

Reporte de Proyecto Grupal 2

Acciones con carpeta de One Drive «

000_ArticulosAgrupar »

Moreno Ledesma Ximena Abigail*, Parras Pecina Maria Fernanda* y Torres Colorado Juan Daniel*

Ingeniería en Tecnologías de la Información

Universidad Politécnica de Victoria

Resumen—En el presente informe se detalla el desarrollo de una interfaz gráfica diseñada para realizar tres tareas específicas. La interfaz permite filtrar los documentos y organizarlos en carpetas separadas según su idioma, clasificando los documentos en español y en inglés, esto es logrado cuando se haya entrenado correctamente el clasificador. Finalmente, la interfaz incluye herramientas para aplicar técnicas de clusterización a los documentos en inglés para agrupar los documentos en función de la similitud de su contenido.

I. INTRODUCCIÓN

El objetivo de este proyecto es desarrollar una interfaz gráfica de usuario (GUI) que facilite la gestión y análisis de documentos PDF en Google Drive. Su importancia radica en automatizar y optimizar tareas que normalmente consumen mucho tiempo, como la clasificación de documentos por idioma y la aplicación de técnicas de clusterización para agrupar los documentos en inglés, mejorando la eficiencia en la organización y análisis de grandes volúmenes de documentos PDF.

La GUI es una interfaz entre la persona y la máquina. El objetivo de esta interfaz gráfica es representar el código del backend de un sistema de la forma más clara posible para el usuario para simplificarle las tareas diarias. Para esto, son muy importantes los iconos y las imágenes, ya que solo estos permiten una aplicación universal e independiente del texto. [1]

Un Dataset, como su nombre lo dice, es simplemente un conjunto de datos, ordenado bajo un sistema de almacenamiento que otorga los lineamientos principales de búsqueda o directorio de la información que se quiere trabajar. Este conjunto de datos por supuesto que se puede utilizar para muchas cosas, dependiendo de la metodología, orientación o tratamiento que se le quiera dar a la información. Su finalidad es hacer mucho más fácil la vida a las personas, automatizar tareas o simplemente analizar información de una manera más ágil. [2]

Este informe detalla el proceso de desarrollo de una interfaz gráfica y sistema con el objetivo de manejar tres funciones principales con un dataset: Filtrar los documentos en documentos en los idiomas de español e inglés, finalmente solo para los documentos en inglés, aplicar técnicas de clusterización. Esto con ayuda de PyQt, una librería de Python para crear aplicaciones GUI utilizando el kit de herramientas Qt, y la

versión más reciente siendo PyQt6 [3] lanzada en el año 2021. Se abordarán conceptos relacionados con *Natural Language Processing: Classification*, conversión de documentos PDF a TXT, técnicas de clusterización para el procesamiento, almacenamiento de datos, específicamente documentos de formato PDF, utilizando tecnologías de almacenamiento en la nube como OneDrive. [4]

La librería PyQt de Python se utiliza para escribir todo tipo de aplicaciones GUI, desde herramientas de visualización utilizadas por científicos e ingenieros, hasta aplicaciones contables.[5] En este caso se hizo uso de la versión PyQt6.

Natural Language Processing (NLP) o Procesamiento del lenguaje natural (PNL) es una rama de la inteligencia artificial (IA) que permite a las computadoras comprender, generar y manipular el lenguaje humano. El procesamiento del lenguaje natural tiene la capacidad de interrogar los datos con texto o voz en lenguaje natural.[6]

Las stopwords son palabras comunes que suelen eliminarse de los textos antes de realizar análisis, ya que no aportan mucho significado (*the, and, is*), mejorando la eficiencia y precisión del análisis eliminando ruido.[7]

Tokenization consiste en dividir el texto en unidades menores como palabras, frases o párrafos. Es el primer paso en muchos procesos de NLP, facilitando la manipulación y análisis del texto.

La clusterización, también conocida como agrupamiento o clustering, es una técnica de análisis de datos que consiste en dividir un conjunto de datos en grupos (o clusters) de tal manera que los datos dentro de un mismo grupo sean más similares entre sí que con los datos de otros grupos.

