



Enriquecimiento automatizado de cuentas en Salesforce mediante entity resolution con
fuentes oficiales colombianas

Juan Pablo Hernández Ortiz

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor

Diego José Luis Botia Valderrama

Doctor en ingeniería electrónica, sistemas de calidad QoS / QoE para TV digital

Profesor asociado del Departamento de Ingeniería de sistemas

Universidad de Antioquia

Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2026

Cita	(Hernández Ortiz, 2026)
Referencia	Hernández Ortiz, J.P. (2026). <i>Enriquecimiento automatizado de cuentas en Salesforce mediante entity resolution con fuentes oficiales colombianas</i> [Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.
Estilo APA 7 (2020)	



Grupo de Investigación Intelligent Information Systems Lab In2Lab

Especialización en Analítica y Ciencia de Datos, Cohorte IX.

Centro de Investigación Ambientales y de Ingeniería (CIA).



Centro de Documentación Ingeniería (CENDOI)

Repositorio Institucional: <http://bibliotecadigital.udea.edu.co>

Universidad de Antioquia - www.udea.edu.co

Rector: Héctor Iván García García.

Decano: Elkin Libardo Ríos Ortiz.

Jefe departamento: Danny Alejandro Múnera Ramírez

Coordinadora del Programa: Maria Bernarda Salazar Sánchez

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

Dedicatoria

Dedico este trabajo a mis padres Jenner Hernández y Elsa Lucía que siempre han estado ahí para apoyarme en todo momento. Los amo.

Agradecimientos

Agradezco a IAC (Ingeniería Asistida por Computador) por darme la oportunidad de crecer como ingeniero, a Guillermo Cañón y a Isabel Ramírez por darme la oportunidad de aprender mientras trabajo, y a mis compañeros Ángel Rey y Camilo Valencia por el compañerismo compartido a lo largo del curso de Analítica

Tabla de contenido

Resumen	7
1. Introducción	8
2. Materiales y Métodos	10
Descripción de los datos	10
Tamaño y estructura	12
Problemas de los datos y análisis exploratorio (EDA)	13
Preprocesamiento y limpieza de datos	13
Métodos de análisis y modelado	13
Evaluación del modelo (criterios y métricas)	13
3. Resultados y Discusión	14
Presentación de resultados	14
Resultado de scrapping de RUES	15
Actualización de industrias y sectores	16
Análisis e interpretación	20
Evaluación crítica (limitaciones y mejoras)	21
4. Conclusiones	21
Referencias	23

Lista de tablas

Tabla 1 <i>Carátula Plenas Individuales (Supersociedades)</i>	10
Tabla 2 <i>Carátula Pymes Individuales (Supersociedades)</i>	11
Tabla 3 <i>10.000 empresas más grandes del país (datos.gov.co)</i>	11
Tabla 4 <i>Fuentes de datos utilizadas y su rol en el sistema</i>	14
Tabla 5 Resultados de matching de cuentas	14
Tabla 6 Resultados del piloto de scraping en RUES (250 consultas).	15
Tabla 7 Desglose de resultados del piloto de scraping en RUES (250 consultas).	15
Tabla 8 Clasificación CIIU de industrias y sectores actualizada en Salesforce.	16
Tabla 9 Fuente de los datos actualizados (solo cuentas actualizadas).	20
Tabla 10 Industria: consistencia vs. sugerencia oficial.	20

Lista de figuras

Figura 1 Estructura del proyecto de actualización de cuentas de SALESFORCE	12
---	----

