Sistema de generación y consulta de un ranking de responsabilidad ambiental empresarial

Benjamin Eduardo Perdomo Morales

Politécnico Grancolombiano

Nota del autor

El presente escrito es un proyecto de grado para el título de pregrado en Ingeniería de Sistemas

Contenido

Resumen

La evaluación de la responsabilidad ambiental empresarial es de gran importancia para las empresas y consumidores. Por medio de técnicas de minería de texto y análisis de sentimiento, se construyó un sistema de información en Spark que procesa noticias de empresas colombianas recopiladas en Common Crawl, cuenta cuántas son y evalúa si su sentimiento es positivo, negativo, neutral o mezclado. Los resultados son consolidados en Excel y visualizados en PowerBI. Se evaluaron empresas colombianas de alimentos (Nutresa, Alpina, Ardilla Lülle y Quala), encontrando que en su mayoría las menciones son neutrales o mezcladas.

Abstract

The assessment of corporate environmental responsibility is of great importance for businesses and consumers. Through text mining techniques and sentiment analysis, an information system was built in Spark, that processes news from Colombian companies compiled on Common Crawl, counts how many they are, and assesses whether their sentiment is positive, negative, neutral or mixed. The results are consolidated in Excel and displayed in PowerBI. Colombian food companies were evaluated (Nutresa, Alpina, Ardila Lülle and Quala), finding that most of their mentions are neutral or mixed.

Palabras clave: Análisis de sentimiento, Spark, Common Crawl, Medio Ambiente

Introducción

En noticias y portales web, las empresas reciben cubrimiento de actividades, tanto negativas como positivas, que realizan afectando el medio ambiente. Por ejemplo: una empresa de cosméticos se puede encontrar en una polémica por el uso de sus productos en animales, u otra empresa puede estar realizando actividades disminuir su huella de carbono. La evaluación y comparación entre empresas son cada vez más solicitados por el público general, quien está empezando a realizar sus compras o negocios según dicha evaluación.

A nivel mundial, el ranking más conocido es el Dow Jones Sustainability Indices / Corporate Sustainability Assesment o DJSI/CSA por sus siglas en inglés (RobecoSAM, s.f.). Este índice utiliza cuestionarios que responden las compañías, documentación, análisis de reputación y discusiones con las compañías, lo que hace que el puntaje sea altamente influenciado por la capacidad de las compañías de obtener la información.

En Colombia, Merco realiza la evaluación de ranking, basado en encuestas a empresas, directivos, expertos y población general (Merco, s.f.). Cada componente del ranking es explicado, pero se basa sobre todo en las valoraciones propias de los encuestados.

Al no encontrar un ranking ambiental que se centre en la información generada en Colombia, se ve la oportunidad de generar uno. Para ello, es necesario el procesamiento de grandes cantidades de información, y analizarlas en paralelo.

Planteamiento del problema

Se busca evaluar la reputación empresarial ambiental de varias empresas colombianas, analizando la información que está publicada sobre ellas en la World Wide Web, mediante técnicas de minería de textos y análisis de sentimiento.

Objetivos

Objetivo General

Desarrollar un sistema de información que permita analizar la información publicada en sitios de noticias Web sobre responsabilidad ambiental de empresas colombianas.

Objetivos Específicos

- Identificar información relevante sobre responsabilidad ambiental de empresas publicadas en la Web.
- Analizar noticias de empresas colombianas a través de técnicas de minería de texto para obtener un indicador sobre la opinión de las empresas en responsabilidad ambiental.
- Establecer un mecanismo de evaluación de responsabilidad ambiental para que pueda ser utilizado por consumidores.

Revisión de literatura

La responsabilidad ambiental empresarial es la parte de la responsabilidad social empresarial que cubre las implicaciones ambientales de las operaciones, productos y facilidades de una compañía, la eliminación de emisiones y basuras, la maximización de la eficiencia y productividad de sus recursos, y la minimización de prácticas que puedan afectar negativamente el disfrute de los recursos naturales de un país por generaciones futuras (Mazurkiewicz, 2004).

Para construir un ranking de responsabilidad ambiental empresarial, es altamente deseable recopilar, indexar y analizar información que esté disponible libremente en Internet. Para esto, se utiliza Minería de Texto. La minería de texto se define como el descubrimiento, por medio de técnicas de computación, de nueva información previamente desconocida, al extraer automáticamente información de distintos recursos escritos (Hearst, 2003). Estos escritos pueden ser libros, correos electrónicos, sitios web, entre otros. La minería de texto tiene diferentes aplicaciones, como el monitoreo de sitios web para propósitos de seguridad nacional, análisis de criptografía, biología molecular y biomédica, buscadores, evaluación de calidad de servicio y relaciones de clientes empresariales, modelos predictivos de mercadeo, entre otros.

Una aplicación específica de la minería de texto es el análisis de sentimiento (también conocido como minería de opinión), donde se busca identificar, extraer, cuantificar y estudiar información subjetiva y estados afectivos de textos, mediante el uso de procesamiento de lenguaje natural, análisis de texto, lingüística computacional y biométricas. El análisis de sentimiento le permite a las empresas y políticos el manejo de su reputación en línea, sistemas de recomendación basados en evaluaciones de productos, entre otros. Una de las fuentes principales de los análisis de sentimiento son las redes sociales, por ejemplo, existen estudios que comparaban la negatividad de las campañas de Trump y Clinton (Raviv, 2016).

