Motor de Búsqueda de Cursos

Pontificia Universidad Javeriana

Implementación práctica de técnicas de recuperación de información aplicadas al catálogo académico universitario

Estudiantes: Esteban Altamiranda - Felipe Morales

Objetivos del Proyecto

Motor de Búsqueda Académico

Desarrollar un sistema de búsqueda especializado para el catálogo de cursos universitarios con capacidades de indexación y recuperación avanzadas.

Web Crawling

Implementar técnicas de rastreo web automatizado para recopilar información estructurada de páginas académicas.

Procesamiento de Texto

Aplicar algoritmos de normalización, tokenización y eliminación de stopwords para optimizar la búsqueda.

Indexación y Similitud

Construir índices invertidos y métricas de similitud para mejorar la precisión en la recuperación de información.

Metodología Técnica Implementada

Tecnologías Utilizadas

- **Lenguaje:** Python 3.x
- **Librerías principales:** requests, BeautifulSoup, html5lib
- **Estructuras de datos:** CSV, JSON

01

Rastreo Web (Crawler)

Navegación automatizada y extracción de contenido académico del sitio web universitario.

02

Construcción de Índice

Generación de estructuras index.csv y courses.json para optimizar las consultas posteriores.

03

Sistema de Búsqueda

Implementación de algoritmos de búsqueda por palabras clave con ranking de relevancia.

04

Medida de Similitud

Cálculo de similitud entre cursos utilizando el coeficiente de Jaccard para recomendaciones.

Implementación del Crawler (crawler.py)

1

Estrategia BFS (Breadth-First Search)

Utiliza una cola FIFO para garantizar un rastreo sistemático y eficiente, visitando hasta *n* páginas dentro del dominio javeriano.

2

Extracción de Metadatos

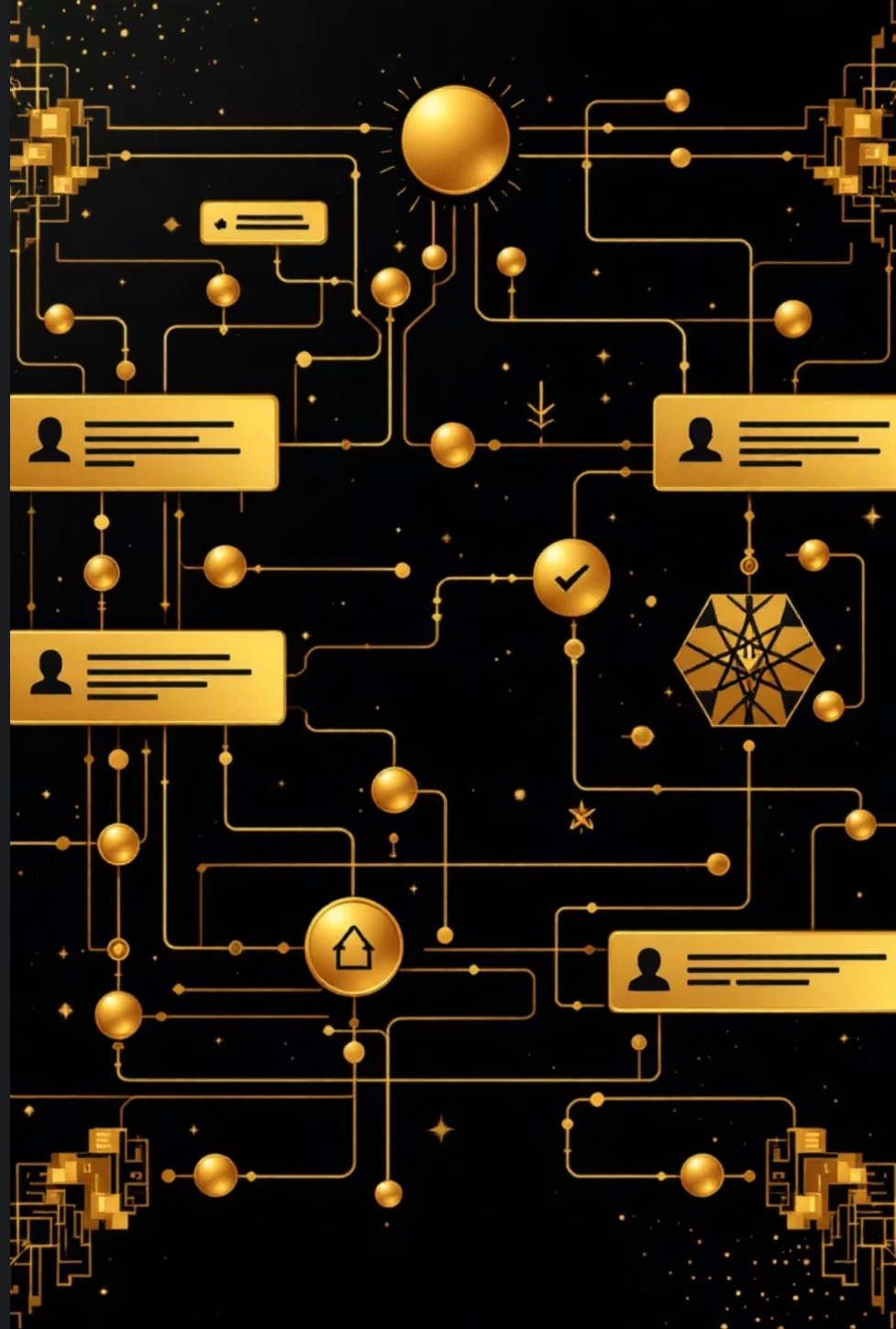
Captura información clave: títulos de cursos, descripciones detalladas, URLs canónicas y contenido textual relevante.