Resumen

Este trabajo desarrolla un sistema de enriquecimiento masivo de cuentas en Salesforce para mejorar la calidad de datos críticos (NIT, CIIU e industria/sector) mediante técnicas de Entity Resolution con fuentes oficiales colombianas. Se construyó un corpus unificado de 30,375 empresas únicas a partir de Supersociedades (pymes y plenas) y una base complementaria de las 10.000 empresas más grandes del país, aplicando normalización de NIT y razones sociales. El enriquecimiento sigue un flujo bifurcado: (i) lookup exacto por NIT cuando está disponible y (ii) matching por similitud de nombre usando TF-IDF con n-gramas de caracteres y similitud coseno. En una prueba real sobre 10.246 cuentas únicas (Cuentas totales IAC), el sistema actualizó 2,164 cuentas (21.1%) con alta confianza: 870 por NIT exacto y 1,294 por TF-IDF automático (umbral 0,85). Del total actualizado, la fuente 10k (10 mil empresas más grandes del país, aportado por datos.gov.co) aportó la mayor cobertura (1,591, 73.5%), seguida por pymes (530, 24.5%) y plenas (43, 2.0%). Adicionalmente, 957 cuentas (9.3%) recibieron industria sugerida donde Salesforce estaba vacío, y 750 (7.3%) mostraron discrepancia frente a la industria sugerida, quedando como candidatas a corrección. En TF-IDF automático (n=1294), el score promedio fue 0.9762 (mín=0.8500, máx=1.0000); 1012 casos tuvieron score=1.0.

Palabras clave: Salesforce, calidad de datos, entity resolution, TF-IDF, NIT, CIIU, Supersociedades.

1. Introducción

IAC – Ingeniería Asistida por Computador gestiona su cartera de clientes y prospectos a través de Salesforce CRM. En el contexto comercial, la calidad y completitud de los datos de cuentas (NIT, CIIU, industria/sector) es un habilitador directo para segmentación, priorización de leads, reporting gerencial y modelamiento analítico. Sin embargo, en la operación diaria estos campos suelen quedar vacíos o inconsistentes porque el registro de cuentas se crea principalmente con el nombre y datos de contacto, mientras que completar NIT/CIIU requiere consultar fuentes externas y aplicar criterios uniformes.

En Colombia, los nombres de empresas presentan alta variabilidad (sufijos legales, tildes, abreviaturas, nombres comerciales), lo que limita los cruces exactos y exige técnicas de similitud robustas, además de un mecanismo de control de calidad (score de confianza y revisión manual). Desde la ciencia de datos, este problema se enmarca candidata a solucionarse con integración de datos y “**Entity Resolution**”; este último es un método y/o herramienta de identificación de datos que permite emparejar y hacer match con información que no tiene necesidad de tener un identificador único. En este proyecto se utilizará para crear matches con cuentas de Salesforce de la empresa con bases de datos públicas, de las cuales se basará la actualización de la información. El mecanismo funciona con dos representaciones textuales que corresponden a la misma entidad, pero sometidos a comparación con la data de la empresa.

El **problema específico** reside cuentas con información crítica faltante o no confiable en Salesforce (o cualquier CRM). Como consecuencia, se dificulta responder preguntas básicas (¿cuántos clientes pertenecen a construcción?, ¿A qué sector debemos abarcar según la información de cuentas de Salesforce?), y se degradan analíticas futuras. Se añade la falta de NIT, que impide usar un identificador único para reduplicación e integración con fuentes oficiales.

En cuanto al **objetivo del estudio**: se desea construir un sistema automatizado que, a partir del nombre y/o NIT de una cuenta en Salesforce, identifique la empresa correspondiente en fuentes oficiales colombianas y enriquezca campos estructurados (NIT, CIIU e

industria/sector) con precisión suficiente para uso operativo, registrando el método y score de confianza para gobernanza de datos.

¿En qué medida un enfoque híbrido (lookup exacto por NIT + similitud por nombre con TF-IDF/embeddings) puede aumentar la completitud de campos críticos en Salesforce manteniendo una tasa controlada de falsos positivos mediante umbrales y revisión manual? Se hace énfasis en que, a pesar de que se creará un lookup de búsqueda de NIT que no requiere creación de modelos sino una simple automatización, se creará modelo de similitud por nombres (debe ser por sector o por otro tipo de datos, dependiendo de las bases de datos disponibles) para asegurar que la actualización quede acorde a lo que se desea en la empresa, maximizando el resultado.

