## línea horizontal

Proyecto prácticas profesionalizantes

**─**

Alumno: Ian Alén Cisneros

Universidad Nacional del Oeste  
Profesor asignado: Enrique Fernández  
2025

# Visión general

El siguiente proyecto trata de un Analizador Histórico de términos con la API de Internet Archive.

# Objetivos

1. Permitir a investigadores sociolingüísticos y usuarios en general analizar las palabras más usadas en documentos digitales en inglés, disponibles en repositorios abiertos (Internet Archive).
2. Proveer un sistema que acepte como parámetro un rango de fechas (desde, hasta) y un número 'N' variable de palabras a mostrar
3. Mostrar la cantidad de ocurrencias de cada palabra dentro del período y permitir análisis en subperiodos seleccionados dentro del rango principal
4. Guardar en memoria temporal (mientras la aplicación esté abierta) el historial de búsquedas recientes para fácil consulta y comparación
5. Presentar un resultado claro: listado de palabras con sus frecuencias y gráficos de barras de manera sencilla y amigable.
6. Limitar la búsqueda a un máximo configurable de documentos (ej. 600) para controlar el tiempo de ejecución (objetivo inferior a 1 hora 30 minutos).
7. Garantizar que el análisis sea solo para documentos en inglés; descartar otros idiomas para evitar complejidades.
8. Excluir palabras comunes (stop words) sin hacer lematización ni agrupamiento por raíz, para preservar las formas originales y conjugadas.
9. No incluir control de accesos ni perfiles de usuario; manejo con rol único y acceso simple.
10. La aplicación debe correr preferentemente en PC, con posibilidad mínima para acceso móvil.
11. Los usuarios deben poder exportar resultados (CSV) para análisis posteriores con otras herramientas (como R).

# Alcances

* Integración con el repositorio Internet Archive para búsqueda y descarga de documentos en inglés, con filtro temporal.
* Procesamiento mediante minería de texto para contabilizar frecuencias de palabras.
* Visualización simple con listas y gráficos tipo histograma de frecuencias.
* Implementación de memoria temporal para almacenar resultados recientes durante la sesión.
* Configuración que permita establecer número máximo de documentos a analizar.
* Interfaz web muy sencilla para ingresar parámetros y obtener resultados rápidamente.
* Exportación en CSV para cada consulta.
* No incluye funcionalidades avanzadas de NLP (ej. lematización).

# Funcionalidades

## 

|  |  |
| --- | --- |
|  | **Funcionalidad** |
| F1 | Definir período histórico con fecha desde y fecha hasta |
| F2 | Permitir definir número variable 'N' de palabras a listar |
| F3 | Conectar con API de Internet Archive y buscar documentos en inglés |
| F4 | Controlar número máximo (configurable) de documentos a analizar |
| F5 | Procesar documentos para extraer palabras y contar ocurrencias |
| F6 | Filtrar stop words (en inglés) sin lematización |
| F7 | Mostrar listado de palabras con número de ocurrencias |
| F8 | Permitir visualización por subperiodos dentro del rango principal |
| F9 | Mantener memoria temporal durante la sesión con historial de búsquedas recientes |
| F10 | Exportar resultados a CSV |