Técnicas de clusterización

- *Tf-idf and Document Similarity.* TF-IDF (Term Frequency-Inverse Document Frequency) es una técnica de vectorización que transforma los textos en vectores numéricos. Calcula la importancia de una palabra en un documento en relación con un corpus de documentos. Ayuda a identificar las palabras más relevantes para cada documento, mejorando el análisis de similitud entre documentos.

- *Kmeans Clustering.* Kmeans es un algoritmo de particionamiento que agrupa los documentos en k clusters basándose en sus características vectoriales. Agrupa do-

- cumentos similares en clusters, facilitando la organización y análisis temático.
- *Multidimensional Scaling (MDS)*. Es una técnica de reducción de dimensionalidad que proyecta datos de alta dimensión en un espacio de menor dimensión, preservando las distancias entre puntos. Facilita la visualización de la estructura de los datos, especialmente útil para visualizar clusters.
- *Visualizing Document Clusters*. Implica usar técnicas de visualización como t-SNE o PCA para representar gráficamente los clusters de documentos. Ayuda a entender la estructura y separación de los clusters, facilitando la interpretación de los resultados.
- *Hierarchical Document Clustering*. Construye un árbol jerárquico (dendrograma) que agrupa documentos en múltiples niveles de granularidad. Útil para explorar la estructura jerárquica de los datos y determinar el número óptimo de clusters.
- *Latent Dirichlet Allocation (LDA)*. Es una técnica de modelado de temas que identifica temas subyacentes en un corpus de documentos y asigna probabilidades de pertenencia a estos temas para cada documento. Facilita la identificación de temas y la organización de documentos en función de sus temas principales.[8]

II. DESARROLLO EXPERIMENTAL

Antes de entrar con detalle a las funcionalidades clave del proyecto, cabe aclarar que el procesamiento de la información se llevó a cabo gracias al uso de la utilidad *pdftotext* [9] de la librería *popper-utils*, la cuál convierte archivos en formato de documento portátil (PDF) en texto sin formato. Además, es necesaria la instalación de librerías como *nltk* y *joblib* para su proceso de clusterización de documentos en inglés, y para su representación visual la instalación de librerías como *numpy*, *matplotlib*, *pandas*, *scipy* y *sklearn*. Caso contrario, la ejecución de este programa no podrá suceder debido a los anteriores factores. Por último, se decidió descartar la idea del renombramiento de documentos con la sintaxis *[Journal]_Titulo.pdf*, esto debido a que, si bien es cierto que se pueden acceder a la lectura de metadatos de los documentos PDFs, cada autor de cada documento puede o no puede haber colocado dicha información en las mismas secciones, causando que para cada documento se puede colocar esta información en secciones diferentes y no tener una certeza asegurada; dicho proceso fue implementado pero al final se decidió optar por no realizar este procedimiento debido a razones anteriormente mencionadas.

Se llevaron a cabo distintos procedimientos para lograr cumplir con el objetivo del proyecto. Dichos procedimientos serán mencionadas a continuación.

II-A. Funcionalidades Clave

Filtrado de Documentos por Idioma:

- Separación automática de documentos en español e inglés.

- Manejo de documentos con resúmenes y contenidos en distintos idiomas.

Clusterización de Documentos en Inglés:

- Aplicación de técnicas de NLP y clusterización para agrupar documentos por temas o contenido similar.
- Visualización de agrupamientos.

II-B. Filtrado de Documentos por Idioma

En esta sección se describe el proceso de filtrado automático de documentos en español e inglés. A continuación, se detallan los pasos y consideraciones clave:

II-B1. Identificación del Idioma: Para identificar el idioma de un documento, se logra mediante el proceso de clustering y el análisis de las palabras más frecuentes en los documentos agrupados. Al agrupar los documentos en clusters basados en la similitud de su contenido, los documentos en el mismo idioma tienden a agruparse juntos, debido a la similitud lingüística y de vocabulario entre ellos. El proceso implica:

1. **Vectorización del Contenido:** Los contenidos de los documentos se convierten en vectores TF-IDF, permitiendo capturar las características importantes del texto.
2. **Clustering con K-Means:** Se aplica el algoritmo de K-Means para agrupar los documentos. Los documentos con contenido similar, es decir, con un vocabulario similar, se agrupan juntos.
3. **Análisis de Términos Principales:** Se examinan los términos más representativos en cada cluster. Esto revela los patrones lingüísticos predominantes en cada grupo, permitiendo inferir el idioma.