La **metodología** se desarrolla a partir de una estructura de proyecto en GitHub, que facilita el análisis de la base de datos de Salesforce. Esta base de datos se conecta mediante APIs propias de la herramienta para la consulta y actualización de cuentas. Actualmente, las pruebas se realizan descargando y almacenando la base de datos de forma manual. De esta información, se normalizan NIT y razones sociales, y se aplica un flujo bifurcado:

(i) match exacto por NIT cuando existe y (ii) matching por similitud de nombre usando TF-IDF con n-gramas de caracteres y similitud coseno. Se definen umbrales de confianza para actualización automática vs. cola de revisión manual, y se evalúa el resultado con métricas de cobertura (porcentaje de cuentas enriquecidas) y análisis de discrepancias de industria.

Por último, tenemos la **estructura del documento**. En la sección de Materiales y Métodos se describen las fuentes de datos, el preprocesamiento, el motor de matching y las herramientas utilizadas. En **Resultados y Discusión** se presentan las métricas del piloto real (cobertura, scores, fuentes del enriquecimiento e industria) y se discuten limitaciones y mejoras. Finalmente, en Conclusiones se sintetizan hallazgos e implicaciones para la actualización en Salesforce.

2. Materiales y Métodos

Se adopta CRISP-DM como marco metodológico para asegurar trazabilidad: **comprensión del problema, comprensión de datos, preparación, modelado, evaluación y despliegue**. El entregable técnico es un proyecto en GitHub que genera archivos enriquecidos y, en fases posteriores, permitirá escritura vía API REST de Salesforce con control de calidad.

Descripción de los datos

La primera base de datos es “**Carátula plenas individuales**” de **Supersociedades**. La calidad de la data es muy alta, sin nulos en campos críticos.

Tabla 1
Carátula Plenas Individuales (Supersociedades)

Metadato	Detalle
Archivo	10000_Carátula_plenas_individuales.xlsx
Origen	Superintendencia de Sociedades — SIIS (Sistema de Información y Riesgo Empresarial)
Tipo de empresa	Sociedades que reportan bajo NIIF Plenas (grandes empresas obligadas a reportar a Supersociedades)
Marco contable	NIIF Plenas — Normas Internacionales de Información Financiera
Actualización	Datos de corte 2024 (entre 2024-03-31 y 2024-12-31)
Volumen	3,116 registros · 41 columnas · 1.1 MB
Cobertura gráfica	Nacional (Colombia)

La segunda base de datos es “**Carátula Pymes individuales**” de **Supersociedades**. La calidad de la data es alta en campos críticos. NIT, razón social y CIU al 100%. Estado: 95% ACTIVA, 4% en Acuerdo de Reorganización. El campo objeto social tiene 4 nulos. La diferencia clave entre la base de datos de las plenas es que hay 7 veces más registros. Cubre el segmento de empresas medianas y pequeñas que no están en los datos de empresas plenas. Es la base más voluminosa del corpus.

Tabla 2
Carátula Pymes Individuales (Supersociedades)

Metadato	Detalle
Archivo	10000_Carátula_pymes_individuales.xlsx
Origen	Superintendencia de Sociedades — SIIS (Sistema de Información y Riesgo Empresarial)
Tipo de empresa	Sociedades que reportan bajo NIIF para Pymes (empresas medianas y pequeñas obligadas a reportar)
Marco contable	NIIF para Pymes — marco simplificado
Actualización	Datos de corte 2024 (entre 2024-03-31 y 2024-12-31)
Volumen	22,729 registros · 41 columnas · 7.3 MB
Cobertura gráfica	Nacional (Colombia)

