METRIKA
Business Analytics

Panorama - Diagnóstico

Estudio de las capacidades tecnológicas y su potencial uso en los objetivos de negocio de Madison

Monterrey, Nuevo León.

15 de mayo de 2020

CONTENIDO

TECNOLOGÍAS DE EXTRACCIÓN DE DATOS DE INTERNET

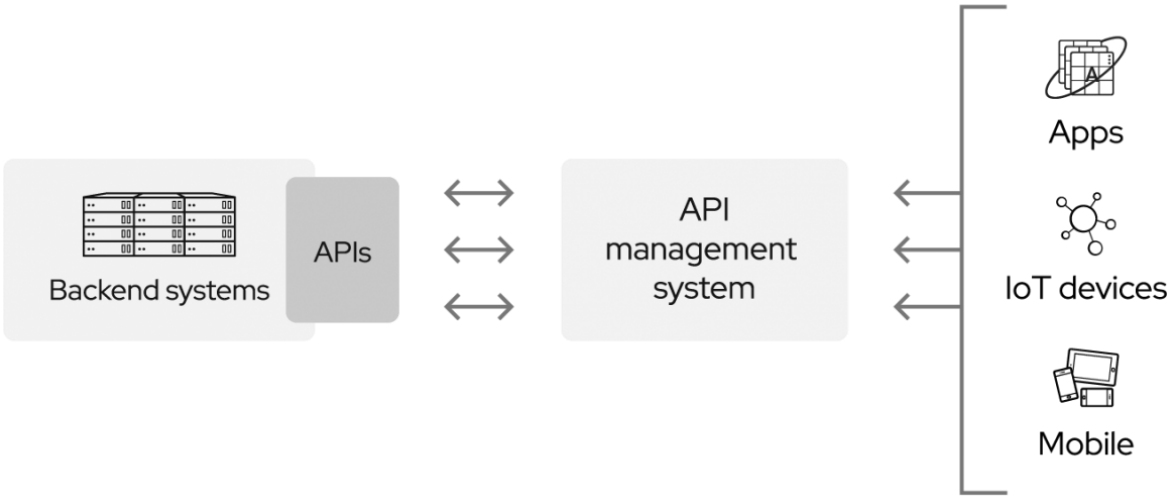
A. API

Una interfaz de programación de aplicaciones (API) es un conjunto de definiciones de subrutinas, protocolos y herramientas para construir software de aplicaciones. En términos generales, es un conjunto de métodos de comunicación claramente definidos entre varios componentes de software.

Las API permiten que sus productos y servicios se comuniquen con otros, sin necesidad de saber cómo están implementados. Esto simplifica el desarrollo de las aplicaciones y permite ahorrar tiempo y dinero. Las API le otorgan flexibilidad, simplifican el diseño, la administración y el uso de las aplicaciones, y proporcionan oportunidades de innovación, lo cual es ideal al momento de diseñar herramientas y productos nuevos (o de gestionar los actuales).

A veces, las API se consideran como contratos, con documentación que representa un acuerdo entre las partes: si una de las partes envía una solicitud remota con cierta estructura en particular, esa misma estructura determinará cómo responderá el software de la otra parte.

Para reducir la complejidad, es mejor tener una herramienta de extracción web con alguna integración de API que pueda extraer y transformar los datos al mismo tiempo sin escribir ningún código.



Para extraer datos con la integración de una API se requiere de dos cosas:

1. Extraiga los datos del sitio web sin la necesidad de esperar la respuesta de un servidor web.
2. Envíe los datos extraídos automáticamente de la nube a sus aplicaciones internas a través de la integración de API.

Una idea errónea es que las APIs pueden extraer datos. No es completamente cierto ya que solo es responsable de buscar los datos de acuerdo con los recursos dedicados. En la mayoría de los casos, obtendrá solo lo que solicita. Sin embargo, no tiene acceso a otra información.

B. "SCRAPING"

Es una técnica que sirve para extraer información de páginas web de forma automatizada, ya sea utilizando el protocolo HTTP manualmente o incrustando a un navegador en una aplicación.

Para realizar la extracción de los datos se requiere:

a) Conocer la estructura de la página donde se realizará la extracción; para ello se debe tener conocimiento en expresión regular **(regex)** para delimitar las búsquedas o hacerlas más precisas y que el filtrado de la información sea mejor,
b) Un lugar que pueda guardar los datos: por ejemplo, una **base de datos**
c) Un **analizador** (software o técnicas) que agregue extraiga la información contenida en la base de datos.

Los elementos que componen la extracción de los datos son,

Crawling: Se refiere al rastreo, esencialmente a seguir los enlaces de las páginas web, tanto internos como externos de los a analizar.
Scraping: Es el acto de la extracción de información de los sitios.
Parsing: Consiste básicamente en dividir en pedazos la información para estructurarla posteriormente.

<u>Herramientas de extracción nivel básico</u>

Orientada en el uso de plugins (complementos de diversos softwares) para la obtención de información. Se requiere un nivel de conocimiento nulo de programación. Regularmente son gratuitos.

ImportHTML
Google Sheets cuenta con una función llamada ImportHTML que importa datos de una tabla o lista dentro de una página HTML. También puede usar esta función para extraer datos automáticamente en una hoja de Google.

Para importar datos de una tabla o lista dentro de un sitio web a una hoja de Google se utiliza la función ImportHTML, los datos deben estar disponibles en la primera carga del sitio web. Los datos también deben estar disponibles públicamente, es decir que no requieren autorización y/o credenciales de inicio de sesión.

La función ImportHTML no actualiza automáticamente la extracción de la información, incluso si los datos en la página web de origen cambian. Se requiere realizar rutinas para actualizar sus datos importados automáticamente, incluso cuando la Hoja de Google está cerrada.

Scraper

Es una extensión de Chrome con funciones de extracción de datos limitadas. También exportar los datos a las hojas de cálculo de Google y CSV. Puede copiar fácilmente los datos al portapapeles o almacenarlos en las hojas de cálculo con estándares abiertos que permite flujos, simples de autorización para sitios web o aplicaciones informáticas (Oauth). Scraper permite construir expresiones que recorren y procesan un documento XML (Xpaths) automáticamente para definir URL.
Puede extraer datos de diversas páginas, lo que la hace más poderosa. Incluso puede extraer datos de páginas dinámicas que usan Javascript y Ajax.

Parsers

Es una extensión confiable para extraer datos de páginas web. Está disponible para Google Chrome y puede realizar una variedad de tareas de extracción de datos en pocos minutos. El Parsers puede extraer información de múltiples páginas al mismo tiempo y tiene capacidades de extracción de datos dinámicas incomparables. También puede manejar páginas con AJAX, cookies, redirecciones y Javascript.

Exporta datos en formato CSV directamente desde el navegador. Utiliza Web Scraper Cloud para exportar datos en formatos CSV, XLSX y JSON, acceder a ellos a través de API, webhooks o exportarlos a través de Dropbox.

- Analiza sitios por hora, día o semana
- API, administrar scrapers a través de una API
- Rotación de IP a través de miles de direcciones IP
- Agiliza el procesamiento posterior de datos

OutWit Hub

Es la extensión de Firefox. Realiza una variedad de tareas de extracción de datos. Outwit Hub mejor conocido por la interfaz fácil de usar y excelentes características de reconocimiento de datos. Adecuado para usuarios sin experiencia en programación y autónomos.

También puede navegar por las páginas y almacenar la información extraída en un formato adecuado. Los contenidos extraídos de una página web se presentan de forma fácil y visual. Los usuarios pueden extraer fácilmente enlaces, imágenes, direcciones de correo electrónico, noticias RSS, tablas de datos, etc. de series de páginas sin tener que ver el código fuente. Los datos extraídos se pueden exportar a bases de datos CSV, HTML, Excel o SQL, mientras que las imágenes y documentos se guardan directamente en su disco duro.

OutWit Hub tiene una interfaz única para extraer pequeñas o grandes cantidades de datos le permite eliminar cualquier página web del navegador. Incluso crear agentes automáticos para extraer datos.

Herramientas de extracción nivel intermedio

Orientada en la implementación de software comercial, para la extracción se requiere un nivel de conocimiento bajo de programación. Regularmente tienen costo.

Import.io

Import.io tiene un gran conjunto de herramientas de extracción web que cubren todos los diferentes niveles. Puede convertir un sitio web en una tabla sin ningún tipo de entrenamiento. Para sitios web más complejos, se debe descargar su aplicación de escritorio que tiene un mayor número de características que incluyen rastreo web, interacciones de sitios web e inicios de sesión seguros.

Una vez que creada la API, ofrece una serie de opciones de integración simples, como Hojas de cálculo de Google, Plot.ly, Excel, así como solicitudes GET y POST. Pago único por el software y soporte. Ofrece una opción de nivel empresarial pagado para empresas que buscan una extracción de datos más compleja o a gran escala.

Webscraper.io

Con la plataforma webscrape.io se crea un plan (mapa del sitio) sobre cómo se debe inspeccionar un sitio web y qué se debe extraer. Con estos mapas de sitio, navega por el sitio en consecuencia y extraerá todos los datos para exportarlos a CSV.

El objetivo de esta plataforma es hacer la extracción de datos web lo más simple posible. Poder configurar la extracción simplemente apuntando y haciendo clic en los elementos deseados de las páginas. No se requiere codificación.

Web Scraper puede extraer datos de sitios con múltiples niveles de navegación. Puede navegar por un sitio web en todos los niveles:

- Categorías y subcategorías
- Paginación
- Páginas de productos

- Ejecución completa de JavaScript
- Esperando solicitudes de Ajax
- Manejadores de paginación
- Desplazamiento de página hacia abajo

Exporta los datos en formato CSV desde el navegador o por medio Scraper Cloud se exportan los datos en formatos CSV, XLSX y JSON, API, webhooks y Dropbox.

Octoparse

Identifica los datos, los Extrae instantáneamente y guarda la información en disco duro a través de múltiples sitios recopila el contenido. Octoparse una opción para programadores y analistas de datos por la tecnología de aprendizaje automático y la exportación de datos a formatos HTML, Excel, CSV y TXT.

