Universidad del Desarrollo

Clasificación de texto con técnicas de procesamiento del lenguaje natural y machine learning para analizar la relación entre noticias publicadas en la web y la variación de los índices de la bolsa de comercio en Chile

Capstone project presentado a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al grado académico de Magíster en Data Science

Profesores Guía:
María Paz Raveau Morales
Cristian Candia Vallejos

Alumno:
Adrián Alarcón Becerra

Santiago, diciembre de 2022

Contenido

1 Introducción

Cuando se produce repercusión de algún suceso a nivel global, en una era donde la informática y la tecnología nos conecta con toda la información en cuestión de segundos, todos los inversores del planeta se muestran temerosos ante dicho suceso en los mercados financieros. Especialmente cuando es un suceso negativo y nuestra posición es alcista, dada la posibilidad de que dicho suceso, por su género, lastre el rendimiento que estábamos obteniendo por nuestra posición en el mercado.

Lo que algunos denominan el análisis fundamental, o análisis macroeconómico, en los mercados financieros, es un tipo de inversión basado, fundamentalmente, en los sucesos que se dan en la economía global, pudiendo anticiparnos al impacto de este en los mercados financieros. Un análisis complejo, pero al que, sin embargo, se suman muchos traders novatos, en busca de generar rendimiento en la era de la información y con el simple hecho de poseer un Smartphone y conexión a internet.

En algunos lugares del mundo, la programación y el avance tecnológico ha llevado a que haya grupos de traders que han configurado robots para invertir de forma automática cuando se produce algún suceso. Incluso, por añadir un caso gracioso, un robot que estaba programado para ejecutar posiciones en los mercados referenciando estas a los tweets del presidente de los Estados Unidos, Donald Trump. En función del carácter del tweet, el robot consideraba si era bueno o malo, y, de forma simultánea, ejecuta posiciones bajistas, o bien, posiciones alcistas.

Si atendemos al caso anteriormente mencionado, más que una estrategia de inversión, programar un robot para ejecutar posiciones en referencia a los tweets del presidente Donald Trump podría parecernos un chiste. Una señal poco común a la que acogerse, pero el cual, de acuerdo con un estudio que ha realizado el Fondo Monetario Internacional (FMI), podría tener más sentido del que, a priori, parece tener. Un estudio en el que, precisamente, se cuestiona el impacto de las noticias en los mercados financieros, así como en los distintos precios de los activos.

Dado lo anterior resultaría interesante revisar si en Chile existe una relación entre las noticias producidas en Chile y la variación diaria de la bolsa de comercio de Santiago.

2 Objetivos

El objetivo general del trabajo será revisar si existe o si se puede encontrar una relación entre el contenido de las noticias asociadas al concepto de “economía” con la variación del índice de Precios Selectivo de Acciones (IPSA), que mide las variaciones de precios de las empresas emisoras chilenas más grandes y líquidas listadas en la Bolsa de comercio de Santiago. Finalmente se pretende evidenciar si existen palabras que se relacionen con variaciones positivas y otras con variaciones negativas del IPSA.

Seguidamente se describen los objetivos particulares de este trabajo:

- Proponer una metodología para obtener el texto de las noticias desde el sitio web de CNN Chile con el objeto de construir un conjunto de datos ordenado y clasificado en categorías que permitan su análisis. Los textos del conjunto de datos comenzarán en noviembre del 2012 hasta septiembre de 2022 en idioma español.

- Se realizará un proceso de limpieza de datos mediante las técnicas de procesamiento del lenguaje natural de tal forma de facilitar la implementación de clasificadores de texto.
- Mediante técnicas de exploración y visualización se revisarán las principales características del conjunto de datos. Se analizarán las diferentes categorías de noticias que se descargarán del sitio web CNN Chile.
- Se realizará una separación de los datos o clusterización mediante métodos de machine learning con el objeto de intentar encontrar una mejor categorización de los datos, comparando esta clusterización con las categorías que traen los datos desde el sitio web. La idea es determinar si existe alguna separación de los datos que se relacione mejor con el concepto de "economía" que la categorización que trae los datos desde el sitio web.
- Se implementará una metodología para relacionar la variación del IPSA con las noticias extraídas desde la web.
- Se probarán distintos clasificadores binarios para poder predecir si existirá una variación positiva o negativa del IPSA en función de las noticias que se vayan generando.
- Se construirán visualizaciones o gráficos que permitan determinar si existen palabras o asociación de palabras que se puedan relacionar con variaciones positivas o negativas del IPSA.

Como base para este trabajo se utilizará el lenguaje de programación Python, utilizando las siguientes herramientas de desarrollo o entorno de desarrollo integrado (IDE, por sus siglas en inglés):

- Júpiter notebook
- Google Colab
- Notepad ++

Para obtener el conjunto de datos se utilizará la librería y/o framework Scrapy (del lenguaje de programación Python), con dicha herramienta se realiza el "raspado" o extracción de los datos desde los sitios web donde se publican las noticias, en particular se utiliza el sitio web https://www.cnnchile.com.

3 Fundamentos Teóricos

3.1 Web

(World Wide Web también conocido simplemente como www), es la agrupación de varios documentos (páginas webs) vinculados entre sí a través de links de hipertexto. El "hipertexto" es la combinación de escrituras, ilustraciones y ficheros de cualquier tipo, todo estos en una misma página web (https://marinolatorre.umch.edu.pe/historia-de-la-web-1-0-2-0-3-0-y-4-0/). En base lo anterior, se entiende que tanto e-mail, juegos, redes sociales como Facebook, Twitter, así como blogs, wikis y lo demás son parte de la Internet, pero no la web en sí. No obstante, existen dos términos que a menudo se utilizan como es el sitio web y página web. El sitio web es definido como el conjunto de páginas web relacionados entre sí, las cuales son identificadas a través de un nombre conocido como el dominio, el cual normalmente es alojado en un servidor HTTP. De ahí que puede ser accedido por medio de un protocolo IP al utilizar un hipervínculo que permite reconocer el sitio web. Por otro lado, una página web, es un documento que por lo general contiene textos, audios, programas, vídeos, etc. A un sitio web se puede ingresar, básicamente utilizando un navegador web a través de su enlace, en este sentido no es necesario acceder hasta el servidor de alojamiento del sitio web objetivo [1]. Las páginas webs se componen de "elementos", que son unidades con significado, como: títulos,

párrafos, lista, tablas, imágenes, etc. Generalmente, los elementos se presentan por medio de etiquetas, estas etiquetas pueden ser de inicio (apertura) y de final (cierre), dentro de ellas está el contenido del documento, que puede ser texto o bien, más elementos. Las etiquetas tienen un nombre y puede abarcar propiedades como las que se puede apreciar en el siguiente ejemplo:

- <h1> Título del artículo </h1>
- Párrafo
- Visita nuestro sitio
- <img src="casa.jpg">

La línea número 1 se refiere a un título. La etiqueta h1 se usa para los títulos principales. Los títulos inferiores se pueden representar con las etiquetas h2, h3 y así sucesivamente. La línea 2 muestra la etiqueta p que usa para describir párrafos de texto. La línea 3 es un enlace, el cual se representa con la etiqueta a. Dentro de este, en el atributo href está indicada la dirección del enlace. Y la línea 4 es la etiqueta img, con la que se representan imágenes. En el atributo src se coloca la dirección del archivo de la imagen (https://blogprog.gonzalolopez.es/articulos/web-scraping.html).

3.2 Protocolo de la web

Para que todas estas técnicas funcionen como cualquier aplicación basados en web debe regirse en ciertas reglas, como es el protocolo de la web. Por tanto, un sitio web para poder brindar o ejecutar las peticiones solicitadas debe usar determinadas tecnologías. Esta acción es conocido como la interoperabilidad entre distintas aplicaciones. Aquí es donde también intervienen ciertos protocolos y normas que determinan el modo de comunicación, la forma de los datos que son enviados y recibidos, al igual que su funcionamiento, entre otros. A continuación, se detallan los protocolos y normas fundamentales para que un servicio web pueda operar y ser utilizado:

Uno de los formatos es el XML (es la sigla de Xtensible Markup Language), en otras palabras, es el lenguaje de marcado que facilita la escritura de los contenidos permitiéndoles dividir de su formato. Cabe mencionar que esta norma contiene a su vez las normas como el DTD o XSD, la cuales permiten la configuración del lenguaje y el XSL-FO y XSLT las cuales sirven para la conversión y presentación de la información [2].

El formato JSON (es sigla de JavaScript Object Notation), es un tipo de lenguaje que permite guardar e intercambiar información, tiene un formato sencillo, de tal manera que puede ser leído y escrito por personas [2].

El estándar SOAP (originalmente las siglas de Simple Object Access Protocol) es un protocolo estándar que define cómo dos objetos en diferentes procesos pueden comunicarse por medio de intercambio de datos XML. Este protocolo se deriva de un protocolo creado por Dave Winer en 1998, llamado XML-RPC. SOAP fue creado por Microsoft, IBM y otros. Está actualmente bajo el auspicio de la W3C. Al igual que el protocolo anterior, este también es utilizado para intercambiar información, funcionando de la siguiente manera: una parte realiza la petición (cliente) y la otra parte responde dicha petición (servicio), básicamente esta norma se utiliza en el ámbito de los servicios web (https://books.google.fr/books?id=wiXOyXdvHO8C&printsec=frontcover&hl=es&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false).

REST (esta sigla corresponde a las siguientes: Representational State Transfer), es una norma de envío de representación de estado, hay que tener en cuenta que este no posee estado (del inglés stateless), lo que significa que, cuando exista dos requerimientos sea cual sea, el servicio tiende a perder toda su información [3].

El protocolo WSDL (sigla que representa a las siguientes: Web Services Description Language), este protocolo o norma se basa en el formato XML, sólo para la interfaz del servicio web, es decir, los métodos y criterios que se muestran, así como la entrada y salida para llamar a los servicios (https://books.google.fr/books?id=wiXOyXdvHO8C&printsec=frontcover&hl=es&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false).

3.3 Web Scraping

Web Scraping (traducido al español sería raspado de páginas web), es una práctica que va tomando fuerza a medida que transcurre el tiempo dentro de las empresas o incluso dentro de las instituciones educativas. En términos simples, es una forma de obtener datos de una o varias páginas web de forma automática, esto incluye tales como redes sociales, repositorios de código, blog, tiendas online, sitios empresariales. Es considerado como una técnica de programación ya que facilita la extracción datos de la World Wide Web, es decir, páginas web. En algunos casos, estos tipos de software van mucho más de allá de una simple programación, dicho de otra manera, los programas son dotados de inteligencia artificial, lo que genera más autonomía posibilitando una navegación continua por Internet y extraer información relevante. Para el desarrollo de esta técnica se puede hacer mediante una variedad de lenguajes de programación que admiten la programación del protocolo de Transferencia de Hipertexto. El objetivo principal de Web Scraping, es conseguir cantidades enormes de información, mediante algoritmos de búsqueda de los cuales pueden rastrear centenares de sitios web, esta actividad se lleva a cabo, normalmente en páginas que utilizan lenguajes de marcado como HTML o XHTML, de tal manera que, es necesario conocer cómo está organizada la información de la página web de la cual se desea extraer información. Bajo este criterio se puede asumir que el web scraping es la solución intermedia entre la recolección manual de datos (marcando, copiando y pegando textos) y el acceso automatizado a los mismos con base en un protocolo predeterminado (API, framework, librerías, etc.). Básicamente, esta práctica se aplica cuando tales protocolos no existen y la cantidad de datos que se desea extraer es demasiado grande para que pueda ser realizada en forma manual (https://www.academia.edu/35895308/Web_scraping). El Web Scraping también ayuda a la automatización de la web de muchas maneras, incluyendo datos meteorológicos, detección de cambios en la web y comparación de precios de sitios web en línea, entre otros. Esta técnica, puede convertir información no estructurado en información estructurado, luego de ello, una vez que sea validado, se procede a guardarlos en una base de datos. Posteriormente, se pueden analizar y conocer el código de HTML devuelto por cualquier sitios tras una petición HTTP:GET [4].

El término “scraping” implica que la extracción de la información puede ser de cualquier fuente como base de datos, archivos CSV, repositorios de códigos como Github, Bitbucket o GitLab, por lo que no necesariamente la fuente de extracción de información debe ser netamente Sitios Web Comerciales o tiendas online.

3.3.1 Web Scraping con librerías

El proceso de Web Scraping se compone de estos sencillos pasos fundamentales:

URL semilla
El primer paso para tener siempre a consideración es tener una url semilla, pues se parte desde aquí. Básicamente una url semilla, es aquel sitio web del cual se hará web scraping.

Request
A través del método request se realiza requerimientos a la url semilla, es decir, se le indica qué información o qué partes de url se quiere extraer información.

Response
Luego de ello se obtiene una respuesta de la url, esta respuesta puede ser en formato XML o HTML, que posteriormente será parseado, es decir, se obtendrán los ítems especificados en el requests.

Populate Items
Los ítems dependen de la página web, según su estructura. Desde ellos se obtiene la información deseada.

More URLs
A partir de la URL semilla se puede ir a más URLs obteniendo así la información deseada. Y de estas urls se repite los pasos indicados anteriormente.

Ventajas y desventajas del Web Scraping con librerías

Ventajas
No se depende de un API: La ventaja más importante que tiene Web Scraping es la independencia del API, por lo cual no existe límite de ningún tipo.
No tiene limitaciones: No tiene límite en el tiempo de extraer información (rate limit), así como tampoco tiene límite sobre qué información se desea extraer.

Desventajas
Siempre dependerá de la estructura de XML o HTML de la página a la cual se quiere realizar scraping.
Pueden banear la IP: Una web scraping al ser una actividad repetitiva, si no se aplica de manera adecuada esta técnica, puede resultar invasivo para el sitio objetivo o incluso puede ser visto como un ataque al sistema. En este sentido y teniendo en cuenta, además que algunos sitios web disponen de mecanismos de seguridad, pueden bloquear la IP del visitante, inhabilitando así la posibilidad de realizar más visitas o peticiones.

3.4 Herramientas y Elementos del Web Scraping

La técnica de Web Scraping requiere de dos actores importantes para llevar a cabo, estos son:

HTML

La sigla HTML corresponde a los siguientes términos HyperText Markup Language (en español conocido como Lenguaje Marcado de Hipertexto) y es el lenguaje de marcado más utilizado para desarrollar aplicaciones web (https://developer.mozilla.org/es/docs/Web/HTML).

“Hipertexto” hace referencia a los hipervínculos que permiten interconectar las diferentes páginas web, de hecho, esto puede existir dentro del mismo o a través de distintos sitios web. El término

“marcado” hace referencia a aquello que permite cargar ya sea textos, imágenes o algún otro contenido que se quiera incluir en una página web con el objetivo de mostrar por medio del navegador web. HTML abarca todo lo que son los elementos especiales, por ejemplo, <head>, <title>, <body>, <header>, <section>, <p>, entre otros (https://www.w3schools.com/html/html_intro.asp).

HTTP

HTTP es una sigla que representa a los siguientes términos HyperText Transfer Protocol, en español denominado Protocolo de Transferencia de Hipertextos. Es una regla para la capa de aplicación que sirve como canal por el cual son transferidos un conjunto de hipertextos y multimedios conocido como hipermedia, esto es el propio HTML, o bien es el protocolo de transmisión de información de la Word Wide Web. Fue creado para la interacción entre los diferentes servidores y navegadores web. De ahí que surge el paradigma cliente-servidor, donde el cliente (usuario) se conecta con el fin de enviar una solicitud hacia el servidor, por la cual recibe una respuesta de parte de este (https://developer.mozilla.org/es/docs/Web/HTTP).

3.4.1 Scrapy

Es un marco de desarrollo enfocada a las aplicaciones que realizan el rastreo de sitios web y extraen información. Es utilizado y aplicado en distintos proyectos orientadas al web scraping, principalmente creado para el lenguaje Python. Con este framework las páginas web se analizan automáticamente y los contenidos web se extraen usando expresiones XPath. En la siguiente ilustración (Figura 1) se puede observar la arquitectura de Scrapy al igual que sus elementos y la secuencia de datos que son parte de la librería.

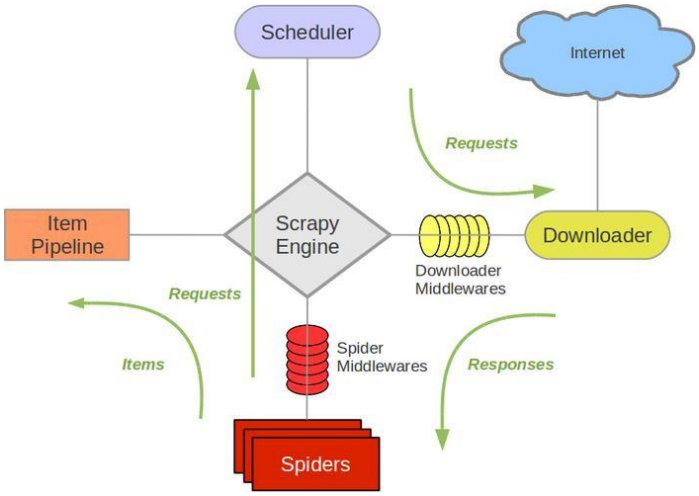


Figura 1: estructura de la librería o framework Scrapy (fuente: https://josefgonzalez.me/es/post/scrapy-en-jupyter/)

Componentes de Scrapy

- Scrapy Engine: Es el componente que se encarga de monitorizar la secuencia de datos dentro del sistema y es capaz de lanzar eventos en caso de que se realiza cierta acción.

- Schedule Planificador: Este componente es el responsable de receptar peticiones del motor, una vez recibido las coloca en cola para luego devolver al motor para cuando éste las necesita.

- Descargador: Este elemento es el encargado de realizar descarga de las páginas web y suministrar al motor y que este a su vez, suministra a las arañas web.

- Spiders o arañas: Este componente es personalizable, es decir, en base un análisis o según lo que se requiera los ítems pueden ser extraídos de las páginas web scrapeadas.

- Item Pipeline o Tubería o canal de elementos: El canal o la tubería de ítems, es la que se encarga del tratamiento de los datos, después de que han sido sacados por los spiders. Las actividades típicas de este elemento es la depuración, verificación o la persistencia, dicho de otra forma, guardarlos en una base de datos.

3.5 Bases de datos

Una Base de Datos (BD) es una colección de archivos interrelacionados creados con un Sistema de Manejo de Bases de Datos (DSMS) (figura 2). El contenido de una BD se obtiene combinando datos de todas las diferentes fuentes en una organización, de tal manera que los datos estén disponibles para todos los usuarios, y los datos redundantes puedan eliminarse, o al menos minimizarse. Los datos almacenados en una BD se encuentran de forma física en una disposición distinta a la perspectiva lógica, además de que los usuarios pueden tener acceso a los datos [5].

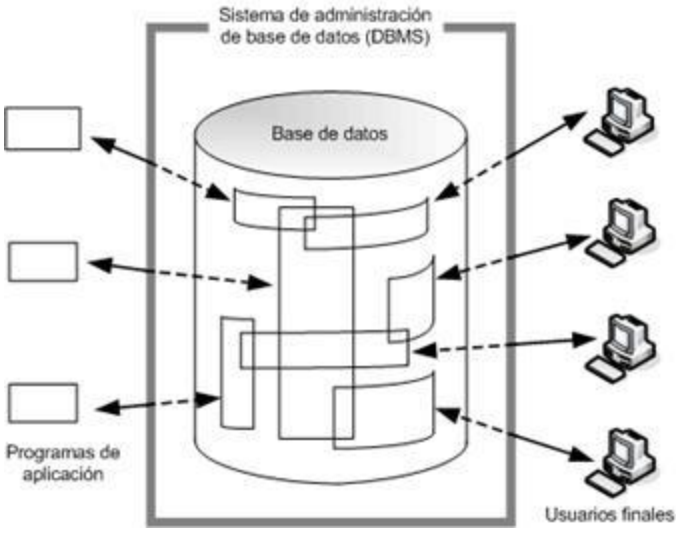


Figura 2: estructura de una base de datos (fuente: https://inquisitivoyanalitico.wordpress.com/2018/08/22/modulo-3-sistemas-manejadores-de-base-de-datos/)

3.5.1 Entidades, Relaciones y Atributos.

En el tema de bases de datos se especifican varios conceptos para el entendimiento de este. La mayoría de los Sistemas de Administración de Bases de Datos (DBMS) comerciales, las tablas almacenan un conjunto de entidades. Las tablas son arreglos de datos en dos dimensiones, además de estar formadas por un encabezado (el primer renglón) y un cuerpo (los otros renglones) que son la muestra del contenido. Las relaciones son las conexiones entre tablas. Mientras que las entidades son un conjunto de datos del mismo tema, y se puede acceder de forma conjunta, la entidad puede

representar una persona, lugar, evento o casa. El atributo es una propiedad de la entidad o relación, cada tributo tiene un tipo de dato que define el valor y las operaciones permitidas, el atributo es sinónimo de campo o columna [6].

3.5.2 Características de una Base de Datos.

- Persistente: Los datos almacenados residen en un almacenamiento estable, tal como un disco magnético. El almacenamiento y el mantenimiento de los datos son costosos, por lo que solo se almacenan los datos relevantes para la toma de decisiones.
- Compartir: La BD puede tener diferentes usos y usuarios, además de proporcionar memoria común para la realización de varias funciones en una organización.
- Interrelación: Los datos almacenados como unidades separadas se pueden conectar para mostrar un cuadro completo. Este concepto demuestra las entidades y las relaciones.

El Sistema de Manejo de Bases de Datos (DBMS) es una colección de numerosas rutinas de software interrelacionadas, cada una de las cuales es responsable de alguna tarea específica. Las funciones principales de un DBMS son:

- Crear y organizar la Base de Datos.
- Establecer y mantener las trayectorias de acceso a la Base de Datos, de tal manera que los datos en cualquier parte de la base puedan acceder rápidamente.
- Manejar los datos de acuerdo con las peticiones de los usuarios.
- Mantener la integridad y seguridad de los datos.
- Registrar el uso de las bases de datos.

El DMBS interpreta y procesa las peticiones del usuario para recobrar información de la base. Además, sirve de interfaz entre las peticiones del usuario y la BD. Las preguntas a la base pueden tener distintas formas, pueden teclearse directamente desde la terminal, o codificarse como programas en lenguajes de alto nivel y presentarse para procesamiento interactivo o por lotes [5] [6].

3.5.3 Ventajas de un Sistema Manejador de Bases de Datos.

El empleo de un DBMS para gestionar los datos tiene muchas ventajas:

- Independencia con respecto a los datos. Ofrece una vista abstracta de los datos, ocultando los detalles de representación y almacenamiento de los datos.
- Acceso eficiente a los datos. Emplean técnicas sofisticadas para almacenar y recuperar los datos de manera eficiente, al igual que el almacenamiento en dispositivos externos.
- Integridad y seguridad de los datos. El acceso a los datos se controla por las restricciones de integridad, además del control de acceso a la información almacenada.
- Administración de los datos. Cuando existen diversos usuarios en un DBMS, y los datos se comparten la centralización de la administración de esos datos ofrece una mejora significativa, además de cuidar la redundancia de los datos y el mejoramiento de almacenamiento de los datos, para obtener una mejor recuperación de datos.
- Acceso concurrente y recuperación en caso de fallo. Un sistema que permita a varias estaciones de trabajo modificar en forma simultánea una misma base de datos, debe tomar precauciones para evitar operaciones concurrentes sobre un mismo registro. Esto es, si un usuario de una

estación de trabajo solicita el registro xxx para ser modificado, el sistema debe advertir a otro usuario que solicite el mismo registro xxx, que está siendo actualizado por otra estación de trabajo.

- Reducción del tiempo de desarrollo de las aplicaciones. El DBMS soporta muchas funciones importantes y aplicaciones que accedan a los datos. Estas aplicaciones cuentan con interfaces de alto nivel, facilitando el fácil acceso a los datos, las aplicaciones suelen ser robustas e independientes debido a su manejo de diferentes funciones [5] [6].

3.5.4 Herramientas de DBMS.

El éxito de los DBMS (https://www.w3schools.com/sql/sql_intro.asp) reside en mantener la seguridad e integridad de los datos. Lógicamente tiene que proporcionar herramientas a los distintos usuarios. Entre las herramientas que proporciona están:

- Herramientas para la creación y especificación de los datos. Así como de la estructura de la base de datos.
- Herramientas para administrar y crear la estructura física requerida en las unidades de almacenamiento.
- Herramientas para la manipulación de los datos de las bases de datos para añadir, modificar, suprimir o consultar datos.
- Herramientas de recuperación en caso de desastre.
- Herramientas para la creación de copias de seguridad.
- Herramientas para la gestión de la comunicación de la base de datos.
- Herramientas para la creación de aplicaciones que utilicen esquemas externos de los datos.
- Herramientas de instalación de la base de datos.
- Herramientas para la exportación e importación de datos.

3.5.5 SQL

Lenguaje de Consulta Estructurado o Structured Query Language (SQL) es un lenguaje de programación diseñado específicamente para el acceso a sistemas de administración a bases de datos relacionales (DBMSR). Actualmente los sistemas son de este tipo utilizando el lenguaje SQL, se puede decir sin ninguna duda, que este lenguaje es empleado mayoritariamente a sistemas existentes hoy en día e indiscutiblemente no tiene rival alguno. Este lenguaje es empleado en sistemas informáticos que van desde ordenadores personales muy básicos hasta los más potentes multiprocesadores y multicomputadores con decenas de procesadores. El lenguaje SQL es un lenguaje de cuarta generación. Es decir, en este lenguaje se indica que información se desea obtener o procesar, pero no como se debe hacer. Es labor interna del sistema elegir la forma más eficiente de llevar a cabo la operación ordenada por el usuario (https://www.w3schools.com/sql/sql_intro.asp).

3.5.6 SQLite

SQLite es un proyecto de dominio público creado por D. Richard Hipp, proyecto que trabaja con SQL y manejar bases de datos, lo que lo hace familiar y fácil de usar para una amplia gama de desarrolladores. SQLite, por otro lado, no es un RDBMS (Relational DataBase Management System, Sistema de gestión de bases de datos relacionales) completo, además se halla contenida en una

pequeña biblioteca escrita en C. El lenguaje de programación Python tiene incluida una librería que trabaja con este gestor de base de datos, haciendo muy sencilla su integración con el Framework Scrapy.