3

Procesamiento Lingüístico

Normalización de tokens, conversión a minúsculas, eliminación de stopwords en español y filtrado de caracteres especiales.

Archivos de salida: index.csv (pares curso-palabra) y courses.json (mapeo curso-URL) para facilitar búsquedas posteriores.



Estructura de Archivos Generados

1000+

Pares Curso-Palabra

Términos indexados en index.csv para búsqueda eficiente

100+

Cursos Mapeados

Enlaces directos almacenados en courses.json

3

Formatos de Salida

CSV, JSON y estructuras tabulares

index.csv

Contiene miles de pares curso-palabra optimizados para búsquedas rápidas y eficientes por términos específicos.

courses.json

Mapeo estructurado de más de 100 cursos universitarios con sus URLs correspondientes para acceso directo.

courses.csv

Listado tabular completo de cursos para análisis estadístico y visualización de datos académicos.

Módulo de Búsqueda (search.py)

Entrada de Consulta

Recibe lista de palabras clave del usuario y prepara el procesamiento de la consulta de búsqueda.

Procesamiento Avanzado

Tokenización inteligente, filtrado de stopwords en español y normalización de términos para optimizar resultados.

Scoring con IDF

Cálculo de relevancia usando Inverse Document Frequency: asigna mayor peso a términos menos frecuentes pero más específicos.

Resultados Ordenados

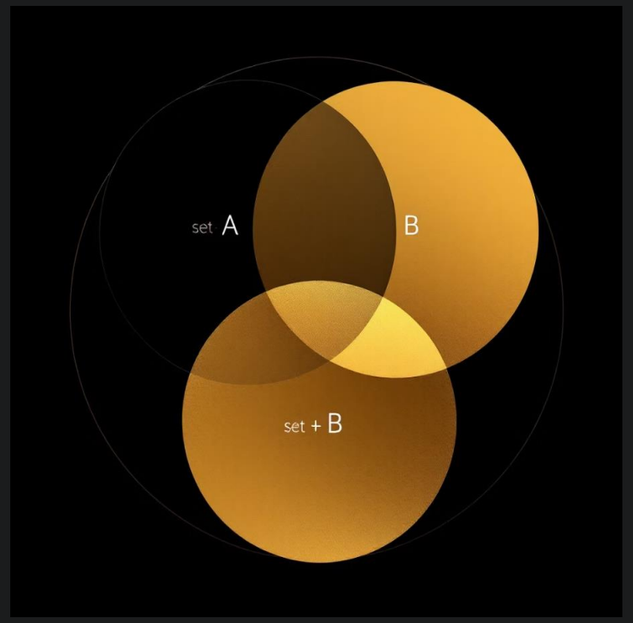
Lista de URLs ordenada por relevancia, presentando los cursos más pertinentes en las primeras posiciones.

Módulo de Comparación (compare.py)

Índice de Jaccard

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

El coeficiente de Jaccard mide la similitud entre conjuntos de palabras de diferentes cursos, proporcionando un rango de valores interpretables:



Ejemplo práctico: Comparación entre "Power BI y Python" vs "Analítica de Datos" muestra alta similitud por vocabulario técnico compartido.

Resultados Obtenidos en el Taller

1 Rastreo Exhaustivo Completado

Se mapearon exitosamente más de 100 cursos del catálogo académico javeriano, construyendo una base de datos comprehensiva del contenido educativo disponible.

2 Índice Invertido Robusto

Construcción de un índice invertido con miles de términos únicos, optimizado para consultas rápidas y recuperación eficiente de información académica.

3 Validación Temática Exitosa

Pruebas exhaustivas con búsquedas especializadas como "Inteligencia Artificial", "Data Science" y "Machine Learning" demostraron alta precisión en los resultados.

4 Medición de Similitud Implementada

Cálculo efectivo de similitud entre cursos con contenidos relacionados, habilitando funcionalidades de recomendación inteligente para estudiantes.

Conclusiones y Proyección Futura

Aplicación Práctica Exitosa

El taller demostró la efectividad de aplicar técnicas avanzadas de analítica de datos en un contexto académico real, consolidando conocimientos teóricos con implementación práctica.

Prototipo Funcional

El sistema desarrollado opera como un prototipo robusto de buscador académico, con capacidades de indexación, búsqueda y análisis de similitud completamente operativas.

Escalabilidad y Extensiones Futuras

Sistema de Recomendación

Implementación de algoritmos de recomendación personalizada basados en perfiles estudiantiles y similitud de contenidos.

Clasificación Temática Automática

Desarrollo de taxonomías automáticas para categorizar cursos por áreas de conocimiento y competencias específicas.

Detección de Redundancias

Identificación inteligente de solapamientos en la oferta académica para optimizar el catálogo educativo institucional.