operativo.
b. Abrir el repositorio con Visual Studio Code y leer detalladamente el fichero README.md. Se pide identificar y comprender la estructura general del proyecto. Esto incluye la organización de carpetas, su contenido y qué rol juega cada una en el flujo de trabajo.
c. Ejecutar uv sync desde la raíz del proyecto y comprobar que se crea el entorno virtual y se instalan las librerías necesarias.

Ejercicio 2. Preparación de los conjuntos de datos

Se pide preparar los datos y recursos necesarios para la experimentación de distintos modelos y técnicas de procesamiento del lenguaje natural:

a. Descargar y descomprimir el dataset de YouTube Spam Collection y el modelo de FastText wiki.simple.bin siguiendo las instrucciones indicadas en el README.md.
b. Ejecutar los scripts de preparación de datos ubicados en scripts/processing/ y analizar el notebook eda.ipnb disponible en la carpeta notebooks. ¿Qué permite esta separación de datos en raw, interim y processed? ¿Y la separación en train, dev y test? ¿Qué tipo de información se explora en el EDA y qué decisiones de preparación de datos sugiere ese análisis?

Universidad Rey Juan Carlos

Ejercicio 3. Ejecución de experimentos

Se proponen distintos experimentos de procesamiento de lenguaje natural y minería de texto sobre los datos anteriormente preparados y analizados.

a. Ejecutar los scripts ubicados en scripts/experiments, los cuales muestran algunos de los modelos y técnicas que se verán más adelante en la asignatura.
b. Comprobar que cada ejecución genera un resultado dentro de la carpeta results. ¿Qué permite esta organización de experimentos y resultados?
c. Analizar los resultados de los experimentos realizados a través del notebook compare_test_results.ipynb ubicado en la carpeta notebooks. ¿Cuándo debería realizarse la experimentación sobre el conjunto test?