3.6 Inteligencia Artificial

La IA se ha estudiado durante décadas y sigue siendo uno de los temas más complejos de abordar en la informática. La IA consiste en simular la inteligencia humana en máquinas programadas para pensar como los humanos e imitar sus acciones. Tiene aplicaciones en casi todos los ámbitos en los que utilizamos los ordenadores en la sociedad. A medida que la tecnología avanza, la investigación en el campo de la IA también crece, por lo que los anteriores puntos de referencia que definen la IA se están quedando obsoletos. Debido a estos avances, surgieron nuevos términos como Aprendizaje Automático, del inglés Machine Learning (ML) y el Aprendizaje Profundo, del inglés Deep Learning (DL). Pero a veces, hay solapamientos entre la IA, el ML y el DL, por lo que la diferencia entre ellos puede ser muy poco clara, en la figura 3 se puede observar una representación de la relación entre los términos que engloba la inteligencia artificial [7] (https://towardsdatascience.com/artificial-intelligence-machine-learning-and-deep-learning-what-the-difference-8b6367dad790).

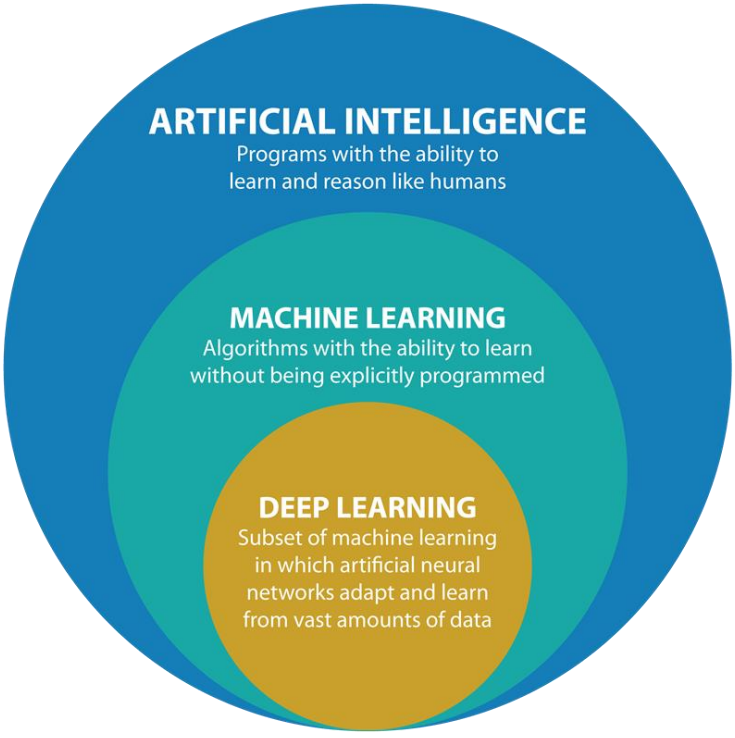


Figura 3: clasificador de inteligencia artificial, machine learning y deep learning (fuente: https://rafneta.github.io/CienciaDatosPythonCIDE/Laboratorios/Lab13/ML.html)

En sus inicios, los enfoques de la IA eran la lógica formal y los sistemas expertos. Estos métodos dominaban la IA en aquella época. Sin embargo, a raíz del desarrollo de la potencia de cálculo de los ordenadores, un mayor énfasis en la solución de problemas específicos, y también nuevos vínculos entre la IA y otros campos, surge un nuevo aprendizaje: el Aprendizaje Automático o ML. En este sentido, ML proporciona a las maquinas la capacidad de aprender y mejorar automáticamente, basándose en la experiencia. Estos sistemas transforman los datos en conocimiento o insights con el

objetivo de mejorar la toma de decisiones en prácticamente cualquier ámbito. El ML está estrechamente relacionado con la estadística computacional, que se centra en hacer predicciones. La minería de datos también está relacionada con este estudio, enfocada más en el análisis exploratorio de datos. Debido a los grandes avances estos últimos años en ML, se ha podido aplicar estas técnicas para mejorar el desempeño en diversas áreas de conocimiento, como por ejemplo en el reconocimiento de imágenes, llamada Visión por Computador (del inglés Computer Vision) o en el campo del PLN (área en la que se centrará este trabajo) entre otros.

3.7 Procesamiento del Lenguaje Natural

El Procesamiento del Lenguaje Natural (PLN o NPL por sus siglas en inglés) es una vertiente de la IA y la Lingüística dedicada a la comprensión por parte de los ordenadores, de los enunciados o palabras escritas en lenguaje natural, con el objetivo de obtener conocimientos a partir de datos en formato de texto [7].

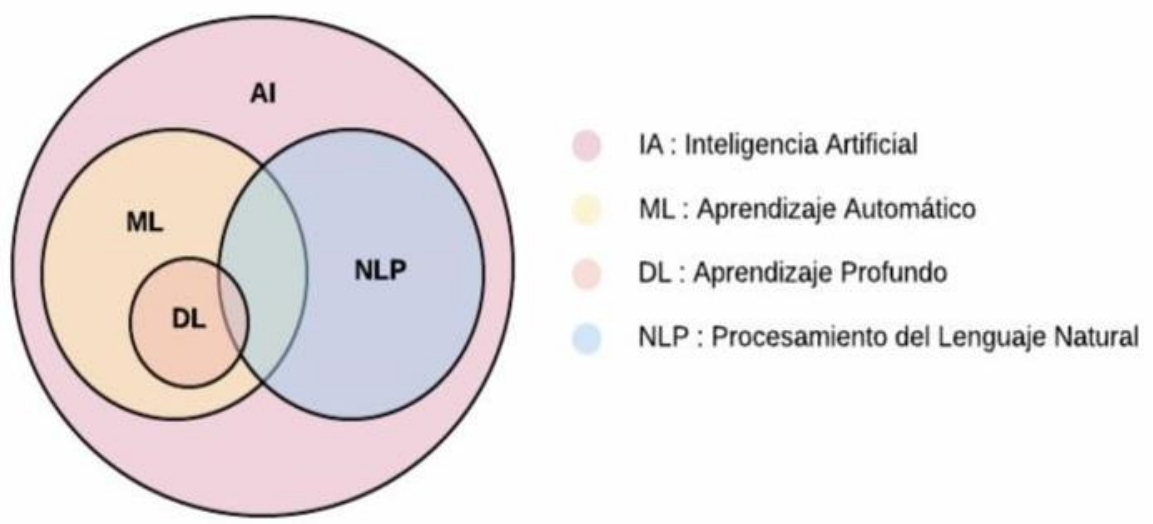


Figura 4: clasificador de inteligencia artificial, machine learning, deep learning y procesamiento del lenguaje natural. (fuente: https://blogs.iadb.org/conocimiento-abierto/es/aplicando-el-procesamiento-del-lenguaje-natural-para-clasificar-articulos-del-coronavirus/)

El PLN solapa con otros campos de estudio dentro de la IA como se puede observar en la figura 4.

La principal finalidad del PNL es recopilar conocimientos sobre el modo en que los seres humanos entienden y utilizan el lenguaje, de modo que puedan desarrollarse las herramientas y técnicas adecuadas para que los sistemas informáticos entiendan e interactúen utilizando los lenguajes naturales para realizar diversas tareas.

Durante las primeras décadas de historia del PLN las computadoras emulaban la comprensión de lenguaje natural aplicando una colección de reglas elaboradas de forma manual. El aumento en el poder de computación y aprendizaje favoreció el uso de métodos estadísticos y probabilísticos, consiguiendo una gran mejora en los resultados. Además, con el crecimiento de la web, smartphones, dispositivos IoT, cantidades ingentes de datos sin etiquetar empezaron a estar disponibles como datos de entrenamiento para modelos de aprendizaje semi y no supervisado. En el presente, gracias

al DL se están produciendo grandes avances en el PLN con la aparición de modelos pre-entrenados como BERT de Google. El PLN se puede clasificar básicamente en dos grupos: Comprensión del Lenguaje Natural y Generación del Lenguaje Natural, que desarrolla la tarea de comprender y generar el texto [7] (https://towardsdatascience.com/artificial-intelligence-machine-learning-and-deep-learning-what-the-difference-8b6367dad790).

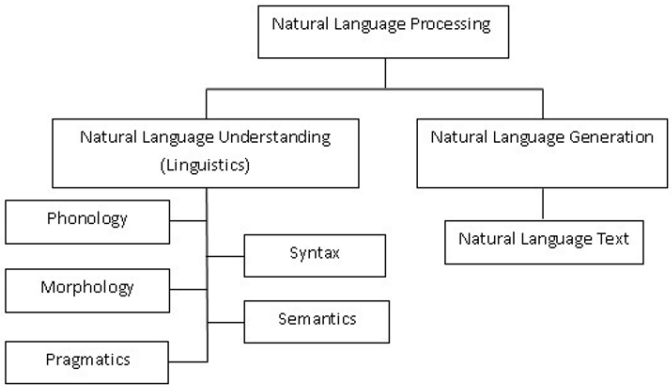


Figura 5: tipos y componentes del procesamiento del lenguaje natural (fuente: https://medium.com/dataman-in-ai/natural-language-processing-nlp-for-electronic-health-record-ehr-part-i-4cb1d4c2f24b)

La lingüística es la ciencia del lenguaje que estudia la fonología (sonido), la morfología (formación de la palabra), la sintaxis (la estructura de la oración), la semántica (significado) y la pragmática (comprensión).

Se puede definir el lenguaje como un conjunto, potencialmente infinito, de oraciones y secuencias de palabras construidas mediante las reglas gramaticales (en sus distintos niveles). Atendiendo su propósito y forma en que se originó, se pueden clasificar los lenguajes en dos grandes grupos: Lenguaje natural y Lenguaje formal.

Noah Chomsky uno de los primeros lingüistas del siglo XX, que inició las teorías sintácticas, marcó una posición única en el campo de la lingüística teórica ya que revolucionó el área de la sintaxis. Dicha área se puede clasificar a grandes rasgos en dos niveles Nivel Superior, que incluye el reconocimiento del habla, y Nivel Inferior, que corresponde al lenguaje natural [8].

Se puede definir la lengua natural como la lengua o idioma que nace espontáneamente de un grupo de hablantes por la mera necesidad de establecer comunicación verbal, ejemplos de lengua natural pueden ser el español, francés, inglés, etc. Está ligado a la cultura de cada civilización y evoluciona según su uso por parte de su comunidad de hablantes, que debe ponerse de acuerdo en la manera de usar el lenguaje para que cualquier hablante del mismo pueda interpretar y producir mensajes con el mismo sentido en que se originó. Es muy importante enmarcar el ámbito de empleo de la lengua natural en la comunicación humana. Por el contrario, los lenguajes formales, difieren de las naturales en que no han surgido espontáneamente, sino que han sido diseñados para un ámbito de aplicación concreto (normalmente en las ciencias) y se definen de manera que sean precisos y libres de cualquier ambigüedad. Ejemplos de lenguas formales son:

- Lenguaje matemático y lógico
- Lenguajes de programación (C, Java, Fortran, ...)
- Lenguaje de la música

Algunas de las tareas investigadas de la PLN son la Resumen Automático, el Análisis de Sentimiento, el Análisis del Discurso, la Traducción Automática, la Segmentación Morfológica, el Reconocimiento de Entidades Nombradas, el Reconocimiento Óptico de Caracteres (OCR), el Etiquetado de Partes del Lenguaje (del inglés Part Of Speech Tagging o POS Tagging), etc. Algunas de estas tareas tienen aplicaciones directas en el mundo real, como la Traducción Automática, el NER, el OCR, etc.

El campo del PLN está relacionado con diferentes teorías y técnicas que tratan de resolver los problemas de la comunicación entre los humanos y las máquinas. La ambigüedad es uno de los principales problemas del lenguaje natural que se suele presentar en el nivel sintáctico, el cual tiene como subtareas el léxico y la morfología, que se ocupan del estudio de las palabras y de su formación. Cada uno de estos niveles puede producir ambigüedades que pueden resolverse mediante el conocimiento de la frase completa.

La corrección de ambigüedades no es una tarea trivial y por tanto es necesario un tratamiento exhaustivo en todas las fases (léxico, estructural, pragmático…). La polisemia de las palabras, diferentes recursos literarios (anáforas, elipsis, hipérbaton...), acentos extranjeros, regionalismos, errores ortográficos y la intención o carga emocional de las sentencias (ironía, sarcasmo, etc.), son algunos de los elementos que hay que tener en consideración al analizar cualquier texto. En algunos casos son difíciles de identificar incluso para las personas, lo que dificulta aún más su formalización y su posterior interpretación por parte de las máquinas. Seguidamente se muestran algunos tipos de análisis que se pueden asociar a las técnicas de PLN (figura 6):

3.7.1 Análisis fonológico

La fonología es la parte de la Lingüística que se estudia la ordenación sistemática del sonido.

3.7.2 Análisis morfológico o “léxico”

Consiste en analizar las palabras que integran las sentencias para obtener sus lemas o raíces, unidades léxicas compuestas, rasgos flexivos, etc. La finalidad de cualquier análisis léxico o morfológico es examinar y dividir el texto que se va a analizar en una serie de componentes léxicos, también conocidos como tokens o símbolos, propios del lenguaje original en el que está escrito. En este sentido, hay que prestar especial atención a los separadores de los componentes léxicos definidos por la lengua en cuestión. Por ejemplo, en lenguajes formales como pueden ser los lenguajes de programación, los separadores léxicos pueden ir desde los espacios en blanco hasta los símbolos no representables, como los saltos de línea. En el caso de las lenguas naturales, los separadores léxicos suelen limitarse a los espacios en blanco y a los símbolos de puntuación. Otra de las tareas realizadas como parte del análisis léxico es el etiquetado morfológico. El etiquetado morfológico, también denominado POS tagging, permite asignar etiquetas morfológicas a los elementos léxicos identificados según su categoría gramatical. Estas etiquetas permiten denotar el tipo (artículo, sustantivo, verbo, adjetivo, adverbio, etc.), género, número, tiempo o modo de cada una de las palabras en cuestión.

3.7.3 Análisis sintáctico

La finalidad de un analizador sintáctico (del inglés parser) es determinar si una secuencia de componentes léxicos o tokens se ajusta a una estructura gramatical, es decir, comprueba que una frase dada está bien construida según las reglas gramaticales del lenguaje analizado.

3.7.4 Análisis semántico

El objetivo del análisis semántico es determinar, de forma inequívoca, el significado de las frases.

3.7.5 Análisis del discurso

Si bien la sintaxis y la semántica trabajan con unidades de longitud de frase, el nivel de discurso del PNL se centra en las unidades de texto más largas que una frase, es decir, no identifica los textos de varias frases como simples secuencias de frases. En su lugar, el discurso se centra en las propiedades del texto en su conjunto que trasmiten un significado al relacionar las conexiones entre las sentencias que lo componen.

3.7.6 Análisis pragmático

Se centra en el análisis del contexto donde se encuentra inmerso el texto analizado y en cómo éste influye en el significado del texto.

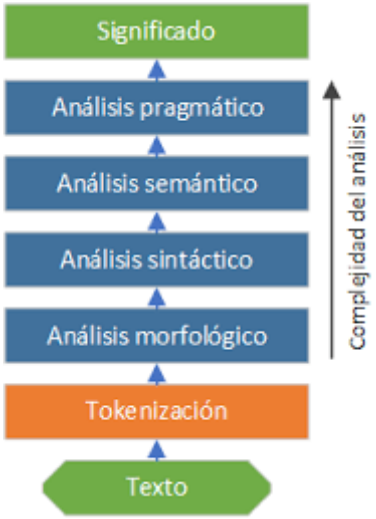


Figura 6: tipos de procesamiento del lenguaje natural (fuente: https://openaccess.uoc.edu/bitstream/10609/81435/6/jsobrinosTFM0618memoria.pdf)

3.7.7 Principales Tareas y Aplicaciones

El PLN agrupa diversas técnicas para interpretar el lenguaje humano, desde métodos estadísticos y de ML hasta los enfoques basados en reglas y algorítmicos.

Las tareas básicas de PLN implican la simbolización y el análisis sintáctico, lematización/derivación, POS, detección del lenguaje e identificación de relaciones semánticas. En general, las tareas de PLN dividen el lenguaje en trozos o piezas elementales más cortas o tokens, con la idea de comprender las relaciones entre dichos tokens y explorar cómo funcionan estas piezas juntas para crear

significado. Dichas tareas implícitas se emplean a menudo en recursos PLN de más alto nivel como [7] (https://www.sas.com/es_ar/insights/analytics/what-is-natural-language-processing-nlp.html):

- Categorización de contenido: Un resumen del contenido del documento basado en la lingüística, incluyendo búsqueda e indización, alertas de contenido y detección de duplicación. Un ejemplo claro de la categorización de contenidos son los filtros de spam.
- Descubrimiento y modelado de temas o del inglés Topic Modeling: Analiza cómo las palabras y frases se relacionan entre sí y automáticamente "selecciona o aprende" grupos de palabras que mejor caracterizan esos documentos. Estos conjuntos de palabras representan un tema o un tópico. Con la introducción de la minería de texto, se han realizado investigaciones para analizar temas y tendencias importantes en la recopilación de documentos. El uso de la Asignación Latente de Dirichlet (en inglés LDA) o la Descomposición en Valores Singulares (en inglés SVD) para el análisis de tendencias en la minería de texto es uno de los métodos de análisis de tendencias más precisos.
- Extracción contextual: Extrae de forma automática datos estructurados de fuentes basadas en texto.
- Clasificación de texto: La clasificación es una las tareas básicas en el análisis de texto y se utiliza ampliamente en una variedad de dominios y aplicaciones. La premisa de la clasificación es sencilla: dada una variable categórica objetivo, la finalidad es extraer los patrones existentes entre las instancias compuestas por variables independientes y su relación con el objetivo. Dado que el objetivo se da por adelantado, se dice que la clasificación es supervisada ya que se entrena un modelo para minimizar el error entre las categorías predichas y las reales en los datos de entrenamiento. Una vez que el modelo de clasificación se ajusta, se asignan las etiquetas categóricas a las nuevas instancias basándose en los patrones detectados durante el entrenamiento.
- Análisis de sentimiento: Identificación de opiniones subjetivas y emociones en grandes volúmenes de texto, incluyendo minería de sentimiento. En la práctica, el Análisis de Sentimientos (AS) es una Clasificación de Texto. Esta última área está muy de moda hoy en día, ya que tiene diversas aplicaciones, como, por ejemplo: medición de la satisfacción del cliente, el sentimiento hacia un producto o la identificación de sentimientos hacia un tema específico en las Redes Sociales, etc.
- Conversión de habla a texto y de texto a habla o del inglés Speech Recognition. Transformación de comandos de voz en texto escrito y viceversa.
- Resumen de documentos: Generación automática de resúmenes de grandes volúmenes de documentos.
- Traducción automática: Traducción automática de texto o habla de un idioma a otro.
- Reconocimiento de Entidades Nombradas o NER: El NER tiene como objetivo localizar y clasificar en categorías predefinidas, como, organizaciones, personas, lugares y cantidades, las entidades nombradas encontradas en un texto dado.
- ChatBots.
- Interfaces en lenguaje natural.
- Reconocimiento óptico de caracteres, del inglés Optical Character Recognition (OCR).

En todos estos casos, la finalidad es convertir el texto en crudo del lenguaje y aplicar técnicas de lingüística computacional y algoritmos para transformar o enriquecer el texto para extraer conocimientos o insights.

3.7.8 Técnicas y Herramientas de Preprocesamiento

En esta sección se presentan las técnicas de preprocesamiento de PLN. Estás técnicas consisten en una serie de tareas enfocadas a preparar o a limpiar el corpus original (datos en bruto) para las diferentes tareas que ocupa la minería de texto. Al conjunto o colección de palabras en forma de texto no estructurado utilizado para entrenar los algoritmos de ML se le denomina corpus. En el contexto del PLN los siguientes términos se utilizan habitualmente y es importante su comprensión (https://www.oreilly.com/library/view/blueprints-for-text/9781492074076/ch04.html y https://www.analyticsvidhya.com/blog/2020/05/what-is-tokenization-nlp/):

- Corpus: Se denomina corpus al conjunto de documentos sobre el que se realiza el análisis. El plural del corpus es corpora, que también es su derivación latina, que significa "cuerpo". Cuando el corpus está etiquetado y estructurado adecuadamente, se denomina corpus etiquetado.
- "Token": Conjunto de fragmentos o trozos de caracteres que representa la mínima unidad en el análisis de texto.
- Documento: La representación escrita de una idea, concepto o diálogo se le denomina documento. Un documento está compuesto por varios tokens. Ejemplos de documentos son un tweet, un artículo de una publicación científica, etc.
- Vocabulario: El conjunto de tokens únicos que se obtienen como resultado al "tokenizar" nuestro corpus completo.

Como se ha mencionado anteriormente, el preprocesamiento de datos consiste en una serie de pasos que pueden aplicarse (o no) a una tarea determinada, pero que generalmente se engloban en las categorías generales de tokenization, normalización y sustitución.

En términos generales, el principal objetivo será obtener un corpus de texto bruto predeterminado y realizar sobre él algunos análisis y transformaciones básicas con el fin de obtener características que sean mucho más útiles para realizar alguna tarea analítica posterior más significativa. Así pues, como se ha mencionado anteriormente, se tres componentes principales en el preprocesamiento de textos:

- Separar palabras del texto en entidades denominadas tokens (Tokenization)
- Normalización
- Sustitución o noise removal

Al establecer un marco para abordar el preprocesamiento, se debe tener presentes estos conceptos de alto nivel.

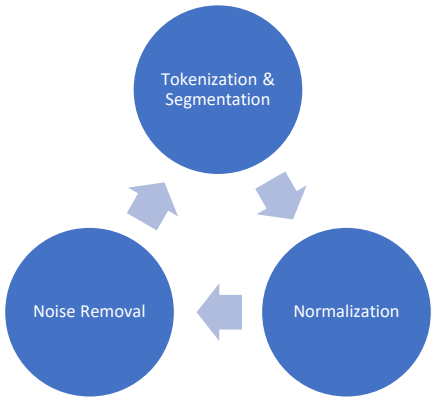


Figura 7: etapas del procesamiento del lenguaje natural

En las siguientes secciones, se describen de forma más detallada las diferentes etapas que componen en el preprocesamiento.

Tokenization

La "tokenización" o tokenization es proceso de separar las cadenas de texto más largas en trozos más pequeños o tokens. Los trozos de texto más grandes pueden dividirse en frases, las frases pueden dividirse en palabras, etc. La manera de seleccionar los tokens del texto depende de la naturaleza del problema que se afronte, en algunas ocasiones una simple "tokenización", como la de separar el texto por los espacios, puede ser suficiente.

El procesamiento posterior se realiza generalmente después de que un trozo haya sido debidamente "tokenizado". El proceso de "tokenización" también es conocido como segmentación del texto o análisis léxico. A menudo, la segmentación se emplea para hacer referencia al desglose de un trozo de texto en piezas más grandes que las palabras (por ejemplo, párrafo u oraciones), mientras que la "tokenización" se limita al proceso de desglose que resulta exclusivamente en palabras. Este proceso de división en pequeños trozos o palabras no es trivial, ya que, dependiendo del separador utilizado para dividir las palabras de una frase, se tendrá un resultado u otro. Por ejemplo, se podría emplear una estrategia de segmentación que identifique (correctamente) un límite particular entre los tokens de la palabra como el apóstrofe en la palabra she's (una estrategia que "tokenice" sólo los espacios en blanco no serían suficiente para reconocer esto). Por lo tanto, dependiendo del idioma en el que se esté trabajando, el proceso de "tokenización" será diferente.

Limpieza de Ruido

Mientras que los procesos de "tokenización" y de normalización se aplican generalmente a casi cualquier trozo de texto o token, la eliminación de ruido es una sección mucho más específica. La eliminación de ruido es una tarea de normalización específica del texto que suele producirse antes de la "tokenización". No es un proceso lineal, cuyos pasos deban aplicarse exclusivamente en un orden determinado. Por lo tanto, la eliminación del ruido puede producirse antes o después de las secciones anteriormente descritas o en algún punto intermedio (https://www.kdnuggets.com/2019/04/text-preprocessing-nlp-machine-learning.html).

Por ejemplo, si obtiene un corpus de internet alojado en un formato web sin procesar, seguramente el corpus contendrá texto envuelto en etiquetas HTML o XML.

Algunas de las tareas más habituales en la eliminación de ruido son:

- Eliminar los encabezados y pies de página de los archivos de texto.
- Eliminar marcas y metadatos de lenguajes como HTML, XML, etc.
- Extraer datos valiosos de otros formatos, como JSON o de bases de datos.
- Mapeo de contracciones/contracciones en expansión: Las contracciones son una versión abreviada de palabras o un grupo de palabras, bastante común tanto en el lenguaje hablado como escrito. En inglés, y sobre todo en las redes sociales, son bastante comunes las contracciones "I'm", "didn't", "haven't", etc. El mapeo de estas contracciones a su forma expandida ayuda en la estandarización de texto.
- Eliminación de signos de puntuación: Los signos de puntuación y caracteres especiales introducen ruido en el corpus, ya que en la mayoría de los mensajes no son relevantes a la hora de determinar la polaridad del mensaje.
- Eliminación números, o en su defecto, conversión de números a texto.
- Eliminación menciones, enlaces y hashtags: Estos tres elementos son específicos de sitios web o redes sociales como la red social Twitter. Las menciones hacen referencia a otros usuarios mediante su nombre de usuario (cuenta en Twitter) precedido del símbolo "@". Los hashtags son cadenas de texto con un significado asociado precedida del símbolo "#". Finalmente, los usuarios pueden añadir enlaces a sus mensajes. Debido a la gran variabilidad en el uso de estos elementos por parte de los usuarios en las redes sociales, se hace necesario realizar un tratamiento específico, ya sea sustituyendo estos elementos por espacios en blanco o por algún tipo de carácter. Para llevar a cabo estas operaciones, suele emplearse de expresiones regulares (en este trabajo se ha empleado el módulo re de Python).

