1 Objetivo

El objetivo de esta tarea es que los estudiantes desarrollen un sistema avanzado de recuperación de información, integrando dos técnicas fundamentales en el manejo de datos masivos y estructurados: el índice invertido y el algoritmo PageRank. Este proyecto busca que los estudiantes investiguen y apliquen una estructura de datos eficiente y un sistema de clasificación robusto, permitiéndoles comprender en profundidad los mecanismos subyacentes de los motores de búsqueda modernos.

2 Descripción

En esta tarea, los estudiantes deberán implementar un motor de búsqueda simplificado que indexe un conjunto de documentos web, donde cada documento contendrá enlaces a otros documentos (simulando hipervínculos). El proyecto constará de dos componentes principales:

- **Índice Invertido**: Los estudiantes implementarán un índice invertido que permita realizar búsquedas rápidas de palabras clave en los documentos.
- **PageRank**: Usarán un grafo dirigido para modelar las relaciones entre los documentos (enlaces entre ellos), implementarán el algoritmo de PageRank para calcular la importancia de cada documento en la web, y devolverán los resultados ordenados por relevancia.

3 Instrucciones detalladas

1. **Carga de documentos**:
   - Los documentos estarán en archivos de texto plano y cada documento puede contener enlaces a otros documentos, representados como `link: docN` en el texto.
   - Los estudiantes deberán procesar estos documentos para extraer las palabras clave y construir el grafo de enlaces entre ellos.
2. **Construcción del Índice invertido**:
   - Cada palabra del documento (exceptuando las "stopwords" comunes como "el", "la", etc.) deberá añadirse al índice invertido junto con el ID del documento.
   - Al realizar una consulta, el sistema debe devolver los documentos donde aparezca la palabra buscada.
3. **Construcción del grafo y cálculo de PageRank**:
   - Los estudiantes deben construir un grafo dirigido donde cada documento es un nodo y un enlace de un documento a otro es una arista dirigida.
   - Implementarán el algoritmo de PageRank para calcular la importancia de cada documento en función de los enlaces entrantes y salientes.
4. **Búsqueda y ordenación de resultados**:

- Al realizar una búsqueda, los documentos que contienen la palabra clave deben ser recuperados del índice invertido.
- Los documentos se ordenarán según su puntaje de PageRank (los documentos más relevantes se mostrarán primero).

Requisitos del Proyecto

- **Estructuras de Datos**:
  - Usar listas enlazadas o tablas hash para implementar el índice invertido.
  - Representar el grafo de enlaces entre documentos mediante listas de adyacencia.
- **Algoritmos**:
  - Implementar el algoritmo de PageRank en iteraciones, usando un factor de amortiguación (damping factor).
  - Algoritmos de búsqueda eficientes en el índice invertido.
- **Resultados**:
  - Al final, el sistema debe mostrar una lista de documentos ordenados por relevancia para una palabra o frase dada.

Formato de Entrega

- **Código fuente** bien documentado.
- **Informe** explicando la implementación, con ejemplos de ejecución.

La tarea es de a cuatro personas.

Importante: para comenzar esta tarea, cada grupo debe presentar un plan de trabajo en el github correspondiente. El plazo de publicación del plan de trabajo es una semana a partir de la fecha de entrega de la tarea.
