Implementar un Buscador de Documentos. Crear un pequeño buscador para recuperar documentos relevantes según una consulta.

Actividades:

1. Preprocesamiento de Documentos PDF: Elegir un conjunto de documentos en formato PDF.
   - Utilizar bibliotecas en Python como PyMuPDF o pdfplumber para extraer el texto.
   - Aplicar técnicas de limpieza, como: Eliminación de caracteres especiales. Conversión a minúsculas. Eliminación de stopwords en español. Tokenización y lematización usando spaCy o NLTK.
   - Guardar el resultado en archivos de texto procesados.
2. Implementar un modelo de búsqueda y la representación vectorial de los documentos basandose en TF-IDF, BM25 o Word2Vec.
3. Permitir que un usuario ingrese una consulta en lenguaje natural.
4. Retornar los documentos más relevantes ordenados por puntaje de similitud. Comparar los documentos calculando la similitud con la métrica de coseno. Permitir que la aplicación recomiende documentos similares a uno seleccionado.
5. Evaluar la calidad de las recomendaciones con métricas
6. Crear una interfaz sencilla (puede ser en Jupyter Notebook, Streamlit, Fllask o algún otro).
7. Entregar un informe con el código, capturas de pantalla y conclusiones.

Realizar 3 grupos. Cada grupo escogerá uno de los siguientes métodos de clustering: K-Means, DBSCAN, Hierarchical Clustering, etc

Recomendaciones:

- Librerías: spaCy, NLTK, scikit-learn, pandas, numpy, matplotlib, seaborn, scipy, gensim, streamlit
- Datasets sugeridos: Artículos académicos, noticias, papers de arXiv, documentos institucionales.
- Utilizar KDD

Comparación de Algoritmos

Algoritmo	Necesita definir clusters	Maneja diferentes densidades	Computacionalmente eficiente	Forma de clusters adecuada
K-Means	✅ Sí	❌ No	✅ Rápido	🔵 Esféricos
DBSCAN	❌ No	✅ Sí	✅ Rápido	🔺 Cualquier forma
Hierarchical Clustering	✅ Sí	✅ Sí	❌ Puede ser lento	🔺 Cualquier forma
Mean-Shift	❌ No	✅ Sí	❌ Lento	🔺 Cualquier forma
GMM	✅ Sí	❌ No	❌ Lento	🔵 Elípticos
Affinity Propagation	❌ No	✅ Sí	❌ Lento	🔺 Cualquier forma
OPTICS	❌ No	✅ Sí	❌ Lento	🔺 Cualquier forma
Spectral Clustering	✅ Sí	❌ No	❌ Lento	🔺 Cualquier forma

¿Cuál elegir?

- Si los datos tienen clusters bien separados y esféricos: ✅ K-Means.
- Si los clusters tienen formas irregulares y densidades variables: ✅ DBSCAN u OPTICS.
- Si no sabemos cuántos clusters hay: ✅ Mean-Shift, Affinity Propagation o DBSCAN.
- Si los datos tienen una estructura compleja: ✅ Spectral Clustering.
- Si queremos probabilidades en la asignación de clusters: ✅ GMM.

Fecha de entrega: 28/03/2025