Normalización del texto

La normalización suele referirse a una serie de tares relacionadas entre sí con el objetivo de equiparar todo el texto, como, por ejemplo: conversión de minúsculas o mayúsculas, eliminación de los signos de puntuación, conversión de números a sus equivalentes en palabras, etc. La normalización establece todas las palabras en igualdad de condiciones y permite que el preprocesamiento sea uniforme. Los caracteres y símbolos especiales contribuyen e la generación de ruido adicional en el texto no estructurado. Se recomienda usar expresiones regulares para eliminarlos como paso previo. A continuación, se lista los principales métodos de normalización del texto (https://www.kdnuggets.com/2019/04/text-preprocessing-nlp-machine-learning.html) [9]:

- Revisión ortográfica: Los documentos de un corpus son propensos a errores ortográficos. Realizar una limpieza del texto es una buena práctica y para ello se puede ejecutar un corrector ortográfico y corregir los errores ortográficos antes de continuar con los siguientes pasos. En determinados corpus, como puede ser el obtenido de una red social, no siempre este paso da buenos resultados ya que esta corrección puede introducir ruido adicional al tener que lidiar con expresiones propias de las redes sociales como 'omg' ('oh my god') o 'lol' ('laughing out loud').
- Radicación o Stemming: Se llama stemming al proceso de conversión de una palabra a su raíz, es decir, permite eliminar afijos (sufijos, prefijos, infijos…) de una palabra para obtener la raíz de la misma. Estas raíces son la parte invariable de palabras relacionadas sobre todo por su forma. De cierta manera se parece a la lematización (como se verá posteriormente), pero los resultados (las

raíces) no tienen por qué ser palabras de un idioma o valores lingüísticos válidos. Por ejemplo, el algoritmo de stemming puede decidir que la raíz de amamos no es “am-” (la raíz) sino “amam-" (cosa que desconcertaría a más de uno). Desde el punto de vista del procesamiento, el stemming es mucho más rápido que la lematización. Además, tiene como ventaja el reconocimiento de relaciones entre palabras de distinta clase. Podría reconocer, por ejemplo, que picante y picar tienen como raíz pic-. En definitiva, el stemming puede reducir el número de elementos que componen nuestras oraciones. La principal ventaja del stemming es que su implantación es más simple que la lematización. Como principal desventaja es que pueden “recortar” demasiado la raíz y encontrar relaciones entre palabras que realmente no existen (overstemming). También puede suceder que deje raíces demasiado extensas o específicas, y que se tenga más bien un déficit de raíces (understemming), en cuyo caso palabras que deberían convertirse en una misma raíz no lo hacen.

- Lematización o Lemmatization: La lematización está relacionada con la derivación, con la diferencia de que la lematización es capaz de capturar formas canónicas basadas en el lema de una palabra. Por ejemplo, la palabra “better” daría como resultado lo siguiente: o Better → Good o Was → Be o Studies → Study. La lematización también ayuda a emparejar sinónimos mediante el uso de un diccionario de sinónimos, de modo que cuando uno busca "hot” (caliente) la palabra “warm” (cálido) también coincide. De la misma manera, una búsqueda de car producirá tanto cars como “automobile”. La técnica de lematización se ha utilizado en varios idiomas para la recuperación de información. Este proceso es más costoso computacionalmente que la radicación o stemming, pero suele dar mejores resultados ya que reduce las formas de las palabras a lemas válidos lingüísticamente hablando.

Otros procesos de normalización de texto:

- Palabras vacías o Stopwords: Las stopwords son aquellas palabras que normalmente se filtran antes de seguir procesando el texto (sobre todo en tareas como el Análisis de n-gramas o Topic Modeling), ya que estas palabras contribuyen poco al significado global, dado que suelen ser las más comunes en un idioma. Existen diversas librerías, como por ejemplo NLTK o Spacy que incorporan una lista bastante amplia de stopwords. Por ejemplo, "el", "y" y "a", aunque son palabras necesarias en un texto concreto no suelen contribuir en gran medida a la comprensión del contenido. Como ejemplo sencillo, el siguiente texto es igual de legible si se eliminan (palabras tachadas) las stopwords: “The quick brown fox jumps over the lazy dog” (“El rápido zorro marrón salta sobre el perro perezoso.”)
- Convertir todos los caracteres en minúsculas. Aunque es fácil identificar que “casa” y “CASA” tienen el mismo significado, los algoritmos de ML las procesan como palabras diferentes. Para evitar esto, todos los textos del corpus serán convertidos a su equivalente en letras minúsculas.
- Tratamiento de la duplicidad de caracteres: Es común, sobre todo en las redes sociales, la repetición de caracteres para dar intensidad a lo que se intenta explicar. Es importante realizar un tratamiento en estos casos con el objetivo de establecer relaciones entre términos que realmente son iguales, como, por ejemplo: “caloooor”, “calor”, “caaaaaaaaalor”, etc., son la misma palabra: ‘calor’.
- Eliminar los espacios en blanco, normalización de jerga, conversión de emojis/emoticonos a texto, etc.

La gran mayoría de estas operaciones se pueden ejecutar con las librerías de Python de expresiones regulares (re), NLTK y Spacy. Gracias al preprocesamiento del texto, la dimensionalidad del conjunto de datos de entrenamiento se verá reducida y el Análisis de n-gramas o el Topic Modeling se verán

beneficiados, ya que el objetivo principal es encontrar términos similares en el corpus para que el desempeño de estas tareas sea mejor, tanto en rendimiento como en tiempo.

3.8 Herramientas y Librerías

A continuación, se listan los principales herramientas y librerías en el lenguaje Python para desarrollar PLN (https://www.cnnchile.com/):

3.8.1 Spacy

Se trata de una biblioteca programada en Python cuyo objetivo es el procesamiento avanzado de lenguaje natural. Proporciona pipelines pre-entrenados y tiene soporte para más de 60 idiomas. Permite realizar etiquetados, "tokenización", análisis o reconocimiento de entidades con nombre (NER) entre otras funcionalidades. En la última actualización se han añadido soporte para trabajar con transfomers, un modelo de DL de vanguardia que utiliza un mecanismo de atención (una técnica que imita la atención cognitiva), sopesando la influencia de diferentes partes de los datos de entrada. Se utiliza principalmente en PNL, pero investigaciones recientes también han desarrollado su aplicación en otras tareas como la comprensión de videos. Para poder utilizarlo en español basta con añadir el modelo para este idioma desde un programa Python (https://spacy.io/usage).

3.8.2 NLTK

Es un conjunto de bibliotecas y programas de PLN, en Python. Cuenta con distintas interfaces para poder trabajar con más de 50 corpus y recursos léxicos que proporciona. Está diseñado para utilizarse principalmente en inglés, pero tiene soporte para varios idiomas bien entrenándolo con archivos que incorpora el propio NLTK (https://www.nltk.org/).

3.8.3 Stanza (The Standford CoreNLP):

Anteriormente conocido como Stanford NLP, es una biblioteca desarrollada en Java que ofrece diversas herramientas para tareas de PLN, así como el acceso a CoreNLP, una librería que proporciona tuberías o pipelines con las que poder realizar operaciones sobre el texto de una forma más automatizada ("concatenando operaciones o métodos") (https://stanfordnlp.github.io/CoreNLP/).

3.8.4 Scikit-Learn

Librería de Python que proporciona métodos de vectorización (conversión de texto a número), como Count Vectorizer o TF-IDF (https://scikit-learn.org/stable/).

3.8.5 Gensim

Es una biblioteca de Python para modelado de temas, indexación de documentos y recuperación de similitudes con grandes corpora. Además, incorpora implementaciones de otros algoritmos como Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), Random Projections (RP), Hierarchical Dirichlet Process (HDP) or word2vec deep learning (https://radimrehurek.com/gensim/).

3.9 Aprendizaje Automático

En secciones anteriores se describieron de forma general la IA y el PLN respectivamente. En esta sección se abordará el ML, repasando brevemente algunos conceptos básicos para posteriormente introducir conceptos clave en el tratamiento de texto previo al entrenamiento de los algoritmos de ML, como la extracción de características, la ponderación de características y la reducción o selección de características.

3.9.1 Métodos de Aprendizaje

Como se aprecia en la siguiente figura, existen básicamente tres grandes grupos de ML:

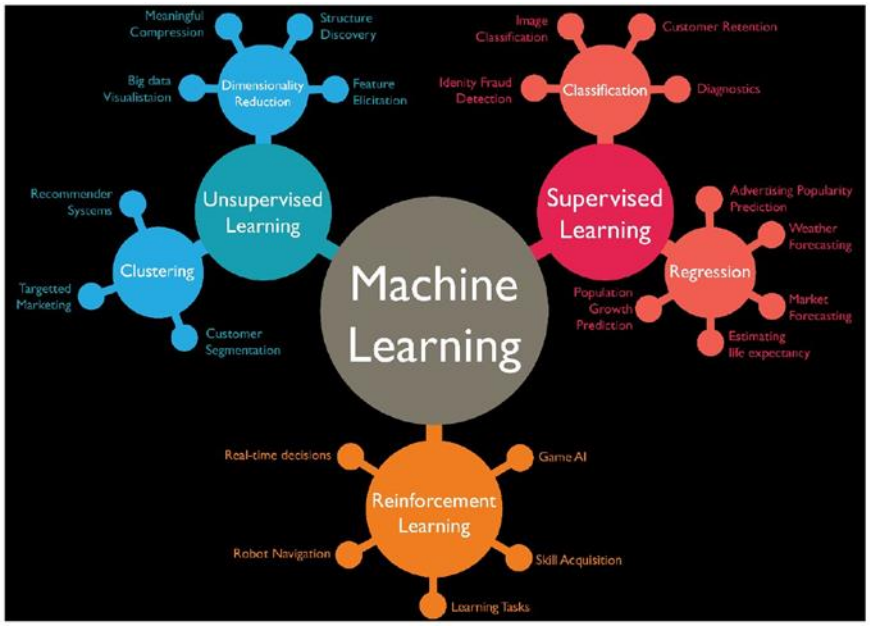


Figura 8: tipo de algoritmos de clasificación en machine learning (fuente: https://www.linkedin.com/pulse/machine-learning-strong-use-case-access-edge-networking-kohli/?trk=public_profile_article_view)

Aprendizaje supervisado: En este tipo de aprendizaje los datos de entrada para el entrenamiento de los algoritmos han sido previamente etiquetados, por lo que existe un conocimiento a priori.
Aprendizaje no supervisado. Este tipo de aprendizaje, por el contrario, los datos de entrada para el entrenamiento no están etiquetados, es decir, no hay un conocimiento a priori (en el caso de clasificación, sin necesidad de estar la muestra etiquetada con su clase).
Aprendizaje por refuerzo. El objetivo de este tipo de aprendizaje es la generalización o extrapolación de situaciones que no estén presentes en los datos de entrenamiento. Para ello, se enfocan en aumentar la señal de recompensa (prueba y error).

3.9.2 Extracción de características

A partir de la etapa de preprocesamiento, los textos que componen el corpus no son computables por los algoritmos de ML (los algoritmos trabajan con datos numéricos), por lo que deben transformarse en datos numéricos, como por ejemplo el modelo de espacio vectorial. Esta transformación se denomina generalmente extracción de características de los documentos. La

extracción de características tiene dos métodos principales: la Bolsa de Palabras o BOW y la incrustación de palabras o Word Embbedings. Ambos son comúnmente utilizados y tienen diferentes enfoques.

3.9.3 Incrustación de Palabras

La incrustación de palabras o Word Embbedings (https://medium.com/@eiki1212/feature-extraction-in-natural-language-processing-with-python-59c7cdcaf064) son una de las posibles representaciones de documentos de textos en el modelo de espacio vectorial. Captura parte del contexto y la semántica de las palabras, a diferencia del modelo BOW. La BOW sólo representa el número de palabras que aparecen en el documento (frecuencia) sin ninguna relación ni contexto. Por otro lado, los Word Embbedings conserva parte del contexto y las relaciones de las palabras, de modo que identifica las palabras similares con mayor precisión. Estos modelos generan incrustaciones que son independientes del contexto, es decir, sólo hay una representación vectorial (numérica) para cada palabra. Los diferentes sentidos de la palabra (si los hay) se combinan en un único vector.

3.9.4 Bolsa de Palabras

Se define Bolsa de Palabras o Bag of Words (BOW) (https://medium.com/@eiki1212/feature-extraction-in-natural-language-processing-with-python-59c7cdcaf064) al conjunto de palabras existente en todo el corpus. Todas las palabras se disponen en una "bolsa" donde se mezclan. La disposición original en el texto se pierde y solo se tiene en cuenta la frecuencia de los términos. En este caso se recomienda eliminar aquellas palabras con poca frecuencia de distribución dentro del corpus (si esté es muy grande), con el objetivo de reducir el vocabulario, mejorando el rendimiento y el tiempo de ejecución del entrenamiento de los modelos, así como reducir el posible sobreajuste (overfitting) del modelo. Hay que diferenciar los "modelos de representación" de datos como BOW de los "métodos de cálculo o vectorización" para ponderar la importancia de las palabras en el documento, como por ejemplo la Frecuencia de Términos o Term Frecuency (TF) o la Frecuencia de Términos-Frecuencia de Documentos Inversa o Term FrequencyInverse Document Frequency (TF-IDF) que se describirán más adelante. A continuación, se muestra un ejemplo partiendo de tres documentos (que en conjunto forman un corpus):

- "Fui a comer tacos de suadero. Juro que es el suadero más delicioso de mi vida. #suadero"
- "Taco de delicioso suadero con bolsa de plástico para que no ensucie el plato."
- "Tengo ganas de comprarme unos tacos de fútbol, ir a la cancha y jugar hasta la noche"

Una vez se tenga "tokenizado" (eliminando stopwords, signos de puntuación, etc.) cada documento se tiene el siguiente vocabulario: bolsa, cancha, comer, comprarme, delicioso, ensuciar, fútbol, ganar, jugar, juro, noche, plato, plástico, suadero, taco y vida.

En la siguiente figura se muestra la BOW del corpus (tres documentos). Cada fila es un documento y cada columna un token:

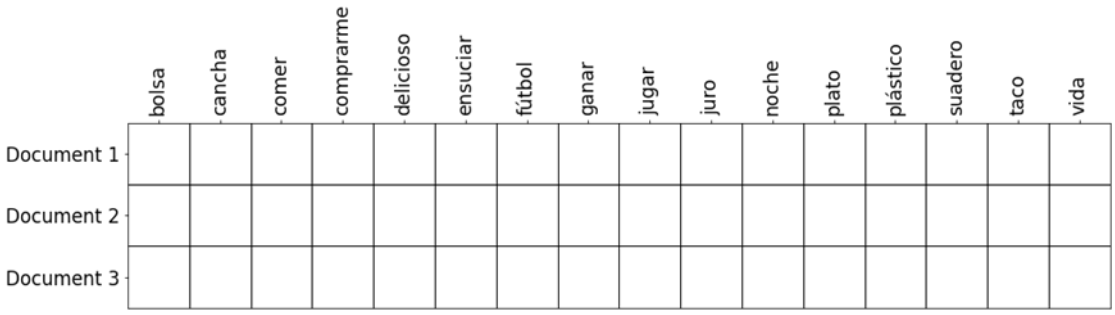


Figura 9: BOW de palabras (fuente: https://old.tacosdedatos.com/texto-vectores)

3.9.5 Bolsa de n-gramas (n-BOW)

Según la Wikipedia, "un n-grama es una secuencia continua de n elementos de una determinada secuencia de texto o discurso". En otras palabras, los n-gramas (https://towardsdatascience.com/another-twitter-sentiment-analysis-with-python-part-4-count-vectorizer-b3f4944e51b5) son simplemente todas las combinaciones de palabras o letras adyacentes de longitud n que se pueden encontrar en el texto fuente. La Figura 2.12 representa bien cómo se construyen los n-gramas a partir de un texto fuente.

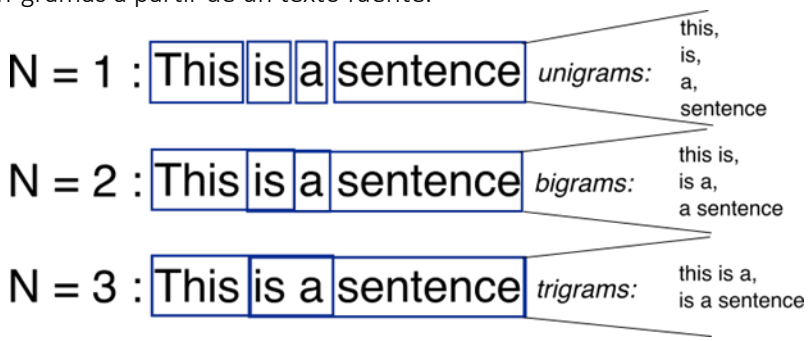


Figura 10: definición de n gramas (fuente: https://deepai.org/machine-learning-glossary-and-terms/n-gram)

Por ejemplo, la frase “el cielo es azul”, los 1-grama serían: “el”, “cielo”, ‘es, “azul”. Un 2grama (o “bigrama”) es una secuencia de dos palabras, siguiendo el ejemplo anterior se tiene: "el cielo", "cielo es", "es azul". y un 3-grama (o “trigrama”) es una secuencia de tres palabras: "el cielo es" y "cielo es azul". Este método en sí mismo no es una representación de un texto, pero puede utilizarse como característica para representar un texto. Extendiendo este concepto para n, se pueden crear modelos tan complejos como se quiera. Sin embargo, este tipo de método puede ser insuficiente en determinadas tareas ya que en el lenguaje hay palabras que dependen de otras que están bastante más atrás en la misma oración.

La BOW es una representación de un texto utilizando un 1-grama, perdiendo el orden original del texto. Es muy fácil de obtener y el texto puede representarse mediante un vector, generalmente de tamaño manejable. Por lo tanto, se puede considerar los n-grama como características y un n-BOW como una representación de un texto utilizando los n-gramas que contiene. La mayoría de las veces, los “bigramas” y “trigramas” pueden capturar más información que un 1-grama. Por ejemplo, "la hamburguesa sabe mal" o "me gustan las hamburguesas" puede transmitir más información, que sólo "hamburguesa". El problema es que en un corpus hay muchos más “bigramas” que 1-gramas, y

es posible que muchos de ellos aparezcan una sola vez, por lo que no es útil comparar elementos entre ellos. Sin embargo, si se trata de clasificar un texto para conocer, por ejemplo, el gusto de la gente hacia una comida, puede que no sea suficiente con utilizar un 1-grama. Para clasificar el sentimiento, tal vez se podría representar un texto con BOW y 2-gramas ("bigramas"), donde una de las palabras es una apreciación de las personas hacia la comida.

3.9.6 Ponderación de características

Las características extraídas en las secciones precedentes pueden tratarse todas con el mismo peso o importancia u asignarles diferentes pesos (weighted) en función de algún tipo de criterio. Aunque existen diversos métodos de ponderación, en este trabajo se emplearán dos modelos muy populares en la Clasificación de Textos y que tienen su origen en el área de la Recuperación de la Información:
Frecuencia de Términos (TF)
Frecuencia de Términos – Frecuencia Inverso de Documento (TF-IDF)

Estos pesos establecen la relevancia de cada característica dentro del texto al que pertenecen y, por tanto, repercutirán en los resultados obtenidos en los entrenamientos de los algoritmos de ML.

3.9.7 Frecuencia de Términos

La frecuencia de términos o Term Frecuency (TF) (https://kumawatrohan.medium.com/tf-idf-nlp-series-part-6-9587db6fe916) se define como el valor de la frecuencia absoluta de ese término o palabra en un documento concreto. Cada característica tendrá un peso igual a su frecuencia absoluta en un documento.

TF es el número de veces que una palabra w aparece en un documento D. Cada documento tiene su propia frecuencia de términos.

A veces también se puede estandarizar o normalizar la frecuencia absoluta utilizando logaritmos o promediando la frecuencia. Partiendo del ejemplo de la sección anterior y utilizando la función Count Vectorizer de la librería Scikit-Learn, se obtiene la matriz de frecuencia de términos que se muestra a continuación:

	bolsa	cancha	comer	comprarme	delicioso	ensuciar	fútbol	ganar	jugar	juro	noche	plato	plástico	suadero	taco	vida
Document 1	0	0	1	0	1	0	0	0	0	1	0	0	0	3	1	1
Document 2	1	0	0	0	1	1	0	0	0	0	0	1	1	1	1	0
Document 3	0	1	0	1	0	0	1	1	1	0	1	0	0	0	1	0

Figura 11: aplicación de la clase CountVectorizer (fuente: https://old.tacosdedatos.com/texto-vectores)

3.9.8 Frecuencia de Términos - Frecuencia Inversa de Documento

Pueden existir algunos problemas potenciales con el modelo BOW cuando se utiliza en corpus grandes. En los modelos basados en TF (frecuencias absolutas de los términos), es posible que haya algunas palabras que aparezcan con frecuencia en todos los documentos y que éstos tiendan a eclipsar otros términos en el conjunto de características, sobre todo palabras que no aparecen con tanta frecuencia pero que podrían ser más interesantes y eficaces como características para identificar ciertas categorías. Term frequency - Inverse document frequency o TF-IDF son las siglas de Frecuencia de Términos - Frecuencia Inversa de Documento. Es una combinación de dos métricas, por un lado, la frecuencia de términos (TF) mencionada anteriormente y por otro la frecuencia inversa de documentos (IDF). Con este modelo, se compensa la frecuencia de cada palabra en el documento con la frecuencia de la palabra en la colección de documentos, lo que permite manejar el hecho de que ciertas palabras sean más comunes que otras. Este método se ideó originalmente como una métrica para clasificar los resultados de los motores de búsqueda en función de las consultas de los usuarios y se ha convertido en una pieza fundamental para de la recuperación de la información y la extracción de características de texto.

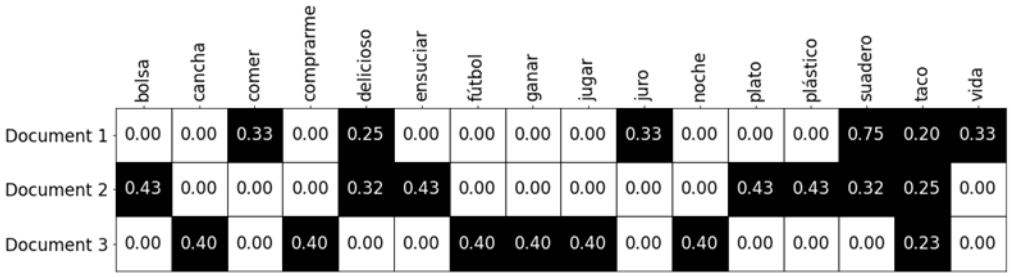
Dependiendo del paquete o librería utilizado, se puede encontrar un parámetro llamado “suavizado” o smooth, que consiste en sumar una cantidad (normalmente 1 ó 0.1) que ayudan a prevenir divisiones entre cero y términos que puedan tener un IDF igual a cero.

En la siguiente figura se muestra la matriz de términos TF-IDF para cada uno de los documentos que forman el corpus del ejemplo de las secciones anteriores:

	bolsa	cancha	comer	comprarme	delicioso	ensuciar	fútbol	ganar	jugar	juro	noche	plato	plástico	suadero	taco	vida
Document 1	0.00	0.00	1.69	0.00	1.29	0.00	0.00	0.00	0.00	1.69	0.00	0.00	0.00	3.86	1.00	1.69
Document 2	1.69	0.00	0.00	0.00	1.29	1.69	0.00	0.00	0.00	0.00	0.00	1.69	1.69	1.29	1.00	0.00
Document 3	0.00	1.69	0.00	1.69	0.00	0.00	1.69	1.69	1.69	0.00	1.69	0.00	0.00	0.00	1.00	0.00

Figura 12: matriz de términos TF-IDF (fuente: https://old.tacosdedatos.com/texto-vectores)

Los algoritmos de ML normalmente funcionan mejor cuando los datos de entrenamiento se encuentran normalizados (para que el entrenamiento del modelo sea menos sensible a la magnitud de las características). Esto permite que los modelos converjan mejor y más rápido logrando un modelo más preciso. La normalización hace que las características sean más coherentes entre sí, lo que permite al modelo predecir los resultados con mayor precisión. Por ello, se utilizará la función TF-IDF de la librería Scikit-Learn para normalizar (normalización euclídea). En la siguiente figura se muestra la Matriz TF-IF normalizada con las herramientas entregadas por la librería Scikit-Learn:



	bolsa	cancha	comer	comprarme	delicioso	ensuciar	fútbol	ganar	jugar	juro	noche	plato	plástico	suadero	taco	vida
Document 1	0.00	0.00	0.33	0.00	0.25	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.75	0.20	0.33
Document 2	0.43	0.00	0.00	0.00	0.32	0.43	0.00	0.00	0.00	0.00	0.00	0.43	0.43	0.32	0.25	0.00
Document 3	0.00	0.40	0.00	0.40	0.00	0.00	0.40	0.40	0.40	0.00	0.40	0.00	0.00	0.00	0.23	0.00

Figura 13: Matriz TF-IF normalizada (fuente: https://old.tacosdedatos.com/texto-vectores)

Como se observa en la figura anterior, en muchas ocasiones se tienen más ceros (valores en blanco) que valores reales, es decir, los vectores (documentos) son dispersos (o en ingles sparse vectors). Esto podría llegar a ser un problema (sobre todo de memoria) cuando se tenga un vocabulario de tamaño considerable.

3.9.9 Descomposición de Valores Singulares

Cada entrada de la matriz de documento-término representa el número de veces que un término aparece en un documento. Para una colección de varios miles de documentos, la matriz de frecuencia término-documento puede contener cientos de miles de palabras (https://support.sas.com/documentation/onlinedoc/txtminer/12.3/tmref.pdf). Se requiere por tanto mucho tiempo y espacio de cálculo para analizar esta matriz de forma eficaz. Además, tratar con datos con alta dimensionalidad es intrínsecamente complejo para el proceso de modelado. Para mejorar el rendimiento, se puede aplicar la descomposición de valores singulares (del inglés Single Value Descomposition o SVD) para reducir las dimensiones de la matriz de frecuencia término-documento, transformando la matriz en una forma de menor dimensión más compacta e informativa. Un número elevado de dimensiones SVD suele resumir mejor los datos, pero cuanto mayor sea el número más recursos informáticos se necesitarán. El cálculo de la SVD es en sí mismo una tarea que requiere mucha memoria. En el caso de problemas con un número de elevado de observaciones y dimensiones, SVD puede realizar automáticamente una muestra aleatoria de los documentos para evitar que se agote la memoria.