La tercera base de datos corresponde a una tomada de **datos.gov.co**, la cual representa a las **10.000 empresas más grandes del país** (téngase en cuenta que esta afirmación no es verificada, pero es el nombre de la base de datos respecto al autor). La calidad de los datos es la mejor de todas: 0 nulos en todos los campos. Sin embargo, el CIU es numérico sin letra de sección, lo que implica que la conversión a industria es aproximada. La diferencia clave entre las bases de Supersociedades es que no tiene razón social en formato CIU con descripción, solo el código. Y al tener 4 años de datos, una empresa aparece hasta 4 veces; el proyecto reduplica por NIT manteniendo el registro más reciente.

Tabla 3
10.000 empresas más grandes del país (datos.gov.co)

Metadato	Detalle
Archivo	10.000_Empresas_mas_Grandes_del_País.xlsx
Origen	Datos Abiertos Colombia (Superintendencia de Industria y Comercio)

Tipo de empresa	Las 10,000 empresas con mayores ingresos operacionales en Colombia
Actualización	Datos de 4 años de corte: 2021, 2022, 2023, 2024. El archivo tiene 40,000 filas (10,000 × 4 años)
Volumen	40,000 filas · 14 columnas · 3.2 MB
Cobertura gráfica	Nacional (Colombia)

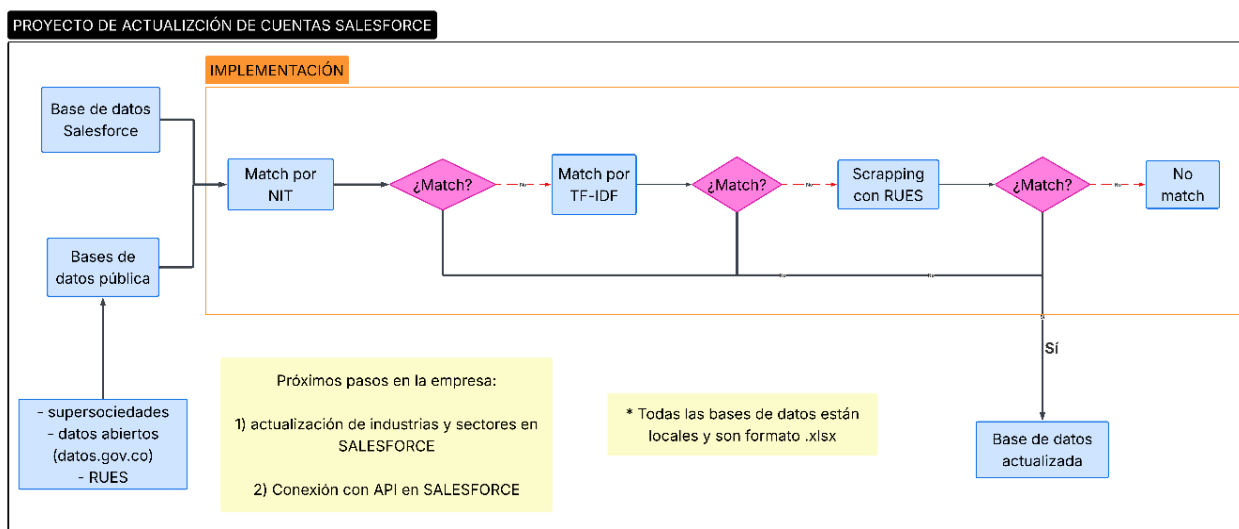
Tamaño y estructura

El corpus unificado contiene 30,375 empresas únicas. Cada registro incluye NIT normalizado, razón social, versión normalizada para matching, CIIU (cuando aplica) y metadatos (departamento/ciudad/estado). El dataset de Salesforce se analiza por cuentas únicas (reduplicadas por nombre) para métricas de cobertura.