Tiene dos tipos de modo de operación: **Modo Asistente** y **Modo Avanzado**, el primero no requiere con un conocimiento en programación y se puede implementar rápidamente. Mediante clic la interfaz puede guiar todo el proceso de extracción. Como resultado, se extraer el contenido del sitio web y se guarda en formatos estructurados como EXCEL, TXT, HTML o sus bases de datos en un corto período de tiempo.

Además, proporciona una **Programación Cloud Extracción** que le permite extraer los datos dinámicos en tiempo real y mantener un registro de seguimiento de las actualizaciones del sitio web. También, puede extraer sitios web complejos con estructuras difíciles mediante el uso de su configuración incorporada de Regex y XPath para localizar elementos con precisión. Cuenta con servidores Proxy IP que automatiza las IP con esto se permite eliminar los bloqueos de IP, con esto no se detecta como sitios web agresivos.

ParseHub

Es un excelente web scraper que admite la recopilación de datos de sitios web que utilizan tecnología AJAX, JavaScript, cookies, etc. Su tecnología de aprendizaje automático puede leer, analizar y luego transformar documentos web en datos relevantes. La aplicación de escritorio de Parsehub es compatible con sistemas como Windows, Mac OS X y Linux. Incluso puede usar la aplicación web que está incorporado en el navegador.

Como programa gratuito, no puede configurar más de cinco proyectos públicos en Parsehub. Los planes de suscripción pagados le permiten crear al menos 20 proyectos privados para scraper sitios web. Se pue acceda a los datos a través de JSON, Excel y API. Los datos son recopilados en la nube.

Herramientas de extracción nivel experto

Orientado a desarrollo a la medida para la extracción de datos, requiere un nivel de conocimiento alto en programación en leguaje como Python, PHP, HTML5, Ajax, javascript,

xml, JSON. Se requiere una mayor inversión de tiempo y personal (con las capacidades) para su implementación.

Scrapy

Herramienta de código abierto, Scrapy es una de las bibliotecas de Python más populares durante años, y es probablemente la mejor herramienta de extracción web de Python para nuevas aplicaciones. Bien documentado y hay muchos tutoriales sobre cómo comenzar. Además, la implementación de los rastreadores es muy simple y confiable, los procesos pueden ejecutarse una vez que se configuran. Como marco de extracción web con todas las funciones, hay muchos módulos de middleware disponibles para integrar varias herramientas y manejar varios casos de uso (manejo de cookies, agentes de usuario, etc.).

BeautifulSoup

El analizador HTML para los desarrolladores de Python, ha existido durante más de una década. Bien documentado, con tutoriales de análisis web que enseñan a los programadores a usarlo para extraer de varios sitios web en Python 2 y Python 3. También proporciona algunos métodos simples y expresiones idiomáticas para navegar, buscar y modificar un árbol de análisis: un juego de herramientas para diseccionar documentos y extraer lo que necesita.

Convierte automáticamente los documentos entrantes a Unicode y los documentos salientes a UTF-8. Por lo cual no se tiene que pensar en codificaciones, a menos que el documento no especifique alguna otra. Asimismo, analiza todo y hace el recorrido del árbol de búsqueda. Puede decirle "Buscar todos los enlaces", o "Buscar todos los enlaces de la clase externalLink ", o "Buscar todos los enlaces cuyas URL coincidan con" foo.com ", o" Buscar el encabezado de la tabla que tiene texto en negrita, luego dar yo ese texto ".

Selenium

Es una opción para automatizar las pruebas realizadas en los navegadores web. Es una herramienta de código abierto, lo que significa que es libre de usar, redistribuir e incluso modificar, por lo que el software está disponible para que cualquiera lo utilice. Para ser específicos: Selenium es un conjunto de herramientas para automatizar los navegadores web en muchas plataformas. Selenium web driver es el software que permite realizar pruebas cruzadas de navegadores y también permite utilizar un lenguaje de programación como python.
Cuenta con la siguientes aplicaciones:

- Para crear suites y pruebas de automatización de regresión robustas basadas en el navegador, escalar y distribuir scripts en muchos entornos, entonces usar Selenium WebDriver, una colección de enlaces específicos del idioma para manejar un navegador.

- Para crear scripts de reproducción rápida de errores, scripts para ayudar en las pruebas exploratorias asistidas por automatización, entonces usar Selenium IDE; un complemento de Chrome y Firefox que hará una simple grabación y reproducción de interacciones con el navegador.
- Para escalar distribuyendo y ejecutando pruebas en varias máquinas y gestionando múltiples entornos desde un punto central, lo que facilita la ejecución de las pruebas en una amplia combinación de navegadores / SO, entonces desea usar Selenium Grid.

Puppeteer

Herramienta de código abierto y respaldado y activamente desarrollado por el propio equipo de Google Chrome. Está reemplazando rápidamente a Selenium como la herramienta de automatización para la extracción web. Cuenta con una API que instala automáticamente un binario Chromium compatible como parte de su proceso de configuración, lo que significa que no tiene que realizar un seguimiento de las versiones del navegador. Es más simple que una biblioteca de rastreo web, a menudo se usa para extraer datos de sitios web de sitios que requieren JavaScript para mostrar información, maneja scripts, hojas de estilo y fuentes como un navegador real. Requiere mucho CPU y memoria y no es necesario un navegador completo.

La web se está volviendo más complicada con html5 y páginas complejas de javascript. Cada vez es más difícil para cualquier herramienta común de extracción web, por ejemplo, recopilar datos como Facebook. Se requieren miles de horas para aprender cómo extraer los datos que desea o copiar y pegar requerirá que miles de personas lo hagan. Por lo cual para realizar dicha tarea se deben tomar en cuenta,

Generalidades

¿Cuándo debo hacer para hacer scraping?

1. Si tienes que extraer datos de una sola página que contenga muchas tablas y, por lo tanto, mucha información.
2. Información dispersa en múltiples bases de datos o sitios.
3. Periodicidad de información, cada cuando se libera cada diaria, semanal o mensual.
4. Analizar la estructura de la paginas a o sitios a realizar la extracción.
5. Cuenta con una API de conexión el sitio.
6. Recibir alertas de cambio en las bases de datos que se usan.

Herramientas de scraping web Nivel básico

Tema	ImportHTML	Scraper Chrome	Parsers	OutWit Hub
Precio X mes	Gratis	Gratis	Gratis hasta 200 Uds	Gratis hasta 200 Usd
Paginación	Una	Múltiples	Múltiples	Múltiples
Soporta seguridad o autorización de los sitios	No	No	No	No
Número de dominios que un proyecto puede funcionar	NA	NA	Si	NA
Complejidad de la estructura del sitio web.	HTML	HTML, Javascript y Ajax	HTML, Javascript y Ajax	HTML, Javascript y Ajax
Trabajo de recopilación de datos recurrentes	No	No	Si	Si
¿Cuántos datos puedo recopilar?	Limitado	Ilimitado	Ilimitado	Ilimitado
Analizador sintáctico	No	No	No	No
Conocimiento de programación	No	No	No	No
Tener el programa puede ejecutarse en sus computadoras	No	No	No	No
Ejecutar en nuestras computadoras	Si	Si	Si	Si
URL	support.google.com	chrome.webstore/web-scraper	parsers.me/	outwit.com

METRIKA
Business Analytics

Herramientas de scraping web Nivel Intermedio

Tema	Import.io	Webscraper.io	Octoparse	ParseHub
Precio X mes	Gratis hasta (cotizar)	Gratis hasta $300 uds	Gratis hasta $500 uds	Gratis hasta $209 uds
Paginación	Multiples	Multiples	Multiples	Multiples
Soporta seguridad o autorización de los sitios	Si	Si	Si	Si
Número de dominios que un proyecto puede funcionar	Si	Si	Si	Si
Complejidad de la estructura del sitio web.	HTML, Javascript, cookies, redirecciones y Ajax	HTML, Javascript,cookies, redirecciones y Ajax	HTML, Javascript, cookies, redirecciones y Ajax	HTML, Javascript, cookies, redirecciones y Ajax
Trabajo de recopilación de datos recurrentes	Si	Si	Si	Si
¿Cuántos datos puedo recopilar?	Ilimitado	Ilimitado	Ilimitado	Limitado
Analizador sintáctico	Si	Si	Si	Si
Conocimiento de programación	No	Si	No	No
Tener el programa puede ejecutarse en sus computadoras	Si	No	No	No
Ejecutar en nuestras computadoras	Si	Si	Si	Si
URL	https://www.import.io/	https://webscraper.io/	https://www.octoparse.com/	https://www.octoparse.com

METRIKA
Business Analytics

Herramientas de scraping web Nivel Avanzado

Tema	scrapy	BeautifulSoup	Selenium	Puppeteer
Precio X mes	Gratis	Gratis	Gratis	Gratis
Paginación	Múltiples	Múltiples	Múltiples	Múltiples
Soporta seguridad o autorización de los sitios	Si	Si	Si	Si
Número de dominios que un proyecto puede funcionar	Si	Si	Si	Si
Complejidad de la estructura del sitio web.	HTML, Javascript, cookies y Ajax	HTML, Javascript, cookies, PHP, Python y Ajax	HTML, Javascript, cookies, PHP, Python y Ajax	HTML, Javascript, cookies y Ajax
Trabajo de recopilación de datos recurrentes	Si	Si	Si	Si
¿Cuántos datos puedo recopilar?	Ilimitado	Ilimitado	Ilimitado	Ilimitado
Analizador sintáctico	Si	Si	Si	Si
Conocimiento de programación	Si (Python)	Si (Python2 , 3)	Si (Rubí, Java, Python, C#, JavaScript)	Si (Python)
Tener el programa puede ejecutarse en sus computadoras	Si	Si	Si	Si
Ejecutar en nuestras computadoras	Si	Si	Si	Si
Url	https://scrapy.org/	https://www.crummy.com/	https://www.selenium.dev/	https://github.com/puppeteer

Obstáculos mientras se somete a web scraping:

Tecnologías anti-scraping:
Tales como Captcha después de iniciar sesión sirven como vigilancia para detener los correos no deseados. Sin embargo, también representan un gran desafío para la implemente web scraper básico. Las tecnologías anti-scraping aplican algoritmos de codificación complejos, se necesita mucho esfuerzo para encontrar una solución técnica para solucionarlo. Algunos incluso pueden necesitar un middleware como 2 CAPTCHA para resolver.