Para cualquier k (en nuestro caso k es el número de temas o tópicos a buscar) dado, el resultado de la transformación será la factorización de la matriz con k dimensiones que mejor se aproxima a la matriz original. Un valor más alto de k da una mejor aproximación a la matriz A. Sin embargo, la elección de un valor demasiado grande para k podría dar lugar a una dimensión demasiado alta para el proceso de modelado. En general, el valor de k debe ser lo suficientemente grande como para preservar el significado de la colección de documentos, pero no tan grande como para capturar el ruido. Los valores entre 10 y 200 son apropiados a menos que la colección de documentos sea pequeña. Para su aplicación específica en minería de textos, es posible que se desee comparar los resultados para varios valores de k. Como regla general, los valores más pequeños de k (de 2 a 50) son útiles para la agrupación, y los valores más grandes (de 30 a 200) son útiles para la predicción o la clasificación. La SVD descompone la matriz de frecuencia término-documento en otras tres matrices. SVD factoriza la matriz de frecuencia término-documento grande y dispersa (sparse matrix) calculando una SVD truncada de la matriz. Suponiendo que A es la matriz de frecuencia término-documento grande y dispersa con entradas ponderadas. La SVD de una matriz A es una factorización de A en tres nuevas matrices U, D y V como se muestra en la siguiente ecuación:

$$A = U * D * V^T$$

Las matrices U y V tienen columnas ortonormales mientras que D es una matriz diagonal de valores singulares. La SVD calcula sólo las primeras k columnas de estas matrices (U, D y V). Una vez calculada la SVD, cada columna (o documento) de la matriz de frecuencias término-documento puede proyectarse sobre las primeras k columnas de U. Matemáticamente, esta proyección forma un subespacio k-dimensional que es el que mejor describe el conjunto de datos. Cada fila de la matriz U (matriz de término-documento) es la representación vectorial del correspondiente documento. La longitud de estos vectores es k, que es el número de temas deseados. Así, la SVD genera los vectores para cada documento y término del conjunto de datos.

3.10 Algoritmos de Clasificación Supervisados

3.10.1 Regresión Logística

La regresión logística es conceptualmente similar a la regresión lineal, donde la regresión lineal estima la variable objetivo a partir de un conjunto de variables independientes xi. Sin embargo, en lugar de predecir valores continuos, como en la regresión lineal, la regresión logística estima las probabilidades de que se produzca un determinado evento.

Los principales parámetros en de este clasificador en su implementación en la librería Scikit-Learn de Python son:

- Penalty: Se utiliza para especificar el criterio de penalización.
- Tol: Tolerancia para el criterio de parada.
- C: Inverso de la "fuerza" de la regularización. Como en las SVM, los valores más pequeños especifican una regularización más fuerte.
- Solver: algoritmo usado en proceso de optimización.
- L1 ratio: parámetro de ajuste de Elastic-Net.

3.10.2 Clasificador Naive Bayes para Modelos Multinomiales

Los clasificadores de la familia de Naive Bayes son muy usados en PLN, en particular en la clasificación de textos, ya que suelen producir resultados comparables con los obtenidos por otros métodos más sofisticados y son bastantes sencillos de implementar (https://medium.com/datos-y-ciencia/algoritmos-naive-bayes-fudamentos-e-implementaci%C3%B3n-4bcb24b307f). Uno de los aspectos desfavorables de estos clasificadores es que presuponen que los términos que figuran en un documento son todos independientes entre sí, lo cual puede no ser totalmente cierto debido a la naturaleza del lenguaje. Muchos estudios han tratado de proporcionar información adicional al clasificador para suavizar las consecuencias de la suposición de independencia. El clasificador Naive Bayes simple considera la probabilidad de ocurrencia de cada término dada la clase de forma binaria, es decir, el término se da o no se da y luego su probabilidad condicional dada la clase es o no considerada. En este sentido, el clasificador Naive Bayes Multinomial suele mejorar su rendimiento ya que considera el número de apariciones del término para evaluar la contribución donde la probabilidad condicional dada la clase con lo que el modelado de cada documento se ajusta mejor a la clase a la que pertenece.

Multinomial NB de Python implementa el algoritmo de Naive Bayes para datos con una distribución multinomial, siendo una de las dos variantes clásicas de Naive Bayes utilizadas en la clasificación de textos (donde los datos se representan típicamente como recuentos de vectores de palabras, aunque también se sabe que los vectores TF-IDF funcionan bien en la práctica).

Los principales parámetros en de este clasificador en su implementación en la librería Scikit-Learn de Python son:

alpha: Parámetro de suavizado aditivo (Laplace/Lidstone) (0 para no suavizar).

3.10.3 Árboles de Decisión

Los árboles de decisión son algoritmos iterativos que consisten en dividir los datos en regiones basadas en intervalos de variables independientes [10]. Además, son capaces de explicar variables de una forma sencilla al tratarse de métodos no paramétricos, no siendo necesaria se cumpla ningún tipo de distribución específica. En términos generales, el proceso se inicia a partir de un nodo raíz que contiene todas las observaciones del conjunto de datos de entrenamiento, posteriormente, se dividen las observaciones en subconjuntos o nodos hijos, con el objetivo de encontrar una segmentación con la mayor homogeneidad posible respecto a la variable objetivo. Cuando se quiere predecir una nueva observación, se recorre el árbol según el valor de sus predictores hasta alcanzar uno de los nodos hoja. La predicción del árbol es la media de la variable respuesta de las observaciones de entrenamiento que están en ese mismo nodo hoja. Los árboles de decisión consisten en:

- Se empieza por encontrar un punto de corte en cada variable (obteniendo dos intervalos) y se observa el error cometido en la predicción fijando valores constantes. Se seleccionan la variable y el punto de corte óptimos. En el caso de las variables categóricas, se establecen agrupaciones de categorías en lugar de puntos de corte.
- Dentro de las divisiones obtenidas en el apartado anterior, se subdivide el árbol hasta alcanzar algún criterio de parada (número de hojas finales, por ejemplo). Además, es necesario establecer una metodología para crear las regiones mencionadas anteriormente, es decir, decidir dónde se establecen las divisiones, sobre qué predictores y sobre qué valores de los mismos.

3.10.4 Random Forest

El algoritmo de Random Forest es una modificación del método de Ensembles y Bagging. Un ensemble es un conjunto de modelos de machine learning, cada modelo produce una predicción diferente. Las predicciones de los distintos modelos se combinan para obtener una única predicción. Bagging consiste en que cada modelo del conjunto de modelos se entrena con subconjuntos del conjunto de entrenamiento. Estos subconjuntos se forman eligiendo muestras aleatoriamente (con repetición) del conjunto de entrenamiento. Esta forma de entrenar los modelos individuales se realiza para conseguir que los errores se compensen entre sí.

Los modelos de Random Forest están formados por un conjunto de árboles de decisión individuales, de modo que los valores de cada árbol son entrenados con una muestra aleatoria extraída de los datos de entrenamiento originales mediante reemplazamiento o bootstrapping. Por lo tanto, cada árbol se entrena con datos ligeramente diferentes, ya que en cada árbol individual las observaciones

se van distribuyendo por nodos, creando la estructura de árbol hasta alcanzar un nodo hoja. La predicción de una nueva observación se obtiene agregando las predicciones de todos los árboles individuales que forman el modelo. El objetivo es lograr un equilibrio entre bias y varianza incorporando dos métodos de variabilidad, por un lado, el remuestreo de las observaciones y por otro la aleatoriedad de las variables, utilizadas para segmentar los nodos de los árboles.

Las principales ventajas y desventajas del Random Forest son:

Ventajas:

- Los árboles son fáciles de interpretar aun cuando las relaciones entre predictores son complejas.
- Encuentran relaciones no lineales y mejoran su capacidad predictiva cuando las los predictores son categóricos, además de evitan el problema de la existencia de variables predictoras muy dominantes. Logran reducir la varianza en el modelo final gracias a la aleatoriedad en las variables de cada nodo, generando árboles diferentes unos de otros.
- Los árboles pueden, en teoría, manejar tanto predictores continuos como categóricos sin tener que crear variables dummies o one-hot-encoding.
- Al tratarse de métodos no paramétricos, no es necesario que se cumpla ningún tipo de distribución específica.
- En general, requieren mucha menos limpieza y preprocesamiento de datos en comparación con otros métodos de ML (no requieren estandarización).
- No se ven muy influenciados por outliers.
- Son muy útiles en la exploración de datos, permiten identificar de forma rápida y eficiente las variables (predictores) más importantes.
- Son capaces de seleccionar predictores de forma automática.
- Pueden aplicarse a problemas de regresión y clasificación.

Desventajas:

- Son sensibles a datos de entrenamiento desbalanceados (una de las clases domina sobre las demás).
- Cuando tratan con predictores continuos, pierden parte de su información al categorizarlos en el momento de la división de los nodos.

Los principales parámetros en de este clasificador en su implementación en la librería Scikit-Learn de Python son:

- Bootstrap: Si se utilizan muestras Bootstrap o reemplazamiento al construir los árboles. Si es Falso, se utiliza todo el conjunto de datos para construir cada árbol.
- Max depth: Máxima profundidad del árbol.
- Max features: Número de variables que se sortean en cada división del árbol.
- Min samples leaf: Número de observaciones que habrá como mínimo en una rama-nodo, es decir, al tamaño mínimo de la hoja.
- Min samples split: Número de observaciones a sortear por nodo.
- N estimators: Número total de árboles.
- Criterion: Función para medir la calidad de una división. Los criterios admitidos son "gini" para la impureza de Gini y "entropía" para la ganancia de información.

3.10.5 Máquinas de Soporte Vectorial

El algoritmo de Maquina de Vector Soporte (SVM) trata de buscar el hiperplano que divida los datos en diferentes grupos de acuerdo con sus características [10]. En la figura 14 se muestra un ejemplo de hiperplano en 2 dimensiones. Se podría pensar que hay múltiples soluciones para dividir un plano con una recta, pero la SVM resuelve este problema cogiendo el punto de cada grupo más alejado, también llamado vector de soporte e intentando maximizar su margen (distancia que lo separa de la recta que divide cada agrupación de datos). Además, se puede tener más de un vector de soporte por grupo, en ese caso se tendrá que elegir el que mayor margen tenga, como el vector H2 que se muestra en la figura 14. Idealmente una SVM separará los grupos de forma perfecta usando los hiperplanos mencionados anteriormente. Esto no siempre es posible, y si lo es, este modelo no podrá ser utilizado en otro conjunto de datos, ya que se consideraría que ha sobre aprendido.

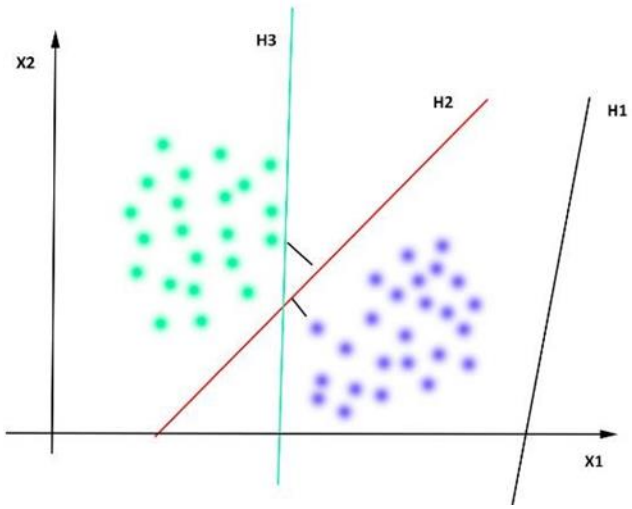


Figura 14: ejemplos de vectores de soporte por grupo (fuente: https://conceptosclaros.com/que-es-maquina-vectores-soporte/)

La separación entre clases en muchos problemas no es lineal. Es posible trabajar, a pesar de ser un algoritmo de separación lineal, en un espacio de alta dimensionalidad. El problema es que este aumento de dimensión hace a menudo impracticables los cálculos. Este problema se soluciona con el llamado “truco Kernel” (“the Kernel trick”): cualquier algoritmo que dependa solo de los productos escalares (como es el caso del SVM) permite trabajar computacionalmente en una dimensión controlada a través de una función llamada Kernel.

Los principales parámetros en de este clasificador en su implementación en la librería Scikit-Learn de Python son:

- RBF: Aumentar gamma en la función RBF implica menor sesgo y mayor sobreajuste.
- Polinomial. Aumentar el grado del polinomio implica menor sesgo y mayor sobreajuste.
- C: Con el fin de contemplar un poco de flexibilidad, SVM cuenta con un parámetro C, es un parámetro de regulación, es el único parámetro que puede ser ajustado a la hora de construir un clasificador con SVM. Su rango depende mucho de los datos. Aumentar C implica menor sesgo y

mayor sobreajuste. Este parámetro regula cómo de precisos se quiere que sea nuestra máquina, por lo tanto puede controlar el nivel de aprendizaje.

- Degree: Grado del polinomio (válido para el kernel polinomial).
- Gamma: Coeficiente para el kernel Radial Basis Function o RBF y Polinomial. Cuando gamma es muy pequeña, el modelo está demasiado restringido y no puede capturar la complejidad o la "forma" de los datos. Cuando gamma es muy grande, el radio del área de influencia de los vectores de soporte sólo incluye el propio vector de soporte y ninguna regularización con C podrá evitar el sobreajuste.
- Max_iter: Límite estricto de iteraciones dentro del solucionador dado, o -1 si no hay límite (por defecto max_iter = -1).

3.11 Algoritmos de Clasificación No Supervisado

3.11.1 Clusterización método KMeans

En lenguaje sencillo, el objetivo de K-Means es colocar puntos de datos con características similares en el mismo grupo (es decir, cohesión interna) y puntos de datos separados con características diferentes en diferentes grupos (es decir, separación externa).

Los pasos para realizar el proceso de clusterización con el método KMeas son los siguientes:

- Paso 1: inicialice los centroides de los clústeres eligiendo aleatoriamente K puntos de inicio
- Paso 2: Asigne cada punto de datos al centroide más cercano. El cálculo de distancia comúnmente utilizado para la agrupación en clústeres de K-Means es la distancia euclidiana, un valor de escala que mide la distancia entre dos puntos de datos.
- Paso 3: actualice los centroides de los clústeres. Un centroide se calcula como el promedio de puntos de datos en un grupo. Los centroides actualizados pueden o no ser los puntos de datos reales. Sería una coincidencia si lo son.
- Paso 4: repita los pasos 2 y 3 (asignar cada punto de datos a nuevos centroides y actualizar los centroides del grupo) hasta que se cumpla una de las condiciones de parada.
    - Los centroides actualizados siguen siendo los mismos que los de la iteración anterior (esta es una situación ideal, pero en la práctica, puede llevar demasiado tiempo)
    - La suma de errores al cuadrado no mejoró en al menos x %
    - Se alcanza el número máximo de iteraciones (elija sabiamente el máximo de iteraciones, de lo contrario, tendríamos clústeres deficientes).

Para la obtención del número óptimo de clústeres se utilizan las siguientes metodologías:

Método de Elbow: utiliza SSE (suma de errores al cuadrado o también conocido como Cluster Inertia) para evaluar la división en clúster. Se ejecutan múltiples clusterizaciones indicando el número de cluster K y se evaluó el SSE, con lo anterior se crea un diagrama de codo de SSE para valores de K que van desde 2 a N. A medida que K aumenta, el SSE correspondiente disminuirá. Observaremos la compensación entre K y SSE, se quiere que SSE sea bajo y mantener un K en un valor razonable. Por lo general, se elige el valor óptimo de K cuando se observa que SSE comienza a aplanarse y toma forma de codo.

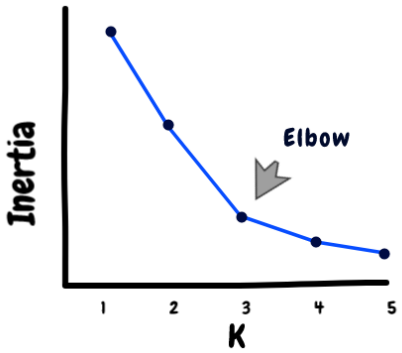


Figura 15: aplicación del método Elbow para encontrar número optimo de clústeres (fuente: https://towardsdatascience.com/k-means-clustering-from-a-to-z-f6242a314e9a)

Podemos definir el coeficiente de la silueta como:

- a(x)= distancia promedio de x a todos los demás puntos en el mismo cluster.
- b(x)= distancia promedio de x a todos los demás puntos en el cluster más cercano.

Dado esto, se dice que el coeficiente de la silueta para x está dado por:

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$$

Donde el valor de s(x) puede variar entre -1 y 1,

- -1 si es un mal agrupamiento
- 0 si es indiferente
- 1 si es un buen agrupamiento

El coeficiente de silueta final se calcula como el coeficiente de silueta promedio de todos los puntos. Luego calculamos los coeficientes de silueta para los valores de K que van de 2 a N. Cuanto mayor sea el coeficiente de silueta, mejor será el agrupamiento.

$$sc = \frac{1}{N}\sum_{i=1}^{N} s(x)$$

4 Metodología

4.1 Recopilación de datos

4.1.1 Inspección de la información o datos a extraer

Para el trabajo se propone extraer los datos o información del sitio web de noticias CNN Chile (https://www.cnnchile.com), en el cual se puede encontrar noticias desde el 2015 hasta la fecha. En dicho sitio web las noticias se encuentran clasificadas o separadas en las siguientes categorías:

- País, URL: https://www.cnnchile.com/pais/
- Mundo, URL: https://www.cnnchile.com/mundo/
- Economía, URL: https://www.cnnchile.com/economia/
- Cultura, URL: https://www.cnnchile.com/cultura/
- Deporte, URL: https://www.cnnchile.com/deportes/
- Tecnología, URL: https://www.cnnchile.com/tecnologias/

Cada noticia tiene una estructura definida, en donde se pueden distinguir los siguientes campos:

Tag: es una o varias palabras que puede representar el contenido de la noticia
Fecha_hora: es un campo que contine la fecha y hora de publicación de la noticia en el sitio web de CNN Chile
Título: es una frase u oración que describe el contenido de la noticia
Subtitulo: este campo contiene un resumen de la noticia, más largo que el título, pero más corto que el texto de la noticia
Texto: contenido completo de la noticia, compuesto por varias oraciones y describe completamente a la noticia

4.1.2 Obtención de los datos

Como se indicó en los puntos anteriores en esta etapa se utiliza la librería o framework Scrapy del lenguaje de programación Python.

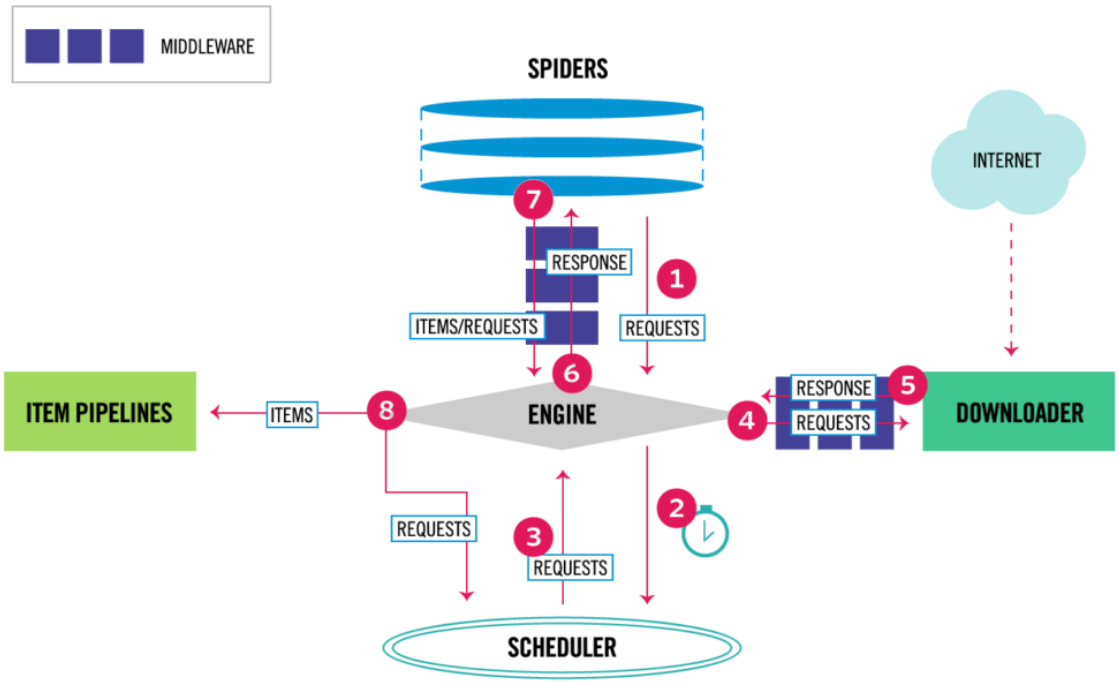


Figura 16: estructura de la librería y/o framework Scrapy (fuente: https://elmundodelosdatos.com/extraccion-datos-sitios-web-productos-zara/)

Para trabajar con Scrapy se debe instalar la librería o framework con el gestor de paquetes de Python denominado pip, el cual se instala normalmente al momento de instalar Python en el computador. Pip debe estar referenciado en las variables de entorno del sistema.

Luego se procede a crear un primer proyecto con Scrapy utilizando el comando startproject de dicho framework. Se ejecuta dicho comando desde el símbolo del sistema o Shell de Windows ubicados en la ruta en donde se desea crear el proyecto.

El comando anterior genera una estructura de carpetas anidadas como las que se muestran a continuación:

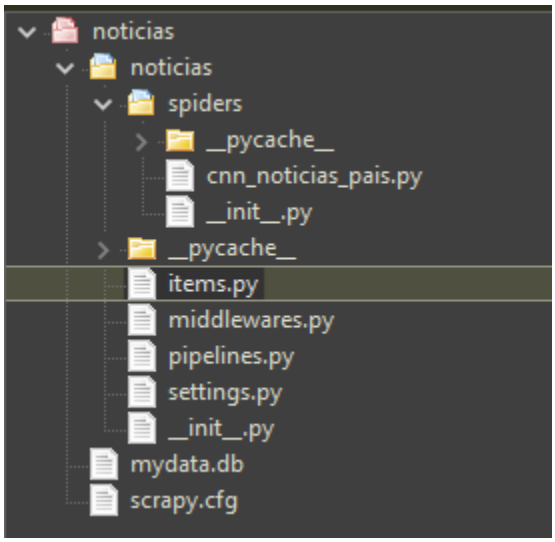


Figura 17: estructura de un proyecto en el framework Scrapy

La descripción de los principales directorios o carpetas y ficheros que son parte del proyecto:

- scrapy.cfg: fichero de configuración que se encuentra en el directorio raíz del proyecto y contiene el nombre del módulo con los ajustes del proyecto.
- noticias: módulo del proyecto
- noticias/items.py: contiene las definiciones de nuestros ítems, que son los objetos estructurados que usaremos para cargar los datos que extraigamos y son muy similares a los diccionarios en Python. Así, este tipo de objeto también almacena la información mediante pares clave-valor.
- noticias/middelwares.py: contiene los middlewares del proyecto.
- noticias/pipelines.py: en este fichero implementaremos los pipelines que se encargarán de procesar los ítems una vez hayan sido rastreados por nuestra araña.
- noticias/settings.py: fichero con los ajustes del proyecto.
- noticias/spiders: directorio donde se encontrarán nuestras arañas.

Pasos que debemos de realizar para poder rastrear y extraer la información de un sitio web mediante la librería y/o framework Scrapy:

- Definir nuestro objeto ítem y los campos que queremos extraer en el fichero items.py.
- Crear nuestra araña que extraerá la información de las páginas.
- Implementar el pipeline para procesar el ítem una vez rastreado y almacenarlo en nuestra base de datos sqlite3.

Definición del ítem

Empezamos definiendo un nuevo ítem al que llamaremos NoticiasItem en nuestro fichero items.py. Además, declaramos objetos scrapy.Field los cinco campos que deseamos extraer de cada noticia: tag, fecha_hora, titulo, subtitulo y texto.

Creación de nuestra araña y extracción de los datos

Las arañas o rastreadores son clases que se encargan de extraer la información de cada página. En este caso, creamos un nuevo fichero en la carpeta spiders llamado cnn_noticias_pais.py. A continuación, la explicación o el uso de algunos atributos de las arañas o spider que se pueden crear en Scrapy:

- name: es el nombre de nuestra araña o rastreador, en este caso cnn_noticias_pais.
- start_urls: cuando se ejecuta un proceso de Scrapy, la araña comienza a hacer peticiones a las URLs que se han definido en el atributo start_urls. En nuestro caso utilizamos las URL asociadas a la categoría de cada noticia, por ejemplo https://www.cnnchile.com/economia/
- rules: este campo contiene reglas que le permiten a Scrapy extraer información desde la pagina web. En este caso se agregan dos reglas, la primera permite extraer cada noticia del sitio web que se indica en el campo start_urls que tenga la estructura que se ingresa en la regla como "restrict_xpaths", permitiendo extraer los link de cada noticia en particular. Luego que se acaban los links que extrae la primera regla se ejecuta la segunda regla la cual busca el link para acceder a la siguiente pagina de grupo de noticias, acción conocida como scrolling o paginación web. Al pasar a la siguiente pagina se vuelve aplicar la primera regla y de esa forma se repite el proceso. Para la definición de las reglas y encontrar los distintos elementos de cada noticia se utiliza el lenguaje XPath, que trata a los documentos XML y HTML como árboles y nos permitirá seleccionar los diferentes nodos como por ejemplo, un atributo, de manera sencilla.
- parse_item: esta es una función que recibe la respuesta de la pagina web de cada link que es extraído por la primera regla antes mencionada. En palabras sencillas esta función tiene acceso al contenido de cada pagina web asociada a los links que extrae la primera regla. Esta función permite separar el contenido en los campos del ítems que se definió anteriormente como NoticiasItem, por lo tanto, la función parse_item genera un nuevo item (objeto iterable) del tipo NotiasItem, objeto que luego es procesado y almacenado.
- contar_pagina: esta función permite identificar el numero de la página en la cual se encuentra luego de hacer scrolling o paginación web, permitiendo definir un número máximo de páginas a revisar de tal forma de terminar con la ejecución del programa.