En cuanto a estructura, las bases de datos, tanto las bases públicas del corpus como la base de datos de Salesforce de la empresa, se encuentran guardadas de manera local en un ordenador, en el cual se genera la implementación de la automatización (match exacto por NIT) y el modelo (TF-IDF), finalizando con scrapping en RUES como un último recurso para aquellas cuentas con las que no se pudo realizar un match.

Figura 1

Estructura del proyecto de actualización de cuentas de SALESFORCE



Problemas de los datos y análisis exploratorio (EDA)

Los principales problemas observados son: (i) ausencia de NIT en una fracción de cuentas, (ii) discrepancias entre nombre comercial y razón social oficial, y (iii) campos de industria inconsistentes. Como EDA del corpus se revisa la distribución por fuente y los CIU más frecuentes.

Preprocesamiento y limpieza de datos

Se aplica una normalización consistente a corpus y cuentas: **eliminación de tildes y puntuación, minúsculas y remoción de stopwords legales (S.A.S., S.A., LTDA, etc.)**. El NIT se limpia a dígitos y se estandariza a 9 caracteres (padding). Para CIU se separa código y descripción cuando está presente. No se realiza partición train/test porque el objetivo operativo es matching; la evaluación se basa en cobertura y análisis de score.

Métodos de análisis y modelado

Ruta 1: lookup exacto por NIT (score=1.0) cuando el identificador está disponible. Ruta 2: matching por nombre con TF-IDF a nivel de caracteres (char_wb, n-gramas 2–4) y similitud coseno para robustez ante variaciones ortográficas. Se definen umbrales: ≥ 0.85 (actualización automática), 0.70–0.84 (revisión manual), < 0.70 (sin match; requiere ampliar corpus con RUES u otras fuentes).

Evaluación del modelo (criterios y métricas)

Métrica 1: Cobertura (cuentas actualizadas automáticamente).

Métrica 2: Tamaño de cola de revisión manual (0,70–0,84).

Métrica 3: Distribución de scores TF-IDF (justificación de umbrales).

Métrica 4: Consistencia de industria (coincide/difiere/SF vacío con sugerencia).

La evaluación preliminar se centra en métricas operativas: (i) cobertura de enriquecimiento (porcentaje de cuentas actualizadas automáticamente), (ii) tamaño de cola de revisión manual, (iii) distribución de scores TF-IDF para justificar umbrales, y (iv) consistencia de industria (coincide/difiere/Salesforce vacío con sugerencia). En fases siguientes se incluirá una muestra etiquetada para estimar precision/recall/F1.

Tabla 4*Fuentes de datos utilizadas y su rol en el sistema*

Fuente	Tipo	Rol en el sistema	Observaciones
Salesforce (Cuentas)	Interna	Cuentas para enriquecer	Export 'Cuentas totales IAC.xlsx' (10.246 cuentas únicas).
Supersociedades (pymes/plenas)	Oficial	Corpus base (NIT, razón social, CIIU)	Fuentes: pymes=15,620, plenas=779.
SIC/datos.gov.co (10.000 empresas más grandes)	Complementaria	Aumentar cobertura (NIT + macrosector)	Registros=13,976.

3. Resultados y Discusión

Se ejecutó la automatización sobre 10,246 cuentas únicas (Cuentas totales IAC). Los resultados se presentan de forma objetiva con métricas y tablas, y se interpretan en el contexto del problema de calidad de datos en CRM.

Presentación de resultados

En total, 2,164 cuentas (21.1%) quedaron listas para actualización automática en Salesforce, mientras que el 13,7% se envía a cola de revisión manual para reducir falsos positivos.

Tabla 5*Resultados de matching de cuentas*

Categoría	Cuentas	% (sobre el total)
NIT Exacto	870	8.5%
TF-IDF automático	1294	12.6%
Revisión manual	1399	13.7%
Sin match	6683	65.2%

La fuente 10k (10 mil empresas más grandes de Colombia, sacada de Datos.gov.co) explica la mayor parte de la cobertura automática porque aporta NITs presentes en Salesforce que no estaban en Supersociedades. Esto confirma que la estrategia de ampliar corpus mejora el recall operativo sin cambiar el motor de matching.