Estas técnicas permiten una clasificación precisa y eficaz de documentos bilingües, asegurando que sean correctamente identificados y gestionados según el idioma predominante.

II-C. Clusterización de Documentos en Inglés

II-C1. Descripción de las Técnicas: Las técnicas de clusterización de documentos permiten agrupar textos similares en conjuntos coherentes, facilitando la organización y el análisis de grandes volúmenes de información. Según el enlace proporcionado en <http://brandonrose.org/clustering>, se destacan varias técnicas comunes:

- **K-means:** Un algoritmo de agrupamiento que asigna puntos de datos a clusters para minimizar la varianza dentro de cada cluster.
- **Clustering Jerárquico:** Organiza los datos en una estructura de árbol o jerarquía, donde los clusters se forman fusionando o dividiendo clusters sucesivamente.
- **Clustering Espectral:** Utiliza la matriz de afinidad de los datos para proyectarlos en un espacio de menor dimensión, donde los clusters se identifican como particiones de la matriz.

II-C2. Aplicación específica a los documentos en inglés: Para documentos en inglés, estas técnicas se aplican utilizando herramientas de Procesamiento de Lenguaje Natural (NLP) que procesan el texto para representarlo en un espacio vectorial donde se pueden aplicar algoritmos de clusterización. Esto

permite identificar temas comunes, categorías o similitudes semánticas entre los documentos.

II-C3. Implementación del Algoritmo: La implementación de estos algoritmos implica los siguientes pasos:

1. **Preprocesamiento de Texto:** Normalización del texto, eliminación de stopwords, lematización o derivación de palabras.
2. **Representación Vectorial:** Convertir el texto preprocesado en vectores numéricos [10].
3. **Aplicación de Clustering:** Utilizar algoritmos como K-means o Clustering Jerárquico sobre los vectores para agrupar los documentos en clusters significativos.

Estas técnicas y herramientas permiten una clusterización eficaz de documentos en inglés, facilitando la exploración y comprensión de grandes conjuntos de datos textuales.

III. RESULTADOS

Los resultados de este proyecto resultan en el filtrado de documentos en español y en inglés en carpetas separadas y la aplicación de clusterización anteriormente mencionadas para documentos en inglés. En la figura 1 demuestra la ventana inicial contando con los elementos:

- **Seleccionar directorio de PDFs:** Tiene la funcionalidad de demostrar una ventana emergente con la capacidad de seleccionar un directorio específico en la cual será usada para realizar el objetivo de este proyecto. Además, si el directorio es válido, entonces se procederá automáticamente a crear 4 carpetas para llevar una clasificación organizada, en la cuál, dentro del directorio proporcionado se creará una carpeta de nombre "documentos" en ella 3 carpetas para organizar de manera precisa, "analysis" usada para almacenar documentos de texto legible, "Inglés" y "Español" que clasifican dichos PDFs.
- **Convertir archivos PDFs a TXTs:** Reliza el procedimiento de convertir los archivos PDFs del directorio proporcionado a TXT en la carpeta "analysis", esto con la finalidad de que el contenido sea legible y sea manipulado.
- **Clasificar documentos:** Una vez convertido los archivos PDFs a TXTs, se realizará la clasificación de documentos en inglés y en español a través de su contenido y los documentos identificados serán transladados a su carpeta correspondiente.
- **Realizar clustering:** Por último, una vez clasificado los documentos, entonces se podrá realizar la aplicación de clustering de documentos en inglés, que posteriormente los resultados serán visualizados en pantalla.

Para cada procedimiento anterior, se realizó la notificación mensajes de advertencias 6 que se pueden producir durante el programa de mensajes al igual de informativos 7 (e.g., la advertencia del congelamiento del programa 8).

En la figura 2 se representa la ventana en la que, gracias a la clase *QFileDialog* de la librería de *PyQt6*, permite que la selección del directorio se proporcione de una manera más fácil y sencilla de obtener.