Para recopilar e indexar información, en ocasiones anteriores era necesario la escogencia de un robot de búsqueda, como Indri, Nutch y Terrier (Mahecha Nieto, 2011). Este enfoque ha cambiado debido a la fundación de Common Crawl. Esta es una organización que realiza una copia de toda la World Wide Web, recopilada por un robot de búsqueda basado en Nutch, procesada por Apache Hadoop (Common Crawl, s.f.). y almacenada en Amazon Web Services (Amazon Web Services, s.f.). El tamaño de la información recolectada mensualmente es de aproximadamente 200 TB. Common Crawl provee un servidor de índices, en el que se pueden obtener qué URLs fueron indexadas para un dominio o subdominio. Mediante programas como cdx-index-client, es posible descargar la información disponible para un dominio (Xu, 2019). Common Crawl provee para la descarga múltiples formatos: archivos WARC, que contienen todo el texto HTML y sus metadatos, archivos WET que almacenan los metadatos únicamente, y archivos WET que tiene el texto extraído únicamente (Common Crawl, s.f.).

Para procesar grandes cantidades de información en paralelo, Apache Hadoop utiliza MapReduce. Este es un modelo de programación y una plataforma para procesar y generar grandes conjuntos de datos. (Dean & Ghemawat, 2004) . Un trabajo de MapReduce usualmente parte los datos en partes independientes, que son procesados por una función de mapeo, que se utiliza para procesar o realizar ordenamientos, extracciones o filtros, en forma totalmente paralela. A los resultados se aplica función de reducción, que produce una operación de resumen (como conteo). Normalmente tanto la entrada como la salida del trabajo son almacenados en el sistema de archivos. La plataforma se encarga de hacer todas las tareas de programación, monitoreo y ejecución (Apache Software Foundation, 2019).

La implementación más popular en el mundo de MapReduce es Apache Hadoop. Se puede instalar en máquinas y clústeres locales, tanto como máquinas normales como servidores.

Consiste en un módulo de almacenamiento (Hadoop Distributed File System), un módulo de programación de trabajos (Hadoop YARN), un módulo de procesamiento (Hadoop MapReduce), un almacén de objetos (Hadoop Ozone) y un módulo de aprendizaje de máquina (Hadoop Submarine). Apache Spark expande la funcionalidad de Hadoop para permitir procesamientos en streaming y reportes parecidos a SQL interactivos. Spark tiene APIs para trabajar en Scala, Java Python y R.

Diseño metodológico

Se realizó el trabajo enmarcado en la metodología CRISP-DM (IBM, n.d.). Esta metodología está compuesta de 6 etapas. Según la necesidad, se puede regresar de una etapa a otra, cuando no se tengan resultados satisfactorios. La secuencia de las fases no es estricta, y se puede personalizar según las necesidades del problema y del negocio. Las etapas son las siguientes:

- Entendimiento de negocio. En esta etapa se explora qué es lo que la organización espera obtener de la minería de datos. En esta etapa se pueden definir algunos de los siguientes elementos:
    - Objetivos de negocio
    - Evaluación de la situación (definición de recursos y datos)
    - Determinación de los objetivos de minería de datos
    - Producción de un plan de proyecto
- Entendimiento de los datos. Busca obtener una vista más específica de los datos disponibles para minería. En esta etapa, se pueden ejecutar algunas de las siguientes tareas:
    - Descarga de datos iniciales
    - Descripción de los datos
    - Exploración de los datos
    - Verificación de la calidad de los datos
- Preparación de los datos. Es una de las más importantes tareas del proceso, y la que a menudo consume la mayoría del tiempo y esfuerzo del proyecto. En esta etapa, se pueden ejecutar algunas de las siguientes tareas:

  - Selección y creación del cuerpo inicial
  - Limpieza de datos
  - Construcción de nuevos datos
  - Integración de datos
  - Formateo de datos

- Modelado. En esta etapa se procesan los datos para obtener resultados de minería. Usualmente se hacen diferentes iteraciones. Se pueden ejecutar algunas de las siguientes tareas:
  - Selección de técnicas de modelado
  - Generación de un diseño de pruebas
  - Creación del modelo de datos
  - Evaluación de modelos
- Evaluación. En esta etapa se determina que el proyecto cumpla con los objetivos de éxito definidos por el negocio. Se pueden ejecutar alguna de las siguientes tareas:
  - Evaluación de resultados obtenidos
  - Revisión del proceso
  - Determinación de los siguientes pasos
- Despliegue. Es el proceso en el que se utilizan la nueva información obtenida para hacer mejoras en la organización. Se pueden encontrar alguna de las siguientes tareas:
  - Planeación del despliegue
  - Planeación del monitoreo y mantenimiento
  - Entrega de resultados obtenidos en un reporte final
  - Ejecución de una evaluación final del proyecto

Resultados

Objetivo del negocio

Evaluar la reputación en línea de distintas empresas referentes a sus responsabilidades ambientales.

Evaluación de la situación

Los recursos humanos disponibles para este proyecto

- el desarrollador y analista de minería de datos,
- la asesora de trabajo de grado.

