Proyecto Bimestral: Sistema de Recuperación de Información basado en Reuters-21578

Prof. Iván Carrera

27 de mayo de 2024

1. Introducción

El objetivo de este proyecto es diseñar, construir, programar y desplegar un Sistema de Recuperación de Información (SRI) utilizando el corpus Reuters-21578. El proyecto se dividirá en varias fases, que se describen a continuación.

2. Fases del Proyecto

2.1. Adquisición de Datos

- **Objetivo**: Obtener y preparar el corpus Reuters-21578.
- **Tareas**:
  - Descargar el corpus Reuters-21578.
  - Descomprimir y organizar los archivos.
  - Documentar el proceso de adquisición de datos.

2.2. Preprocesamiento

- **Objetivo**: Limpiar y preparar los datos para su análisis.
- **Tareas**:
  - Extraer el contenido relevante de los documentos.
  - Realizar limpieza de datos: eliminación de caracteres no deseados, normalización de texto, etc.
  - Tokenización: dividir el texto en palabras o tokens.
  - Eliminar stop words y aplicar stemming o lematización.
  - Documentar cada paso del preprocesamiento.

2.3. Representación de Datos en Espacio Vectorial

- **Objetivo**: Convertir los textos en una forma que los algoritmos puedan procesar.
- **Tareas**:
  - Utilizar técnicas como Bag of Words (BoW) y TF-IDF para vectorizar el texto.
  - Evaluar las diferentes técnicas de vectorización.
  - Documentar los métodos y resultados obtenidos.

2.4. Indexación

- **Objetivo**: Crear un índice que permita búsquedas eficientes.
- **Tareas**:
  - Construir un índice invertido que mapee términos a documentos.
  - Implementar y optimizar estructuras de datos para el índice.
  - Documentar el proceso de construcción del índice.

2.5. Diseño del Motor de Búsqueda

- **Objetivo**: Implementar la funcionalidad de búsqueda.
- **Tareas**:
  - Desarrollar la lógica para procesar consultas de usuarios.
  - Implementar algoritmos de similitud como similitud coseno o Jaccard.
  - Desarrollar un algoritmo de ranking para ordenar los resultados.
  - Documentar la arquitectura y los algoritmos utilizados.

2.6. Evaluación del Sistema

- **Objetivo**: Medir la efectividad del sistema.
- **Tareas**:
  - Definir un conjunto de métricas de evaluación (precisión, recall, F1-score).
  - Realizar pruebas utilizando el conjunto de prueba del corpus.
  - Comparar el rendimiento de diferentes configuraciones del sistema.
  - Documentar los resultados y análisis.

2.7. Interfaz Web de Usuario

- **Objetivo**: Crear una interfaz para interactuar con el sistema.
- **Tareas**:
  - Diseñar una interfaz web donde los usuarios puedan ingresar consultas.
  - Mostrar los resultados de búsqueda de manera clara y ordenada.
  - Implementar características adicionales como filtros y opciones de visualización.
  - Documentar el diseño y funcionalidades de la interfaz.

3. Entrega Final

- **Documentación Completa**: Incluyendo los procesos, decisiones tomadas, y resultados de cada fase.
- **Código Fuente**: Organizado y bien comentado.
- **Informe de Evaluación**: Análisis detallado de la evaluación del sistema.
- **Demostración del Sistema**: Presentación funcional del sistema a través de la interfaz web.

4. Requisitos Técnicos

- **Lenguajes de Programación**: Python (preprocesamiento y modelado), JavaScript (para la interfaz web).

5. Evaluación del Proyecto

- **Funcionamiento**: (35 %) Efectividad y eficiencia en la recuperación de información.
- **Documentación**: (35 %) Claridad en la documentación de cada fase.
- **Innovación y Creatividad**: (15 %) En la implementación de técnicas y la interfaz de usuario.
- **Presentación Final**: (15 %) Calidad y claridad de la demostración del sistema.