Velocidad de carga lenta:
Cuantas más páginas web se necesiten analizar (scraper), más tardará en completarse. Es obvio que el scraping a gran escala requerirá muchos recursos en una máquina local. Una carga de trabajo más pesada en la máquina local puede provocar fallos.

Almacenamiento de datos:
Una extracción a gran escala genera un gran volumen de datos. Esto requiere una infraestructura sólida en el almacenamiento de datos para poder almacenar los datos de forma segura. Se necesitará muchos recursos y tiempo para mantener dicha base de datos.

ALMACENAMIENTO

El almacenamiento de datos es el proceso mediante el cual la tecnología de la información archiva, organiza y comparte los bits y bytes que conforman los sistemas de los que dependemos todos los días, desde las aplicaciones hasta los protocolos de red, los documentos, el contenido multimedia, las libretas de direcciones y las preferencias del usuario.

Tipos de almacenamiento de datos:

A. **Almacenamiento definido por software**

El almacenamiento definido por software (SDS) usa sistemas de software de gestión por extracción para separar los datos del hardware antes de cambiar su formato y organizarlos para su uso en la red. En particular, el SDS es útil para las cargas de trabajo de contenedores y microservicios que utilizan datos sin estructurar, ya que puede expandirse a un nivel que las soluciones de almacenamiento conectadas simplemente no pueden alcanzar.

B. **Almacenamiento en la nube**

Es la organización de los datos almacenados en cierto lugar al que puede acceder cualquier persona que tenga los permisos adecuados, a través de Internet. No es necesario que esté conectado a una red interna (conocida como almacenamiento adjunto a la red o NAS) ni que acceda a los datos desde un sistema de hardware conectado directamente a la

computadora. Algunos de los proveedores de almacenamiento en la nube más conocidos son Microsoft, Google e IBM.

Almacenamiento adjunto a la red

El almacenamiento adjunto a la red (NAS) facilita el acceso a los datos por parte de las redes internas instalando un sistema operativo liviano en un servidor que lo convierte en una caja, unidad o cabezal de NAS. La caja de NAS se convierte en una parte importante de las intranets porque procesa todas las solicitudes de almacenamiento.

C. Almacenamiento de objetos

El almacenamiento de objetos divide los datos en unidades independientes y las combina con los metadatos para brindar contexto sobre su contenido. Los datos almacenados en estos objetos no están comprimidos ni cifrados, lo cual permite que las cargas de trabajo que cambian rápidamente, como los contenedores, accedan a ellos a gran escala.

D. Almacenamiento de archivos

El almacenamiento de archivos organiza los datos como archivos jerárquicos que los usuarios pueden abrir y explorar en su totalidad. Dado que los archivos se almacenan de la misma forma en backends y frontends, los usuarios pueden solicitarlos con identificadores únicos, como el nombre, la ubicación o la URL. Es el formato de almacenamiento legible por el ojo humano más usado.

E. Almacenamiento en bloques

En el almacenamiento en bloques, se dividen los volúmenes de almacenamiento en instancias individuales conocidas como bloques. Cada bloque es independiente, por lo que los usuarios tienen la autonomía total sobre la configuración. Dado que los bloques no tienen los mismos requisitos de identificador único que los archivos, constituyen un sistema de almacenamiento más rápido. Esto los convierte en el formato ideal para las bases de datos de contenido multimedia.

BASES DE DATOS

Se llama base de datos, o también banco de datos, a un conjunto de información perteneciente a un mismo contexto, ordenada de modo sistemático para su posterior recuperación, análisis y/o transmisión. Existen actualmente muchas formas de bases de datos, que van desde una biblioteca hasta los vastos conjuntos de datos de usuarios de una empresa de telecomunicaciones.

El manejo de las bases de datos se lleva mediante sistemas de gestión (llamados DBMS por sus siglas en inglés: Database Management Systems o Sistemas de Gestión de Bases de Datos), actualmente digitales y automatizados, que permiten el almacenamiento ordenado

y la rápida recuperación de la información. En esta tecnología se halla el principio mismo de la informática.

A. Sistemas Gestores de bases de datos Relacionales

Este modelo se basa fundamentalmente en establecer relaciones o vínculos entre los datos, imaginando una tabla aparte por cada relación existente con sus propios registros y atributos**.**

MySQL
Es un SGBD multihilo y multiusuario utilizado en la gran parte de las páginas web actuales. Es el más usado en aplicaciones creadas como software libre.
Se ofrece bajo la GNU GPL aunque también es posible adquirir una licencia para empresas que quieran incorporarlo en productos privativos (Desde la compra por parte de Oracle se está orientando a este ámbito empresarial).
Las principales ventajas de este Sistema Gestor de Bases de datos son:

- Facilidad de uso y gran rendimiento
- Facilidad para instalar y configurar
- Soporte multiplataforma
- Soporte SSL

La principal desventaja es la escalabilidad, es decir, no trabaja de manera eficiente con bases de datos muy grandes que superan un determinado tamaño.

MariaDB
Nace a partir de la adquisición de MySQL por parte de Oracle para seguir la filosofía Open Source y tiene la ventaja de que es totalmente compatible con MySQL.

Entre las principales características de este Sistema Gestor de Bases de datos se encuentran:

- Aumento de motores de almacenamiento
- Gran escalabilidad
- Seguridad y rapidez en transacciones
- Extensiones y nuevas características relacionadas con su aplicación para Bases de datos NoSQL.

No tiene desventajas muy aparentes salvo algunas pequeñas incompatibilidades en la migración de MariaDB y MySQL o pequeños atrasos en la liberación de versiones estables.

PostgreSQL
Este sistema gestor de base de datos relacional está orientado a objetos y es libre, publicado bajo la licencia BSD.

Sus principales características son:

- Control de Concurrencias multiversión (MVCC)
- Flexibilidad en cuanto a lenguajes de programación
- Multiplataforma
- Dispone de una herramienta muy fácil e intuitiva para la administración de las bases de datos.
- Robustez, Eficiencia y Estabilidad.

La principal desventaja es la lentitud para la administración de bases de datos pequeñas ya que está optimizado para gestionar grandes volúmenes de datos.

Microsoft SQL Server

Es un sistema gestor de bases de datos relacionales basado en el lenguaje Transact-SQL, capaz de poner a disposición de muchos usuarios grandes cantidades de datos de manera simultánea.

Es un sistema propietario de Microsoft. Sus principales características son:

- Soporte exclusivo por parte de Microsoft.
- Escalabilidad, estabilidad y seguridad.
- Posibilidad de cancelar consultas.
- Potente entorno gráfico de administración que permite utilizar comandos DDL y DML.
- Aunque es nativo para Windows puede utilizarse desde hace ya un tiempo en otras plataformas como Linux o Docker.

Su principal desventaja es el precio. Cuenta con un plan gratuito (Express) pero lo normal es la elección de alguno de los planes de pago disponibles (Standard, Developer, Enterprise o SQL Azure, la versión de SQL Server en la nube).

Oracle

Tradicionamente, Oracle ha sido el SGBD por excelencia para el mundo empresarial, considerado siempre como el más completo y robusto, destacando por:

Soporte de transacciones.

Estabilidad.

Escalabilidad.

Multiplataforma.

La principal desventaja, al igual que SQL Server, es el coste del software ya que, aunque cuenta con una versión gratuita (Express Edition o XE), sus principales opciones son de pago.

Las opciones de pago disponibles son:

1. Standard Edition (SE)
2. Standard Edition One (SE1)
3. Standard Edition 2 (SE2)
4. Personal Edition (PE)

5. Lite Edition (LE)
6. Entreprise Edition (EE)

B. Sistemas Gestores de bases de datos No Relacionales

Una base de datos no relacional (NoSQL) es aquella base de datos que:

- No requiere de estructuras de datos fijas como tablas
- No garantiza completamente las características ACID
- Escala muy bien horizontalmente.

Se utilizan en entornos distribuidos que han de estar siempre disponibles y operativos y que gestionan un importante volumen de datos.

MongoDB

Estamos ante el Sistema Gestor de Bases de Datos no relacionales (SGBD NoSQL) más popular y utilizado actualmente.

MongoDB es un SBGD NoSQL orientado a ficheros que almacena la información en estructuras BSON con un esquema dinámico que permite su facilidad de integración.

Empresas como Google, Facebook, eBay, Cisco o Adobe utilizan MongoDB como Sistema Gestor de Bases de datos.

Las principales características de MongoDB son:

- Indexación y replicación
- Balanceo de carga
- Almacenamiento en ficheros
- Consultas ad hoc
- Escalabilidad horizontal
- Open Source

Como desventaja principal, MongoDB no es un SGBD adecuado para realizar transacciones complejas.

Redis

Redis está basado en el almacenamiento clave-valor. Podríamos verlo como un vector enorme que almacena todo tipo de datos, desde cadenas, hashses, listas, etc.

El principal uso de este SGBD es para el almacenamiento en memoria caché y la administración de sesiones.

Las características principales son:

- Atomicidad y persistencia
- Gran velocidad
- Simplicidad
- Multiplataforma

Cassandra

Al igual que Redis, Cassandra también utiliza almacenamiento clave-valor. Es un SGBD NoSQL distribuido y masivamente escalable. Facebook, Twitter, Instagram, Spotify o Netflix utilizan Cassandra.

Dispone de un lenguaje propio para las consultas denominado CQL (Cassandra Query Languaje).
Las principales características de este SGBD NoSQL son:

- Multiplataforma
- Propio lenguaje de consultas (CQL)
- Escalado lineal y horizontal
- Es un SGBD distribuido
- Utiliza una arquitectura peer-to-peer

Es importante entender que, para elegir el SGBD más adecuado, se debe comenzar por el **estudio del tipo de datos** que se van a almacenar y **cómo se van a administrar**.