En cuanto a recursos informáticos, están disponibles

- Un PC de escritorio con procesador i7 de tercera generación, 8 GB de RAM, 500 GB de almacenamiento de estado sólido y Windows 10 Professional.
- Un portátil con procesador i5 de sexta generación, 4 GB de RAM, 128 GB de almacenamiento de estado sólido y Windows 10 Professional.
- Un portátil con procesador i7 de octava generación, 16 GB de RAM, 1 TB almacenamiento de disco duro tradicional y Windows 10 Professional.
- Cuentas de Azure y Amazon Web Services estudiantiles, para el uso de créditos gratuitos y pagos.

Entendimiento de los datos

Descarga de datos iniciales.

Se intentó instalar Apache Spark en Windows, en el ambiente Windows Subsystem for Linux versión 2. Después se ejecutaron los ejemplos de CommonCrawl con Spark y Python, en específico el de conteo de servidores (Common Crawl, 2019). En específico:

- Se instalaron los requerimientos

- Se descargaron los datos de ejemplo, ejecutando el comando get-data.sh.
- Se ejecutó localmente el job server_count.py, que realiza un conteo de los servidores web utilizados por las páginas referenciadas. Esto es almacenado en el sistema de archivos en el formato Apache Parquet.
- Mediante el comando Pyspark, se consultó el parquet para comprobar la ejecución correcta.

Descripción de los datos.

Common Crawl.

Los datos de Common Crawl están almacenados en Amazon S3. Para los conjuntos de datos a partir de abril de 2017, se almacenan comprimidos en tres formatos:

- WARC: almacenan todos los datos de la página obtenidos por el robot, tanto los metadatos, las respuestas HTTP y todo el contenido de la respuesta HTTP (incluido el contenido HTML)
- WAT: almacenan los metadatos de la página consultada, únicamente
- WET: Almacena el contenido del texto plano extraído únicamente

Adicionalmente proveen un índice para cada uno de los repositorios disponibles, que se puede acceder desde https://index.commoncrawl.org/

El formato que más se adecúa a lo que se necesitaba analizar en el proyecto es el WARC.

Medio ambiente.

Para detectar que en las páginas el tema sea sobre medio ambiente, se utilizaron las siguientes palabras clave: 'medio ambiente', 'tierra', 'planeta', 'calentamiento global', 'contaminación', 'basura', 'basuras', 'agua', 'rio', 'ríos', 'océano', 'océanos', 'reciclaje', 'aire', 'clima'.

Estas 18 palabras claves se obtuvieron mediante sinónimos de Medio Ambiente, y mediante una revisión de intereses sobre el tema en medios de comunicación.

Sector empresarial y empresas.

Se decidió hacer el análisis para el sector de alimentos, para el Grupo Nutresa S.A. (Nutresa), Quala S.A. (Quala), Organización Ardila Lülle (Ardila) y Alpina Productos Alimenticios (Alpina).

Para cada conglomerado, se escogieron palabras clave que tuvieran que ver con el nombre de la empresa y los productos. Este listado se revisó al ingresar al portal web y la página de Wikipedia de cada una de las empresas

Para Nutresa, se utilizaron las siguientes 52 palabras clave: 'Nutresa', 'Zenú', 'Ranchera', 'Rica', 'Pietrán', 'Setas Colombianas', 'Galletas Noel', 'Saltín Noel', 'Ducales', 'Galletas Festival', 'Tosh', 'Dux', 'Compañía Nacional de Chocolates', 'ChocoLyne', 'Jet', 'Jumbo', 'Chocolisto', 'Corona', 'Café Sello Rojo', 'Colcafé', 'Café La Bastilla', 'Matiz Café', 'Leños & Carbón', 'Beer Station', 'Leños Gourmet', 'El Corral', 'Crem Helado', 'Polet', 'Aloha', 'Bocatto', 'Pastas Doria', 'Pastas Monticello'

Para Quala, se utilizaron las siguientes 42 palabras clave: 'Quala', 'InstaCrem', 'BatiCrema', 'Batilado', 'Quipitos', 'Hogareña', 'La Sopera', 'Frutiño', 'Gelatina Frutiño', 'Doña Gallina', 'FamiliaYá', 'Bon Ice', 'Activade', 'Del Fogón', 'LightYá', 'Ricostilla', 'Gelagurt', 'Savital', 'Popetas', 'PulpiFruta', 'Fortident', 'Gustiarroz', 'Sasóned', 'Boka', 'SunTea', 'Vive 100', 'Aromatel', 'Don Gustico', 'BioExpert', 'Nutribela', 'Triangulito', 'Saviloe', 'Disfrutaloe', 'Ricompleto ya'

Para la Organización Ardila Lülle, se utilizaron las 26 siguientes: 'Organización Ardila Lülle', 'Postobón', 'Hipinto', 'Jugos Hit', 'Agua Cristal', 'Agua Oasis', 'Néctar Hit', 'Jugos Tutti Frutti', 'Bretaña', 'Gaseosa Lux', 'Squash', 'Té Hatsu', 'Sr. Toronjo', 'H2OH!', 'Natumalta'.

Y para Alpina se utilizaron las siguientes 20: 'Alpina', 'Alpin', 'Alpinito', 'Avena Alpina', 'Bon Yurt', 'Leche Alpina', 'Regeneris', 'Yox', 'Alpinette', 'Boggy', 'Finesse', 'Queso Sabana', 'Fruper', 'Frutto', 'Soka'

Fuentes de datos a obtener.