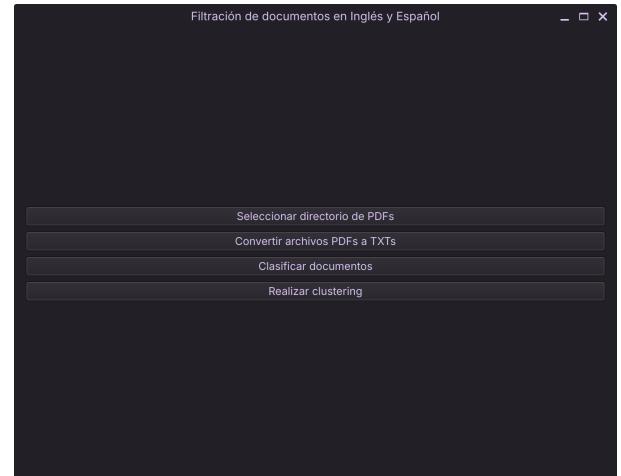


Figura 1: Interfaz de funcionales del sistema.

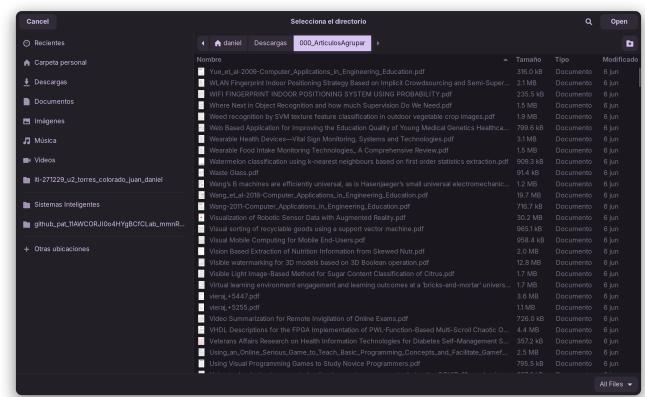


Figura 2: Selección del directorio que contiene los documentos

Una vez terminado el proceso de la clasificación, se moverán los documentos PDF hacia cada grupo (inglés o español) 3, exceptuando por los documentos que no cuenten con un mínimo de 100 palabras. Debido a la cantidad grande de documentos, se optó por solo demostrar una vista general.

En las figuras 4 y 5 se representa la visualización del resultado de la clusterización, mostrando las categorización obtenida para cada documento.

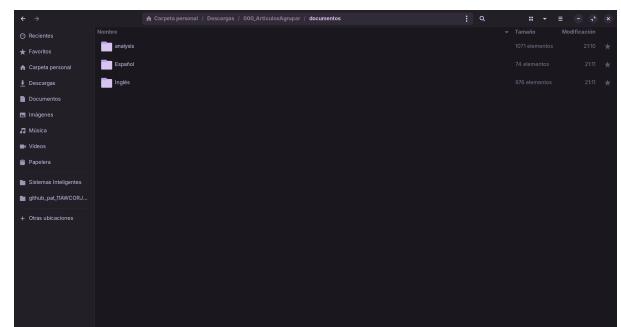
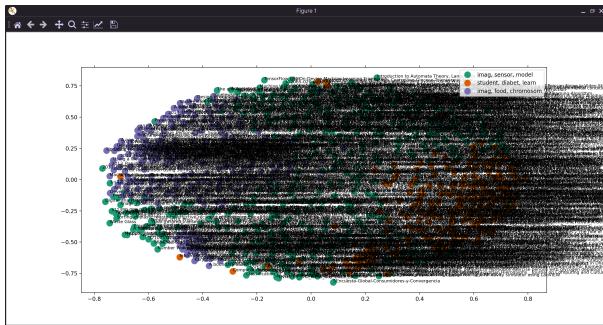
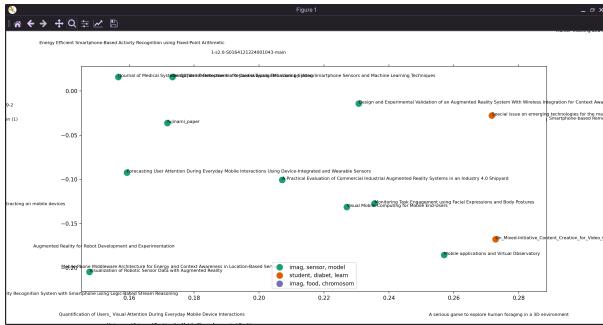