Almacenando los ítems

El siguiente paso consiste en implementar un pipeline que se encargue de almacenar cada ítem en una base de datos con sqlite3 (Scrapy también permite exportar los datos a ficheros CSV o JSON mediante los Feed Exports).

De esta manera, para almacenar la información de cada ítem a una base de datos sqlite3 añadiremos al fichero pipelines.py el código necesario para crear una base de datos, crear los campos de dicha base de datos y almacenar en cada iteración el items.

Dentro del fichero pipeline.py se crea la clase NoticiasPipeline, la cual será ejecutada cada vez que se procese un ítem, ósea cada vez que sea llamada la función parse_item, dicha función es llamada a su vez cuando la regla uno definida en la clase CnnNoticiasPaisSpider encuentra un link de noticia que se ajuste al patron xpath utilizado. En el archivo settings.py se debe indicar cual es el pipeline que se debe ejecutar cada vez que se procesa un ítem.

Descripción de las principales funciones de la clase NoticiasPipelines

- __init__: este método se ejecuta cada vez que se llama a la calse NoticiasPipelines y en dicha función se llama a su vez a las siguientes funciones: create_conn y create_table.
- create_conn: esta función crea la conexión a la base de datos, se le debe pasar el nombre de la base de datos. Si no existe esta función y sus métodos la crean.
- create_table: esta función permite crear la tabla con todos los campos que componen a una noticia. En resumen esta función ejecuta una sentencia sql en la base de datos para crear la tabla con su nombre y campos correspondientes. Si existe una tabla con el mismo nombre se borra.
- process_item: esta es la función base que define a la clase pipeline y es la que permite procesar los distintos ítems que se le entregan desde la clase spider. Esta función llama a otra función con el nombre de putitemsintable.
- Putitemsintable: esta función ejecuta una sentencia sql que permite agregar los campos de cada noticia (item) que es procesada por la clase spider, de tal manera de agregar los distintos campos a la tabla de la base de datos.

Este procedimiento se repite para cada una de las categorías de noticias que se indicaron en las secciones anteriores.

Para poder responder la pregunta si es que existe una relación entre las noticias publicadas por el sitio web CNN Chile y la variación de la bolsa de valores de Chile se requiere también datos de este último organismo. Para este fin se utilizará el indicador IPSA.

El IPSA (Índice de Precio Selectivo de Acciones) es el principal índice bursátil de Chile, elaborado por la Bolsa de Comercio de Santiago. Corresponde a un indicador de rentabilidad de las 30 acciones con mayor presencia bursátil, siendo dicha lista revisada anualmente. En su cálculo el índice considera todas las variaciones de capital de cada acción incluida en el índice, ponderada por el peso relativo de cada una de ellas, siendo dicho peso calculado a partir de una fórmula que considera, tanto la capitalización bursátil, como el número de transacciones y el free float (https://es.wikipedia.org/wiki/%C3%8Dndice_de_Precio_Selectivo_de_Acciones).

En el sitio web "https://es.investing.com" (https://es.investing.com/indices/ipsa-historical-data) se encontró una base de datos con los valores históricos del IPSA como se observa subsiguientemente:

Datos históricos S&P CLX IPSA

Plazo:

Diario | Descargar datos | 16.10.2022 - 16.11.2022

Fecha	Último	Apertura	Máximo	Mínimo	Vol.	% var.
15.11.2022	5.239,13	5.302,44	5.302,44	5.218,71		-1.20%
14.11.2022	5.302,89	5.350,79	5.378,52	5.281,41		-0.90%
11.11.2022	5.350,79	5.339,54	5.368,82	5.334,72		+0.21%
10.11.2022	5.339,54	5.350,36	5.421,27	5.311,11		-0.20%
09.11.2022	5.350,36	5.406,12	5.406,12	5.315,98		-1.03%
08.11.2022	5.406,12	5.289,55	5.406,12	5.289,55		+2.20%
07.11.2022	5.289,55	5.220,04	5.290,16	5.212,16		+1.33%

Figura 18: tabla de datos del IPSA (fuente: https://es.investing.com/indices/ipsa-historical-data)

En dicho sitio web se puede descargar un histórico en un formato csv, para este trabajo se descargarán los datos desde el 01-01-2015 hasta la fecha y se utilizará el índice "% var." el cual muestra como varia el IPSA del día actual con respecto al día anterior.

4.2 Procesamiento de datos y limpieza de datos

Luego del trabajo realizado con la librería o framework Scrapy descrito en la sección anterior, se creó una base de datos con 6 tablas, cada una de ellas asociada una categoría de noticias. Para poder explorar la base de datos se utilizó una herramienta de escritorio llamada "DB Browser for SQLite" con una interfaz gráfica:

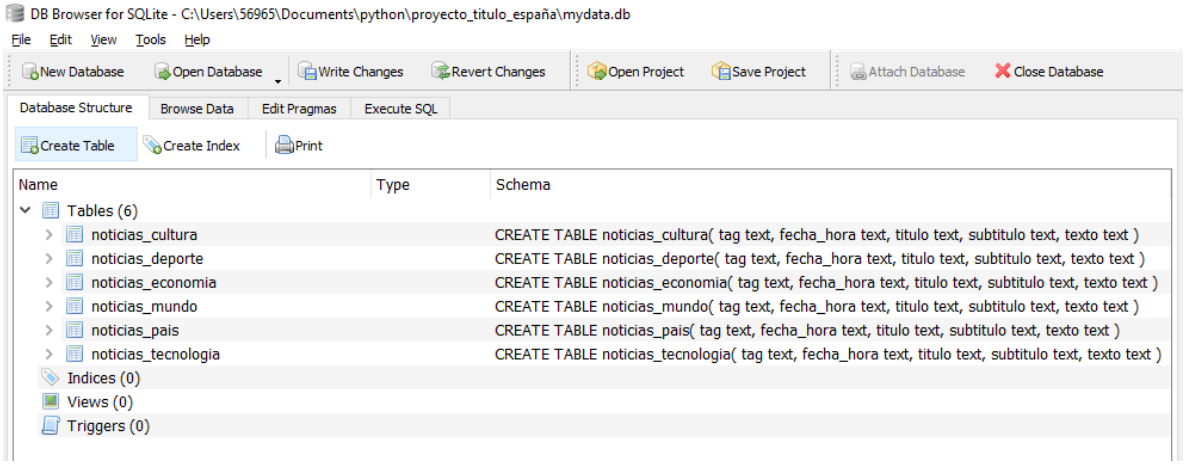
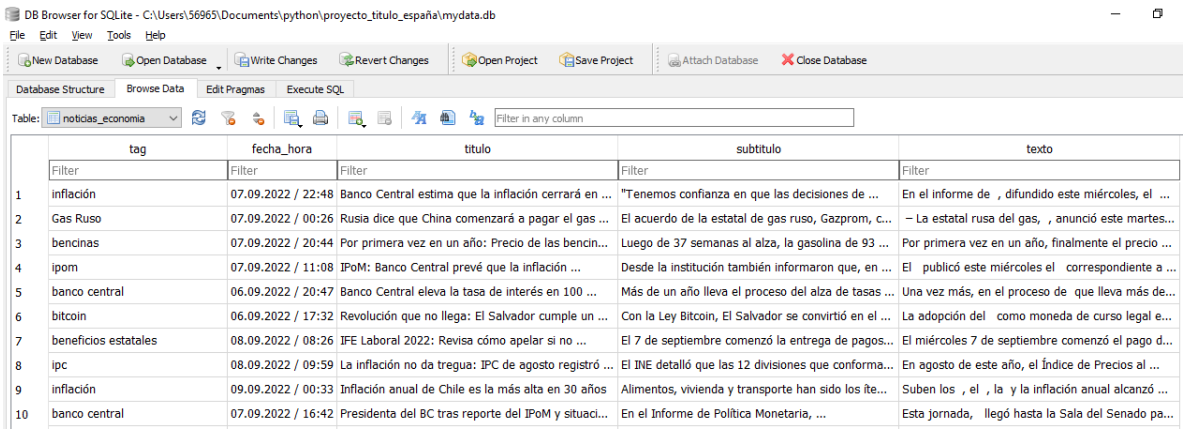


Figura 19: tabla de datos de noticias visualizada en la herramienta DB Browser

Con esta herramienta se puede no solo explorar la base de datos, también se puede navegar por cada tabla y explorar sus campos o columnas y sus registros:



	tag	fecha_hora	titulo	subtitulo	texto
1	inflación	07.09.2022 / 22:48	Banco Central estima que la inflación cerrará en ...	"Tenemos confianza en que las decisiones de ...	En el informe de , difundido este miércoles, el ...
2	Gas Ruso	07.09.2022 / 00:26	Rusia dice que China comenzará a pagar el gas ...	El acuerdo de la estatal de gas ruso, Gazprom, c...	– La estatal rusa del gas, , anunció este martes...
3	bencinas	07.09.2022 / 20:44	Por primera vez en un año: Precio de las bencin...	Luego de 37 semanas al alza, la gasolina de 93 ...	Por primera vez en un año, finalmente el precio ...
4	ipom	07.09.2022 / 11:08	IPoM: Banco Central prevé que la inflación ...	Desde la institución también informaron que, en ...	El publicó este miércoles el correspondiente a ...
5	banco central	06.09.2022 / 20:47	Banco Central eleva la tasa de interés en 100 ...	Más de un año lleva el proceso del alza de tasas ...	Una vez más, en el proceso de que lleva más de...
6	bitcoin	06.09.2022 / 17:32	Revolución que no llega: El Salvador cumple un ...	Con la Ley Bitcoin, El Salvador se convirtió en el ...	La adopción del como moneda de curso legal e...
7	beneficios estatales	08.09.2022 / 08:26	IFE Laboral 2022: Revisa cómo apelar si no ...	El 7 de septiembre comenzó la entrega de pagos...	El miércoles 7 de septiembre comenzó el pago d...
8	ipc	08.09.2022 / 09:59	La inflación no da tregua: IPC de agosto registró ...	El INE detalló que las 12 divisiones que conforma...	En agosto de este año, el Índice de Precios al ...
9	inflación	09.09.2022 / 00:33	Inflación anual de Chile es la más alta en 30 años	Alimentos, vivienda y transporte han sido los íte...	Suben los , el , la y la inflación anual alcanzó ...
10	banco central	07.09.2022 / 16:42	Presidenta del BC tras reporte del IPoM y situaci...	En el Informe de Política Monetaria, ...	Esta jornada, llegó hasta la Sala del Senado pa...

Figura 20: Visualización de la tabla de noticias asociadas a la categoría economía

La cantidad de registros por cada tabla o categoría se muestra a continuación:

- País: 66270
- Mundo: 20688

- Economía: 8508
- Cultura: 3528
- Deporte: 17736
- Tecnología: 2484

El paso siguiente es comenzar con el procesamiento y/o limpieza de los datos. En una etapa temprana se decidió para este trabajo realizar una pre-limpieza directamente sobre la base de datos utilizando la librería sqlite3 y el lenguaje de base de datos sql. La principal razón de lo antes expuesto es un tema de velocidad y recursos de procesamiento, la librería sqlite está escrita en C, implementando el lenguaje sql, por lo cual, se encuentra optimizada para trabajar con bases de datos, es decir, realizar operaciones de texto, numéricas, procesamiento de datos, etc., con grandes volúmenes de información, facilitando la limpieza del conjunto de datos de noticias de manera más rápida y con menos recursos computacionales.

4.2.1 Descripción de los componentes del código de limpieza que se aplica en esta etapa en la base sqlite:

- Función limpiar_texto: a esta función se le pasa una variable tipo string, se le pasará el contenido de la noticia para realizar la limpieza. Los componentes de esta función son los siguientes:
    - Se utiliza la librería BeautifulSoup para quitar todos los tags HTML que pudieran quedar desde el proceso de scrapy
    - Se utiliza el método lower de la clase string de Python para dejar todo el texto en minúscula
    - Se utiliza el método replace de la clase string de Python en conjunto con el método findall de la librería re o regex (expresiones regulares) para reemplazar los puntos apartes por espacios en blanco
    - Se utiliza el método replace de la clase string de Python en conjunto con el método findall de la librería re o regex (expresiones regulares) para reemplazar las comas por espacios en blanco
    - Se utiliza el método sub de la librería re o regex (expresiones regulares) para reemplazar los números o dígitos por espacios en blanco
    - Se utiliza el método translate de la clase string de Python para reemplazar los signos de puntuación por espacios en blanco. Se utiliza una constante string.punctuation de la clase string la cual contine todos los signos de puntuación y algunos signos como el de exclamación, el de pregunta, el de adición, etc. Adicionalmente se concatena a dicha constante los signos “¡¿” dado que son propios del idioma español.
    - Se utiliza el método sub de la librería re o regex (expresiones regulares) para reemplazar cualquier carácter que sea distinto de caracteres de palabra.
    - Se utiliza el método sub de la librería re o regex (expresiones regulares) para reemplazar caracteres especiales de utf-8
    - Se utiliza el método sub de la librería re o regex (expresiones regulares) para reemplazar varios espacios por un único espacio en blanco
    - Se utiliza el método strip de la clase string de Python para eliminar espacios al inicio y fin de un texto
- Se crea el objeto connect de la librería sqlite: con este objeto se ejecuta una conexión a la base de datos previamente creada desde el “raspado” de la información de las noticias con la librería Scrapy. Se ejecuta dentro de un try por si ocurre algún error de conexión.

- A partir del objeto conexión a la base de datos (objeto connect) se crea un cursor (objeto cursor) el cual permite ejecutar sentencias del lenguaje sql sobre la base de datos.
- Mediante el método create_function del objeto conexión a la base de datos (objeto connect) se crea una función dentro de la base de datos. Los parámetros que se le pasan al método create_function son:
    - Nombre de la función en la base de datos. Este nombre permite invocar la función mediante sentencias sql desde el objeto cursor y su método execute.
    - Cantidad de parámetros que se le pasa a la función. En este caso se pasa un único parámetro que es el texto a limpiar o procesar.
    - Función previamente definida en Python. En este caso se pasa la función limpiar_texto previamente definida en el código.
      En resumen, el método create_function traduce una función de Python a una función sql y la incorpora en la base de datos para luego ser invocada.
- Se definen dos listas, la primera con el nombre tablaNueva que contine el nombre de las nuevas tablas a crear en la base de datos. La segunda con el nombre tablaAntigua que contine el nombre de las tablas existentes en la base de datos
- Mediante un ciclo for se recorre la lista de las tablas nuevas y se ejecutan las siguientes sentencias del lenguaje sql con el método execute del objeto cursor previamente definido:
    - Se ejecuta la sentencia CREATE TABLE, complementado con otras sentencias, para crear la nueva tabla con los mismos campos que la tabla antigua
    - Se ejecuta la sentencia INSERT INTO, complementando con otras sentencias, para insertar todos los registros de la tabla antigua en la tabla nueva
    - Se ejecuta la sentencia ALTER TABLE, complementado con otras sentencias, para incorporar un nuevo campo texto2
    - Se ejecuta la sentencia UPDATE, complementado con otras sentencias, para llamar a la función LIMPIAR_TEXTO, pasándole como parámetro el campo texto de la tabla antigua y asignando el resultado de la ejecución de la función al nuevo campo texto2
    - Se ejecuta la sentencia DELETE FROM, complementando con otras sentencias, para borrar aquellos registros en donde el campo texto2 este vacío
- Se ejecuta el comando commit del objeto connect para que las actualizaciones de transacciones o comandos en la base de datos sean permanentes
- Se cierra el objeto cursor con el método close()
- Se cierra el objeto connect con el método close(), de esta manera se cierra la conexión a la base de datos

El siguiente paso es la limpieza de los datos, pero directamente desde Python sin utilizar sql sobre una base de datos, lo anterior se realiza leyendo los datos con la librería pandas.

Una vez cargados los datos a un dataframe de pandas, que es un símil de una base de datos en dicha librería, se pueden aplicar distintas funciones para limpiar y procesar los datos.

4.2.2 Descripción de las funciones implementadas para la limpieza y procesamiento de los datos:

- Función cambiar_letra_punto_letra: esta función detecta el patrón palabra – punto – palabra. En un análisis de los datos se detectó que en muchas ocasiones por errores ortográficos del redactor

de las noticias no se incluye el espacio luego de un punto seguido, por lo tanto esta función permite borrar el punto e introducir un espacio en su lugar.

- Función cambiar_letra_3punto_letra: esta función es similar a la del punto anterior, con la salvedad que detecto el patrón palabra – puntos suspensivos o tres puntos – palabra, reemplazando el puntos suspensivos por un espacio
- Función limpiar_texto: esta función es muy similar a la función implementada directamente en la base de datos para procesar el texto, sus principales métodos se pueden resumir:
    - Se utiliza el método lower de la clase string para transformar el texto a minúscula
    - Se utiliza el método translate de la calse string en conjunto con la contante string.punctuation para reemplazar los signos de puntuación por espacios en blanco
    - Se utiliza el método strip de la clase string para borrar los espacios al inicio y fin de un texto
    - Se utiliza el método replace de la clase string para eliminar los saltos de carro o carácter "\n"
    - Se utiliza los métodos compile y sub de la librería re o regex para eliminar las posibles subcadenas de texto que tengan un patrón de tipo URL
    - Se utiliza los métodos compile y sub de la librería re o regex para eliminar los dígitos del 0 al 9, para eliminar los caracteres que no sean parte de una palabra, para eliminar los espacios múltiples y remplazarlos por un único espacio
- Función remplazar_stop_word: esta función utilizando la librería re o regex, en conjunto con la constante stopwords de la librería nltk en español (from nltk.corpus import stopwords) y aplicando las listas por comprensión de Python para borrar las stopwords que se puedan encontrar en los documentos que son parte del corpus de nuestros datos
- Función quitarEmoticones: esta función permite quitar algunos caracteres especiales como emoticones y signos mediante la utilización de la librería unicodedata la cual dispone de varias categorías de caracteres, la idea es que borre aquellos caracteres que son parte de categorías especiales (emoticones o signos especiales), para más detalles sobre las categorías de las que dispone la librería antes mencionada se puede consultar la siguiente URL: https://www.fileformat.info/info/unicode/category/index.htm
- Función lematizarTexto: esta función permite llevar esta función permite llevar una palabra a su forma lema o raíz si se puede decir de alguna forma. En general las palabras son formas flexionadas (es decir, en plural, en femenino, conjugada, etc.) de una palabra lema o raíz y esta función permite reemplazar la forma flexionada a su lema. Se utiliza la librería spacy que tiene un núcleo o core de palabras en español y permite dada una palabra encontrar su lema

Para aplicar las distintas funciones previamente definidas se utiliza el método apply de pandas el cual permite aplicar una función sobre una columna completa de valores. En el dataframe que contiene la base de datos se crea un nuevo campo llamado "texto3" que contiene la información luego de aplicar las funciones de limpieza y procesamiento de datos.

El dataframe resultante queda como se muestra en la siguiente figura:

In [17]:
```
datos.head()
```

Out[17]:

	index	tag	fecha_hora	titulo	subtitulo	texto	texto2	categoria	texto3
0	0	"18 más seguro"	14.09.2015 / 07:06	Inician campaña "18 más seguro"	Iniciativa es promovida por el Instituto de Se...	\n Cerca de 50 trabajadores sufrieron algún t...	cerca de trabajadores sufrieron algún tipo de ...	pais	cerca trabajadores sufrieron algún tipo accide...
1	1	"20 reglas de oro de la seguridad minera en ch...	03.08.2015 / 06:47	Lanzan campaña "20 reglas de oro de la segurid...	Luis Urzúa, uno de "Los 33" conversó con CNN C...	El Sernageomin está lanzando la campaña , inic...	el sernageomin está lanzando la campaña inicia...	pais	el sernageomin lanzando campaña iniciativa bus...
2	2	"abrazo de amor"	15.02.2016 / 08:08	"Glasko" triunfó en el Clásico Ciudad de Santiago	Comentario con todas las noticias de la hípica...	\n – sorprendió a todo en el estelar del sába...	sorprendió a todo en el estelar del sábado en ...	deporte	sorprendió estelar sábado además comentario hí...
3	3	"abuelas de plaza de mayo"	07.08.2013 / 21:54	Identificaron a hijo de chilenos secuestrados ...	El matrimonio de chilenos fue secuestrado en 1...	La organización argentina de derechos humanos ...	la organización argentina de derechos humanos ...	mundo	la organización argentina derechos humanos anu...
4	4	"adiós al lenguaje". crítica	18.08.2017 / 12:01	Transición 2017: Jornada clave para "los grand...	Se jugará la cuarta fecha del Campeonato Trans...	Este viernes a las 20 horas en el Estadio Bice...	este viernes a las horas en el estadio bicente...	deporte	este viernes horas estadio bicentenario florid...

Figura 21: tabla de datos de noticia con texto limpio y procesado

Adicionalmente se revisa si existen valores nulos o NAN en las columnas categoría y texto3, sin encontrar registros nulos o NAN.

4.3 Exploración de datos

Para explorar los datos se utilizan visualizaciones en base a la cantidad de palabras o frecuencia de las palabras en las distintas noticias. En las siguientes secciones se mostrarán dichas visualizaciones para el conjunto de datos como para las distintas categorías de noticias.

4.3.1 General

En este apartado se muestran visualizaciones para el conjunto de datos completo, sin distinguir por categoría.

Dentro de las palabras más utilizadas destaco las siguientes:

- Chile
- Gobierno
- Presidente

Queda claro que dentro del periodo en que se publicaron las noticias que son par del conjunto de datos se llevó a cabo un proceso de elección de presidente, La mayor parte de las noticias se enfocan en Chile.

En cuanto a los bigramas puedo destacar los siguientes:

- Sebastian Piñera
- Michelle Bahelet

Estructuras muy relacionadas con el gobierno de Chile y el proceso de elecciones que se llevo a cabo. En el caso de los trigramas también se observan referencias al presidente Sebastián Piñera reforzando lo indicado anteriormente.

Finalmente, la nube de palabras es una forma más visual de mostrar las palabras más usadas, en donde si descartamos las palabras más comunes (como decir, hacer, etc.), podemos encontrar palabras como Chile, presidente, gobierno, etc.

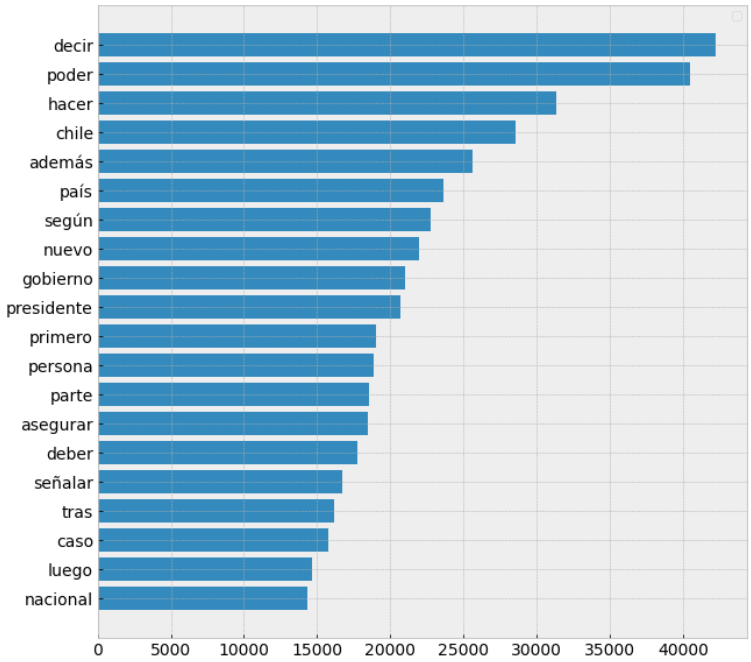


Figura 22: histograma de frecuencia de palabras para el conjunto de datos completo

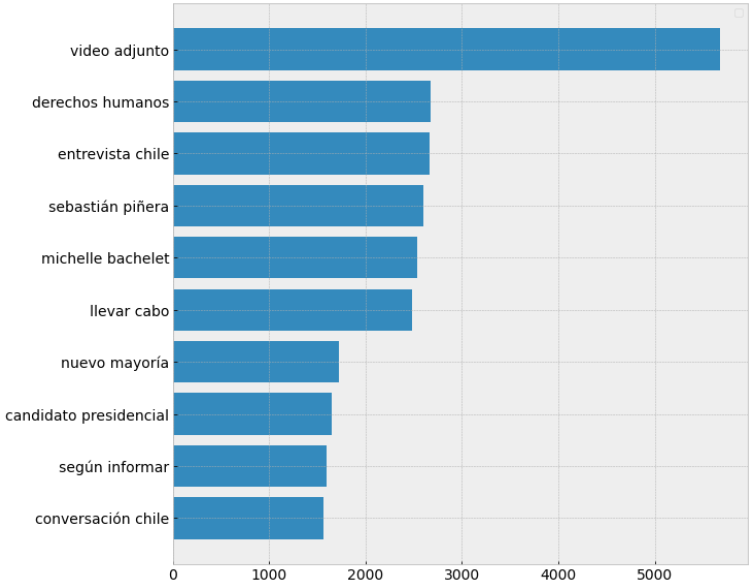


Figura 23: histograma de frecuencia de bigramas para el conjunto de datos completo

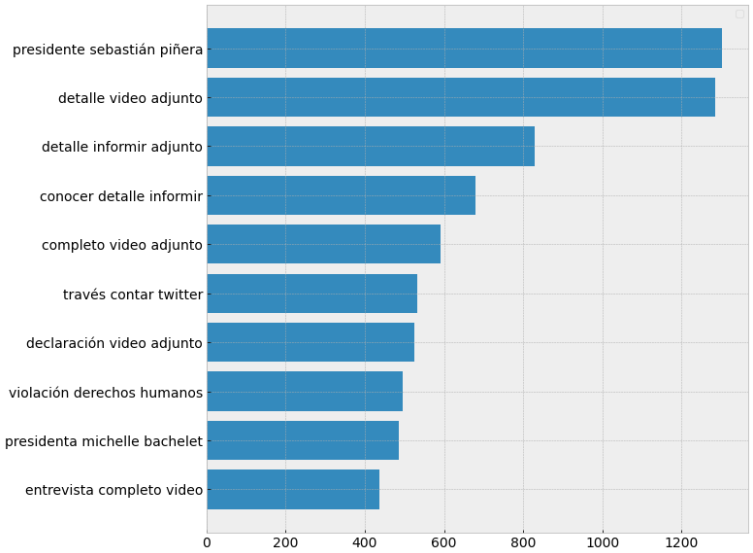


Figura 24: histograma de frecuencia de trigramas para el conjunto de datos completo

Figura 25: nube de palabras del conjunto de datos completo

4.3.2 Categoría País

En este apartado se muestran visualizaciones para el conjunto de datos asociado a la categoría País.