Se buscaron portales web de noticias que estuvieran basados en Colombia, en español, que tuvieran foco en noticias empresariales o de medio ambiente. Así, se escogieron los siguientes:

- Portafolio.co (Portafolio), periódico diario y portal de noticias de economía y negocios.
- Dinero.com (Dinero), revista semanal y portal de noticias de economía y negocios.
- LaRepublica.co (La República), periódico diario y portal económico, empresarial y financiero.
- sostenibilidad.semana.com (Sostenibilidad), revista y portal web de noticias de sobre desarrollo sostenible.

Preparación de los datos

Creación del cuerpo inicial.

Para descargar el contenido desde Common Crawl de los últimos 3 años, de cada uno de los portales, se utiliza un cliente del índice de Common Crawl, llamado cdx-index-client, que indica dónde están almacenados los archivos WARC a descargar. Está desarrollado en Python y disponible en GitHub (Xu, 2019).

Posteriormente a la descarga de los índices, se descargan los archivos WARC correspondientes. Esto se hace con otro programa en Python, llamado cdx-index-retrieval, y

también está disponible en GitHub (Xu, 2019). Los programas cdx-index-client y cdx-index-retrieval se ejecutaron para cada uno de los dominios referenciados.

De las 496818 páginas indexadas, se descargaron en total 214.499 artículos, con un peso de 17,66 GB. El detalle de la información descargada se puede ver en la Tabla 1.

Tabla 1

Información descargada de Common Crawl

Sitio	Número de páginas en el índice	Número de páginas descargadas	Diferencia	% diferencia	Tamaño de archivos descargados (GB)
Dinero	175615	53929	121686	69,29%	5,4
La República	164998	87016	77982	47,26%	6,8
Portafolio	148034	65395	82639	55,82%	4,9
Sostenibilidad	8171	8159	12	0,15%	0,56
Total	496818	214499	282319	56,83%	17,66

Se analizó cómo funcionaba el ejemplo de conteo de servidores de CommonCrawl, creándose un fork de ese desarrollo para su posterior modificación. El código se almacenó en un repositorio de GitHub, disponible en la URL https://github.com/benjaminperdomo/rrae

Se diseñó un archivo batch (get-data-index.sh) para que exportara a un archivo txt global, y a uno por sitio, el listado de los archivos descargados de CommonCrawl. Se modificó el ejemplo para que llamara este listado y se realizó el conteo de servidores. Con esto se comprobó el correcto funcionamiento de la infraestructura.

Construcción de nuevos datos.

Posteriormente se procedió a hacer un script en Python (rrae_count.py), para que realizara el conteo de las páginas descargadas, cuáles tenían relación con las palabras clave, y cuántas veces se detectaba una de estas. Las palabras clave se almacenaron directamente en el

código (palabrasclaves.py). Inicialmente se hizo comparación con el título (en minúscula) de la página y las empresas, y con el contenido de la página y las palabras claves de Ambiente. Esto dio una cantidad de falsos positivos, así que se decidió cambiar la comparación de ambiente con el título de la página únicamente. Los datos se almacenaron en parquets y exportados a Excel.

De las páginas descargadas, el 5,93% tienen que ver con temas de ambiente, según lo detallado en la Tabla 2.

Tabla 2

Información procesada y relacionada con palabras clave de ambiente

Sitio	Descargado	Ambiente	Ambiente %	No Ambiente	No procesado
Dinero	8159	2433	4.51%	50398	185
La República	87016	5519	6.34%	75064	8
Portafolio	53929	3706	5.67%	61264	115
Sostenibilidad	65395	1053	12.91%	7096	6
Total	214499	12711	5.93%	193822	314

Por empresa, 116 tienen relación con Alpina, 174 con Ardila, 7074 con Nutresa y 22 con Quala, según lo detallado en la Tabla 3.

Tabla 3

Información por empresa

Sitio	Alpina	Ardila	Nutresa	Quala
Dinero	24	42	1907	5
La República	56	96	2836	6
Portafolio	36	34	2113	11
Sostenibilidad	0	2	218	0
Total	116	174	7074	22

Relacionados con ambiente y por empresa, se detectaron 37 por Alpina, 13 por Ardila, 320 por Nutresa y 0 por Quala, según lo relacionado en la Tabla 4.

Tabla 4

Información relacionada con el medio ambiente por empresa

Sitio	Alpina	Ardila	Nutresa	Quala
Dinero	2	3	70	0
La República	3	10	156	0
Portafolio	1	0	94	0
Sostenibilidad	31	0	0	0
Total	37	13	320	0

Para la evaluación de sentimiento, se encontraron 4 posibles librerías a utilizar:

- Textblob, disponible para Python, gratuita, se ejecuta en el clúster de Spark.
- VaderSentiment, disponible para Python, gratuita, se ejecuta en el clúster de Spark.
- Amazon Comprehend, disponible para Python mediante el AWS SDK for Python (Boto), gratuita para los primeros 5 millones de caracteres. Se ejecuta como servicio.
- Azure Text Analytics, disponible para Python mediante un cliente REST, gratuita para las primeras 5000 transacciones por mes. Se ejecuta como servicio.

Al revisar con detenimiento, tanto TextBlob como VaderSentiment están diseñadas para el idioma inglés únicamente. Cuando se especifica un idioma diferente, ambas librerías van a un servicio web de traducción, y trabajan contra esta. Así, fueron descartadas.

