ntic master
revolucionamos la comunicación

UNIVERSIDAD COMPLUTENSE DE MADRID

Módulo
Text Mining

Autor: Luis Gascó Sánchez

Actualizado 2024

Ejercicio práctico (Tiempo estimado de realización: ~5-7 horas)

Se propone a los alumnos la realización de dos ejercicios prácticos que cubren las áreas principales expuestas en la formación teórica.

La entrega del ejercicio se realizará subiendo en la plataforma los archivos *.ipynb" generados en Google Collab. El nombre de los notebooks entregados deben seguir la estructura:

- ***Apellido1_Apellido2_Nombre_Ejercicio1.ipynb***
- ***Apellido1_Apellido2_Nombre_Ejercicio2.ipynb***

Además, ambos notebooks se entregarán comprimidos dentro de un archivo zip o rar.

Para comenzar a desarrollar los ejercicios en el enunciado se proporcionan links a notebooks de Google Colab de referencia para comenzar a trabajar. Estos archivos incorporan unas funciones para descargar y preparar el dataset que se utilizará en la tarea. Recordad que no todas las librerías que hemos visto están disponibles en Colab, así que tendréis que descargar algunas de las librerías y o módulos necesarios en cada caso.

Si tenéis alguna duda, por favor no dudéis en preguntarla en el foro o en la plataforma. Estaré encantado de ayudaros a resolverla.

Enunciado ejercicio 1 (8 puntos)

El objetivo de este ejercicio es comprobar los conocimientos que habéis adquirido en el área de análisis exploratorio de datos textuales, su pre-procesado y la generación de modelos de clasificación utilizando técnicas de **ingeniería de características**.

Para el ejercicio contáis con este Google Colab inicial, en el que encontraréis la descripción de los datos que se utilizarán, así como su descarga y carga en el entorno de ejecución. En el ejercicio deberéis hacer un análisis exploratorio, preprocesar los texto, entrenar modelos de clasificación y evaluar los modelos. Para el desarrollo de los ejercicios os recomiendo que copiéis el *notebook* de referencia y realicéis el ejercicio dentro de la plataforma de Google para evitar problemas con la instalación de librerías e incompatibilidades.

Para que os sirva de orientación, los criterios de evaluación del ejercicio serán los siguientes:

- **Análisis exploratorio, pre-procesado y normalización de los datos (30%):**
  - El ejercicio deberá contener un análisis exploratorio de los datos como número de documentos, gráficas de distribución de longitudes y/o wordclouds, entre otros análisis que se os pudieran ocurrir. Vuestros ejercicios deberán incorporar al menos los análisis exploratorios vistos en clase.
  - También tendréis que tener funciones para normalizar textos que permitan eliminar palabras vacías, quitar símbolos de puntuación, normalizar tokens (si fuera necesario) y lematizar.
- **Vectorización de textos (40%)**

  En clase hemos visto diferentes estrategias de vectorización como TF-IDF y *Word Embeddings*. Será necesario incorporar características adicionales como el sentimiento o características léxicas.
- **Entrenamiento y validación del sistema (30%)**
  - En el proceso de entrenamiento del modelo tendréis que testear al menos 3 modelos de clasificación. El procedimiento debe ser similar al visto en clase, en el que primero estimábamos el rendimiento de varios algoritmos de forma general, para posteriormente seleccionar el mejor para ajustar los hiperparámetros.

Enunciado ejercicio 2 (2 puntos)

El objetivo de este ejercicio es comprobar los conocimientos que habéis adquirido en el área de análisis exploratorio de datos textuales, su pre-procesado y la generación de modelos de clasificación utilizando **modelos de lenguaje**.

Para el ejercicio contáis con este Google Colab inicial, en el que encontraréis la descripción de los datos que se utilizarán, así como su descarga y carga en el entorno de ejecución. Dado que el análisis exploratorio ha sido realizado en el ejercicio anterior, en este caso podréis centraros en entrenar el modelo utilizando la librería Transformers, seleccionando un modelo pre-entrenado adecuado, entrenando el modelo y llevando a cabo la evaluación.

Se valorará positivamente un análisis comparativo entre la calidad del modelo resultante del ejercicio 1 y de este ejercicio.

Nota 1: Los ejercicios propuestos son similares a los vistos en clase, por lo que en ambos ejercicios es muy importante que documentéis y expliquéis adecuadamente (con vuestras palabras) los procesos llevados a cabo en el ejercicio. No hacerlo puede llegar a penalizar hasta 1 punto en la calificación final.

Nota 2*:* Cualquier cálculo adicional a los vistos en clase (visualizaciones, nuevas características añadidas al modelo, distribuciones por clase...) será valorado positivamente en la calificación.

Nota 3*:* Insisto en seguir la estructura de análisis en tres fases vista en clase. Esto os facilitará llevar un orden en el análisis, explicar los pasos de forma estructurada y clara, y conseguir una mejor calificación.

Nota 4: El segundo ejercicio requiere el uso de las GPUs de Google Colab. El link de Google Colab inicial está configurado para correrse con este tipo de hardware. Si alguno tuviera problemas para la ejecución que me contacte a través de Moodle para buscar una solución alternativa.

Instrucciones para subir los archivos:

- Nombrar el archivo del notebook con el código: *Apellido1_Apellido2_Nombre_Ejercicio1.ipynb*
- Nombrar el ejercicio adicional con el código: *Apellido1_Apellido2_Nombre_Ejercicio2.ipynb*
- Comprimir los notebooks, con las celdas ejecutadas, en un archivo tipo zip o rar.
- Subir el archivo a la plataforma