Dentro de las palabras más utilizadas destaco las siguientes, descartando las palabras más comunes (como decir, hacer, el, él, etc.):

- Chile
- Gobierno
- Presidente
- País

La conclusión es similar a la que se dio para el conjunto de datos completo, se observan grades referencias a un proceso de elección de un presidente y que gran parte del conjunto de noticas de la categoría de país hace referencia al gobierno.

En cuanto a los bigramas puedo destacar los siguientes:

- Sebastian Piñera
- Michelle Bahelet
- Derechos humanos

Se repiten los bigramas que predominaban en el conjunto de datos completo, sin embargo, se observa la presencia del bigrama derechos humanos, frase que es parte habitual de las noticias en Chile debido a su historia no tan reciente. En el caso de los trigramas también se observan referencias al presidente Sebastián Piñera y a la presidenta Michelle Bachelet.

La nube de palabra hace referencia a Chile, de forma consecuente con la categoría de noticias, al poder y gobierno, pudiendo concluir que el conjunto de datos tiene una fuerte relación con el gobierno y la política.

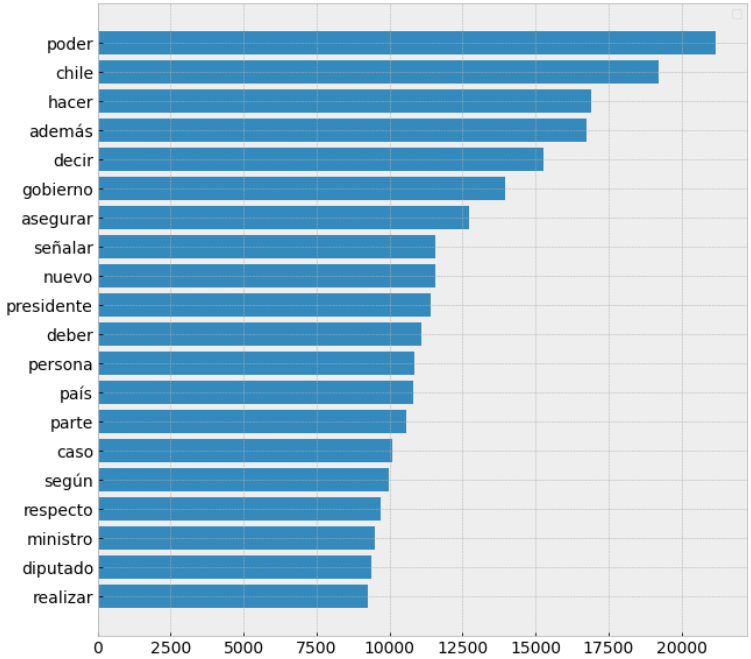


Figura 26: histograma de frecuencia de palabras para los datos asociados a la categoría país

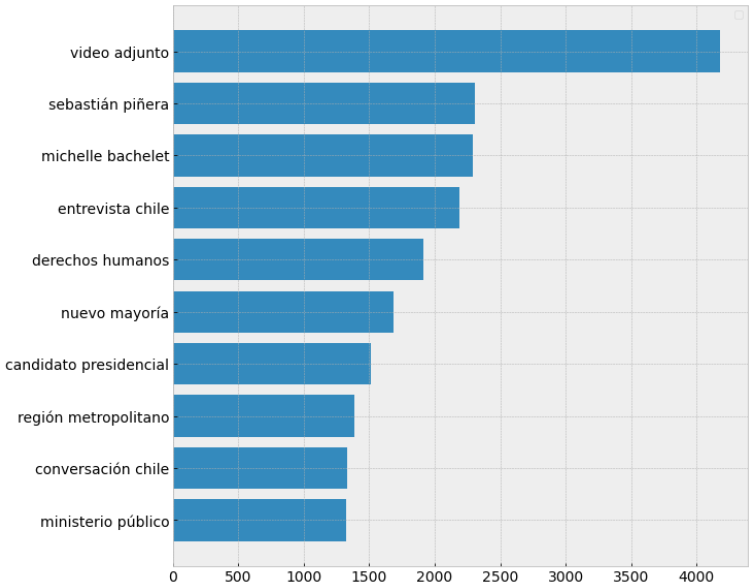


Figura 27: histograma de frecuencia de bigramas para los datos asociados a la categoría país

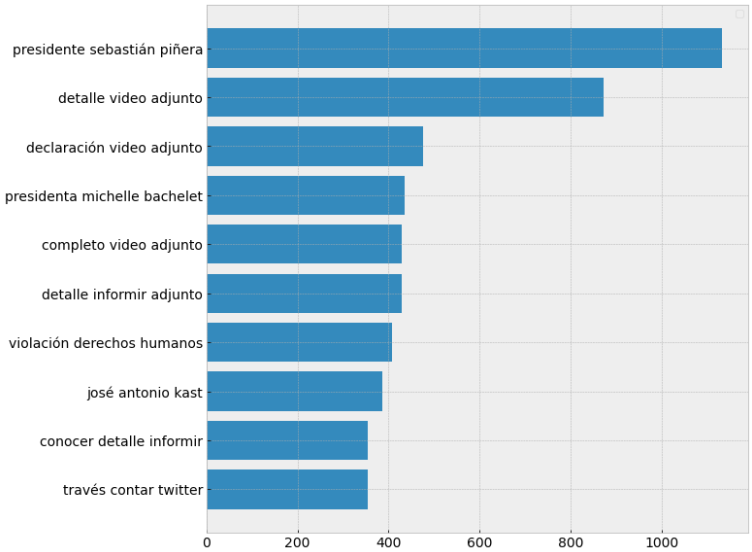


Figura 28: histograma de frecuencia de trigramas para los datos asociados a la categoría país

Figura 29: nube de palabras para los datos asociados a la categoría país

4.3.3 Categoría Mundo

En este apartado se muestran visualizaciones para el conjunto de datos asociado a la categoría Mundo.

Dentro de las palabras más utilizadas destaco las siguientes, descartando las palabras más comunes (como decir, hacer, el, él, etc.):

- Gobierno
- Presidente
- País
- Trump

En general la frecuencia de cada palabra por separado no entrega gran información. Se puede destacar que Chile ya no aparece como una palabra con gran frecuencia lo que es consecuente con la categoría de los datos analizados en este apartado Mundo, aparecen palabras como Trump lo que da cuenta de la presencia de EEUU en las noticias.

En cuanto a los bigramas puedo destacar los siguientes:

- Primer ministro
- Donal Trump
- Reino Unido
- Casa blanca
- Corea del Norte

Se observa que los bigramas tienen mayor cantidad de información, se observa la presencia de palabras asociadas a países o naciones como EEUU, Reino Unido, Corea del Norte, etc., en general las principales naciones que participan en la contingencia internacional y que producen la mayor cantidad de noticias en el periodo en que se publicaron los datos (2015 al 2022).

En el caso de los trigramas ya se observan conceptos como organización mundial de la salud y centro control prevención, conceptos muy asociados a la pandemia que se produjo dentro del periodo en que se genero el centro de datos.

En la nube de palabras se observa la presencia de varias palabras comunes como decir, hacer, año, etc., pero también se observan palabras como Trump, presidente, país, etc., palabras que tienen relación con la contingencia internacional.

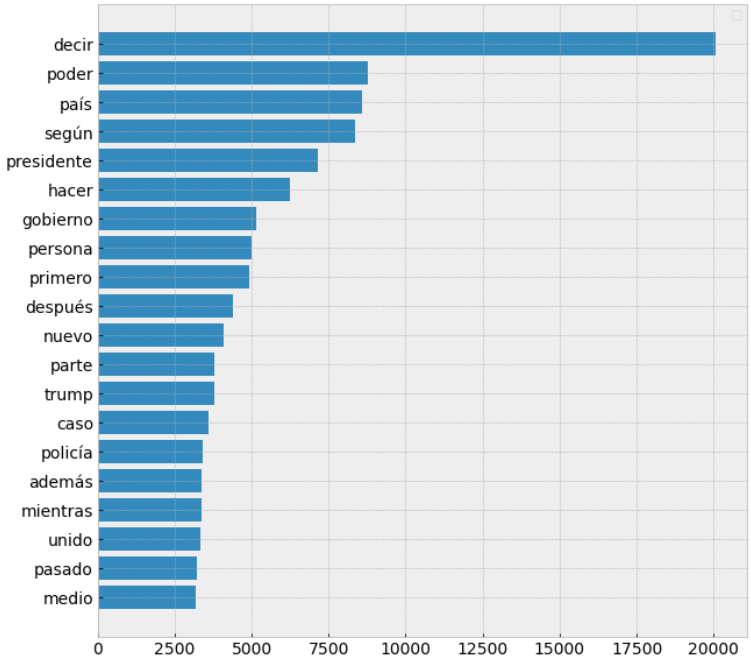


Figura 30: histograma de frecuencia de palabras para los datos asociados a la categoría mundo

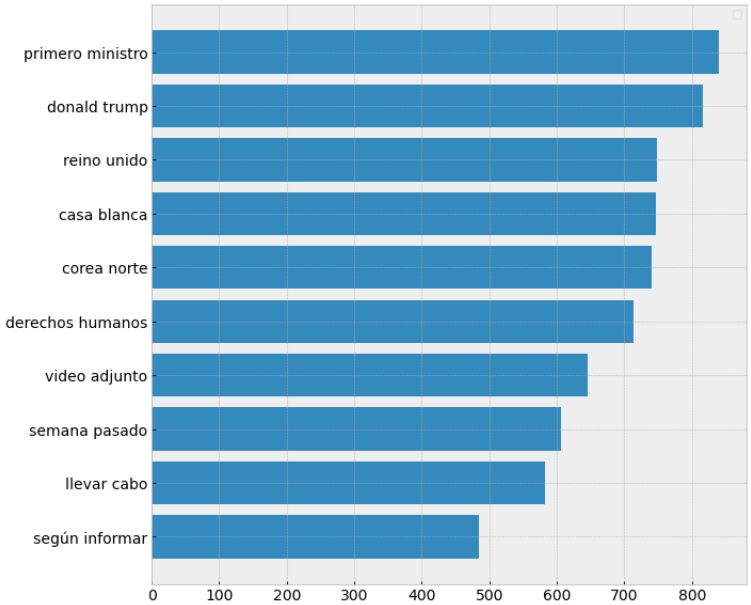


Figura 31: histograma de frecuencia de bigramas para los datos asociados a la categoría mundo

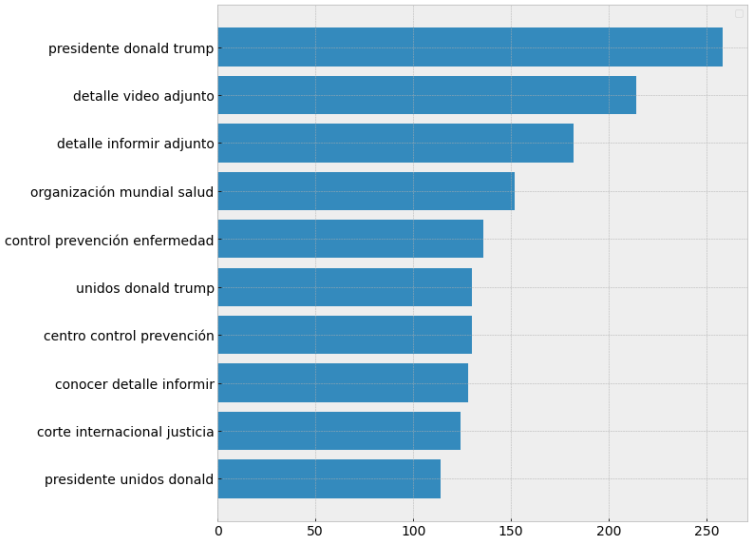


Figura 32: histograma de frecuencia de trigramas para los datos asociados a la categoría mundo

Figura 33: nube de palabras para los datos asociados a la categoría mundo

4.3.4 Categoría Economía

En este apartado se muestran visualizaciones para el conjunto de datos asociado a la categoría Economía.

Dentro de las palabras más utilizadas destaco las siguientes, descartando las palabras más comunes (como decir, hacer, el, él, etc.):

- Chile
- Empresa
- Económico
- Mercado
- Gobierno

Para esta categoría se observa que la frecuencia de cada palabra por separado tiene bastante relación con la categoría economía, destacando palabras como empresa, mercado y económico, distinguiendo la idea central de este subconjunto de datos.

En cuanto a los bigramas puedo destacar los siguientes:

- Banco de central
- Ministro hacienda
- Agenda económica
- Tasa interés

Se observa que los bigramas tienen mayor cantidad de información, se observa la presencia de palabras asociadas fuertemente al sector económico

En el caso de los trigramas no muestran gran cantidad de información predominando conceptos más genéricos, se observan conceptos asociados a economía, pero predominan palabras comunes como informe, entrevista, etc.

En la nube de palabras se observa la presencia de varias palabras comunes como decir, hacer, año, etc., pero también se observan palabras como Chile, económico, precio, fondo, etc.

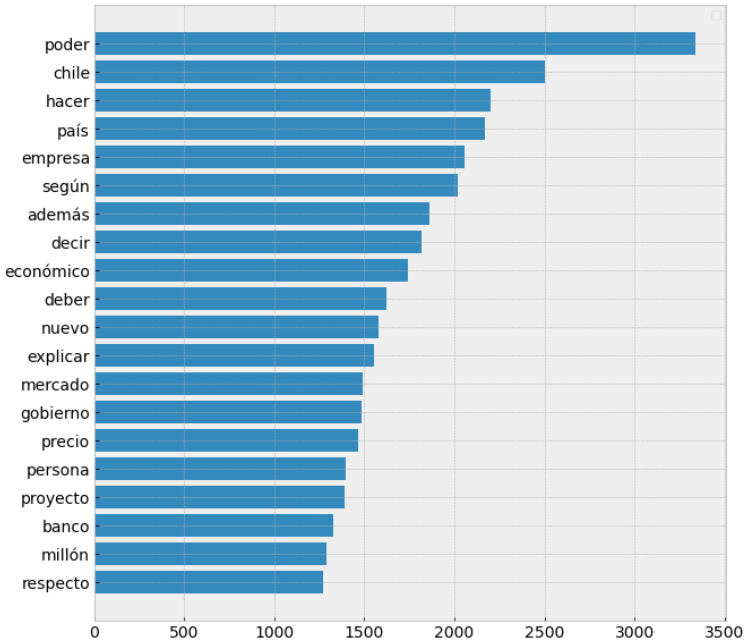


Figura 34: histograma de frecuencia de palabras para los datos asociados a la categoría economía

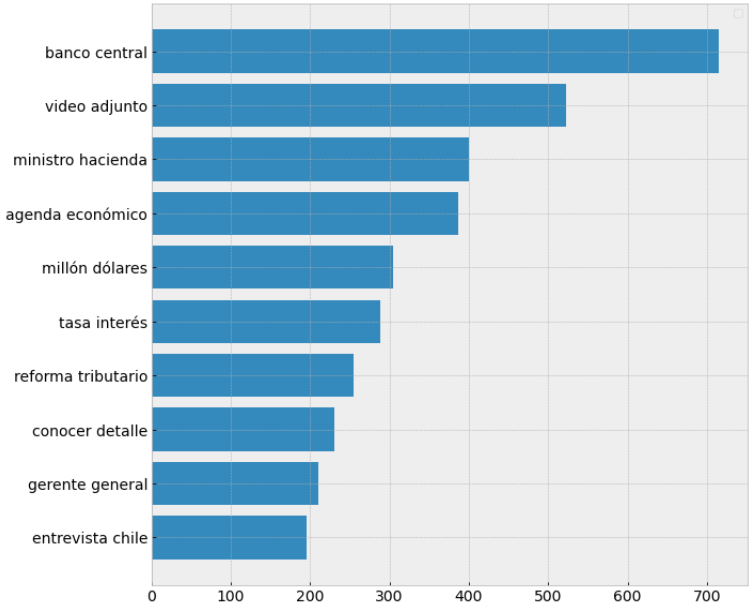


Figura 35: histograma de frecuencia de bigramas para los datos asociados a la categoría economía

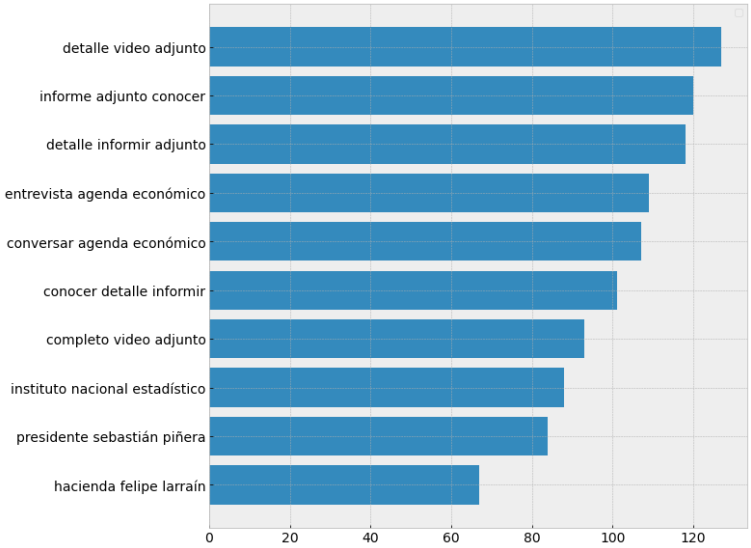


Figura 36: histograma de frecuencia de trigramas para los datos asociados a la categoría economía

Figura 37: nube de palabras para los datos asociados a la categoría economía

4.3.5 Categoría Deporte

En este apartado se muestran visualizaciones para el conjunto de datos asociado a la categoría Deporte.

Dentro de las palabras más utilizadas destaco las siguientes, descartando las palabras más comunes (como decir, hacer, el, él, etc.):

- Partido
- Equipo
- Jugador
- Futbol

Las frecuencias de cada palabra por separado evidencia claramente que la idea central del subconjunto de datos se relaciona con los deportes, en particular se observa la aparición de la palabra futbol lo cual refleja el predominio de este deporte a nivel nacional.

En cuanto a los bigramas puedo destacar los siguientes:

- Colo Colo
- Universidad de Chile
- Universidad Católica
- Futbol Chileno
- Copa América

Los bigramas reflejan la influencia del futbol chileno en las noticias que se asocian al concepto de deporte, mostrando los principales equipos de futbol de Chile. También se observa referencias a la copa américa suceso ocurrido durante el periodo en que se generaron los datos (2015 al 2022).

En el caso de los trigramas también se observan conceptos muy asociados al futbol chileno, como San Carlos de Apoquindo o Juan Pinto Duran, lo que refuerza lo indicado en los párrafos anteriores.

En la nube de palabras se observa la presencia de varias palabras comunes como decir, hacer, año, etc., pero también se observan palabras como partido, equipo jugador, etc.

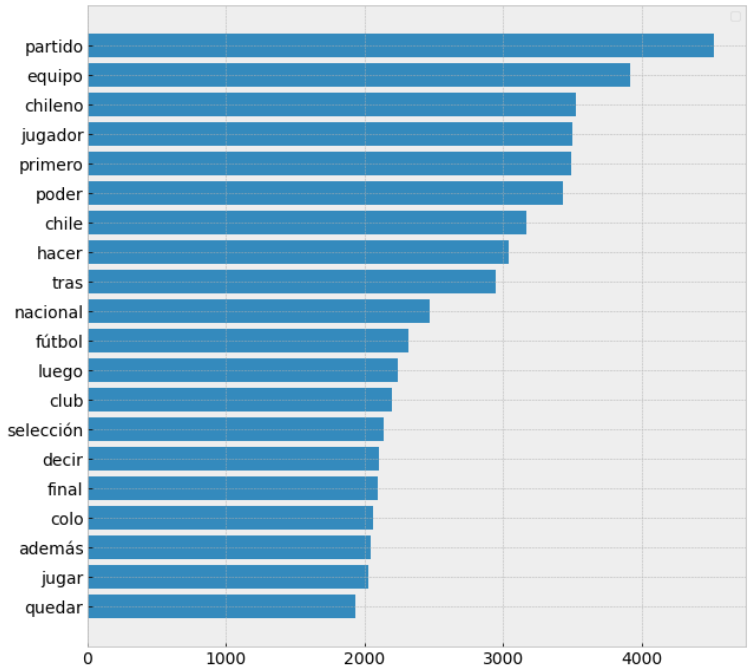


Figura 38: histograma de frecuencia de palabras para los datos asociados a la categoría deporte

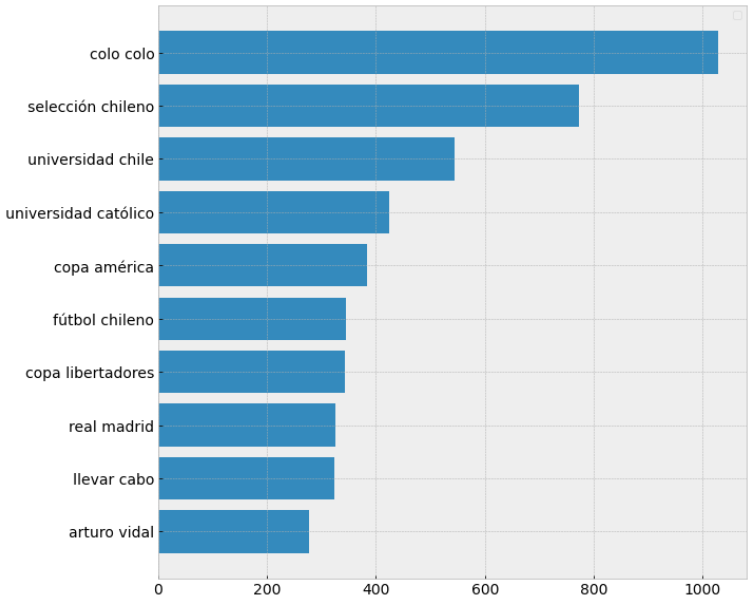


Figura 39: histograma de frecuencia de bigramas para los datos asociados a la categoría deporte

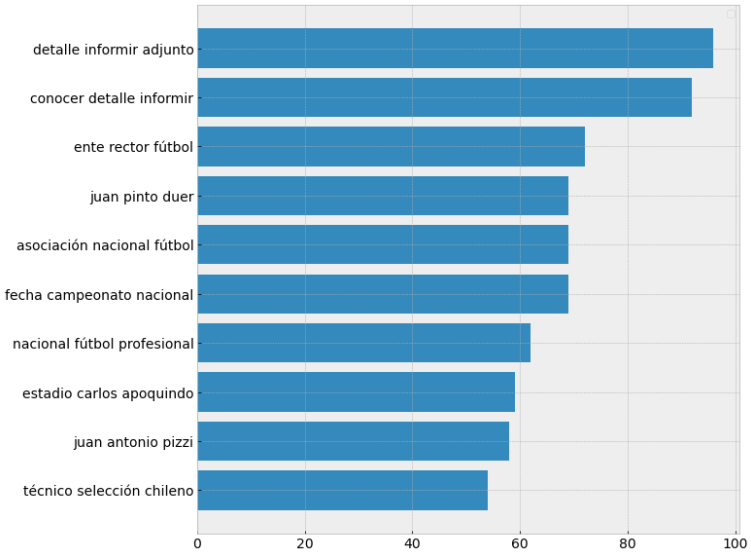


Figura 40: histograma de frecuencia de trigramas para los datos asociados a la categoría deporte

Figura 41: nube de palabras para los datos asociados a la categoría deporte

4.3.6 Categoría Cultura

En este apartado se muestran visualizaciones para el conjunto de datos asociado a la categoría Cultura.

Dentro de las palabras más utilizadas destaco las siguientes, descartando las palabras más comunes (como decir, hacer, el, él, etc.):

- Canción
- Artista
- Cine

Las frecuencias de cada palabra por separado muestran una relación con temas de culturas y arte, como es el cine y la música.

En cuanto a los bigramas puedo destacar los siguientes:

- Mejor pelicula
- Nuevo disco
- Nueva cancion

Los bigramas no reflejan directamente una relación con el concepto de cultura, quizás pueden acercarse al concepto de espectáculos más que cultura como la entendemos.