Con esto en mente, se construyó un programa en Python (rrae_score.py), para que almacenara la evaluación de sentimiento. Se decidió inicialmente hacer el análisis de sentimiento

contra el título del artículo, tanto en Amazon como en AWS, y promediar el resultado. Se ejecutó contra un conjunto de datos reducido de cuatro artículos, y luego contra todo el conjunto de datos. Los resultados se exportaron a Excel y se consolidaron.

Esto dio que, en promedio, un análisis de sentimiento neutral.

Con este resultado, se actualizó el programa para que hiciera análisis de sentimiento del título y del contenido de la página, y para que contara la categoría a la cual el servicio web había asignado el sentimiento. Las cuatro categorías son (Microsoft, 2019):

- Positivo: Cuando al menos hay una frase positiva en el documento, y el resto de las frases son neutrales.
- Negativo: Cuando al menos hay una frase negativa en el documento, y el resto de las frases son neutrales.
- Mezclado: Cuando al menos hay una frase negativa, y al menos una frase positiva, en el documento.
- Neutral: Cuando todas las frases en el documento son neutrales.

Se ejecutó de nuevo contra un conjunto de datos reducido de cuatro artículos, y luego contra el conjunto de datos, exportándose al final a Excel para su consolidación.

Modelado

Creación del modelo de datos.

Los diferentes componentes y sus datos se pueden observar en la Ilustración 1.

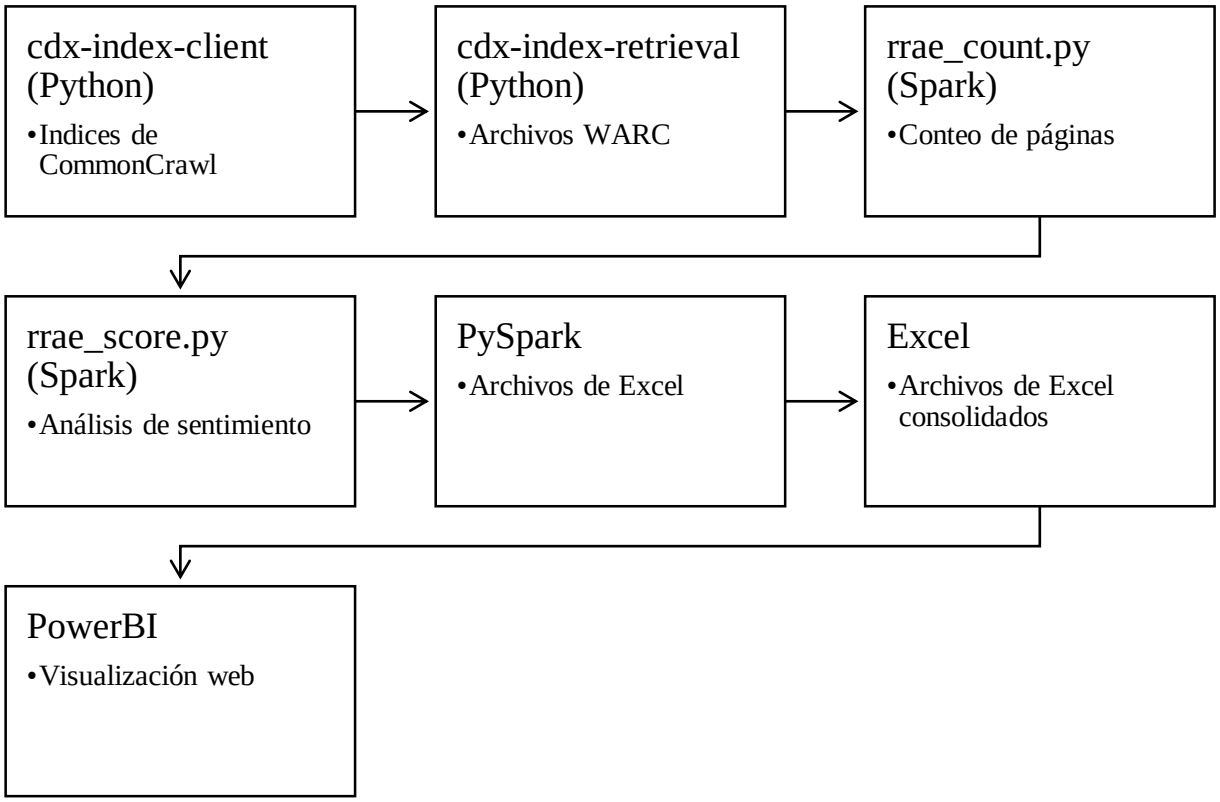

Ilustración 1. Diagrama del proceso

Evaluación

Evaluación de resultados obtenidos.

Para el caso de Alpina, al analizar por título, la mayoría dio un resultado positivo para Amazon, según lo visto en la Tabla 5; y para Amazon todas tuvieron un resultado neutral, según lo visto en la tabla 6.

Tabla 5

Análisis de sentimiento para títulos de Alpina, mediante Azure Cognitive Services

Sitio	Positivo	Negativo	Neutral	Mezclado
Dinero	1		1	
La República	2		1	
Portafolio		1		
Total	3	1	2	0
Porcentaje del total	16,67%	33,33%	50%	0%

Tabla 6

Análisis de sentimiento para títulos de Alpina, mediante Amazon Comprehend

Sitio	Positivo	Negativo	Neutral	Mezclado
Dinero			2	
La República			3	
Portafolio			1	
Total	0	0	6	0
Porcentaje del total	0%	0%	100%	0%

Al analizar todo el contenido de las páginas, para Azure todo el contenido fue Mezclado (Tabla 7), y para Amazon, todo el contenido fue Neutral (Tabla 8).