Computo en la nube

El computo en la nube, conocida también como servicios en la nube, nube de conceptos, es un paradigma que permite ofrecer servicios de computación a través de una red, que usualmente es Internet.

Amazon Web Services

Amazon Web Services es considerado un líder del mercado de almacenamiento en la nube. Es el "punto de referencia de la industria" en cuanto precios. Simple Storage Service (S3) es el objeto básico de almacenamiento, mientras que Elastic Block Storage es para los volúmenes de almacenamiento.
Cuenta con herramientas para vincular datos que se encuentran en las instalaciones de la empresa con la nube, llamada AWS Storage Gateway, la capacidad de crear arquitecturas híbridas de almacenamiento que se extiendan al almacenamiento en las instalaciones y la nube de AWS todavía están en progreso.

AT&T

El servicio de almacenamiento en la nube AT&T Synaptic está alineado estrechamente con el servicio de almacenamiento Atmos de EMC, que se utiliza como sistema de almacenamiento en las instalaciones. Esto crea una oportunidad para que AT&T pueda venderle a la sólida base de clientes de EMC, y ofrecerles a sus clientes capacidades de nube híbrida con un proveedor líder de almacenamiento. AT&T Synaptic ya se extiende por varias

regiones, cuyos clientes pueden optar por aprovechar los planes de AT&T para ampliar el servicio a nivel mundial, siendo Europa la próxima parada. Los clientes que utilizan el servicio de AT&T VPN son liberados de los costos de entrada y salida al usar el servicio en la nube de la empresa.

Google Cloud Storage

Google Cloud Storage es el servicio de almacenamiento subyacente para otros productos y servicios de nube de la empresa, incluyendo a Google App Engine la plataforma de desarrollo de aplicaciones, Google Compute Engine, y BigQuery, que son máquinas virtuales basadas en la nube y una herramienta de análisis de big data, respectivamente. Los clientes acceden a Google Cloud Storage a través de una API.

Google Cloud Storage como ideal para clientes que deseen crear y gestionar el despliegue, y para desarrolladores específicos en busca de gran capacidad de almacenamiento para las aplicaciones de Google.

HP

HP anunció la versión beta pública de su plataforma de almacenamiento en la nube que debutó en mayo del 2012 y que está destinada a trabajar en conjunto con su red de cómputo y de entrega de contenido (CDN), que se asoció recientemente con Akamai. La plataforma de almacenamiento se basa en la tecnología OpenStack, y HP ofrece soporte de chat gratuito 24/7 con una garantía de disponibilidad del 99,95%. "Entre los proveedores de almacenamiento en la nube basado en OpenStack, HP está bien posicionada para entender las necesidades de almacenamiento empresarial de TI, debido a su extenso hardware, software y opciones de servicio. El sistema replica los datos automáticamente a través de tres zonas de disponibilidad de capacidad de recuperación (que los clientes pueden optar por hacer en la nube de Amazon), y HP dice que al hacer que la información se ejecute en su hardware, nube pública y en los locales de los clientes hace que la configuración de la nube híbrida sea más fácil.

IBM

El almacenamiento en la nube de IBM es parte de su oferta empresarial SmartCloud, que incluye otros servicios, como el desarrollo de aplicaciones basadas en la nube y la infraestructura. IBM comercializa su nube para copias de seguridad y recuperación, pero esos servicios no utilizan IBM Object Storage SmartCloud en su servidor. Parte de esto podría deberse a que IBM está asociado con Nirvanix, otro proveedor de almacenamiento en la nube, para ejecutar el almacenamiento de objetos SamrtCloud. La heterogeneidad de estos servicios bajo el paraguas de IBM SmartCloud podría crear "silos de capacidades" para varios servicios. Sin embargo, IBM se ha comprometido a integrar más estrechamente sus

productos y servicios. Su experiencia en la venta a los principales departamentos de TI empresariales le da una ventaja significativa para convertirse en un jugador importante en el mercado de almacenamiento empresarial en la nube.

Microsoft

Detrás de Amazon Web Services, Windows Azure Blob Storage de Microsoft puede ser el segundo servicio de almacenamiento en la nube más utilizado. Soporta una amplia gama de características que incluyen almacenamiento de objetos, almacenamiento de tablas, SQL Server y una red de entrega de contenido (CDN). El almacenamiento de Azure Blob está en una carrera por el menor precio, con Amazon y Google que bajaban sus precios consistentemente durante el año pasado para ofrecer los precios más competitivos entre los tres competidores. Sus opciones de soporte apelan a los clientes de las grandes empresas, proporcionándoles un práctico equipo de apoyo basado en cuotas. Microsoft ha ampliado reciente mente su almacenamiento con la compra del vendedor de almacenamiento en la nube, StorSimple.

Nirvanix

Un proveedor de almacenamiento en la nube, Nirvanix se dedica exclusivamente a este mercado. Ideal para empresas que buscan necesidades de almacenamiento intensivo de datos, pero podría ser un inconveniente para los clientes que buscan un proveedor de todo que ofrece servicios de cálculo sobre una plataforma de almacenamiento. Sin embargo, Nirvanix tiene algunas características atractivas, incluyendo la capacidad de tener servicios de almacenamiento público, mixto o en las instalaciones de la empresa, y un ciclo de facturación todo incluido con opciones de soporte de alta calidad, un objetivo claro para los clientes empresariales, pero que puede alejar a empresas pequeñas y medianas que prefieren la fijación de precios a la carta.

Rackspace

Rackspace es otro jugador importante en el ecosistema de almacenamiento en la nube, con su servicioCloud Files aumentado por un robusto conjunto de servicios de acompañamiento, incluida la infraestructura de cómputo y una red CDN impulsada por Akamai. Para las necesidades de almacenamiento de alto rendimiento, tiene Cloud Storage Block, que tiene altas capacidades de ingreso-salida. Rackspace trabaja fuertemente en el proyecto de código abierto OpenStack y sus servicios siguen de cerca la evolución del proyecto. Debido a su trabajo en el entorno de OpenStack, los servicios públicos de almacenamiento en la nube de Rackspace se integran muy bien con las nubes potenciadas por OpenStack, pudiendo crear servicios de nube híbridos para los clientes.

MÉTODOS ANALÍTICOS DE TEXTOS

Desde la aparición de Internet, los periódicos digitales se han convertido en un referente importante entre los medios de comunicación. Los periódicos digitales tienen dos características que los separan del resto de los medios; por un lado, está la inmediatez, ya que, a diferencia de medios tradicionales como la televisión o la radio, los periódicos digitales no requieren de un espacio al aire para lanzar una nueva noticia o actualizar contenido. La segunda característica sustancial de los periódicos digitales es la interactividad, lo que nos permite a los lectores volvernos partícipes activamente no solo dando opiniones en los foros y blogs, sino generando información como estado del tráfico, clima, o dando a conocer eventos locales.

Es esta interactividad la que provoca que hoy en día exista un debate entre la opinión pública y la opinión publicada. La opinión pública es aquella que describe las tendencias de pensamiento de la comunidad o de la mayoría de ella, mientras que la opinión publicada es simplemente la realización de una expresión de una idea o pensamiento a través de la difusión en los diferentes medios de comunicación social. La información publicada puede ser o no ser cierta, y mucha de las cosas ciertas no se publican, al considerarse que no son noticia. El debate que gira en torno a la opinión pública y la opinión publicada radica en que no es lo mismo lo que opinan y creen los ciudadanos que lo expresado por los periodistas en los medios; y más aún, en que si por el hecho de que algo esté publicado se convierta, tarde o temprano, en opinión pública.

Aunque lograr que un mensaje aparezca en diferentes medios no crea opinión pública, de alguna manera sí puede contribuir a su conformación. Mucho se ha señalado que si lo que publica un medio lo repiten otros, entonces la opinión publicada adquiere mayor fuerza y trata de presentarse como opinión pública por las múltiples voces que repiten la misma información, como si los medios fabricaran una opinión, actuando como jueces de lo que pasa en la sociedad al pronunciarse sobre algo, bien sea la simple opinión del periodista, lo que dicen los actores públicos, o las distorsiones o prejuicios que se generan en torno a los hechos y opiniones.

A. RSS

Dado que el interés es analizar la opinión publicada a partir de las noticias de los periódicos digitales, es muy natural pensar en aprovechar los canales RSS que ofrecen los sitios web. Un canal RSS (“Really Simple Syndication”) contiene datos en formato XML para distribuir contenido en la web, siendo usualmente información actualizada de alguna subscripción a alguna fuente de contenido, en nuestro caso, noticias. La ventaja de utilizar los canales RSS es que se puede acceder al contenido sin la necesidad de un navegador.

Como se mencionó anteriormente, se accede a las noticias en formato XML, por lo que hace falta realizar un procesamiento para obtener el texto de la noticia y los metadatos de interés. A continuación, se muestra una metodología ejemplo para la extracción y almacenamiento de las noticias:

1. Se recolectaron los canales RSS de diferentes periódicos.
2. Utilizando Python se extrae la noticia del canal.
3. Se asigna un id (nombre del artículo) con la finalidad de no duplicar la extracción de las noticias.
4. Se extraen los metadatos (url, fecha, título).
5. Como la noticia fue extraída en HTML, se utiliza BeautifulSoup para extraer el texto de los tags de interés. Se ejecutan los pasos anteriores por cada noticia que se encuentre en el canal RSS.
6. Se guardan los textos resultantes en un archivo JSON (incremental) o en algún otro formato según la base de datos.

Respecto a la automatización del proceso, se puede crear un bash que ejecuta la extracción de noticias. Este script se programa para correr una sola vez diariamente, semanalmente o a una frecuencia deseada.

Limpieza de los datos

La limpieza de los datos es una parte fundamental en el proceso de análisis de texto pues los resultados obtenidos por los métodos de análisis dependen en gran medida de la calidad de los textos obtenidos mediante Scraping. Ahora bien, es importante mencionar que a pesar de que por lo general los textos de las noticias están bien redactados, la limpieza a la que nos referimos en este contexto tiene que ver con el preproceso de los textos generados a partir de los archivos JSON que obtuvimos de los canales RSS. A grandes rasgos, la estructura de estos textos es la siguiente:

- Leyenda: URL del texto
- Día, fecha y hora de la noticia.
- Texto de la noticia.