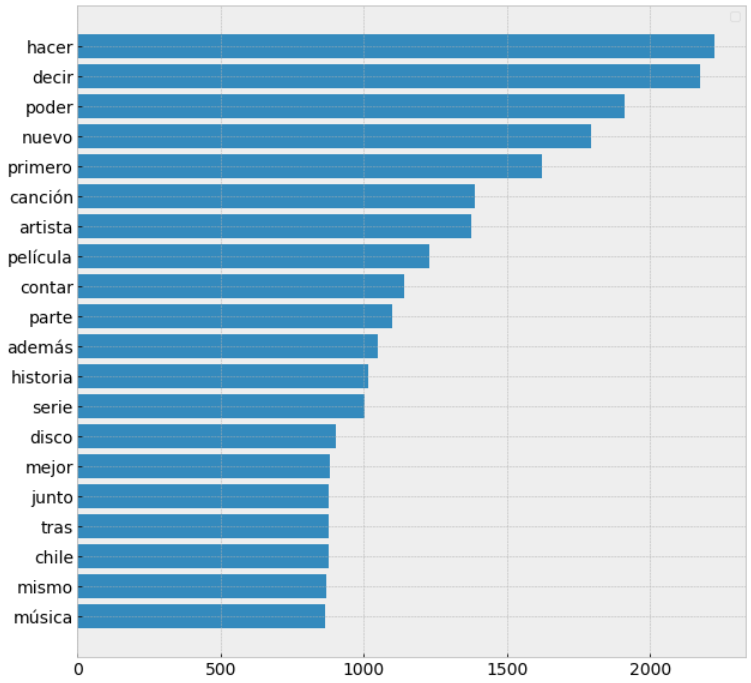


Figura 42: histograma de frecuencia de palabras para los datos asociados a la categoría cultura

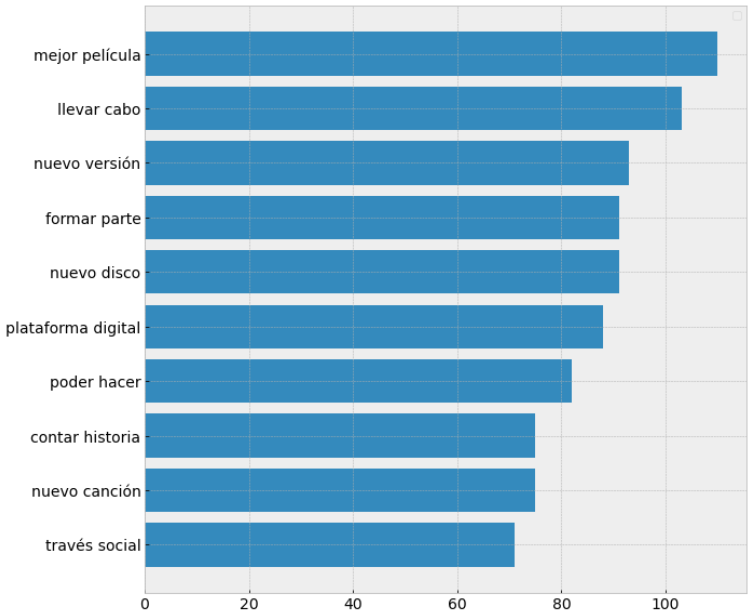


Figura 43: histograma de frecuencia de bigramas para los datos asociados a la categoría cultura

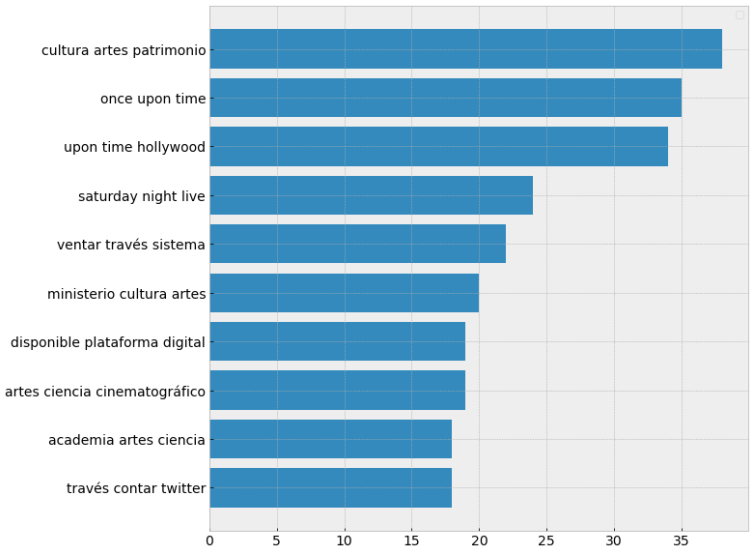


Figura 44: histograma de frecuencia de trigramas para los datos asociados a la categoría cultura

Figura 45: nube de palabras para los datos asociados a la categoría cultura

4.3.7 Categoría Tecnología

En este apartado se muestran visualizaciones para el conjunto de datos asociado a la categoría Tecnología.

Dentro de las palabras más utilizadas destaco las siguientes, descartando las palabras más comunes (como decir, hacer, el, él, etc.):

- Aplicación
- Nuevo
- Empresa

Las frecuencias de cada palabra por separado no muestran una gran relación con el concepto de tecnología, se pudieran relación con otros conceptos, no es evidente la relación.

En cuanto a los bigramas puedo destacar los siguientes:

- Sistema operativo
- Inteligencia artificial

Los bigramas si reflejan una relación directa con el concepto de tecnología, el tema de las redes sociales o la inteligencia artificial muestran que el subconjunto de datos está enfocado al concepto de tecnología.

En el caso de los trigramas también se observan conceptos muy asociados a la tecnología, inclusive se observa conceptos asociados a la investigación que realiza la universidad católica, lo que refuerza lo indicado en los párrafos anteriores.

En la nube de palabras se observa la presencia de varias palabras comunes como decir, hacer, año, etc., pero también se observan palabras como aplicación, usuario, apple, etc.

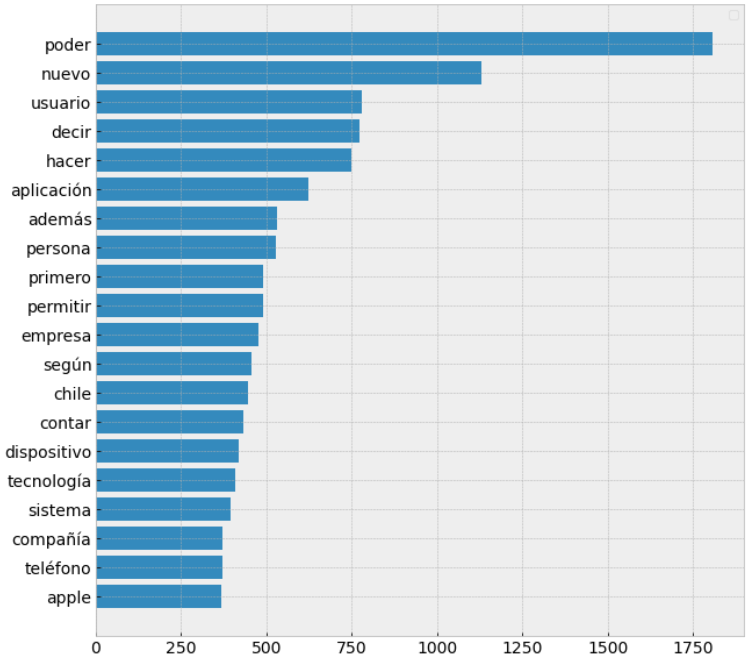


Figura 46: histograma de frecuencia de palabras para los datos asociados a la categoría tecnología

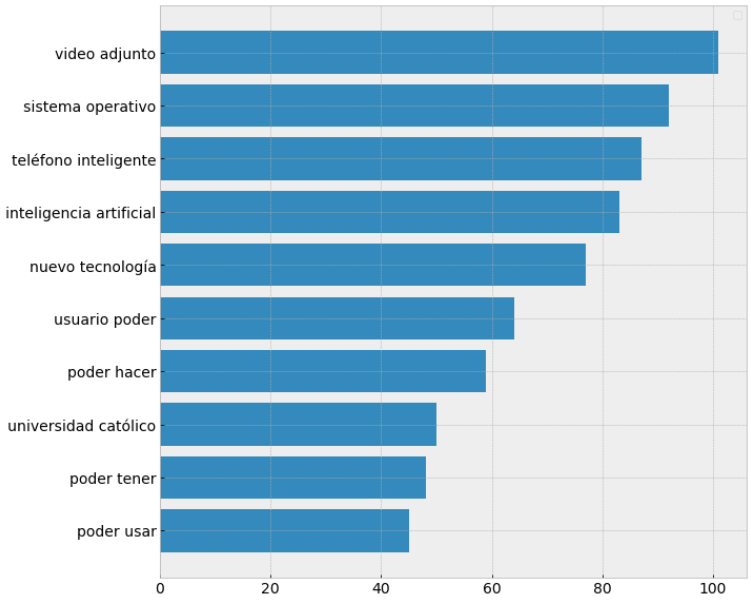


Figura 47: histograma de frecuencia de bigramas para los datos asociados a la categoría tecnología

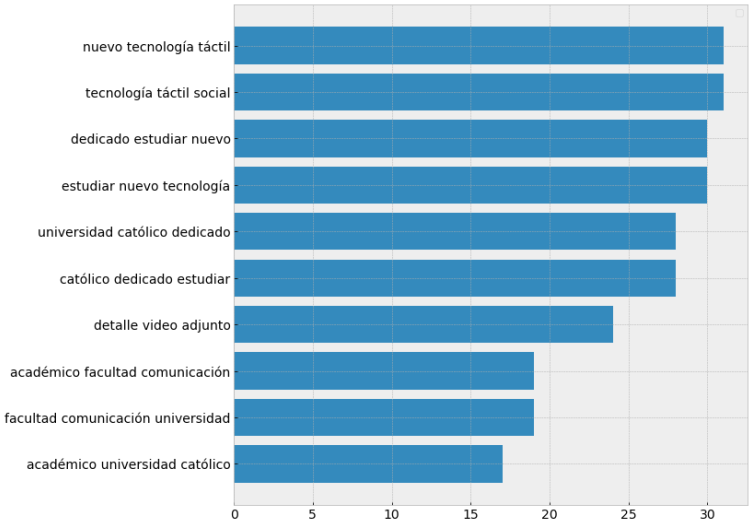


Figura 48: histograma de frecuencia de trigrama para los datos asociados a la categoría tecnología

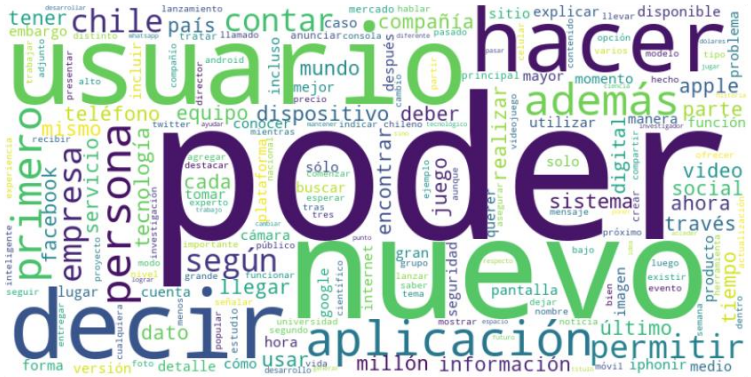


Figura 49: nube de palabras para los datos asociados a la categoría tecnología

Exploración datos históricos IPSA

Como se definió anteriormente el IPSA es el Índice de Precio Selectivo de Acciones, el principal indicador de la bolsa de valores de Chile y se utilizaran sus valores históricos desde el 2012 a la fecha para determinar si una noticia influye de manera positiva o negativa en los resultados de la bolsa de comercio, la idea es evidenciar si existe una relación entre las noticias publicadas y la variación del IPSA.

La hipótesis para relacionar el IPSA con las noticias se basa en el siguiente artículo (https://towardsdatascience.com/a-step-by-step-tutorial-for-conducting-sentiment-analysis-cf3e995e3171), por lo tanto, cada vez que se presente una variación diaria positiva del IPSA se asignara a las noticias de ese día un valor de 1 correspondiente a una influencia positiva, en caso que la variación diaria del IPSA sea negativa o igual a cero se asignara a la noticias de ese día un valor de 0 correspondiente a una influencia negativa.

Seguidamente se encontrará una visualización de los datos históricos del IPSA:

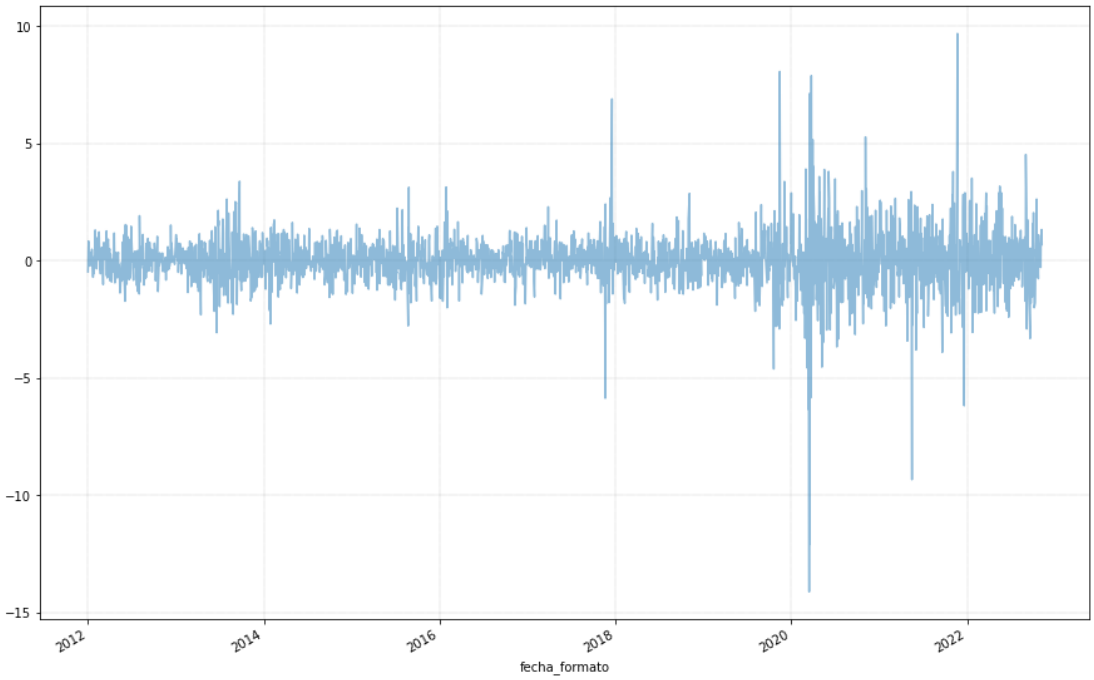


Figura 50: evolución del IPSA desde el 2012 a la fecha

En la siguiente imagen se realizó un resampleo de los datos con un muestreo mensual para permitir una mejor representación de los datos de manera grafica, en color azul se muestran los valores máximos mensuales, con color rojo se muestran los mínimos mensuales y en verde los valores medios mensuales:

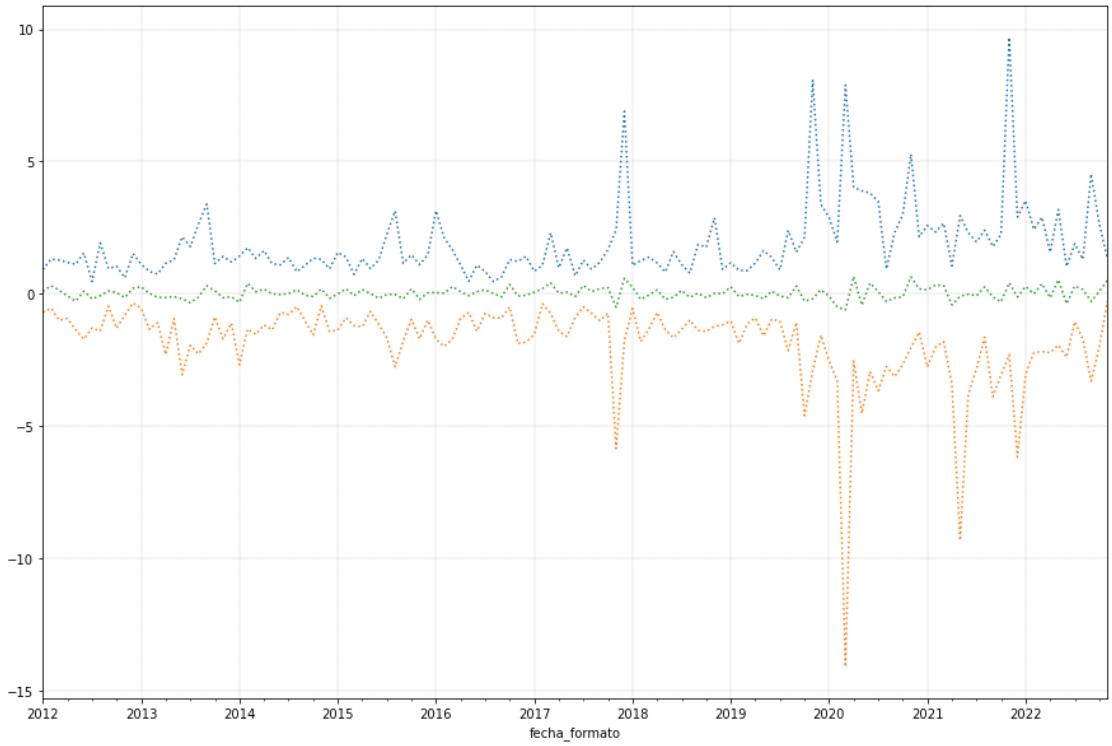


Figura 51: valor máximo, valor medio y valor mínimo durante cada mes para el IPSA

Para evaluar la hipótesis en principio se utilizará las noticias asociadas a la categoría "Economía", pero antes se realizará un análisis de clusterización para determinar si se encuentra otra forma de agrupar los datos, distinto a la clasificación que trae desde el sitio web CNN Chile, que pueda representar mejor el concepto de "Economía".

4.4 Análisis de clusterizacion

De acuerdo con lo indicado en la sección anterior antes de realizar la relación entre el IPSA y las categorías de las noticias, se realizará un análisis de clusterizacion para ver si existe una mejor manera de agrupar los datos. Para lo anterior se utilizará la librería sklearn y en particular su clase KMeans que permite agrupar datos como se explicó en la sección de fundamentos teóricos.

Para el cálculo optimo del número de clúster se utilizaron las dos metodologías descritas en la sección de fundamentos teóricos, el método de Elbow y el método de la silueta.

Continuamos definiendo una lista con las stopwords en español utilizando la librería nltk. Definimos la ruta en donde se guardarán los resultados y se crea un dataframe de la librería pandas para guardar dichos resultados.

Se utiliza la clase CountVectorizer transformar las palabras de cada documento del corpus o conjunto de datos en una representación numérica. Con el método fit_transform se convierten los datos de texto en una representación numérica. Se imprimen las dimensiones de la matriz de representaciones numéricas de los datos de texto y los nombres de sus características o tokens utilizados para contabilizar la frecuencia de cada palabra.

Se definen listas para almacenar los resultados de calcular los índices de suma de errores al cuadrado y del coeficiente de silueta.

Se inicia un ciclo for para buscar el numero óptimo de clústeres, ciclo que itera desde el valor 2 al 19. En cada iteración se define un numero de clústeres asociad al número de la iteración con la clase KMeas y se evalúa la suma de errores al cuadrado y el coeficiente de silueta almacenándolo en la lista previamente definida. Se imprimen resultados preliminares y se envía un mensaje indicando el número de clúster evaluado.

De igual forma, se utiliza la clase TfidfVectorizer para transformar las palabras de cada documento del corpus o conjunto de datos en una representación numérica. Con el método fit_transform se convierten los datos de texto en una representación numérica.

Se definen listas para almacenar los resultados de calcular los índices de suma de errores al cuadrado y del coeficiente de silueta.

Se inicia un ciclo for para buscar el numero óptimo de clústeres, ciclo que itera desde el valor 2 al 19. En cada iteración se define un numero de clústeres asociad al número de la iteración con la clase KMeas y se evalúa la suma de errores al cuadrado y el coeficiente de silueta almacenándolo en la lista previamente definida. Se imprimen resultados preliminares y se envía un mensaje indicando el número de clúster evaluado.

Se crea un diccionario con los resultados de los cálculos (listas) de la suma de errores al cuadrado y el coeficiente de silueta que se definió para las dos clases de transformación de texto en número (CountVectorizer y TfidfVectorizer). Con el diccionario se crea un dataframe de la librería pandas y se guardan los resultados en un archivo con extensión csv.

En las siguientes imágenes se puede observar los gráficos de suma de errores cuadrados y coeficiente de silueta para la clusterización realizadas con las clases CountVectorizer y TfidfVectorizer:

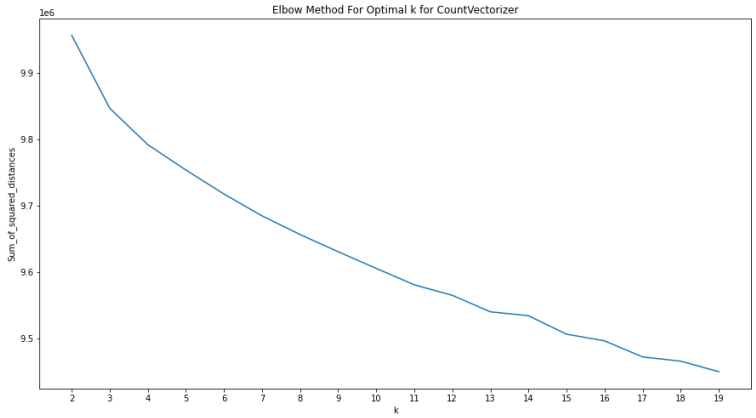


Figura 52: Grafica de suma de errores cuadráticos versus número de clusteres con CountVectorizer:

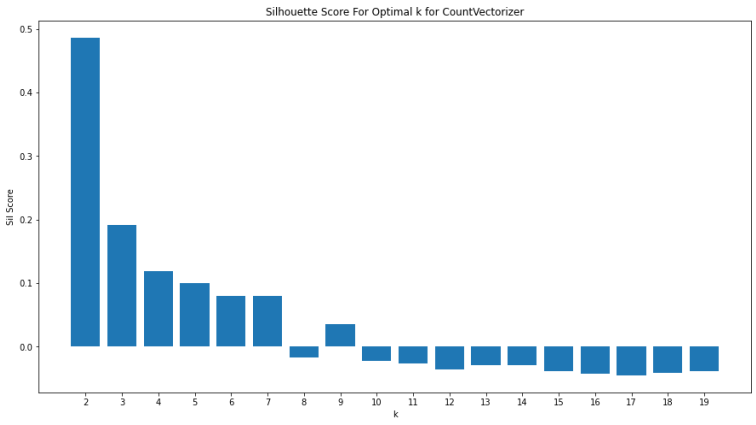


Figura 53: Grafica de coeficiente de silueta versus número de clústeres con CountVectorizer

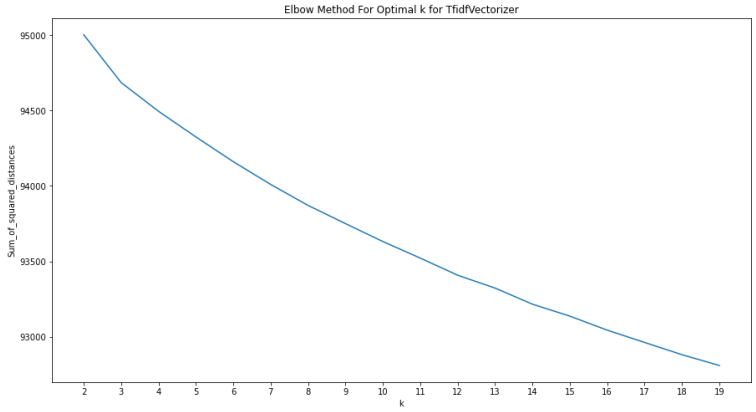


Figura 54: Grafica de suma de errores cuadráticos versus número de clusteres con TfidfVectorizer:

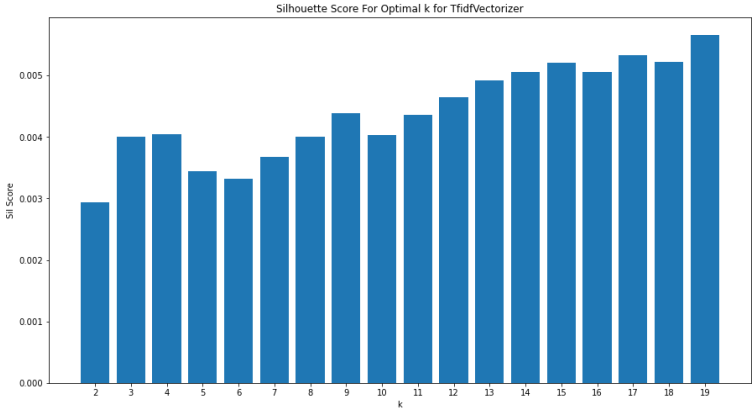


Figura 55: Grafica de coeficiente de silueta versus número de clústeres con TfidfVectorizer

Analizando las gráficas se determina utilizar como número óptimo de clústeres a cuatro (4) y se utilizará la clase TfidfVectorizer para la representación de los documentos. Se decide utilizar la representación TfidfVectorizer por sobre la representación CountVectorizer, dado que esta última

considera la frecuencia de cuanto se repite un token (en este caso una palabra) en cada documento, sin tomar en consideración la frecuencia del token en el corpus completo, en cambio TfidfVectorizer tienen en consideración pondera el token en base a su frecuencia en cada documento y en base a la frecuencia en el corpus completo.

Se utilizará la librería KMeans para realizar el proceso de clusterizacion. Se inicia el proceso importando las librerías y clases requeridas, en este caso se destaca la clase TfidfVectorizer para convertir el texto en una representación numérica. Con la librería nltk se descarga la lista de stopwors en español. Este proceso se debiera realizar una única vez y luego se llama a esta lista de palabras sin problemas.

Se crea un dataframe con el conjunto de datos previamente procesado (limpieza de texto, remover stopwords, lematización, etc.). Se procede a crear la representación numérica de los documentos del conjunto de datos mediante la clase TfidfVectorizer, creando una matriz de valores numéricos la cual servirá de entrada para el proceso de clusterizción. Seguidamente se define una constante con el número de clústeres deseado (en este caso 4) y se llama a la clase KMeans para generar el proceso de clusterización. Se le pasa como parámetro la matriz de representación numérica de los documentos y se entrena para obtener la clusterización. Finalmente se crea una columna en el conjunto de datos (dataframe) con la asignación de los números de clústeres que se encontraron.

4.4.1 Visualización de clusterización nuevas categorías

Clúster 0

En las gráficas de nube de palabras, bigramas y trigramas no se visualiza un concepto claro, se repiten palabras como adjunto, video, entrevista, etc., pudiendo relacionarse el clúster con el concepto de periodismo o noticias.