Tabla 7

Análisis de sentimiento para páginas de Alpina, mediante Azure Cognitive Services

Sitio	Positivo	Negativo	Neutral	Mezclado
Dinero				2
La República				3
Portafolio				1
Total	0	0	0	6
Porcentaje del total	0%	0%	0%	100%

Tabla 8

Análisis de sentimiento para páginas de Alpina, mediante Amazon Comprehend

Sitio	Positivo	Negativo	Neutral	Mezclado
Dinero			2	
La República			3	
Portafolio			1	
Total	0	0	6	0
Porcentaje del total	0%	0%	100%	0%

Para el caso de Ardila, al analizar por título, la mayoría dio un resultado positivo, tanto para Azure (Tabla 9) como para Amazon (Tabla 10).

Tabla 9

Análisis de sentimiento para títulos de la organización Ardila Lülle, mediante Azure Cognitive Services

Sitio	Positivo	Negativo	Neutral	Mezclado
Dinero		1	2	
La República	1		9	
Total	1	1	11	0
Porcentaje del total	7,69%	7,69%	84,62%	0%

Tabla 10

Análisis de sentimiento para títulos de la organización Ardila Lülle, mediante Amazon Comprehend

Sitio	Positivo	Negativo	Neutral	Mezclado
Dinero			3	
La República			10	
Total	0	0	13	0
Porcentaje del total	0%	0%	100%	0%

Al analizar por contenido, la mayoría del contenido fue Mezclado en Azure (Tabla 11) o Neutral en Amazon (Tabla 12).

Tabla 11

Análisis de sentimiento para páginas de la organización Ardila Lülle, mediante Azure Cognitive Services

Sitio	Positivo	Negativo	Neutral	Mezclado
Dinero				3
La República	1			9
Total	1	0	0	12
Porcentaje del total	7,69%	0%	0%	92,31%

Tabla 12

Análisis de sentimiento para páginas de la organización Ardila Lülle, mediante Amazon Comprehend

Sitio	Positivo	Negativo	Neutral	Mezclado
Dinero			3	
La República	1		9	
Total	1	0	12	0
Porcentaje del total	7,69%	0%	92,31%	0%

Para el caso de Nutresa, al analizar por título, la mayoría fue neutral, tanto para Azure (Tabla 13) como para Amazon (Tabla 14).

Tabla 13

Análisis de sentimiento para títulos de Nutresa, mediante Azure Cognitive Services

Sitio	Positivo	Negativo	Neutral	Mezclado
Dinero	10	13	47	
La República	30	26	115	
Portafolio	5	13	75	1
Sostenibilidad	5	14	12	
Total	50	66	249	1
Porcentaje del total	9,97%	18,80%	70,94%	0,28%

Tabla 14

Análisis de sentimiento para títulos de Nutresa, mediante Amazon Comprehend

Sitio	Positivo	Negativo	Neutral	Mezclado
Dinero	2	9	58	1
La República	5	18	129	4
Portafolio		9	84	1
Sostenibilidad		10	21	
Total	7	46	292	6
Porcentaje del total	1,99%	13,11%	83,19%	1,71%

Al analizar por contenido, la mayoría fue Mezclada en Azure (Tabla 15) y Neutral en Amazon (Tabla 16).

Tabla 15

Análisis de sentimiento para páginas de Nutresa, mediante Azure Cognitive Services

Sitio	Positivo	Negativo	Neutral	Mezclado
Dinero	1	5		64
La República	12	24	2	118
Portafolio			1	93
Sostenibilidad		1		30
Total	13	30	3	305
Porcentaje del total	3,70%	8,55%	0,85%	86,89%

Tabla 16

Análisis de sentimiento para páginas de Nutresa, mediante Amazon Comprehend

Sitio	Positivo	Negativo	Neutral	Mezclado
Dinero			70	
La República	30	15	111	
Portafolio		1	93	
Sostenibilidad			31	
Total	30	16	305	0
Porcentaje del total	8,55%	4,56%	86,89%	0%

Es de destacar que el análisis de título y de página en Amazon es predominante Neutral, con unas pocas variaciones. En cambio, en Azure si vemos que los valores cambian entre el análisis de título y de página.

Despliegue

Para la visualización de los datos por parte del público general, se dispuso de un PowerBI, que se puede acceder desde la URL http://bit.ly/rrae_poli, con los datos consolidados del ranking.

En la primera página del PowerBI, se muestran los resultados de la información descargada, y las menciones por sitio, como se puede ver en la ilustración 3.

