Trabajo Práctico 3 - Lenguaje

El objetivo de este trabajo práctico es construir un modelo de lenguaje basado en n-gramas capaz de generar texto de buena calidad.

Este TP debe realizarse en grupos de **dos (2) a cuatro (4) integrantes**. Será puntuado con una nota binaria, de 0 puntos, o bien de 100 puntos. **La entrega no es obligatoria**; quienes no entreguen tendrán nota 0. (Recomendamos revisar las reglas de aprobación de la materia.)

Para lograr 100 puntos en este TP, se deberá:

1. Conseguir y limpiar un corpus de textos en algún idioma y dominio de interés. Por ejemplo, pueden descargar *El Quijote de la Mancha* o las obras completas de William Shakespeare del Proyecto Gutenberg, una colección de subtítulos de películas de OpenSubtitles, etc.

2. Programar un modelo de n-gramas, que implemente dos funciones principales:

   - **construcción**: a partir de una colección de textos, cargar tablas con la frecuencia de cada n-grama;
   - **generación**: a partir de una cadena de $n-1$ palabras, elegir la palabra siguiente.

3. Analizar los textos generados por este modelo. Estudiar al menos las siguientes cuestiones, mostrando ejemplos concretos:

   - ¿Cómo es la calidad de los textos generados, a medida que aumentan n y/o la cantidad de datos de entrenamiento? ¿Qué tipos de errores se producen?
   - ¿Cuánto se parecen los textos generados a los textos originales, a medida que aumentan n y/o la cantidad de datos de entrenamiento?
   - ¿Qué grado de creatividad ven en estos modelos? ¿Y de inteligencia?

La **entrega** consiste en un informe en formato PDF, con una descripción del trabajo realizado y con las respuestas al punto 3 de arriba. Opcionalmente, pueden entregar además un programa en Python que permita usar el mejor modelo construido.

Fecha límite: viernes 29/11 a las 23:59hs. Este TP no tiene recuperatorio.

Comentarios adicionales:

- El enunciado es intencionalmente difuso, a un alto nivel de abstracción, para incentivar la exploración.
- Hay muchas ideas para intentar mejorar los resultados del modelo de n-gramas: usar POS tags, matchear palabras en forma difusa, elegir al azar entre k posibles palabras siguientes, implementar algún mecanismo de atención, usar algún tipo de *embedding*, etc.
- Vale usar bibliotecas de Python, ChatGPT o lo que quieran, pero es obligatorio reportar todas las fuentes y herramientas que usen.