La complejidad resulta del hecho que cada una de las fuentes de información considera un formato distinto para la publicación de las noticias, asimismo del hecho de que todas las fuentes tienen, en mayor o menor medida, contenido que resulta irrelevante para la noticia. Entre este tipo de contenido podemos encontrar leyendas de: copyright, slogans de los periódicos, texto de hipervínculos a noticias que no necesariamente están relacionadas, pies de foto y videos, urls hashtags y nombres de usuarios de twitter, nombre del autor de la noticia, código html y javascript de widgets, textos de botones de navegación. Incluso algunos periódicos incluyen acuerdos de privacidad y sección de comentarios de los lectores como parte del texto de la noticia. Adicionalmente algunas fuentes de información pueden tener un etiquetado del lugar donde proviene una noticia.

Al hacer Scraping es importante considerar la estructura de los JSON resultantes de cada fuente, puesto que cada uno puede tener una estructura específica que puede meter ruido como la información de pie de la noticias con la información relacionada a la fuente como “copyright” etc.

B. PLN y Normalización

El análisis de textos, a través del Procesamiento del Lenguaje Natural (**PLN**), se ha vuelto un elemento de mucha utilidad en distintos campos y para estudiar diferentes fenómenos, las aplicaciones abarcan una gran variedad de ejemplos como el análisis de expedientes clínicos, currículos, detección de personas, entidades o lugares en textos y la minería de opinión.

El PLN es un campo de las ciencias de la computación, inteligencia artificial y lingüística que estudia las interacciones entre las computadoras y el lenguaje humano. Algunas aplicaciones que tiene el PLN es el reconocimiento de voz, la traducción entre idiomas, la comprensión de oraciones completas, la corrección de ortografía, entre otros.

Antes de comenzar a trabajar los textos con modelos de lenguaje natural, es conveniente transformar el texto a una forma estándar. La normalización o preprocesamiento del texto consiste en aplicar una serie de transformaciones en los datos originales para dejar todos los textos en el mismo formato. La normalización de los textos dependerá de la tarea que se esté abordando, entre las transformaciones más usuales están el convertir todos los textos a minúsculas o mayúsculas, eliminar los signos de puntuación o convertir los números a sus equivalentes de palabras.

Por ejemplo, en todos los idiomas es posible encontrar, en mayor o menor medida, palabras que se derivan de otras. En gramática, se conoce como “flexión” a la alteración que experimentan las palabras para expresar diferentes categorías como tiempo, número o género. A través de las flexiones es posible expresar una o más categorías gramaticales (con un prefijo, sufijo o infijo, o un cambio de vocal).

La “derivación” (Stemming) y la “lematización” (Lemmatization) son técnicas de normalización de texto que se utilizan para reducir las formas de flexión de cada palabra a una base o raíz común, con el fin de homologar términos.

El proceso de Stemming es con el que principalmente eliminamos los sufijos del final de la palabra para obtener su raíz. Esta es la manera más básica para hacer análisis morfológico de las palabras. Por su parte, la “lematización” consiste en, dada una forma flexionada (plural, femenino, conjugada, etc), encontrar el “lema” correspondiente. El “lema” es la forma que por convenio se acepta como representante de todas las formas flexionadas de una misma palabra.

Palabra	Stemming	Lema
cantábamos	cant	cantar

canción	cancion	cantar
cantantes	cantant	cantar

El uso de la “lematización” o “stemming” puede ser más o menos útil, de acuerdo al idioma en el que se trabaje. El inglés es una lengua que presenta pocas flexiones (lengua analítica), mientras que en el español se cuenta con gran cantidad de morfemas por palabra (lengua sintética).

La conversión a minúsculas es otro tipo de normalización. En algunas tareas de PLN, como algunos casos de recuperación de información, el convertir a minúsculas no afecta tanto el análisis cómo puede afectarlo para tareas de análisis de sentimientos o reconocimiento de Entidades, ya que el uso de mayúsculas puede aportar pistas en estas tareas (generalmente se escriben los nombres propios en mayúsculas o se escriben palabras completas en mayúsculas para dar énfasis).

Existe además un conjunto de palabras denominadas stop words, que son palabras que no contienen un significado importante para ser utilizadas en el modelo de lenguaje. Por lo general, estas palabras se eliminan del texto, ya que agregan una gran cantidad de información innecesaria. Cada idioma tiene su propia lista de stop words, pero generalmente contienen las palabras que se usan comúnmente en el idioma. En el caso del inglés son palabras como “as, the, be, are”, y para español incluye palabras como “a, un, el, la, en”.

C. Segmentación/Tokenización

La normalización también incluye la separación del texto en unidades más pequeñas. La tarea de separación o tokenización de los textos se puede realizar de diferentes maneras. Normalmente, las palabras se encuentran separadas entre sí por espacios en blanco, pero los espacios en blanco no siempre son suficientes. Existen casos particulares, como la palabra “Nueva York”, que representan un solo concepto a pesar del hecho de que contiene un espacio. Para el procesamiento de textos se utilizan diferentes técnicas de tokenización. Se puede considerar a los “tokens” como representación de palabras, o secuencias de palabras (n-gramas de palabras), o extender el enfoque a secuencias de caracteres (n-gramas de caracteres), e incluso a skip-gramas.

N-gramas

Los n-gramas se refieren al proceso de combinar las palabras que se encuentran juntas con fines de representación, donde n representa el número de palabras que se desea combinar. Por ejemplo, si consideramos la oración:

El coche circulaba a gran velocidad

Un modelo de 1-grama, o unigrama, tokenizará la oración palabra por palabra, dando como resultado:

El	coche	circulaba	a	gran	velocidad

Similarmente, en la tokenización en bigramas se combinan de 2 palabras para formar un token, para nuestro ejemplo esto es:

El coche	coche circulaba	circulaba a	a gran	gran velocidad

De manera análoga, se puede tokenizar un texto mediante secuencia de caracteres. Por ejemplo, al tokenizar la oración "El coche circulaba" por pentagramas de caracter, se tendrá:

el co	l coc	coch	coche	oche	che c	he ci
e cir	circ	circu	ircul	rcula	culab	ulaba

K-skip-n-gramas

El proceso de tokenización en k-skip-n-gramas consiste en tomar las cadenas de texto que se forman entre "saltos" de palabras, es decir, omitiendo palabras que se encuentren entre ellas. Formalmente, un k-skip-n-grama es una subsecuencia de longitud n donde los componentes ocurren a una distancia máxima de k entre sí. Un 2-skip-1-grama incluye todos los bigramas que se forman al saltar una palabra, siguiendo el ejemplo anterior tendremos el siguiente resultado:

el circulaba	coche a	circulaba gran	a velocidad

D. Métodos de representación de documentos

Para que un algoritmo de clasificación pueda capturar relaciones entre datos requiere pasar de un conjunto de textos a datos estructurados. Existen diversas formas de representar los documentos en un esquema estructurado, particularmente en este trabajo se abordan diversos métodos desde dos enfoques: el **léxico** y el **semántico**. A continuación, se abordan de manera general, ambos enfoques.

Enfoque léxico

El léxico es el conjunto de palabras que conforma un determinado idioma. En el Procesamiento del Lenguaje Natural, a través de la normalización de los textos, es posible

representar con unidades léxicas (los tokens) el contenido de un conjunto de documentos. Las características que se pueden extraer de la representación de los documentos en los tokens se conocen como características léxicas.

La “Bolsa de Palabras” (“Bag of Words”) es una de las formas más comunes de representar los datos en un formato tabular donde en las columnas se representa el vocabulario total del corpus y cada fila representa un documento. La celda (intersección de la fila y la columna) representa el recuento de la palabra representada por la columna en ese documento en particular. De esta manera, se pueden aplicar algoritmos sobre los datos para construir modelos predictivos.

Esta representación funciona muy bien en algunas tareas de aprendizaje automático como la detección de spam, el clasificador de sentimientos, entre otros. Sin embargo, hay dos grandes inconvenientes de esta representación:

- Ignora el orden y la gramática de los documentos, por lo que se pierde el contexto en el que se usa una palabra.
- La matriz generada por esta representación es altamente dispersa y sesgada hacia las palabras más comunes; las palabras con mayor frecuencia son las más comunes, y no siempre aportan información del significado del texto.

Para atender el segundo punto, surge la representación **TF-IDF**, la cual es una forma de ponderar las palabras del vocabulario para dar un peso en proporción al impacto que tiene en el significado de un documento. La puntuación (score) es un producto de 2 medidas independientes: la frecuencia de término (TF) y la frecuencia inversa de documento (IDF). La ponderación para un token, o palabra, i en el documento j se calcula de acuerdo con la siguiente ecuación:

$$w_{ij} = \underbrace{tf_{ij}}_{TF} \times \underbrace{log\left(\frac{N}{df_i}\right)}_{IDF}$$

donde tfij es el número de ocurrencias del término i en el documento j, dfi es el número de documentos que contienen el término i y N es el número total de documentos. La IDF es una medida de cuánta información proporciona el token, es decir, si es común o raro en los documentos.

Existen, además, modelos basados en rasgos o características léxicas, los cuales siguen la idea de caracterizar a los documentos mediante una lista de atributos que resumen los rasgos más significativos o relevantes para la tarea que se quiere estudiar. Estos modelos convierten los textos a una representación numérica para cada documento a través de estadísticas de texto (por ejemplo, la longitud de palabra promedio, proporción de palabras

largas/cortas, entre otras) y por características sintácticas (a nivel de gramática, como el uso de verbos, adjetivos o adverbios, por nombrar algunos ejemplos).

La construcción de estas características depende del fenómeno de estudio, por ejemplo, en la tarea de análisis de sentimiento en textos, una variable (característica) importante es el número de palabras positivas, el número de palabras negativas o alguna relación entre estas; sin embargo, esta variable puede no ser significativa para la segmentación de contenidos

Se pueden representar textos mediante palabras o frases claves a través de los keywords y las keyphrases, los cuales son ampliamente usadas en las colecciones de documentos. Escriben el contenido de documentos y proporcionan un tipo de metadatos semánticos que se pueden usar para una amplia cantidad de propósitos. La tarea de asignar keyphrases a un documento se le llama keyphrase indexing. Algunos métodos para seleccionar a los candidatos son la extracción por n-gramas y shallow parsing.

