**Clasificación y búsqueda semántica de documentos en SharePoint**

**🔹 Flujo propuesto mejorado**

**1. Ingesta y preprocesamiento**

* Conectar con la API de **Microsoft Graph** para listar documentos de SharePoint.
* Crear un **índice de control** (en SQL o en una lista de SharePoint) donde se guarde:
  + ID del archivo
  + Fecha de última modificación
  + Estado de etiquetado
* Así solo procesas documentos **nuevos o modificados** (evitas reprocesar lo ya hecho).

**2. Verificación de etiquetas**

* Revisar si el archivo ya tiene etiquetas (metadatos de SharePoint).
* Si tiene → **lo saltas**.
* Si no tiene → lo envías al flujo de etiquetado.

**3. Flujo de etiquetado con IA**

* Para cada archivo no etiquetado:
  1. Extraer las **primeras 500–1000 palabras** (dependiendo del tipo de archivo: PDF, DOCX, PPTX, TXT, etc.).
  2. Usar un **modelo de clasificación** con contexto:
     + Lista de *tickers* de empresas que tú definas.
     + Lista de *categorías generales* (ej. "Finanzas", "Legal", "Comercial", etc.).
  3. El modelo asigna una o varias etiquetas:
     + Ticker: SAN
     + Tipo: Reporte Financiero
     + Tema: Riesgo / Mercado
* Esas etiquetas se guardan como **metadatos en SharePoint** (para que queden disponibles en la búsqueda).

**4. Indexación semántica**

* Además del etiquetado, puedes crear un **embedding vectorial** de cada archivo (con Azure OpenAI, OpenAI o modelos locales).
* Guardar esos embeddings en un **almacén vectorial** (ej. Azure Cognitive Search, Pinecone, Weaviate, o en una tabla optimizada en SQL).
* Esto permitirá búsquedas mucho más naturales (no solo por etiquetas).

**5. Buscador en Teams / SharePoint**

* Crear un **bot en Teams** o un **copilot personalizado en SharePoint** que use:
  1. **Consulta en lenguaje natural** (ej: "muéstrame reportes del Banco Santander sobre riesgo de crédito en 2023").
  2. IA convierte la consulta → embeddings.
  3. Se buscan coincidencias en el almacén vectorial + filtrado por etiquetas (tickers).
  4. Devuelve los documentos más relevantes (con snippet o link directo a SharePoint).

**🔹 Mejoras respecto a tu idea inicial**

* **No procesar todo de nuevo siempre** → ahorro de cómputo.
* **Combinar etiquetas + embeddings** → resultados más precisos y flexibles.
* **Separar ticker y categoría** → te da doble nivel de búsqueda.
* **Bot en Teams** → más amigable para el usuario final.
* **Automatización** → se puede correr 1 vez al día con Azure Functions, n8n o Power Automate.