Resultado de scrapping de RUES

Como complemento al proceso de matching por NIT y TF-IDF, se implementó un piloto de web scraping sobre el Registro Único Empresarial y Social (RUES) para enriquecer las cuentas que no alcanzaron el umbral de similitud mínimo en el pipeline interno. De las cuentas procesadas, aproximadamente 6.799 no alcanzaron el umbral de similitud interno y fueron candidatas al fallback RUES. Como prueba de concepto se ejecutaron 250 consultas al portal RUES, limitadas por el tope configurado en el script. La Tabla 6 resume los resultados de este piloto.

Tabla 6

Resultados del piloto de scraping en RUES (250 consultas).

Métrica	Valor
Consultas RUES ejecutadas	250 (tope del script)
Hits RUES (scraping con datos útiles)	10
Tasa de éxito sobre las 250 consultas	$10 / 250 = 4,0\%$
Sin éxito en esas 250 consultas	240

La tasa de éxito del piloto fue del 4,0% sobre las 250 consultas ejecutadas. La gran mayoría de los fallos se concentraron en errores HTTP, lo que sugiere restricciones de velocidad (rate limiting) por parte del portal del RUES al recibir peticiones automatizadas en volumen. La Tabla 7 desglosa los tipos de fallo detectados.

Tabla 7

Desglose de resultados del piloto de scraping en RUES (250 consultas).

Tipo (agrupado)	Cantidad	Interpretación breve
error_http	219	Rate limit o bloqueo por volumen al automatizar peticiones

sin_resultados	21	Página cargó pero sin enlace "Ver información" disponible
Éxito	10	Se abrió detalle y se parsearon datos de la empresa

Los 10 aciertos del piloto RUES representan únicamente el 0,15% de la cola total de cuentas sin match (~6.799). Para futuras iteraciones se recomienda implementar espaciado entre peticiones, reintentos con backoff exponencial y rotación de agentes de usuario para mitigar el bloqueo por volumen.

Actualización de industrias y sectores

A partir del código CIIU (Clasificación Industrial Internacional Uniforme) extraído durante el proceso de matching y enriquecimiento de cuentas, fue posible actualizar los campos de industria y sector en el Salesforce de IAC. Esta clasificación estandariza la categorización de empresas según su actividad económica principal, permitiendo segmentación y análisis más precisos.

Los datos de industria y sector presentes en el corpus de empresas están codificados según la Clasificación Industrial Internacional Uniforme (CIIU), adoptada en Colombia. Esta clasificación permite categorizar empresas por su actividad económica principal de forma estandarizada, facilitando segmentación y análisis comparativos. La siguiente tabla presenta la estructura completa de sectores e industrias del CIIU utilizada en el proceso de enriquecimiento del CRM de IAC.

La Tabla 8 presenta la estructura de divisiones CIIU por sector utilizada como base para la actualización masiva de registros.

Tabla 8

Clasificación CIIU de industrias y sectores actualizada en Salesforce.

Industria	Sector
Agricultura, ganadería, caza, silvicultura y pesca	Agricultura, ganadería, caza y actividades de servicios conexas
	Silvicultura y extracción de madera

	Pesca y acuicultura
Explotación de minas y canteras	Extracción de carbón de piedra y lignito
	Extracción de petróleo crudo y gas natural
	Extracción de minerales metalíferos
	Extracción de otras minas y canteras
	Actividades de servicios de apoyo para la explotación de minas
Industrias manufactureras	Elaboración de productos alimenticios
	Elaboración de bebidas
	Elaboración de productos de tabaco
	Fabricación de productos textiles
	Confección de prendas de vestir
	Curtido y fabricación de calzado
	Transformación de madera
	Fabricación de papel y cartón
	Actividades de impresión
	Coquización y refinación de petróleo
	Fabricación de sustancias químicas
	Productos farmacéuticos
	Productos de caucho y plástico
	Productos minerales no metálicos
	Productos metalúrgicos básicos
	Productos elaborados de metal
	Productos informáticos y electrónicos
	Aparatos eléctricos
	Maquinaria y equipo
	Vehículos automotores
	Otros equipos de transporte
	Fabricación de muebles