Figura 3: Resultado de clasificación

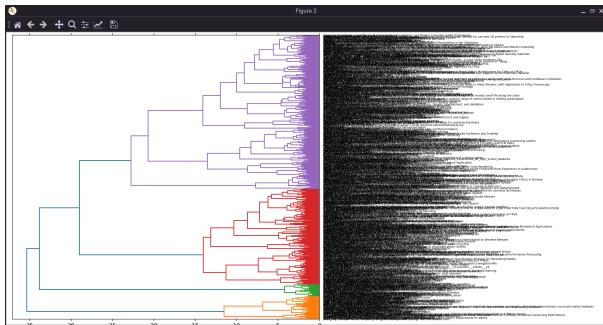


(a) Visualización completa

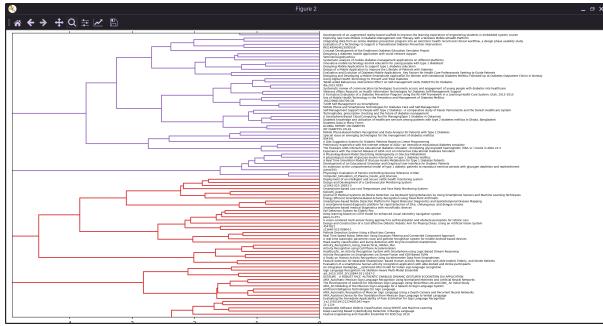


(b) Acercamiento de la visualización

Figura 4: Visualización de la cluterización, mostrando en qué categoría se encuentran



(a) Visualización completa



(b) Acercamiento de la visualización

Figura 5: Visualización de la cluterización jerárquico

IV. CONCLUSIÓN

En este proyecto se ha desarrollado una interfaz gráfica de usuario (GUI) que facilita la gestión y análisis de documentos PDF almacenados en Google Drive. La herramienta implementa varias funcionalidades clave que optimizan tareas comunes y mejoran la eficiencia en la organización de grandes cantidades de documentos. Se tomaron como objetivo tres funciones: Primero, se la implementación de un filtrado automático que clasifica los documentos por idioma, separando los documentos en español y en inglés. En segundo lugar, un método para acceder y extraer metadatos de los documentos, permitiendo renombrar los archivos según una sintaxis específica que incluye el nombre del journal y el título del documento. Esto no solo facilita la identificación y búsqueda de documentos, sino que también estandariza su nomenclatura para una mejor gestión. Finalmente, se han aplicado técnicas avanzadas de clusterización a los documentos en inglés, proporcionando una visión clara de la estructura y los temas predominantes en el conjunto de datos.

La implementación de estas funcionalidades con la ayuda de PyQt6 y diversas librerías de procesamiento de lenguaje natural (NLP) y machine learning demuestra la versatilidad de las herramientas de software para la automatización y análisis de datos.

REFERENCIAS

- [1] IONOS. *¿Qué es una interfaz gráfica de usuario (GUI)?* <https://www.ionos.mx/digitalguide/paginas-web/desarrollo-web/que-es-una-gui/>. Consultado el 5 de junio de 2024.
- [2] Diego Caceres Solis. *Datasets: Qué son y cómo acceder a ellos.* <https://openwebinars.net/blog/datasets-que-son-y-como-acceder-a-ellos/>. Consultado el 17 de junio de 2024.
- [3] Martin Fitzpatrick. *PyQt6.* <https://www.pythonguis.com/pyqt6-tutorial/>. Consultado el 5 de junio de 2024.
- [4] Microsoft. *Microsoft OneDrive.* <https://www.microsoft.com/es-mx/microsoft-365/onedrive/online-cloud-storage>. Consultado el 5 de junio de 2024.
- [5] Mark Summerfield. *Rapid GUI Programming with Python and Qt: The Definitive Guide to PyQt Programming.* Prentice Hall, 2007.
- [6] Oracle. *What Is Natural Language Processing (NLP)?* <https://www.oracle.com/ph/artificial-intelligence/what-is-natural-language-processing/>. Consultado el 5 de junio de 2024.
- [7] Emmanuel Ameisen. *Building Machine Learning Powered Applications.* O'Reilly Media, 2020.
- [8] Brandon Rose. *Document Clustering with Python.* <http://brandonrose.org/clustering>. Consultado el 5 de junio de 2024.
- [9] LLC Glyph & Cog. *pdftotext.* <https://www.xpdfreader.com/pdftotext-man.html>. Consultado el 5 de junio de 2024.