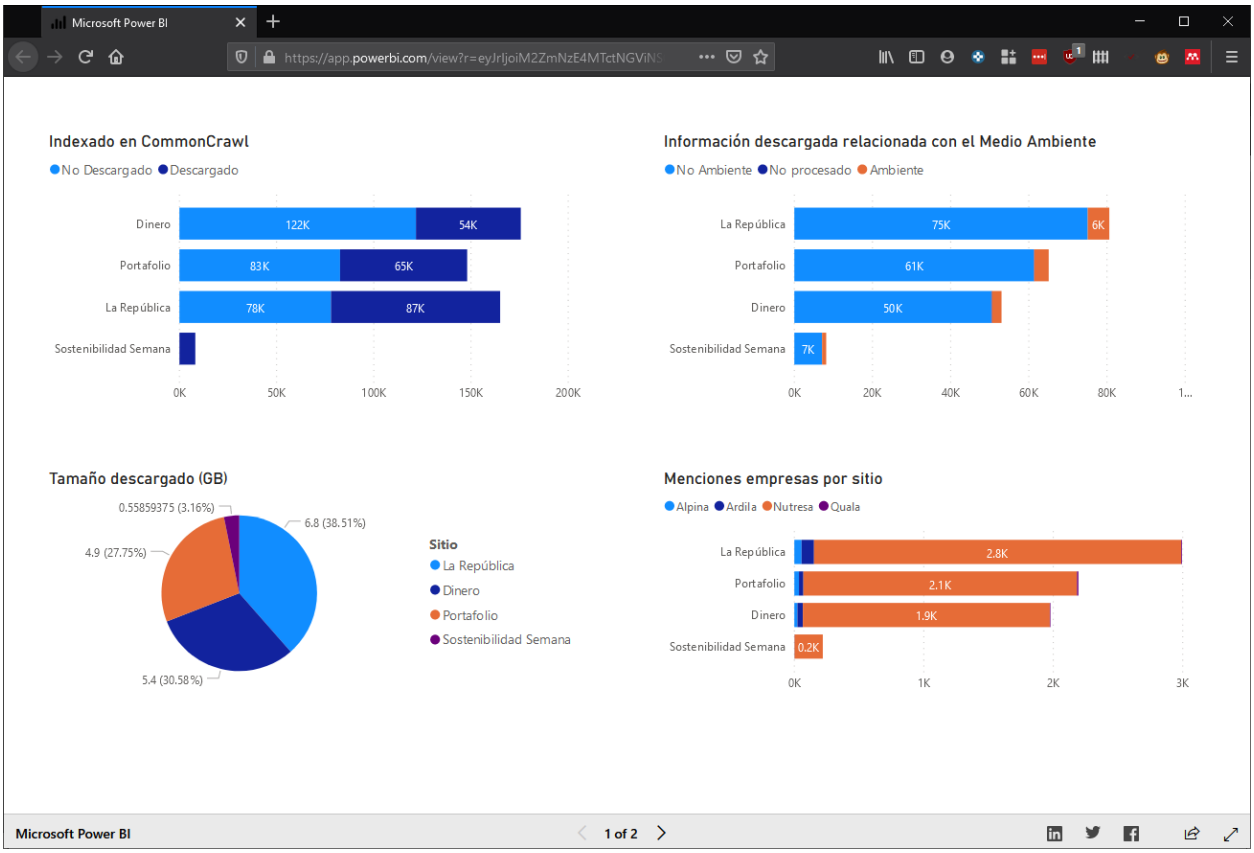

Ilustración 2. Captura de pantalla de la página 1 de Power BI

En la segunda página del PowerBI, se muestran los resultados del sentimiento en texto y página, por proveedor y empresa, como se puede ver en la ilustración 2.