	Otras industrias manufactureras
	Instalación y mantenimiento de maquinaria
Suministro de electricidad, gas, vapor y aire	Suministro de electricidad, gas y vapor
Distribución de agua; evacuación y tratamiento de aguas residuales, gestión de desechos y descontaminación	Captación y distribución de agua
	Evacuación y tratamiento de aguas residuales
	Recolección y tratamiento de desechos
	Actividades de saneamiento ambiental
Construcción	Construcción de edificios
	Obras de ingeniería civil
	Actividades especializadas para la construcción
Comercio al por mayor y al por menor; reparación de vehículos automotores y motocicletas	Comercio y reparación de vehículos
	Comercio al por mayor
	Comercio al por menor
Transporte y almacenamiento	Transporte terrestre
	Transporte acuático
	Transporte aéreo
	Almacenamiento y actividades complementarias
	Correo y mensajería
Alojamiento y servicios de comida	Alojamiento
	Servicios de comidas y bebidas
Información y comunicaciones	Actividades de edición
	Producción audiovisual
	Programación y transmisión
	Telecomunicaciones
	Desarrollo de sistemas informáticos
	Servicios de información

Actividades financieras y de seguros	Servicios financieros
	Seguros
	Actividades auxiliares financieras
Actividades inmobiliarias	Actividades inmobiliarias
Actividades profesionales, científicas y técnicas	Actividades jurídicas y contabilidad
	Administración empresarial y consultoría
	Arquitectura e ingeniería
	Investigación científica
	Publicidad y estudios de mercado
	Otras actividades profesionales
	Actividades veterinarias
Actividades de servicios administrativos y de apoyo	Alquiler y arrendamiento
	Actividades de empleo
	Agencias de viajes
	Seguridad e investigación privada
	Servicios a edificios y paisajismo
	Actividades administrativas de oficina
Administración pública y defensa; planes de seguridad social de afiliación obligatoria	Administración pública
Educación	Educación
Actividades de atención de la salud humana y de asistencia social	Salud humana
	Atención residencial medicalizada
	Asistencia social sin alojamiento
Actividades artísticas, de entretenimiento y recreación	Actividades creativas y artísticas
	Bibliotecas, museos y cultura
	Juegos de azar y apuestas
	Actividades deportivas y recreativas

Otras actividades de servicios	Asociaciones
	Reparación de computadores y efectos personales
	Otras actividades de servicios personales
Actividades de los hogares individuales como empleadores	Hogares como empleadores
	Actividades no diferenciadas de hogares
Actividades de organizaciones y entidades	Organizaciones extraterritoriales

Análisis e interpretación

De las bases de datos disponibles, se hizo una caracterización sobre cuáles son más efectivas. En la tabla 9 se muestra como desde las bases de datos públicas de “Datos Abiertos” de Colombia (datos.gov.co), parte de la SIC (Superintendencia de Industria y Comercio) se derivan 3 de los 4 casos exitosos de match en la implementación.

Tabla 9

Fuente de los datos actualizados (solo cuentas actualizadas).

Fuente	Cuentas	% (sobre actualizadas)
SIC/datos.gov.co	1,591	73.5%
Pymes (Supersociedades)	530	24.5%
Plenas (Supersociedades)	43	2.0%

Y de este porcentaje de casos exitosos (21,1%) el 4.5% coincide con el corpus. Como muestra la tabla 10, los demás casos son candidatos a actualización de la información, que probablemente esté errónea en Salesforce, o son candidatas a poner información completamente nueva referida de las bases de datos.