La filtración se realiza por métodos simbólicos, donde se les aplica un esquema de pesos a las frases para asignarles un score. También se puede proponer un modelo estadístico y una función de rankeo a partir del conjunto de datos de entrenamiento (de donde se extrajeron las keywords manualmente).

Finalmente se crea una matriz indicadora, donde cada renglón pertenece a una noticia, y cada fila pertenece a una keyword. Si la noticia contiene la keyword se asigna el valor de 1, y se asigan 0 en otro caso. Con esto se puede crear una representación matemática para utilizar cualquier método de clasificación que pudiera funcionar con este tipo de datos.

Keyphrase Extraction Algorithm

La lógica del algoritmo es que su funcionamiento emula la forma en la que los humanos realizamos la extracción de keyphrases. Aunque un profesional no esté familiarizado con algún tema en particular, éste es capaz de identificar las keyphrases basándose en sus propiedades, tales como frecuencia en el documento, presencia en partes significativas del documento, etc.

Para extraer las keywords, se remueven los stop words, se realiza el stemming de las palabras resultantes y se ordenan alfabéticamente. Se usa la conflación de términos semánticos, gracias a los tesauros. El resultado es un conjunto de términos gramaticales relacionados al contenido del documento.

Para construir el modelo, se utiliza un conjunto de documentos de entrenamiento, de donde previamente se identificaron manualmente las keywords. Se calculan 4 características: el score tf-idf, la posición de la primera ocurrencia de la frase, la logitud de la frase y el node degree (número de enlaces del tesauro que conectan al término a otros candidatos). La razón de usar el node degree es que, si un documento describe un tópico en particular, entonces cubrirá la mayor cantidad de términos del tesauro de ese tema.

Cada candidato de los datos de entrenamiento se marca con una variable binaria, que indica si es index term (de acuerdo a la identificación manual). Esta variable binaria es la clase que será utilizada por el modelo naive Bayes. Para seleccionar index terms de nuevos documentos, se determinan a los candidatos y sus características, y se aplica el modelo construido durante el entrenamiento.

Enfoque semántico

El término semántica se refiere a los aspectos del significado, sentido o interpretación de signos lingüísticos como símbolos, palabras, expresiones o representaciones formales. A pesar de que los modelos desarrollados desde el enfoque semántico se basan también en los tokens, estos modelos buscan capturar la mayor cantidad de información posible del contexto, en una palabra, tratando de capturar información semántica.

Desde el enfoque semántico, los word embeddings son modelos utilizados para transformar las palabras a una representación estructurada. En estos modelos las palabras se mapean a un espacio vectorial n-dimensional, de manera que palabras semánticamente similares se encontrarán cerca entre ellas.

Existe una gran gama de propuestas de word embeddings que han mejorado el comportamiento de muchas tareas de Procesamiento de Lenguaje Natural, como es la identificación de nombres, identificación de idiomas, traducción automática, entre otros.

Word embeddings

En el lenguaje natural, las palabras no aparecen de manera aislada o aleatoria, dependen de otras palabras y van formando una secuencia de acuerdo con una estructura gramatical. Los modelos del lenguaje tratan de capturar este condicionamiento, con el objetivo de poder entender y predecir la estructura lingüística en elementos como frases u oraciones.

Además, estos modelos buscan capturar la relación existente entre las palabras. Así, un modelo de lenguaje puede ser representado por la probabilidad condicional de la siguiente palabra dada las anteriores, como

$$\hat{P}(w_1^T) = \prod_{i=1}^{T} \hat{P}(w_t|w_1^{t-1}), \tag{5.5}$$

donde w_t es la t-ésima palabra, y la subsecuencia $w_i^j = (w_i, w_{i+1}, \ldots, w_{j-1}, w_j)$.

Con esta representación, se puede expresar la función de probabilidad como un producto de probabilidades condicionales de la siguiente palabra dada las anteriores, es decir que para determinar la probabilidad tomamos en cuenta un contexto de T palabras. Esta función tiene parámetros que se pueden ajustar de forma iterativa para maximizar la probabilidad de los datos de entrenamiento o algún criterio de regularización.

Dentro de los modelos del lenguaje, la incorporación del aprendizaje profundo ha generado un importante avance para el Procesamiento del Lenguaje Natural. El modelo del lenguaje natural propuesto por Bengio et al., 2003, es un modelo basado en una red neuronal profunda para estimar las probabilidades de transición a partir de n-gramas. A grandes rasgos, el enfoque que propone se resume en tres puntos principales:

- Representar cada palabra del vocabulario a partir de un vector de características (un vector con entradas en R m).
- Expresar la función de probabilidad conjunta de secuencia de palabras en términos de los vectores de características de las palabras en la secuencia.
- Aprender simultáneamente los vectores de características y los parámetros de la función de probabilidad.

E. Clasificadores

Una vez extraídas todas las características deseadas de los textos se procede a utilizar algoritmos de aprendizaje máquina (ML) para hacer la clasificación de los textos.

En el esquema tradicional de un periódico existen secciones como: Política, Economía, Seguridad, entre otros, para ello se usan diferentes metodologías de modelación de tópicos, con el objetivo de encontrar grupos de noticias con características similares y comparar con las categorías de un periódico convencional.

Se distinguen métodos de agrupación no supervisados y clasificadores supervisados. Los métodos no supervisados no requieren ni permiten que de entrada se especifiquen las secciones o clases que esperamos obtener, de manera que los agrupamientos que se forman serán de acuerdo a la información disponible; mientras que los métodos de

clasificadores supervisados utilizan información de la categoría de los textos para clasificar textos futuros. Esta metodología puede ser más precisa puesto que puedes buscar categorías específicas de interés, pero requiere de trabajo manual previo para entrenar los modelos.

En términos generales, no existe un clasificador o tipo de agrupación que sea mejor para todas las tareas, depende siempre del problema que se esté estudiando. El desempeño de cada algoritmo puede variar debido a factores como el tamaño de la muestra, número de características, la naturaleza de los datos, cantidad de categorías, entre otros.

Entre los clasificadores supervisados más utilizados en el estado del arte (SVM y Random Forest), así como otros algoritmos que tienen buen desempeño en diversas tareas de aprendizaje supervisado (Regresión Logística y AdaBoost). Particularmente, los modelos de Regresión Logística y SVM son útiles cuando la relación entre la variable dependiente (clases) y las independientes (características) se aproxima a un modelo lineal. Sin embargo, si la relación es compleja y altamente no lineal, entonces los Árboles de Decisión tendrán mejores resultados de que un método lineal.

A continuación, se hace una descripción de algunas de ellas:

- **SVM:** Las SVM fueron introducidos como algoritmo de aprendizaje por Cortes y Vapnik, 1995. Es un algoritmo de clasificación lineal, el cual para un conjunto de datos de k clases, en un espacio N dimensional, induce separadores lineales o hiperplanos (N–1) dimensionales para separar esos puntos, ya sea en el espacio original o en una transformación del mismo (mediante Kernels), en k grupos
- **Regresión logística:** Es un modelo de regresión que permiten estudiar si una variable, generalmente binomial, depende de un conjunto variables o características. Desde la perspectiva de aprendizaje supervisado, la regresión logística es un algoritmo de gran utilidad en los problemas de regresión y de clasificación.
- **Arboles de decisión y modelos de ensamble:** Los árboles de decisión son un tipo de algoritmo de aprendizaje supervisado principalmente usados en problemas de clasificación. La construcción del árbol sigue un enfoque de división binaria recursiva, donde la tasa de error de clasificación se utiliza como criterio para la división binaria. En cada proceso de división analiza la mejor variable para ramificación sólo en el proceso actual, y se divide la muestra en conjuntos basados en la variable de entrada más significativa.

  Como mejora a los árboles de decisión, surgen los modelos de ensamble, que como su nombre lo dice, consisten en combinan varios árboles de decisión para producir un mejor rendimiento predictivo que utilizar un solo árbol de decisión. El principio fundamental detrás de estos modelos es que la unión de muchos clasificadores débiles puede formar un clasificador más fuerte.

Al construir un árbol de decisión pequeño se obtiene un modelo con baja varianza y alto sesgo. Se puede reducir el sesgo de predicción al incrementar la complejidad del modelo, pero haciendo que la varianza aumente. El uso de modelos de ensamble es una forma de lidiar con el ajuste entre varianza y sesgo. Los principales modelos de ensamble son Bagging, Random Forest y Boosting.

