UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS
POSTGRADO EN CIENCIAS DE LA COMPUTACIÓN

RECUPERACIÓN, EXTRACCIÓN Y CLASIFICACIÓN DE INFORMACIÓN DE SABER UCV

Trabajo de Grado de Maestría presentado ante la
ilustre Universidad Central de Venezuela por el
Econ. José Miguel Avendaño Infante para optar al título de
Magister Scientiarum en Ciencias de la Computación

Tutor: Dr. José Francisco Mirabal

Caracas - Venezuela
Marzo 2024

Resumen

Se presenta la investigación *Recuperación, Extracción y Clasificación de Información de Saber UCV*, donde se ejecutan procesos de clasificación, almacenamiento y recuperación de información sobre las tesis y trabajos de grado que se encuentran publicados en el repositorio institucional Saber UCV.

En tal sentido, se implementa un sistema que clasifica, según el área académica donde cursó estudios el autor, el 96% de las 9.982 investigaciones publicadas. Adicionalmente, con los textos de los resúmenes de los trabajos y con las clasificaciones obtenidas, se conforma un corpus al cual se le aplican técnicas de procesamiento de lenguaje natural, de minería de texto y con modelos de inteligencia artificial preentrenados se crean *embeddings* desde los documentos. Finalmente, con toda la información procesada se alimenta una base de datos indexada que contiene un índice invertido.

Por otra parte, el sistema cuenta con una aplicación web para hacer procesos de recuperación de información donde el usuario puede explorar el corpus, mediante la búsqueda semántica y la búsqueda de texto completo, indicando los siguientes valores: texto a buscar, rango de fechas, área en la cual se generó la investigación y nivel académico; posteriormente se recuperan los trabajos de mayor relevancia, enriqueciendo la experiencia con la presentación de los resultados en tablas interactivas, mapas de conocimiento y recomendaciones de documentos que puedan ser de interés.

La implementación se hace bajo un sistema distribuido con la arquitectura cliente-servidor y se soporta en el uso de contenedores orquestados.

Palabras Clave: recuperación de información, procesamiento del lenguaje natural, inteligencia artificial, embeddings, búsqueda semántica, mapas de conocimiento.

Abstract

The research *Recovery, Extraction and Classification of Information from Saber UCV*, is presented, where processes of classification, storage and retrieval of information on theses and degree works published in the institutional repository Saber UCV are executed.

In this sense, a system is implemented that classifies 96% of the 9,982 research papers to be categorized according to the academic area where the author of the research studied. Additionally, with the texts of the abstracts of the papers and the classifications obtained, a corpus is formed to which natural language processing and text mining techniques are applied, and with pre-trained artificial intelligence models, embeddings are created from the documents. Finally, all the processed information is fed into an indexed database containing an inverted index. On the other hand, the system has a web application for information retrieval processes where the user can explore the corpus, through semantic search and full text search, indicating the following values: text to search, date range, area in which the research was generated, academic level; subsequently, the most relevant works are retrieved, enriching the experience with the presentation of the results in interactive tables, knowledge maps and recommendations of documents that may be of interest.

The system is implemented under a distributed system with client-server architecture and is supported by the use of orchestrated containers.

Keywords: information retrieval, natural language processing, artificial intelligence, embeddings, semantic search, knowledge maps.

Dedicatoria:

A fulana,

perencejo.

Todos los que andan por ahí.

Agradecimientos:

- A mi mismo,

Como todos los hombres de la Biblioteca, he viajado en mi juventud; he peregrinado en busca de un libro, acaso del *catálogo de catálogos*; ahora que mis ojos caso no pueden descifrar lo que escribo, me preparo a morir a unas pocas leguas del hexágono en que nací.
— Jorge Luis Borges, *La Bibioloteca de Babel*, Ficciones

Every important aspect of programming arises somewhere in the context of sorting or searching.

— Donald Knuth, *The Art of Computer Programming*, Volume 3

Tabla de contenidos

1 Contenido

Esto es un **demo** para un *libro* Quarto book

Más info sobre documentación Quarto https://quarto.org/docs/books.

