# Proyecto 2:
# Recuperación de Documentos de Texto

## 1- Introducción

El logro del estudiante está enfocado a entender y aplicar los algoritmos de búsqueda y recuperación de la información basado en el contenido.

Este proyecto está enfocado a la construcción óptima de Índice Invertido para tareas de búsqueda y recuperación en documentos de texto.

## 2- **Backend**: Implementación del Índice Invertido

Implementar el índice invertido para recuperación de texto usando el modelo de recuperación por ranking para consultas de texto libre. Considere los siguientes pasos generales:

- Preprocesamiento:
  - Tokenization
  - Filtrar Stopwords
  - Reducción de palabras *(Stemming)*
- Construcción del Índice
  - Estructurar el índice para obtener fácilmente los pesos TF-IDF.
  - Manejo del índice en memoria secundaria para soportar grandes colecciones de datos.
    - Blocked Sort-Based Indexing (slide 43-48)
    - Puede ayudarse de las siguientes lecturas: referencia 1, referencia 2.
- Consulta
  - La consulta es está formado por una o más palabras en lenguaje natural.
  - El scoring está basado en la similitud de coseno y retorna una lista ordenada de documentos que se aproximan a la consulta.

## 3- **Frontend**: Motor de Búsqueda

Para probar el desempeño del índice invertido, se debe construir una aplicación frontend que permita interactuar con las principales operaciones del índice invertido:

- Carga e indexación de documentos en tiempo real
- Búsqueda textual relacionado a ciertos temas de interés
- Presentación de resultados de búsqueda de forma amigable e intuitiva.

Se proveerá una colección de aproximadamente 20mil tweets de Twitter (carpeta “clean”). En donde el diccionario de términos puede construirse usando el contenido del atributo “text”, y el docID vendría a ser el Id del tweet. En la carpeta también se provee un código para extraer datos de Twitter (tracker.py). Enlace del repositorio

El grupo tiene la libertad de escoger cualquier tópico de enteres para realizar la recolección de Tweets. Por ejemplo, COVID-19 pandemic.

## 4- Entregable

Los alumnos formaran grupos de máximo de tres integrantes.
El proyecto estará alojado enteramente en GitHub, GitLab o Bitbucket.
En el Canvas se subir solo el **enlace público** del proyecto.
La fecha límite de entrega es el 09/06/2020 antes del mediodía (no habrá prorroga).
Los resultados deben visualizarse de forma amigable e intuitiva.

## 5- Informe del proyecto

- Archivo Readme en Markdown
- El archivo debe describir todos los aspectos importantes de la implementación.
- Se debe acompañar de imágenes o video de resultados.
- Ortografía y consistencia en los párrafos.
- Trabajar de forma colaborativa, se considerará para su nota individual.

## 6- Rúbrica

**Versión Beta**

| **Criterio** | **Puntos** |
|---|---|
| Construcción del Índice invertido | 5 |
| Manejo de memoria Secundaria | 4 |
| Ejecución Optima de Consultas | 3 |
| Presentación amigable al usuario. | 3 |
| Limpieza de código e informe. | 2 |
| Sustentación individual | 3 |