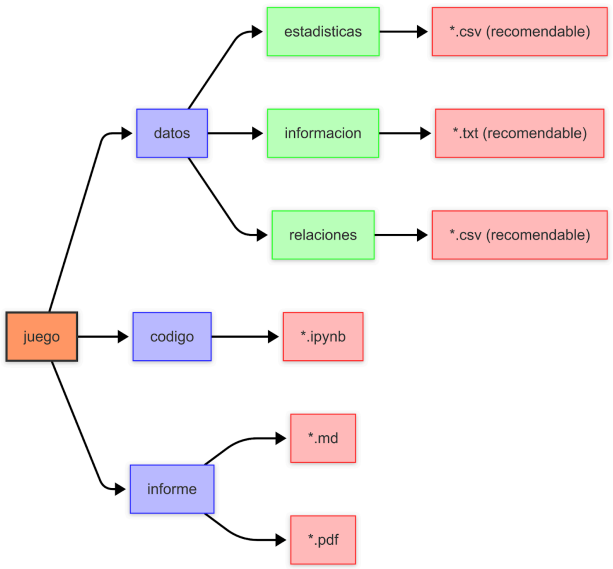
TUIA NLP 2025
TRABAJO PRÁCTICO 1 - Parte 2

Pautas generales:

- El trabajo deberá ser realizado de manera individual.
- Las librerías que utilicen deben estar (la primera vez que se cargan) sobre el código donde se las utiliza, no todas juntas al inicio.
- Se debe entregar un informe en el cual se incluya las justificaciones y un vínculo a los archivos que permitan reproducir el proyecto. Recomendamos **gitlab o github** para tal fin. Debe realizarse en **colab** y ser entregado en el formato de Jupyter Notebook **.ipynb**, dentro de un repositorio. Guardar una vez ejecutado.
- Para la solución del ejercicio puede utilizar todas las herramientas presentadas en las unidades 1, 2 y 3 de la materia.
- La entrega de la misma tendrá fecha límite el miércoles **21 de mayo a las 23:59**.
- En el repositorio de entrega deben ser compartidos con los docentes de la cátedra en el rol de editor:
  jpmanson@gmail.com
  alan.geary.b@gmail.com
  constantinoferrucci@gmail.com
  dolores.sollberger@gmail.com

Notas previas (*EJERCICIO 1*):
En este apartado solamente se solicita cargar el repositorio asignado a cada estudiante, el cuál se podrá ver cuál corresponde **en este enlace**.

Recordamos que el repositorio a extraer luce de la siguiente manera:



IMPORTANTE: Para poder acceder a cada repositorio cada estudiante lo debe hacer público (solo lectura) para que los alumnos que fueron asignados a esa fuente de datos puedan comenzar a trabajar.

El repositorio es la base de la información a utilizar, en el caso de no encontrar la información necesaria se puede agregar contenido extra, siempre y cuando sea previo a la realización de los ejercicios en función de mantener el orden.

La carga en **colab** puede ser mediante un archivo ZIP cargado manualmente o mediante código.

EJERCICIO 2
Apoyándose en la sección de **información**. Separa en fragmentos un texto extenso extraído y vectoriza cada fragmento con alguno de los modelos de embedding vistos en clases.

Luego realiza un análisis de similitud de texto ingresando varias frases a buscar semánticamente, compare distintas técnicas de distancias vistas en clases, elija la mejor y justifique la razón por la que esa técnica se ajusta para este tipo de búsquedas.

OPCIONAL: Visualizar en 3D aplicando PCA o t-SNE la ubicación de los fragmentos y la query ingresada vectorizada en el espacio. Realizar una observación sobre la visualización.

EJERCICIO 3
Apoyándose nuevamente en la sección de **información**. Recoge un texto extenso extraído, divídelos en fragmentos, luego realiza extracciones de sustantivos (POS) y categoriza estos sustantivos (NER), a continuación realiza una búsqueda de similitud filtrando por sustantivos, compara las distintas técnicas de distancias vistas en clases, elija la mejor y justifique la razón por la que esta técnica se ajusta para este tipo de búsquedas.

EJERCICIO 4
Mediante detección de idioma, separar los archivos en distintos lenguajes y guardar esa información en un dataframe.

EJERCICIO 5
En el caso de las reseñas realizadas por usuarios, utiliza análisis de sentimientos con modelos pre entrenados y guarda la clasificación predecida de cada reseña.

Luego, crea un sistema de búsquedas por similitud semántica y que permita filtrar por sentimiento para obtener.

EJERCICIO 6
Crea un set de datos de consultas (más de 300 preguntas en total) y categorizalas entre la fuente de datos que pueda llegar a responder esa pregunta entre **estadísticas**, **información** y **relaciones**.

Por ejemplo:
- ¿Cómo gano en el ajedrez? -> **Información**
- ¿Quién trabajó para el ta-te-ti? -> **Relaciones**
- ¿Qué puntaje tienen las damas? -> **Estadística**

A continuación, transforma esas consultas en vectores y entrena un modelo de clasificación (a gusto del estudiante) en donde pueda predecir la categoría a través de la consulta ingresada.

Agregar métricas y análisis durante todo el desarrollo, trabaje en varios modelos y comparelos.