- **AdaBoost:** es una forma de aprendizaje secuencial, la idea es entrenar un modelo con todo el conjunto de entrenamiento, y los modelos posteriores se construyen ajustando los valores de error residual del modelo inicial. De esta manera, se intenta dar mayor peso a aquellas observaciones que el modelo anterior estimó pobremente, con lo que el algoritmo se adapta y logra obtener mejores resultados. Esta clasificación podría ser relevante para productos como **Monitoreo semanal de negocios** o **Reporte de contexto internacional**. Esto para poder asignar automáticamente las noticias de negocios y la sección internacional

F. Modelo STM.

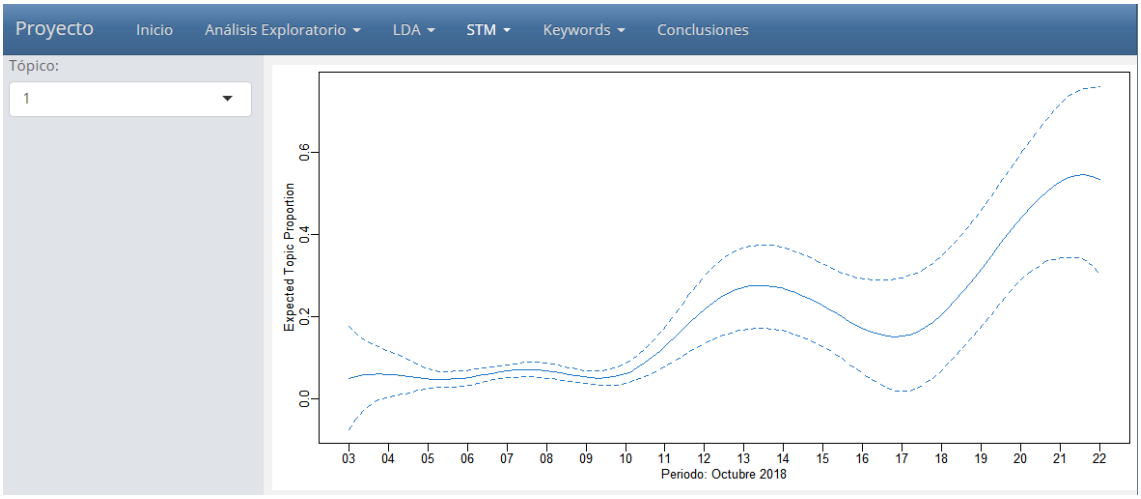
El Modelo de tópicos estructural (STM) es una generalización de LDA (Latent Dirichlet Allocation) que se adapta a la estructura del corpus a través de covariables a nivel de documento que afectan la prevalencia tópica y/o el contenido tópico. La idea central es especificar las *a priori* como modelos lineales generalizados a través de los cuales podemos condicionar los datos observados arbitrariamente. El modelo generaliza varios enfoques existentes en la literatura y se pueden incorporar a la estructura específica del corpus sin desarrollar nuevos modelos desde cero.

El modelo combina y extiende tres modelos existentes: el modelo de tópico correlacionado (CTM), el Modelo de Tópico de Regresión Multinomial (DMR) de Dirichlet y el Modelo de Tópico de Generación Aditiva Sparse (SAGE). La normal logística utilizada en CTM se reemplaza por un modelo lineal logístico-normal. La matriz de diseño para las covariables X permite formas funcionales arbitrariamente flexibles de las covariables originales utilizando funciones de base radial. La distribución sobre las palabras se reemplaza con un logit multinomial, de modo que la distribución de un token es la combinación de tres efectos (tema, covariables, interacción tema-covariable) que se operan como desviaciones dispersas de una frecuencia de palabra de referencia (m).

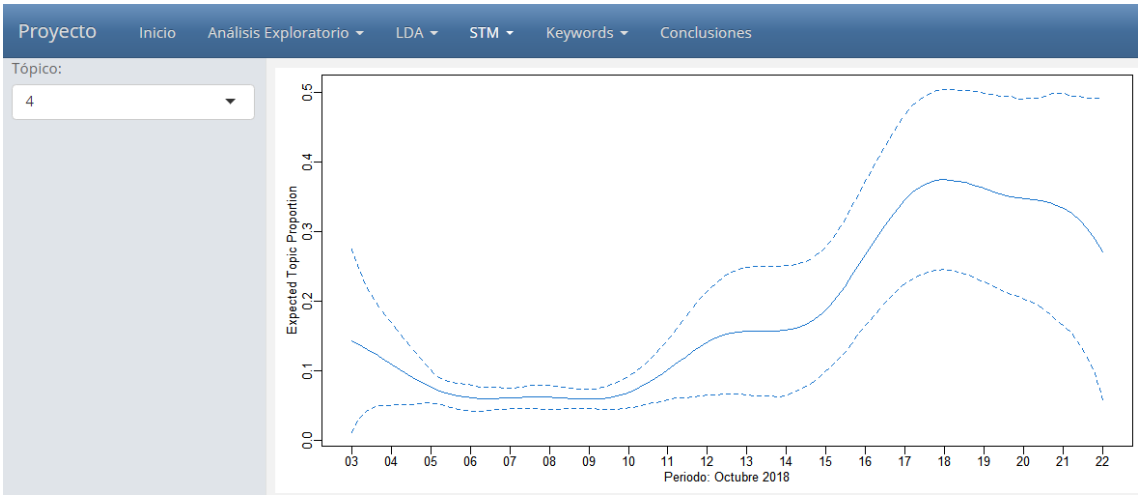
El atractivo principal de STM es que permite incorporar información de covariables (metadatos) en el modelado de los tópicos. Los metadatos se pueden ingresar en el modelo de dos maneras: prevalencia y contenido tópicos. Las covariables de metadatos para la prevalencia tópica permiten que los metadatos observados afecten la frecuencia con la que se discute un tema. Las covariantes en el contenido tópico permiten que los metadatos observados afecten el uso de la tasa de palabras dentro de un tema dado, es decir, cómo se discute un tema en particular. En un caso particular, los metadatos podrían consistir en

la fecha de publicación de la noticia y el periódico que la emite. El objetivo es usar la fecha para medir la prevalencia de los tópicos y la fuente de información (el periódico) y para analizar el contenido del tópico.

Nuevo Aeropuerto Internacional de la Ciudad de México (NAIM)
Año: 2018



Caravana de migrantes
Año: 2018



G. Georreferenciación de noticias

Para identificar el estado al que pertenece la noticia, se puede crear un vector de estados y capitales, y buscar la coincidencia de palabras del vector en cada noticia. Posteriormente, se filtra por las noticias que sí contenían una o más palabras del vector. Finalmente, se conserva la primera palabra del vector que es identificada, bajo la lógica de que primero aparece el estado más relevante.

También de cada sección se puede hacer un análisis de tópicos relevantes alrededor de una región geográfica por la frecuencia de aparición y su permeo en la sociedad, combinando los resultados de clasificadores y el modelo STM. Esto podría ser relevante dado que podremos concentrarnos en ver información relevante en regiones geográficas específicas. Por ejemplo, poder clasificar la procedencia de las noticias podría facilitar el estudio del contexto de alguna región geográfica. También podría ser útil análisis más complicados de la popularidad de algún actor u organización en distintas regiones geográficas. De igual forma, puede ser relevante para evaluar la popularidad de candidatos de elecciones presidenciales a lo largo de la república, evaluar la relevancia de una organización en cierta región, hacer comparaciones de las palabras clave de temas de seguridad en distintas regiones, etc.

H. Reconocimiento de entidades.

La relevancia del reconocimiento de entidades de las noticias recae en el hecho que se pueden hacer grafos de actores principales en determinada región, por ejemplo, identifica actores que son mencionados en las noticias y puedes hacer una especie de listado de personalidades y su relación con temas alrededor de ellos.

Se puede hacer un grafo de relación de personajes que aparezcan en una misma noticia, analizando las interacciones entre personajes y organizaciones.

Una de las ventajas de esta metodología es que las entidades pueden ser dadas, es decir, queremos saber todas las noticias relacionadas con el nombre “Andrés Manuel López Obrador” y analizar cuáles son los tópicos relevantes que le rodean. Tener este tipo de desarrollos listos también podrían facilitar el desarrollo de una herramienta de análisis de sentimiento alrededor de un personaje. Por ejemplo, con relación a seguridad, ¿cuál es el sentimiento en relación a AMLO?. Las complicaciones podrían venir en el sentido de buscar cual es el titular o personaje principal de la nota, es decir, identifica de actor A o actor B es al que se hace referencia la nota. Pero de cualquier manera podría ser un buen indicador del sentimiento hacía el personaje.

Esto mismo para las entidades que no son una personas, por ejemplo, “Banco de México”, y ver todos los actores o entidades relacionadas al banco de México y el sentimiento alrededor de esta identidad.

Cabe señalar que es necesario también utilizar la georreferenciación ya que se requiere hacer un diagnóstico de un municipio, estado o región con relación a temas como Gobierno, Seguridad, Social o “trending topics” y la prevalencia de estos temas. Para esto último se requiere la clasificación de tópicos y el modelo STM. Con relación a actores de interés en un lugar se puede utilizar la frecuencia de aparición de actores en diferentes noticias. Esto con el fin de identificar la popularidad (positiva o negativa) del actor en una región de interés.

TECNOLOGÍAS BIG DATA

Teniendo bases de datos muy grandes de noticias, se pueden hacer los análisis previos para incorporar variables como: Entidades participantes (nombres de las entidades relacionadas a las notas), región geográfica a la que pertenece (país, estado, municipio, localidad), fecha, fuente y tópico. Esto facilitaría hacer un sistema de consultas basado en estos parámetros. Por lo que entrenar modelos de reconocimiento de entidades geográficas y de actores ayudaría a robustecer los sistemas de información y automatizar la búsqueda. Esto evita hacer búsquedas masivas de información constantemente (ej. buscar en cada .txt las notas donde aparezca tales personajes), propone entonces una sola ejecución y se automatiza cada ocasión que identifique una nueva entidad o especificación, marcándose como variables relacionadas al texto de la noticia. Enseguida, las consultas se pueden hacer buscando específicamente las variables relacionadas a la nota y no el texto de la noticia *per se*. Esto es importante puesto que al automatizar la obtención de noticias inevitablemente ampliaría considerablemente la cantidad de noticias en la base de datos.

Utilización de plataformas como Azure, Google, WATSON

Existen una serie de plataformas que proveen de estos productos. Esto facilita el uso de algoritmos que serían complicados de programar o muy laboriosos y da la disponibilidad de ellos en forma de APIs que están listas para ser consumidas por el usuario. Similarmente puede hacer buena sinergia con algún otro producto de bases de datos que se utilice con dicho proveedor de servicios.

La desventaja principal es que por sí solos no brindan la capacidad de análisis deseada, a menos que haya un esfuerzo previo de pre-procesamiento de las notas descargadas (como observaciones en la sección de limpieza de datos). Estos productos funcionan óptimamente cuando ya tienes a tu disposición la automatización del web Scraping y el tratamiento de las noticias, por lo que no podemos considerar que estas plataformas son suficientes para el objetivo de automatización de la obtención de información de medios digitales.

Como ejemplo, mostramos (abajo) las tareas que realiza el producto de Cognitive Services – Análisis de texto de Microsoft Azure:

El director general del Instituto Mexicano del Seguro Social (IMSS), Tuffic Miguel, supervisó la obra del Hospital General Regional (HGR) No. 2, en el municipio de El Marqués, la cual presenta un avance de 99 por ciento.

