Instituto Tecnológico y de Estudios Superiores de Monterrey

Maestría en Inteligencia Artificial:

Spin Voyager & Spin Compass: Avance 1. Análisis exploratorio de datos

Tania Alicia Caballero Saavedra - A01795957

Oscar Enrique García García - A01016093

Dante Rosas Fragoso – A01795850

Profesor Titular: Dra. Grettel Barceló Alonso

Asesor: Dr. Horacio Martínez Alfaro

Proyecto Integrador

Enero, 2026

Índice

1. Introducción

El presente documento establece la estrategia técnica, el análisis exploratorio y los protocolos de limpieza de datos necesarios para el desarrollo de **Spin Compass**, el asistente conversacional inteligente diseñado para la organización Spin. Este proyecto tiene como objetivo fundamental transformar la manera en que los colaboradores («*Spinners*») acceden a la información institucional, migrando de un modelo de gestión manual y dispersa a una arquitectura de **Generación Aumentada por Recuperación (RAG)** centralizada y eficiente.

1.1. Contexto del Problema y Justificación

Actualmente, el área de Finanzas de Spin enfrenta una carga operativa crítica derivada de la atención constante a consultas repetitivas sobre políticas de viajes, viáticos, telefonía y equipo de cómputo. La información oficial reside en formatos no estructurados y se encuentra fragmentada en múltiples plataformas como Google Drive, Canvas, La Órbita y canales de Slack, lo que provoca respuestas inconsistentes, fricción operativa y una dependencia excesiva del conocimiento tribal del equipo humano.

Para resolver esto, **Spin Compass** se integrará directamente en Slack como un agente capaz de recuperar información precisa y generar respuestas naturales. Sin embargo, la eficacia de este agente no depende únicamente del modelo de lenguaje (LLM) elegido, sino principalmente de la calidad y estructura de los datos que lo alimentan.

Como establece el principio fundamental de RAG:

Si los datos de entrada están desordenados, son redundantes o carecen de estructura, las respuestas del agente reflejarán esas mismas deficiencias.

1.2. Alcance del Análisis Exploratorio de Datos (EDA) para RAG

A diferencia de un EDA tradicional enfocado en estadísticas descriptivas de valores numéricos, este análisis se centra en la calidad semántica y la recuperabilidad del texto. El objetivo es transformar documentos estáticos (PDFs, manuales) en una base de conocimiento dinámica («Trusted Source») que el agente pueda consultar de manera confiable.

Este documento aborda los siguientes desafíos críticos para **Spin Compass**:

1. **Estandarización de Formatos:** Los LLMs pueden malinterpretar documentos visualmente complejos. Se define la estrategia para convertir archivos PDF y documentos de Word a formato Markdown, preservando la jerarquía de títulos y tablas para asegurar que el agente comprenda la estructura lógica de las políticas (ej. distinguir entre una restricción de viáticos nacional vs. internacional).
2. **Integridad Semántica (Chunking):** Los métodos tradicionales de división de texto (**fixed-size chunking**) a menudo fragmentan ideas completas, rompiendo el contexto necesario para una respuesta correcta. Basándonos en literatura reciente, este documento propone estrategias de segmentación semántica (como **Max-Min** o **Growing Window**) que agrupan oraciones por su significado, asegurando que **Spin Compass** recupere reglas de negocio completas y coherentes, reduciendo el riesgo de alucinaciones.
3. **Limpieza y Enriquecimiento:** Se establecen protocolos para eliminar «ruido» (pies de página, descargos legales irrelevantes) que entorpece la búsqueda vectorial, y se define una estrategia de metadatos (autor, fecha de vigencia) para garantizar que el bot siempre priorice la información más actual.
4. **Privacidad y Seguridad:** Dado que el sistema procesará consultas reales de los empleados, se incluyen lineamientos para la detección y enmascaramiento de Información de

Identificación Personal (PII) antes de la indexación, cumpliendo con los requisitos de seguridad de la organización.

En resumen, este documento no solo describe los datos actuales de Spin, sino que prescribe el tratamiento de ingeniería necesario para que **Spin Compass** actúe como un «compañero de trabajo» competente, capaz de descargar al equipo de Finanzas y ofrecer una experiencia ágil y precisa a todos los Spinners.

2. Diferencias Clave: EDA para RAG vs. EDA para Machine Learning (ML)

Es fundamental distinguir el Análisis Exploratorio de Datos (EDA) que realizaremos para Spin Compass de un EDA tradicional. Mientras que en el aprendizaje automático clásico (ML) el objetivo es entender distribuciones estadísticas para predecir valores, en RAG el objetivo es evaluar la coherencia semántica y la «recuperabilidad» de la información no estructurada.

A continuación, se detallan las diferencias críticas en cuatro dimensiones: **Naturaleza de los Datos**, **Unidad de Análisis**, **Limpieza** y **Métricas de Éxito**.