Voy a crear referencia a un paquete (Wickham et al., 2023) que hemos usado ampliamente

1.1 Sesiones Programadas

7-10 Config YML, Build- Render pdf y/o html

9-10 Capítulos, Preview

16-10 configurar GIT

21-10 Citas: zotero, DOI,packages. formato APA

4-11 Referencias cruzadas, imágenes

6-11 Tablas y gráficos, chunks según formato destino

11-11 Git Sync

13-11 usar Netlify para hospedaje, Publicarlo en una servidor gratuito (sin publicidad)

2 Introducción a la Economía

Bienvenidos a este curso introductorio de economía. La economía es el estudio de cómo las sociedades gestionan sus **recursos escasos**. Esta definición simple abarca un campo de estudio inmenso y complejo.

En este libro, dividiremos nuestro estudio en dos ramas principales:

1. **Microeconomía**: El estudio de cómo los hogares y las empresas toman decisiones e interactúan en los mercados.
2. **Macroeconomía**: El estudio de los fenómenos que afectan al conjunto de la economía, como la inflación, el desempleo y el crecimiento económico.

Comenzaremos explorando los fundamentos de la oferta y la demanda en el Capítulo 3. Posteriormente, analizaremos los grandes agregados en el Capítulo 4.

3 Microeconomía: Oferta y Demanda

Como mencionamos en la Capítulo 2, la microeconomía se centra en las decisiones individuales. El modelo más fundamental en microeconomía es el de la **oferta y la demanda**.

3.1 El Modelo de Mercado

Un mercado es un grupo de compradores y vendedores de un bien o servicio en particular.

- La **demanda** (D) representa la cantidad de un bien que los compradores están dispuestos y son capaces de comprar.
- La **oferta** (S) representa la cantidad que los vendedores están dispuestos y son capaces de vender.

3.2 El Equilibrio

El punto donde ambas curvas se cruzan se llama **equilibrio de mercado**. En este punto, el precio ha alcanzado un nivel en el que la cantidad ofrecida equivale a la cantidad demandada.

Este equilibrio se ilustra claramente en el gráfico de oferta y demanda, como se ve en la Figura 3.1.

Cualquier precio por encima de P^* resultará en un excedente, mientras que cualquier precio por debajo causará escasez.

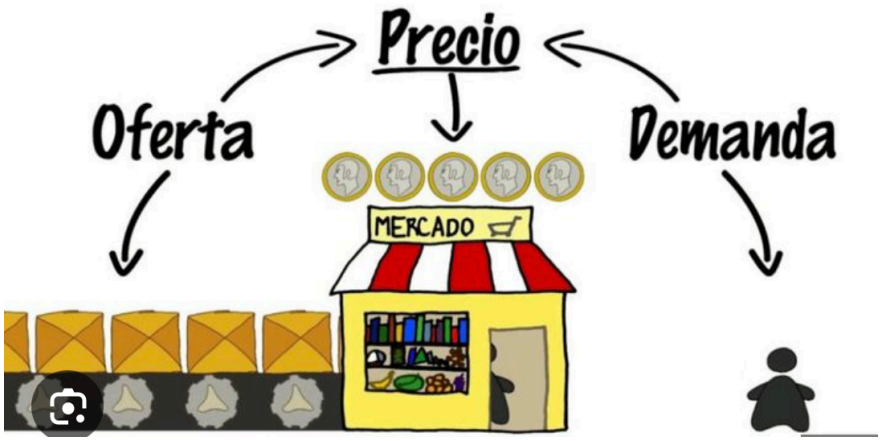

Figura 3.1: Un gráfico estándar de oferta y demanda que muestra el precio de equilibrio (P^*) y la cantidad de equilibrio (Q^*).

4 Macroeconomía: Medición del Ingreso

Ahora pasamos a la macroeconomía. A diferencia del análisis de mercados individuales que vimos en Capítulo 3, aquí nos interesa la economía en su conjunto.

4.1 El Producto Interno Bruto (PIB)

La estadística más importante para medir la salud económica de un país es el **Producto Interno Bruto (PIB)**.

> El PIB es el valor de mercado de todos los bienes y servicios finales producidos dentro de un país en un período de tiempo determinado.

El PIB (Y) se compone de cuatro elementos:

$$Y = C + I + G + (X - M)$$

Donde: * C = Consumo * I = Inversión * G = Gasto del Gobierno * $(X - M)$ = Exportaciones Netas

Mientras que nuestro análisis en la Figura 3.1 se centraba en el precio y la cantidad de un solo bien, el PIB suma el valor de *todos* los bienes.

5 Summary

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

Bibliografía

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation*. https://doi.org/10.32614/CRAN.package.dplyr