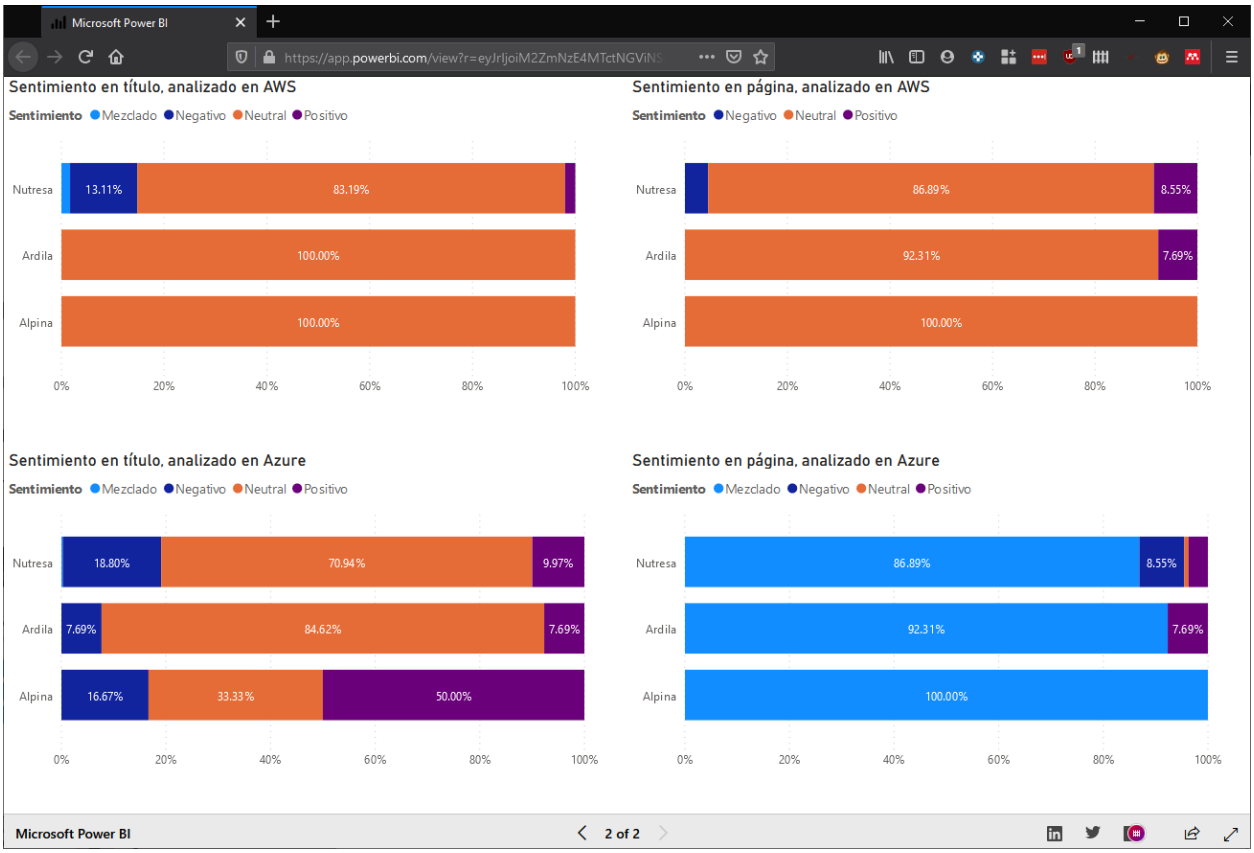

Ilustración 3. Captura de pantalla de la página 2 de Power BI

Conclusiones y recomendaciones

El software y plataformas disponibles para minería de datos, minería de texto y análisis de sentimiento han avanzado considerablemente, permitiendo que con software que corre en PCs, y con servicios en la nube, se puedan realizar análisis de grandes cantidades de datos. Esto llevó al cumplimiento de los objetivos de este trabajo.

En cuanto a las dificultades, se dieron principalmente en la configuración de las herramientas, ya que estás están usualmente diseñadas para Linux, y en la escogencia del título como único factor de comparación de sentimiento. Los portales web seleccionados tienden a escribir los títulos de forma neutra, así que la acumulación del resultado iba a ser neutral también.

A futuro, se puede mejorar la herramienta de las siguientes formas:

- Manejando la infraestructura de Spark en la nube, sea en Amazon (EMR) como en Azure (HDInsight), para que el procesamiento sea más rápido, o se puedan utilizar todo el conjunto de datos de Common Crawl, filtrado al idioma español.
- Manejar dinámicamente (mediante un archivo de texto) las palabras claves al sistema
- Cambiar la visualización para que consulte directamente la información desde el clúster, y no se exporte a Excel.

Referencias

Amazon Web Services. (n.d.). *Common Crawl*. Retrieved from Registry of Open Data on AWS: https://registry.opendata.aws/commoncrawl/

Apache Software Foundation. (2019, septiembre 10). *MapReduce Tutorial*. Retrieved from Apache Hadoop: https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html

Common Crawl. (2019, octubre 31). *Common Crawl PySpark Examples*. Retrieved from GitHub: https://github.com/commoncrawl/cc-pyspark

Common Crawl. (n.d.). *Frequently Asked Questions*. Retrieved from Common Crawl: https://commoncrawl.org/big-picture/frequently-asked-questions/

Common Crawl. (n.d.). *So you're ready to get started*. Retrieved from Common Crawl: https://commoncrawl.org/the-data/get-started/

Dean, J., & Ghemawat, S. (2004). MapReduce: Simplied Data Processing on Large Clusters. *OSDI'04: Sixth Symposium on Operating System Design and Implementation*, (págs. 137-150). San Francisco. Obtenido de https://research.google/pubs/pub62/

Hearst, M. (2003, Octubre 17). *What is Text Mining?* Retrieved from http://people.ischool.berkeley.edu/~hearst/text-mining.html

IBM. (n.d.). *CRISP-DM Help Overview*. Retrieved from IBM Knowledge Center: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.crispdm.help/crisp_overview.htm

Mahecha Nieto, I. A. (2011). Sistema de generación, administración y consulta de una librería digital de documentos para un portal web. Bogota.

Mazurkiewicz, P. (2004). *Corporate environment responsability: Is a common CSR framework possible?* Retrieved from The World Bank: http://documents.worldbank.org/curated/en/577051468339093024/Corporate-environmental-responsibility-Is-a-common-CSR-framework-possible

Merco. (n.d.). *Ranking Merco Empresas Colombia*. Retrieved from merco: http://merco.info/co/ranking-merco-empresas

Microsoft. (2019, diciembre 16). *How to: Detect sentiment using the Text Analytics API*. Retrieved from Microsoft Azure Docs: https://docs.microsoft.com/en-us/azure/cognitive-services/text-analytics/how-tos/text-analytics-how-to-sentiment-analysis?tabs=version-3

Raviv, G. (2016, octubre 16). *Clinton vs Trump – Whose campaign is more negative? Opponent-Mentions and Sentiment Analysis in Excel.* Retrieved from Datachant: https://datachant.com/2016/10/16/opponent-mention-effect-sentiment-analysis-in-excel/

RobecoSAM. (n.d.). *RobecoSAM*. Retrieved from The SAM Corporate Sustainability Assessment: https://www.robecosam.com/csa/csa-resources/about-csa.html

Xu, L. (2019, enero 19). *How to Retrieve Archived Pages of Specific Domain Using CommonCrawl Index*. Retrieved from Just Chillin': https://liyanxu.blog/2019/01/19/retrieve-archived-pages-using-commoncrawl-index/