De esta forma, la infraestructura será inaugurada en las siguientes semanas, informó el director del IMSS.

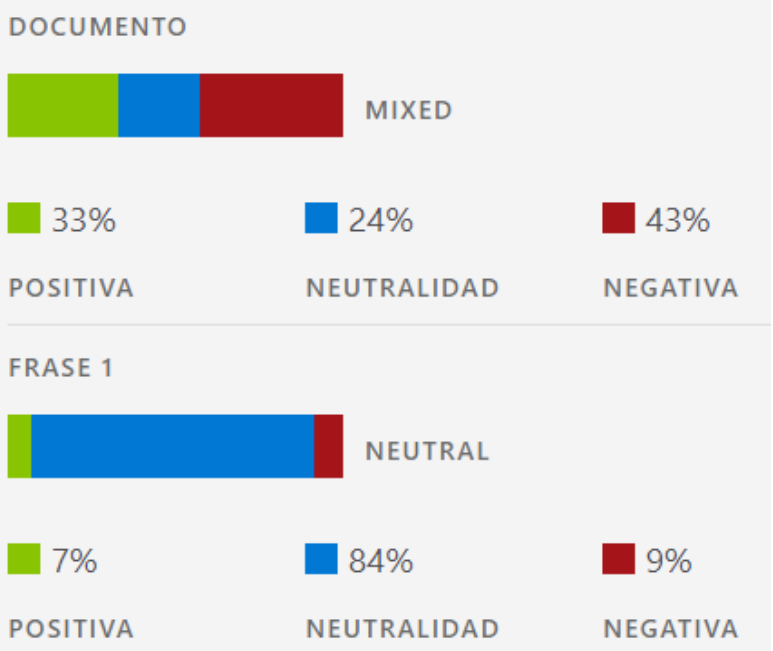
El nuevo hospital, que representa una inversión de mil 500 millones de pesos en obra y equipamiento, beneficiará a más de 800 mil derechohabientes con 53 especialidades, más de las que ofrece el HGR No. 1 del IMSS, en la capital del estado.

Entre las nuevas especialidades, de acuerdo con el organismo, se encuentran la atención en inmunología y gineco–oncología; mientras que en el área de pediatría se otorgarán endocrinología, cardiología, nefrología, neumología, neonatología, gastroenterología, hematología y oncología.

Las autoridades del Instituto han destacado la importancia de este proyecto, toda

FRASES CLAVE: IMSS, Seguro Social, director general, HGR, derechohabientes de Querétaro, Hospital General Regional, obra, gobernador de Querétaro, oncología, Campaña de Vacunación, nuevo hospital, organismo, Tuffic Miguel, secretario de Salud nacional, Instituto Mexicano, millones de acciones, millones de pesos, Semana Nacional, nuevas especialidades, distribución de vida suero oral, Manuel Ruiz López, neumología, neonatología, aplicación de vacunas, ácido fólico, José Narro, enfermeras, endocrinología, cardiología, nefrología, gastroenterología,

OPINIÓN:



ENTIDADES CON NOMBRE:

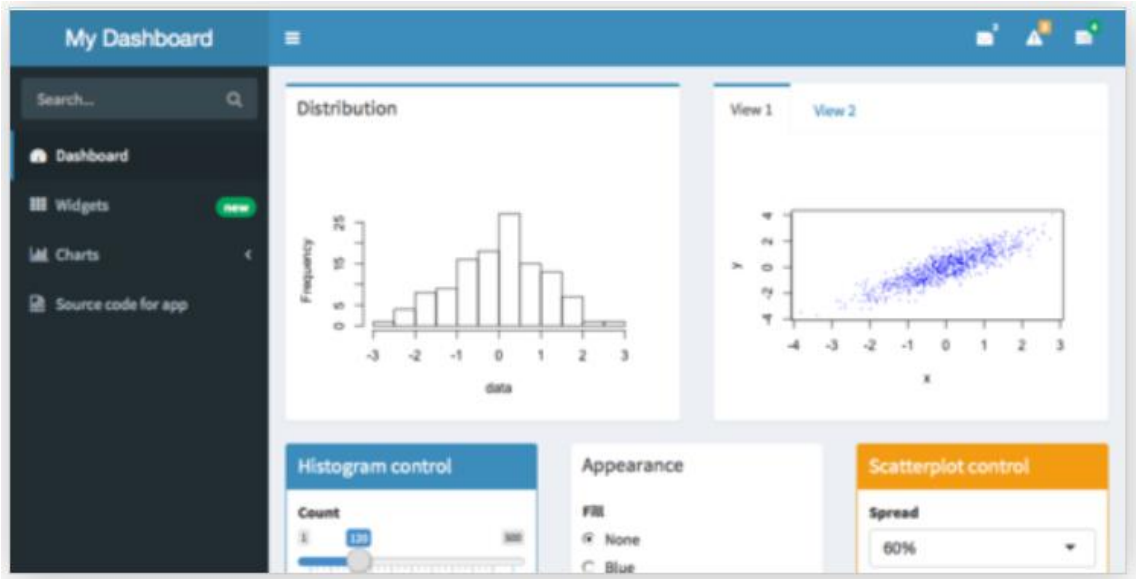
Instituto Mexicano del Seguro Social [Organization]
IMSS [Organization]
Tuffic Miguel [Person]
2 [Quantity-Number]
un [Quantity-Number]
99 por ciento [Quantity-Percentage]
IMSS [Organization]
una [DateTime-Time]
mil [Quantity-Number]
500 millones de pesos [Quantity-Currency]
800 mil [Quantity-Number]

En general las diferentes plataformas ofrecen los mismos productos. La diferencia principal se presenta en los precios y la robustez de sus modelos entrenados para el idioma deseado.

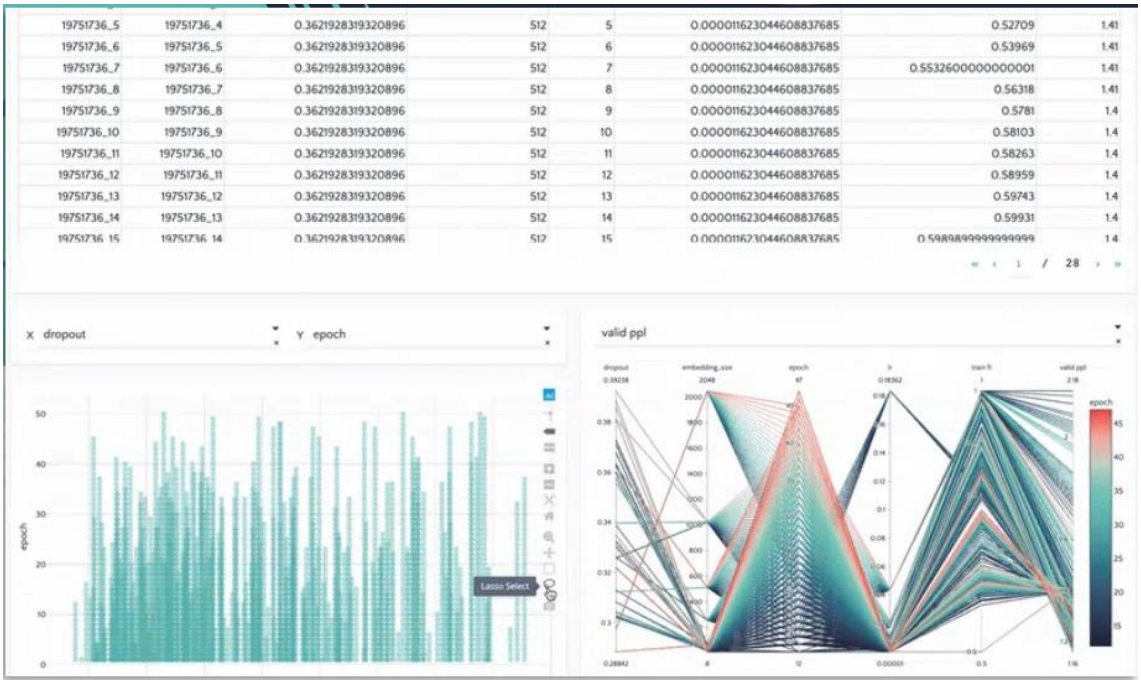
VISUALIZACIÓN

A. **Tableau:** Plataforma de análisis y diseño de Dashboards con interfaz gráfica intuitiva.

B. R-shiny: Puedes montar dashboards fácilmente con las estructuras de datos de R y subirlas a una IP expuesta en un servidor. Puedes utilizar Plotly. Este involucra más programación.



C. Plotly: Son recursos gráficos pero se levantan en servicios como Shiny, Dash, etcétera. Involucra más programación y funciona mejor en Python. Esto porque Python tiene mejor manejo de diferentes estructuras de datos y es más fácil desplegar APIs.



19751736_5	19751736_4	0.3621928319320896	512	5	0.00001162304460837685	0.52709	1.41
19751736_6	19751736_5	0.3621928319320896	512	6	0.00001162304460837685	0.53969	1.41
19751736_7	19751736_6	0.3621928319320896	512	7	0.00001162304460837685	0.5532600000000001	1.41
19751736_8	19751736_7	0.3621928319320896	512	8	0.00001162304460837685	0.56318	1.41
19751736_9	19751736_8	0.3621928319320896	512	9	0.00001162304460837685	0.5781	1.4
19751736_10	19751736_9	0.3621928319320896	512	10	0.00001162304460837685	0.58103	1.4
19751736_11	19751736_10	0.3621928319320896	512	11	0.00001162304460837685	0.58263	1.4
19751736_12	19751736_11	0.3621928319320896	512	12	0.00001162304460837685	0.58959	1.4
19751736_13	19751736_12	0.3621928319320896	512	13	0.00001162304460837685	0.59743	1.4
19751736_14	19751736_13	0.3621928319320896	512	14	0.00001162304460837685	0.59931	1.4
19751736_15	19751736_14	0.3621928319320896	512	15	0.00001162304460837685	0.5989899999999999	1.4

« ‹ 1 / 28 › »