- [10] Sitio Big Data. *Machine Learning Procesamiento de texto*. <https://sitiodataglobal.com/2019/12/23/machine-learning-procesamiento-de-texto/>. Consultado el 5 de junio de 2024.

V. ANEXOS

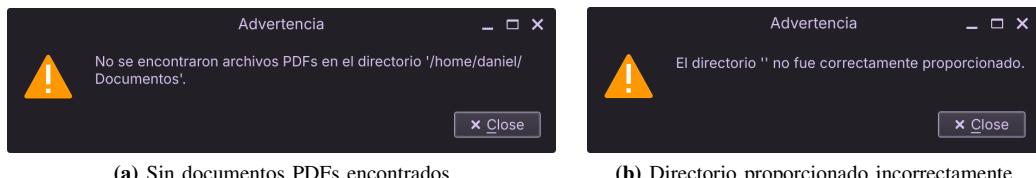


Figura 6: Mensajes de advertencia

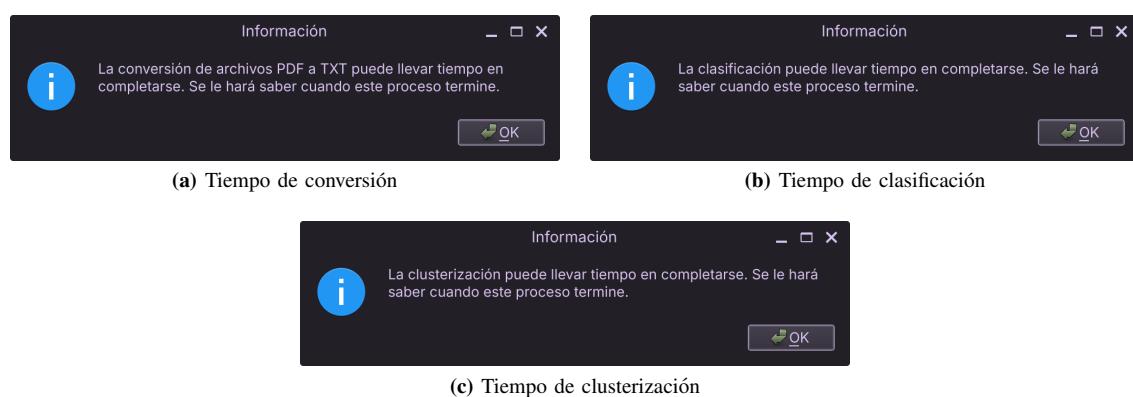


Figura 7: Mensajes informativos

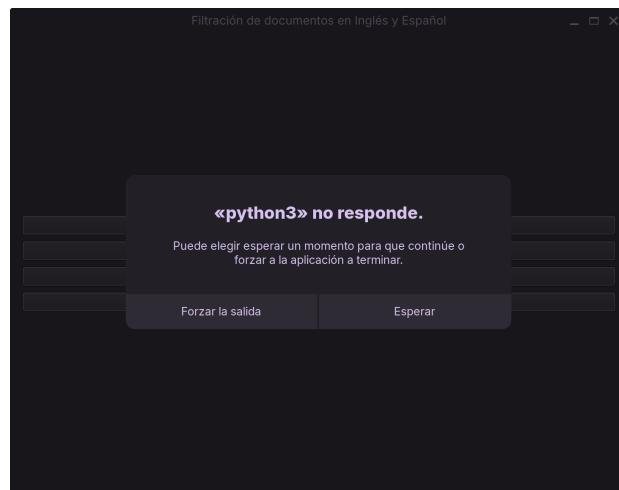


Figura 8: Congelamiento de programa