2.1. Naturaleza de los Datos y Estructura

En un proyecto de ML estándar, los datos suelen ser tabulares y estructurados (filas y columnas). El EDA se centra en encontrar valores nulos, outliers numéricos o correlaciones entre variables. Para Spin Compass, la materia prima es distinta:

- **Datos No Estructurados:** Trabajamos con texto libre proveniente de documentos PDF principalmente y otros formatos de texto.
- **Jerarquía vs. Tabular:** En lugar de filas, analizamos jerarquías de información. Un documento de política de viajes no es una «fila»; es una estructura lógica con títulos, subtítulos y tablas. Un EDA para RAG debe verificar si esta estructura es legible para

un LLM. Si un título se procesa como texto plano, el modelo pierde la capacidad de categorizar la información correctamente.

- **Formato de Archivo:** Mientras que en ML se prefieren CSV o Parquet, para RAG debemos priorizar formatos convertibles a Markdown, ya que este preserva la estructura semántica (encabezados, listas) que ayuda al agente a entender el contexto.

2.2. Unidad de Análisis: El «Chunk» vs. La «Observación»

En ML, la unidad atómica es la observación (ej. un cliente). En RAG, la unidad es el Fragmento (Chunk). El desafío principal no es la ingeniería de características (feature engineering), sino la estrategia de segmentación.

- **Riesgo de Ruptura Semántica:** En un EDA tradicional, no «cortamos» una fila a la mitad. En RAG, si utilizamos métodos tradicionales de corte por caracteres (fixed-size chunking), corremos el riesgo de romper la coherencia semántica, separando una regla de viáticos de su excepción.
- **Coherencia:** El análisis debe enfocarse en determinar si los fragmentos resultantes mantienen una «idea completa». Estudios recientes sobre Max-Min Semantic Chunking demuestran que agrupar oraciones basándose en similitud semántica (usando embeddings) supera significativamente a los cortes estáticos, asegurando que el agente recupere contextos completos y no frases cortadas.

2.3. Tratamiento de Ruido y Limpieza

La definición de «ruido» cambia drásticamente entre ambos paradigmas:

- **En ML:** Ruido suele referirse a varianza inexplicada o errores de medición.
- **En RAG (Spin Compass):** Encabezados repetitivos, números de página, pies de página y descargos de responsabilidad legales (disclaimers) en los PDFs de Spin. Estos elementos «ensucian» la búsqueda vectorial y deben ser eliminados durante el preprocesamiento.

2.4. Métricas de Evaluación: Recuperación vs. Precisión Predictiva

En ML evaluamos con Accuracy, RMSE o AUC. Para la base de conocimientos de Spin Compass, estas métricas son insuficientes. Debemos adoptar métricas de Recuperación de Información (IR) y Calidad de Generación:

- **Evaluación de Recuperación (Retrieval Evaluation):** No basta con saber si el modelo responde bien; debemos saber si encontró el documento correcto.
  - **Hit Rate (HR@K):** Mide la frecuencia con la que el fragmento correcto aparece en los primeros K resultados recuperados.
  - **MRR (Mean Reciprocal Rank):** Evalúa qué tan arriba en la lista de resultados aparece la respuesta correcta. Es crucial para Spin Compass, ya que el usuario de Slack rara vez mira más allá de la primera respuesta.
  - **Métricas Semánticas (SRA y RCC):** Investigaciones recientes proponen métricas como Same Relevant Article (SRA) y Relevant Chunk Contained (RCC), que miden si el sistema recuperó fragmentos que pertenecen al mismo documento oficial o si el fragmento contiene la respuesta completa, respectivamente.

2.5. Resumen Comparativo

Característica	**EDA Tradicional (Machine Learning)**	**EDA para RAG (Spin Compass)**
Datos	Estructurados (Tablas, Numéricos).	No Estructurados (Texto, PDFs, Slack).
Objetivo	Correlación y Predicción.	Recuperabilidad y Coherencia Semántica.
Preprocesamiento	Imputación de nulos, Scaling, One-Hot.	Conversión a Markdown, *Chunking* Semántico, Limpieza de Headers/Footers.
Métrica Clave	Accuracy, F1-Score, MSE.	Hit Rate, MRR, Semantic Coherence (AMI).
Riesgo Principal	Overfitting / Underfitting.	Alucinaciones por contexto irrelevante o fragmentado.

Tabla 1: Comparación entre EDA tradicional vs EDA para RAG

En Spin Compass, no buscamos correlaciones estadísticas, sino garantizar que si un «Spinner» pregunta sobre un gasto, el sistema recupere el párrafo exacto de la política vigente y no un texto fragmentado o irrelevante.

3. Referencias

[1] L. Visengeriyeva, A. Kammer, I. Bär, A. Kniesz, y M. Plöd, «CRISP-ML(Q): The ML lifecycle process». Accedido: 22 de enero de 2026. [En línea]. Disponible en: https://ml-ops.org/content/crisp-ml

[2] R. Hernández-Sampieri y C. Mendoza, *Metodología de la investigación: Las rutas de la investigación cuantitativa, cualitativa y mixta*, 3.ª ed. McGraw-Hill, 2023.

[3] A. Moreno-Cediel, E. Garcia-Lopez, D. De-Fitero-Dominguez, y A. Garcia-Cabot, «Optimising retrieval performance in RAG systems: A new growing window semantic chunking strategy to address weak semantic boundaries». Alcalá de Henares, Madrid, Spain, 2025.

[4] C. Kiss, M. Nagy, y P. Szilagyi, «Max–Min semantic chunking of documents for RAG application». Budapest, Hungary, 2025.