Tabla 10

Industria: consistencia vs. sugerencia oficial.

Resultado	Cuentas	% (sobre total)
Coincide con corpus	457	4.5%

Difiere (candidata a corrección)	750	7.3%
Salesforce vacío, industria sugerida	957	9.3%

Evaluación crítica (limitaciones y mejoras)

Limitaciones identificadas:

- 1) Cobertura incompleta del universo empresarial: 65,2% queda sin match porque no aparece en las fuentes actuales (entidades públicas, universidades, microempresas, liquidadas, etc.).
- 2) La evaluación aún es principalmente operativa; falta estimar precision/recall/F1 con muestra etiquetada para calibrar umbrales con respaldo estadístico.
- 3) El matching por nombre basado en TF-IDF es robusto ante variaciones ortográficas, pero puede fallar cuando la similitud es semántica (sin coincidencia léxica). Los embeddings podrían mejorar estos casos.

Entre las mejoras propuestas específicas de la implementación se identifica una conexión por API/MCP a través de Salesforce: La actualización solo se hace de manera local en esta implementación. La posibilidad de crear la actualización de cuentas y espacios de Salesforce de manera automática debe ser el siguiente paso. También se debe validar otros modelos diferentes a TF-IDF respecto a la efectividad en resultados. Sin embargo, en un enfoque más administrativo y menos técnico, el uso de mejores bases de datos puede ser clave para unos mejores resultados.

4. Conclusiones

El sistema de enriquecimiento demostró viabilidad y valor inmediato: actualizó automáticamente 2,164 cuentas (21.1%) sobre una base real de 10,246 cuentas únicas, incorporando trazabilidad por método (NIT exacto vs. TF-IDF) y score de confianza.

El enfoque híbrido es apropiado para un CRM: el NIT resuelve de forma determinística cuando está disponible, y TF-IDF por caracteres aporta robustez frente a variaciones ortográficas típicas en nombres de empresas. La cola de revisión manual actúa como control de calidad para minimizar falsos positivos.

El piloto de scraping sobre el RUES arrojó una tasa de éxito del 4,0% en las 250 consultas de prueba. Si bien la cobertura inicial fue limitada, el enfoque valida la viabilidad técnica de integrar RUES como fuente complementaria. Sin embargo, la prueba de RUES también confirma que los principales avances propuestos para una nueva implementación no se deben concentrar principalmente en una reforma del modelo más que en nuevas bases de datos más robustas y con data de mejor calidad.

El modelo también habilitó la actualización sistemática de los campos de industria y sector en Salesforce mediante la clasificación CIIU extraída del enriquecimiento. Esto permite a IAC ejecutar segmentaciones sectoriales precisas, priorizar cuentas según actividad económica y generar reportes alineados con sus propios estándares.

Como impacto de negocio, el sistema habilita segmentación y reporting por industria, y provee una lista priorizada de cuentas candidatas a corrección (industria difiere) y a completitud (Salesforce vacío con sugerencia).

Para trabajo futuro: ampliar cobertura de bases de datos (ver la posibilidad de usar bases de datos pagas), evaluar formalmente precisión/recall/F1 con muestra etiquetada, y automatizar escritura a Salesforce vía API con reglas de gobernanza.

Referencias

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS.

Salton, G., & McGill, M. J. (1983). Introduction to Modern Information Retrieval. McGraw-Hill.

Datos.gov.co. (s. f.). *10.000 empresas más grandes de Colombia* (dataset). Portal de Datos Abiertos de Colombia.

Superintendencia de Sociedades (Colombia). (s. f.). *Información empresarial – Descarga de datos (SIIS)* [conjuntos de datos: “Carátula pymes individuales” y “Carátula plenas individuales”].

Salesforce. (2024). Salesforce Object Reference: Account Object.