Tipología y fuentes de datos

PEC 2

[NOMBRE y APELLIDOS]

Universitat Oberta de Catalunya

Fecha de entrega :
2 de noviembre de 2021

Tipología y fuentes de datos
PEC2

Esta PEC consiste en codificar un Sistema de Recuperación de Información paso a paso. El *jupyter notebook* que acompaña a esta PEC contiene código python incompleto para que lo completes adecuadamente según las especificaciones. Los fragmentos de código a completar están marcados con comentarios #TODO que deberán ser sustituidos por código válido python.

El notebook está pensado para ejecutarse en el entorno Colab de Google (https://colab.research.google.com/).

Parte 1 [4 puntos]. Tareas básicas de tratamiento de texto y de recuperación de información. Hay que completar el código para obtener las funciones necesarias para el tratamiento de texto que podrán ser utilizadas en partes posteriores del ejercicio.

Parte 2 [4 puntos]. Usando cuestiones planteadas en Stack Overflow que es sitio de preguntas y respuestas para programadores, completar el código para resolver las cuestiones planteadas sobre:

- Tratamiento del texto
- Vectorización con tf-idf
- Similitud entre documentos
- Clustering

Parte 3 [2 puntos]. Con el texto de la Declaración Universal de Derechos Humanos en varios idiomas, escribir el código necesario para tratar las siguientes cuestiones:

- Número de palabras en la Declaración Universal de Derecho Humanos
- Número de palabras únicas
- Longitud media de las palabras
- Número de sentencias incluidas
- Número medio de palabras por sentencia
- La Ley Zipf

Se pide:

- Entregar el notebook con el código y la ejecución.
- Responder a las preguntas (usando el código necesario para ello) que están en el notebook en celdas de texto.

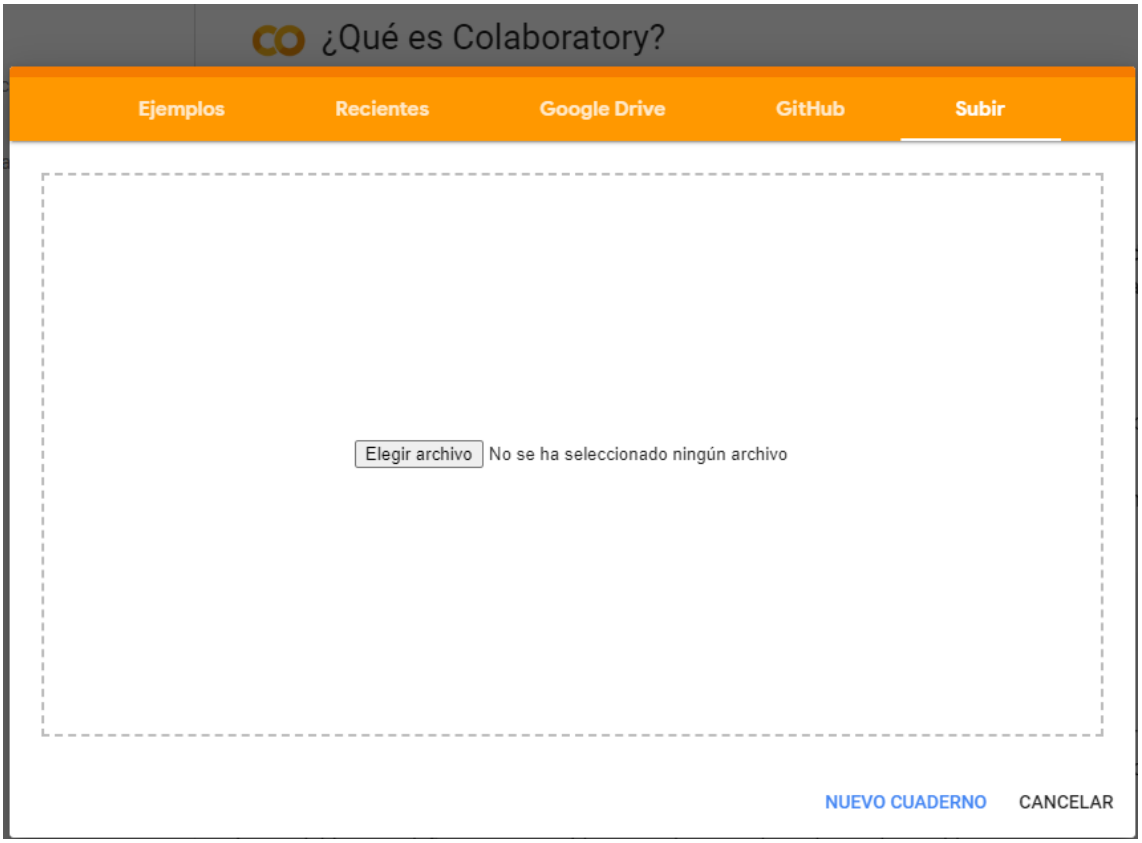
Anexo

Descarga en tu equipo el notebook `TyFdD-PEC2-2021.ipynb` así como los archivos de datos en csv. El notebook está pensado para que lo ejecutes en Google Colab.
Google Colab es un servicio en la nube que permite ejecutar Jupyter Notebooks accediendo con un navegador web. Tiene además las siguientes ventajas:

- Posibilidad de ejecución mediante GPUs
- Basado en jupyter notebook pudiendo crear y ejecutar libros en Python 2 o 3
- Tiene preinstaladas las librerías comunes usadas en ciencia de datos y la posibilidad de instalar otras.
- Enlaza con cuentas de Google Drive y desde github

Primero hay que entrar en sesión (login) con una cuenta de Google (la de la uoc debería funcionar).

