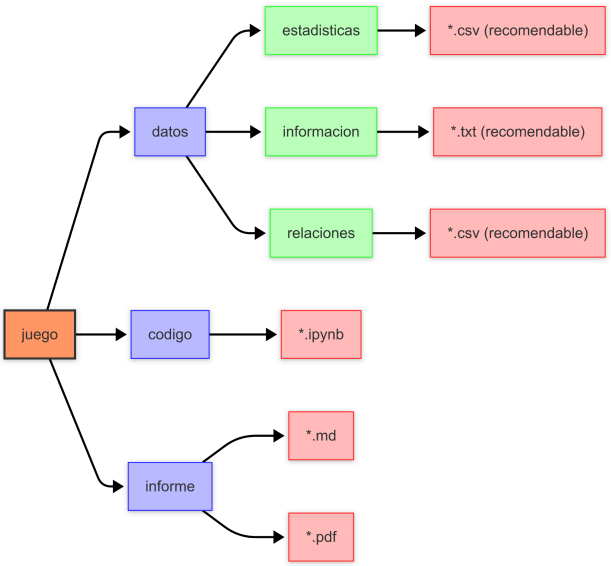
TUIA NLP 2025
TRABAJO PRÁCTICO 1 - Parte 1

Pautas generales:

- El trabajo deberá ser realizado en grupos de 5 integrantes como máximo.
- 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 la primera unidad de la materia.
- Toda la información se debe contener en un repositorio. Es obligatorio que el repositorio cuente con la siguiente estructura:



- La entrega de la misma tendrá fecha límite el miércoles **7 de mayo a las 23:59**.

EJERCICIO 1:
Según los grupos conformados en relación al juego cada grupo deberá
generar un repositorio en drive donde guardar los recursos obtenidos a continuación enumerados. Este repositorio deberá ser compartido por el grupo con los profesores de la cátedra en el rol de editor.

jpmanson@gmail.com
alan.geary.b@gmail.com
constantinoferrucci@gmail.com
dolores.sollberger@gmail.com

(nota: en las siguientes tareas considerar la extracción de recursos en varios idiomas)

- Extraer texto de documentos tipo pdf, txt, word o formatos de imágenes con las librerías adecuadas que se consideren relevantes.
- Extraer texto de videos de reviews y tutoriales.
- Evaluar diferentes tutoriales teniendo en cuenta que existen en varios idiomas.
- Extraer texto de foros con Web Scraping.
- Extraer estadísticas. La idea es que generen un formato tabular de información que crean relevante sobre datos del juego.
- Extraer relaciones entre juegos, creadores y otros sujetos implicados (ejemplo en la sección créditos).

A continuación, se detallan los requerimientos mínimos para el armado del set de datos. El mismo debe contar con:

- Mínimo de información por tema:
  - **Estadísticas**: Al menos 30 estadísticas del juego (guardado de manera tabular, es decir un csv, xlsx o hasta SQL).
  - **Información**: Al menos 30000 caracteres (alrededor de 20 páginas).
  - **Relaciones**: Mínimo de 25 relaciones siendo el juego el centro de las mismas (de momento almacenarlas en formato tabular con las columnas SUJETO1 - RELACION - SUJETO2, además cada relación puede extenderse en otras relaciones).
- Requerimientos en general:
  - **Idiomas**: Al menos 2 idiomas.
  - **Extracción**: Debe existir al menos 2 fuentes de datos (sitios web) de donde se extraiga información.
  - **Análisis post-extracción**: Es importante agregar un código que al ejecutar nos devuelva una serie de gráficas que puedan resumir el contenido extraído anteriormente (Como cantidad de caracteres, de palabras, de categorías, etc).
  - **Modularización del código**: Se debe respetar cada proceso de extracción de información por lo que, salvo que sea código compartido (que se puede apoyar en una función en común), es importante que cada tipo de extracción contenga su propio módulo que partirá desde llamar y ejecutar la función hasta la carga de datos en el directorio correspondiente.
  - **Documentación clara**: Cada celda de código debe estar acompañada de información correspondiente para comprender la misma.