Ahora ya se puede subir el notebook de la PEC a colab:

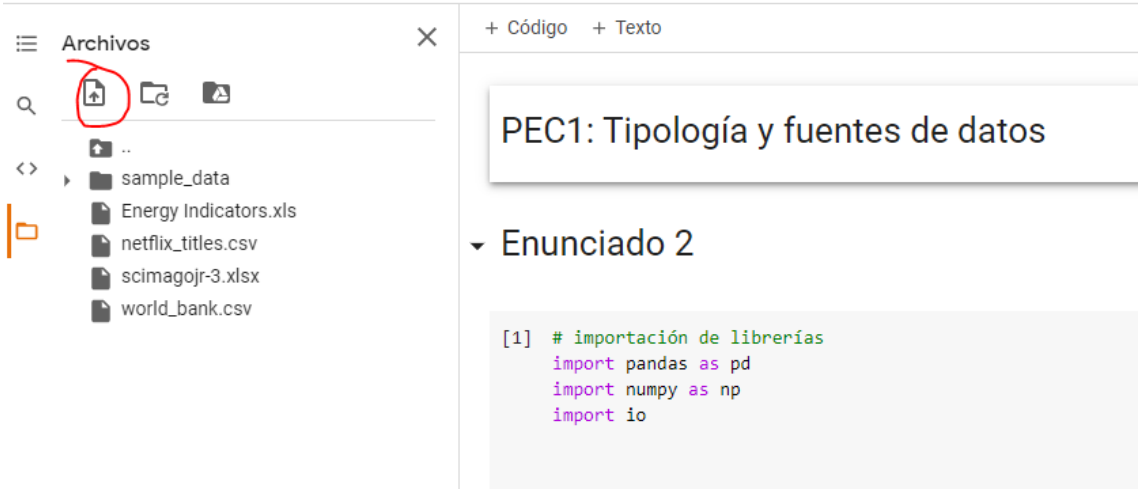


La ejecución del libro es exactamente igual que en cualquier jupyter notebook. Hay que pulsar Shift + Enter para que el código (python) se ejecute.
El dataset necesario para la PEC se cargan mediante código (hay que seleccionar el archivo .json adjunto):

```
[ ]  # subir todos los archivos de la PEC a colab
     from google.colab import files
     uploaded = files.upload()
```

```
Elegir archivos  Ningún archivo seleccionado  Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Saving netflix_titles.csv to netflix_titles.csv
Saving Energy Indicators.xls to Energy Indicators.xls
Saving world_bank.csv to world_bank.csv
Saving scimagojr-3.xlsx to scimagojr-3.xlsx
```

Alternativamente también es posible subirlos con la herramienta para subir archivos del propio colab:



No olvides que hay que subir **el archivo json**.

A partir de ahí debes completar el código python que falta que está marcado con `# TODO` y responder a las cuestiones planteadas escribiendo tanto las respuestas como el código con el que obtienes las respuestas.

Criterios de valoración
Cada uno de los apartados tiene un peso asignado en el total de la PEC. Se valorará, para cada apartado, la validez de la solución y la claridad de la argumentación.

Formato y fecha de entrega
Tenéis que enviar la PEC al buzón de Entrega y registro de EC disponible en el aula (apartado Evaluación). El formato del archivo que contiene vuestra solución puede ser .pdf,. odt, .doc y .docx. Para otras opciones, por favor, contactar previamente con vuestro profesor colaborador. El nombre del fichero debe contener el código de la asignatura, vuestro apellido y vuestro nombre, así como el número de actividad (PEC2). Por ejemplo *apellido1_nombre_tyfdd_pecX.pdf*. No olvidéis poner vuestro nombre y apellidos en el documento.

La fecha límite para entregar la PEC es **la indicada en la portada.**

Propiedad intelectual
Al presentar una práctica o PEC que haga uso de recursos ajenos, se tiene que presentar junto con ella un documento en que se detallen todos ellos, especificando el nombre de cada recurso, su autor, el lugar donde se obtuvo y su estatus legal: si la obra está protegida por el copyright o se acoge a alguna otra licencia de uso (Creative Commons, licencia GNU, GPL etc.). El estudiante tendrá que asegurarse que la licencia que sea no impide específicamente su uso en el marco de la práctica o PEC. En caso de no encontrar la información correspondiente tendrá que asumir que la obra está protegida por el copyright.

Será necesario, además, adjuntar los ficheros originales cuando las obras utilizadas sean digitales y su código fuente, si así corresponde.