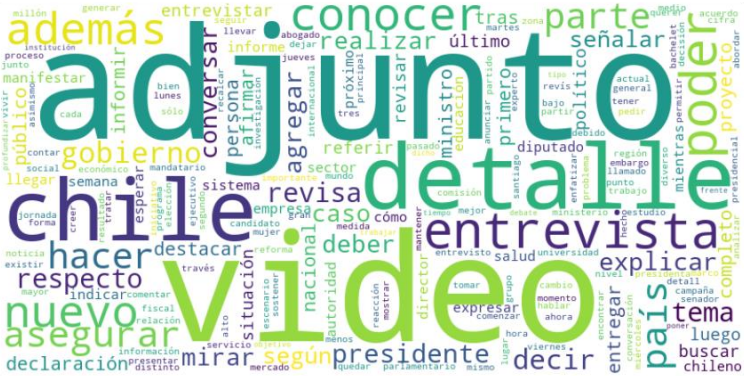


Figura 56: nube de palabras asociada a los datos del clúster 0

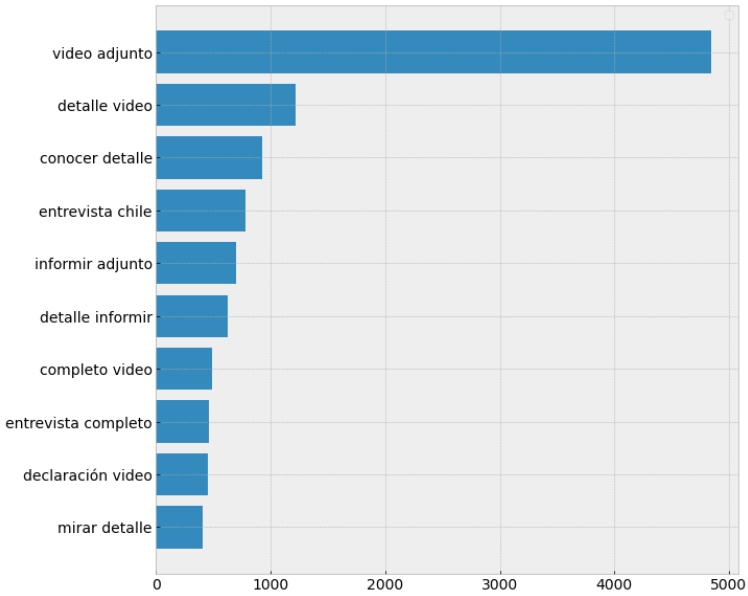


Figura 57: histograma de la frecuencia de bigramas asociada a los datos del clúster 0

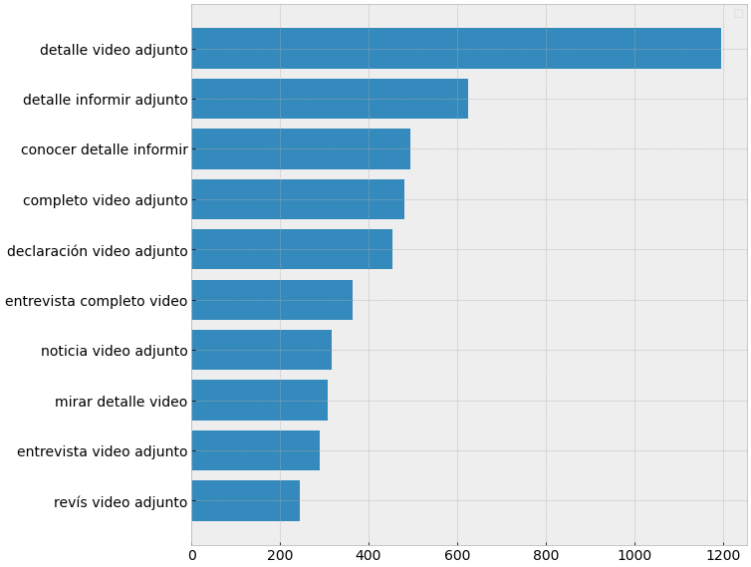


Figura 58: histograma de la frecuencia de trigramas asociada a los datos del clúster 0

Clúster 1

En este clúster se puede observar que la idea central ronda alrededor de la presidencia y el gobierno, mencionándose de manera repetitiva al ex presidente Sebastián Piñera, a algunos candidatos. También se observan ideas asociadas a la política, como diputado, político, senado, ministro, etc.

Figura 59: nube palabras asociada a los datos del clúster 1

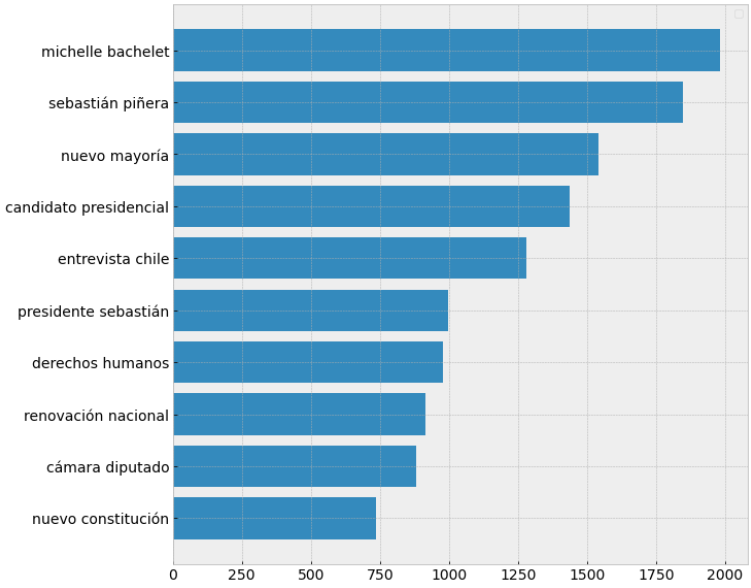


Figura 60: histograma de la frecuencia de bigramas asociada a los datos del clúster 1

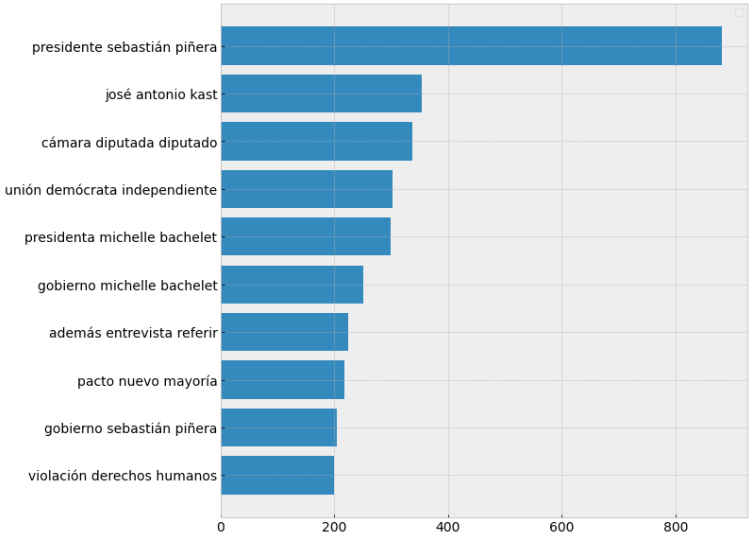


Figura 61: histograma de la frecuencia de trigramas asociada a los datos del clúster 1

Clúster 2

En este caso no se evidencia una idea central o predominante, se observan conceptos asociados a la política, a la salud, a las redes sociales, a entrevistas, etc., no se puede relacionar a un concepto o idea central.

Figura 62: nube de palabras asociada a los datos del clúster 2

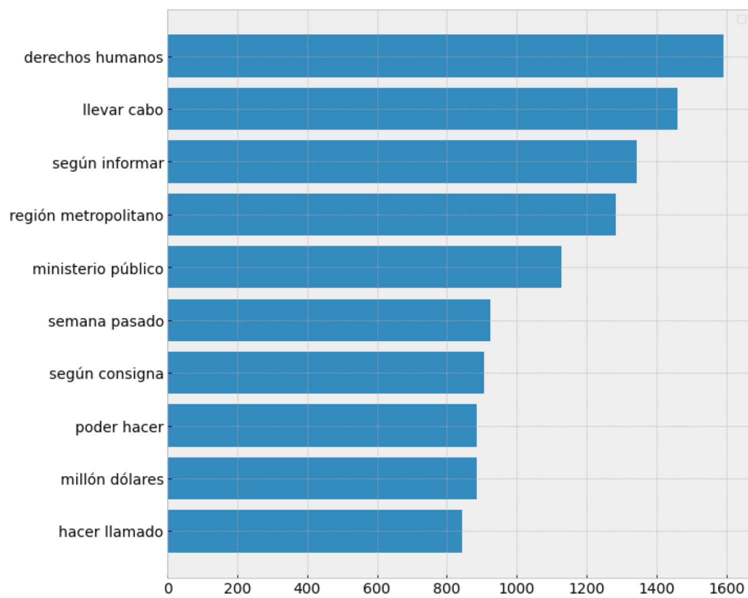


Figura 63: histograma de la frecuencia de bigramas asociada a los datos del clúster 2

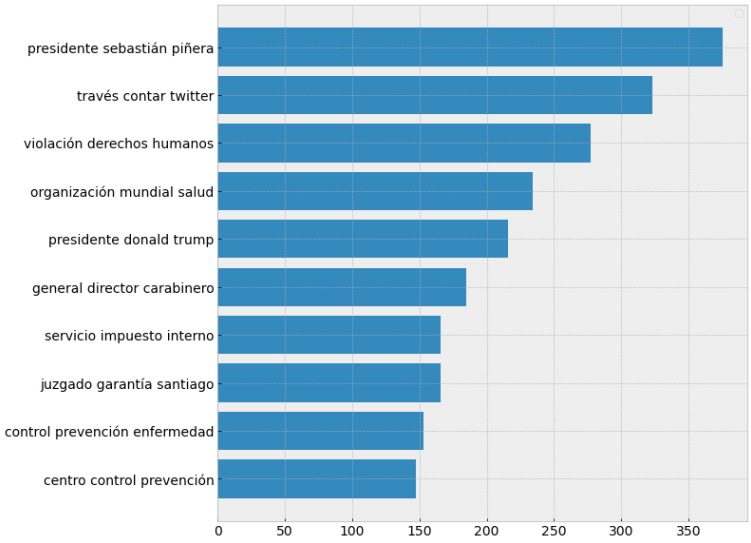


Figura 64: histograma de la frecuencia de trigramas asociada a los datos del clúster 2

Clúster 3

En el caso de este clúster se puede evidenciar una idea central, asociada al deporte nacional en Chile, claramente predominan conceptos como el futbol y el campeonato de dicho deporte en Chile.

Figura 65: nube de palabras asociada a los datos del clúster 3

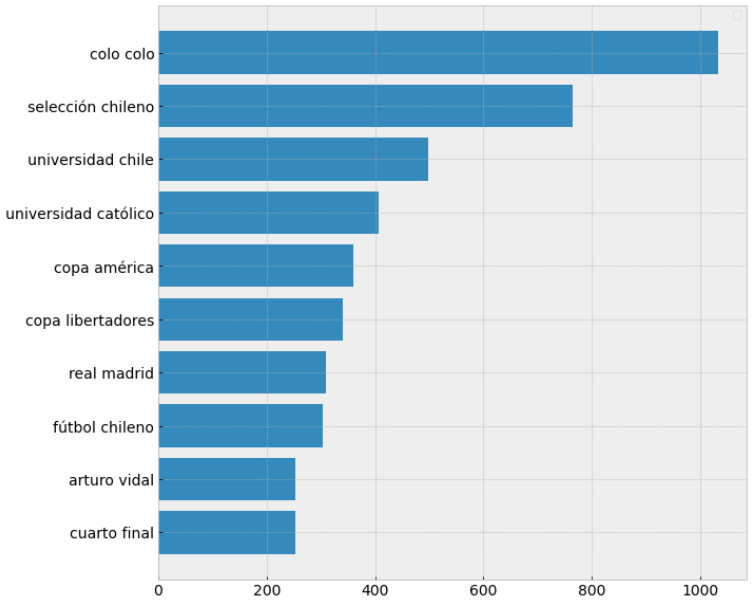


Figura 66: histograma de la frecuencia de bigramas asociada a los datos del clúster 3

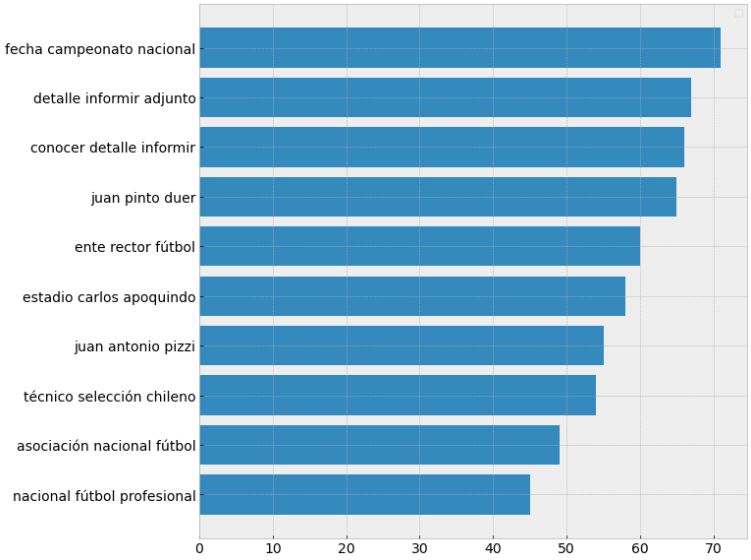


Figura 67: histograma de la frecuencia de trigramas asociada a los datos del clúster 3

Como conclusión de este análisis se puede observar que no se encontró una clusterización que permita agrupar noticias asociadas a la economía de mejor manera que las categorías originales del sitio web CNN Chile. Al comparar las visualizaciones de las categorías originales y de la clusterizacion probada, se evidencia que la categoría económica representa de mejor manera ese concepto, por lo tanto, se utilizará esa categoría como set de datos para revisar y analizar la relación entre las noticias de la categoría economía con la variación del índice IPSA.

4.5 Construcción de modelos

En primera instancia se combinan los conjuntos de datos asociados al IPSA y a las noticias.

Se define una función que asigna un 1 si la variación del IPSA es positiva y asigna un 0 si la variación es negativa o cero. La asignación se realiza a cada día del periodo bajo análisis. Esta función se denomina "clasificador" y se aplica a la columna "% var." del conjunto de dato IPSA creando una nueva columna llamada "ipsa". Se cambia el formato de la columna fecha_formato del conjunto de datos IPSA a un formato datatime de pandas. Se crea una función "combinarIpsa" para unificar los datos de variación del IPSA (1 o 0) con los datos de las noticias. La función busca una cierta fecha en los datos del IPSA y devuelve el valor de clasificación de la variación del IPSA (1 o 0). Se aplica la función "combinarIpsa" creando una columna nueva en el conjunto de datos de noticias. En dicha columna se muestra si la variación del IPSA fue positiva (1) o fue negativa o cer0 (0). Luego los datos son almacenados para poder trabajar con los distintos modelos de clasificación.

En esta sección se explicará la distinta construcción de varios modelos de clasificación que se han evaluado para analizar la relación entre las noticias y el IPSA.

En primera instancia se importan todas las librerías necesarias para construir los modelos de clasificación (sklearn), las librerías para cargar los datos (pandas) y la librería para graficar los datos (matplotlib).

Se importan los datos, en este caso se trabajó en Google Colab para acelerar el proceso de entrenamiento de los modelos. Se asigna las variables de entrada "X" que son las noticias y las variables de salida "y" que es la variación del IPSA, filtrando por la categoría "economía". Se asigna una variable random_state la cual permite separar los datos (entrenamiento y prueba) de manera aleatoria, si fijamos su valor la separación aleatoria se podrá reproducir cada vez que se ejecute la sentencia que genera los sets de datos de entrenamiento y prueba.

Se crea un pipeline que es una forma manejable de aplicar una serie de transformaciones de datos seguidas por la aplicación de un estimador. Este pipeline integra las funciones de CountVectorizer y TfidfTransformer que transforman los documentos de texto a una representación numérica utilizando la medida numérica "frecuencia de término – frecuencia inversa de documento". Este mismo paso se repetirá en cada clasificador que se utilice más adelante.

Se incluye en el pipeline un clasificador en este caso RandomForestClassifier, clasificador 1, de la librería sklearn explicado en la sección de fundamentos teóricos.

Se realiza el entrenamiento del clasificador previamente definido con el pipeline. Se imprimen los resultados y medidas o métricas del modelo:

```
[ ] y_pred1 = clf1.predict(X_test)
    print(metrics.classification_report(y_test, y_pred1))
    print(confusion_matrix(y_test, y_pred1))
    print(precision_score(y_test, y_pred1))
```

```
              precision    recall  f1-score   support

           0       0.54      0.58      0.56       575
           1       0.51      0.46      0.48       539

    accuracy                           0.52      1114
   macro avg       0.52      0.52      0.52      1114
weighted avg       0.52      0.52      0.52      1114

[[335 240]
 [291 248]]
0.5081967213114754
```

```
[ ] print(clf1.score(X_train, y_train))
    print(clf1.score(X_test, y_test))
```

```
1.0
0.5233393177737882
```

Figura 68: resultados de clasificador 1

Se utiliza la clase pipeline para generar un nuevo clasificador, clasificador 2. La transformación del texto se realiza nuevamente con las clases CountVectorizer y TfidfTransformer (como se explicó anteriormente). Se incluye en el pipeline un clasificador en este caso LogisticRegression de la librería sklearn explicado en la sección de fundamentos teóricos, luego se entrena el modelo. Se imprimen los resultados y medidas o métricas del modelo:

```
[ ] y_pred2 = clf2.predict(X_test)
    print(metrics.classification_report(y_test, y_pred2))
    print(confusion_matrix(y_test, y_pred2))
    print(precision_score(y_test, y_pred2))
```

```
              precision    recall  f1-score   support

           0       0.54      0.53      0.54       575
           1       0.51      0.51      0.51       539

    accuracy                           0.52      1114
   macro avg       0.52      0.52      0.52      1114
weighted avg       0.52      0.52      0.52      1114

[[307 268]
 [265 274]]
0.5055350553505535
```

```
[ ] print(clf2.score(X_train, y_train))
    print(clf2.score(X_test, y_test))
```

```
0.8602875112309074
0.5215439856373429
```

Figura 69: resultados del clasificador 2

Se utiliza la clase pipeline para generar un nuevo clasificador, clasificador 3. La transformación del texto se realiza nuevamente con las clases CountVectorizer y TfidfTransformer (como se explicó anteriormente). Se incluye en el pipeline un clasificador en este caso LinearSVC de la librería sklearn explicado en la sección de fundamentos teóricos, luego se entrena el modelo. Se imprimen los resultados y medidas o métricas del modelo:

```
[ ] y_pred3 = clf3.predict(X_test)
    print(metrics.classification_report(y_test, y_pred3))
    print(confusion_matrix(y_test, y_pred3))
    print(precision_score(y_test, y_pred3))
```

```
                  precision    recall  f1-score   support

               0       0.54      0.53      0.53       575
               1       0.50      0.51      0.51       539

        accuracy                           0.52      1114
       macro avg       0.52      0.52      0.52      1114
    weighted avg       0.52      0.52      0.52      1114

    [[303 272]
     [262 277]]
    0.5045537340619308
```

```
[ ] print(clf3.score(X_train, y_train))
    print(clf3.score(X_test, y_test))
```

```
    0.9773135669362084
    0.5206463195691203
```

Figura 70: resultados del clasificador 3

Se utiliza la clase pipeline para generar un nuevo clasificador, clasificador 4. La transformación del texto se realiza nuevamente con las clases CountVectorizer y TfidfTransformer (como se explicó anteriormente). Se incluye en el pipeline un clasificador en este caso MultinomialNB de la librería sklearn explicado en la sección de fundamentos teóricos, luego se entrena el modelo. Se imprimen los resultados y medidas o métricas del modelo:

```
[ ] y_pred4 = clf4.predict(X_test)
    print(metrics.classification_report(y_test, y_pred4))
    print(confusion_matrix(y_test, y_pred4))
    print(precision_score(y_test, y_pred4))
```

```
                  precision    recall  f1-score   support

               0       0.52      0.50      0.51       575
               1       0.48      0.50      0.49       539

        accuracy                           0.50      1114
       macro avg       0.50      0.50      0.50      1114
    weighted avg       0.50      0.50      0.50      1114

    [[285 290]
     [267 272]]
    0.48398576512455516
```

```
[ ] print(clf4.score(X_train, y_train))
    print(clf4.score(X_test, y_test))
```

```
    0.9602425876010782
    0.5
```

Figura 71: resultados del clasificador 4

Se utiliza la clase pipeline para generar un nuevo clasificador, clasificador 5. La transformación del texto se realiza nuevamente con las clases CountVectorizer y TfidfTransformer (como se explicó anteriormente). Se incluye en el pipeline un clasificador en este caso LogisticRegression utilizando el parámetro solver asignando su valor a “lbfgs”, clase de la librería sklearn explicado en la sección de fundamentos teóricos, luego se entrena el modelo. Se imprimen los resultados y medidas o métricas del modelo:

```
[ ] y_pred5 = clf5.predict(X_test)
    print(metrics.classification_report(y_test, y_pred5))
    print(confusion_matrix(y_test, y_pred5))
    print(precision_score(y_test, y_pred5))
```

```
                precision    recall  f1-score   support

             0       0.54      0.53      0.54       575
             1       0.51      0.51      0.51       539

      accuracy                           0.52      1114
     macro avg       0.52      0.52      0.52      1114
  weighted avg       0.52      0.52      0.52      1114

  [[307 268]
   [265 274]]
  0.5055350553505535
```

```
[ ] print(clf5.score(X_train, y_train))
    print(clf5.score(X_test, y_test))
```

```
  0.8602875112309074
  0.5215439856373429
```

Figura 72: resultados del clasificador 5

Se utiliza la clase pipeline para generar un nuevo clasificador, clasificador 6. La transformación del texto se realiza nuevamente con las clases CountVectorizer y TfidfTransformer (como se explicó anteriormente). Se incluye en el pipeline un clasificador en este caso SVC de la librería sklearn explicado en la sección de fundamentos teóricos, luego se entrena el modelo. Se imprimen los resultados y medidas o métricas del modelo:

```
[ ] y_pred6 = clf6.predict(X_test)
    print(metrics.classification_report(y_test, y_pred6))
    print(confusion_matrix(y_test, y_pred6))
    print(precision_score(y_test, y_pred6))
```

```
                precision    recall  f1-score   support

             0       0.51      0.51      0.51       575
             1       0.47      0.47      0.47       539

      accuracy                           0.49      1114
     macro avg       0.49      0.49      0.49      1114
  weighted avg       0.49      0.49      0.49      1114

  [[293 282]
   [284 255]]
  0.4748603351955307
```

```
[ ] print(clf6.score(X_train, y_train))
    print(clf6.score(X_test, y_test))
```

```
  0.993486073674753
  0.4919210053859964
```

Figura 73: resultados clasificador 6

Se utiliza la clase pipeline para generar un nuevo clasificador, clasificador 7. La transformación del texto se realiza nuevamente con las clases CountVectorizer y TfidfTransformer (como se explicó anteriormente). Se incluye en el pipeline un clasificador en este caso MLPClassifier de la librería sklearn, esta clase implementa una red neuronal, luego se entrena el modelo. Se imprimen los resultados y medidas o métricas del modelo:

```
[ ] y_pred7 = clf7.predict(X_test)
    print(metrics.classification_report(y_test, y_pred7))
    print(confusion_matrix(y_test, y_pred7))
    print(precision_score(y_test, y_pred7))
```

```
              precision    recall  f1-score   support

           0       0.54      0.52      0.53       575
           1       0.50      0.52      0.51       539

    accuracy                           0.52      1114
   macro avg       0.52      0.52      0.52      1114
weighted avg       0.52      0.52      0.52      1114

[[298 277]
 [258 281]]
0.503584229390681
```

```
[ ] print(clf7.score(X_train, y_train))
    print(clf7.score(X_test, y_test))
```

```
1.0
0.5197486535008977
```

Figura 74: resultados clasificador 7

Para el octavo clasificador se utiliza la clase pipeline para generar un nuevo clasificador. La transformación del texto se realiza nuevamente con las clases CountVectorizer y TfidfTransformer (como se explicó anteriormente) y se vuelve a utilizar el clasificador RandomForestClassifier, sin embargo, para intentar mejorar su desempeño se utilizará GridSearchCV, la cual permite evaluar y seleccionar de forma sistemática los parámetros de un modelo y se encuentra disponible en la librería sklearn:

Primero se define la clase pipeline utilizando las clases CountVectorizer, TfidfTransformer y RandomForestClassifier. Se define el parámetro verbose = True para imprimir los resultados de las distintas iteraciones que realizara el modelo

Seguidamente se definen los distintos parámetros del modelo, para las clases TfidfTransformer y RandomForestClassifier (https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html) integradas en el pipeline:

- TfidfTransformer:
  - norm, puede tomar dos valores:
    - l2: La suma de los cuadrados de los elementos vectoriales es 1. La similitud del coseno entre dos vectores es su producto escalar cuando se ha aplicado la norma l2.
    - l1: La suma de los valores absolutos de los elementos vectoriales es 1.
  - use_idf: habilita la reponderación de frecuencia de documento inversa, puede ser True o False.
  - smooth_idf: suaviza los pesos de la matriz idf agregando un uno a las frecuencias de los documentos, como si se viera un documento adicional que contiene todos los términos de la colección exactamente una vez, evita las divisiones por cero, puede ser True o False.
  - sublinear_tf: aplica la escala tf sublineal, es decir, reemplaza tf con 1 + log(tf).
- RandomForestClassifier:
  - n_estimators: número de árboles de clasificadores en el bosque de clasificadores.

- max_depth: La profundidad máxima del árbol. Si es None, los nodos se expanden hasta que todas las hojas sean puras o hasta que todas las hojas contengan menos de min_samples_split samples.
- min_samples_split: el número mínimo de muestras requeridas para dividir un nodo interno.
- min_samples_leaf: el número mínimo de muestras requeridas para estar en un nodo hoja. Un punto de división a cualquier profundidad solo se considerará si deja al menos min_samples_leaf muestras de entrenamiento en cada una de las ramas izquierda y derecha. Esto puede tener el efecto de suavizar el modelo, especialmente en regresión.
- class_weight: pesos asociados a clases en el formulario. Si no se proporciona, se supone que todas las clases tienen peso uno.

Se define la clase GridSearchCV con el pipeline y los parámetros previamente definidos. Adicionalmente se asigna al parámetro verbose el valor de 3 para imprimir los distintos resultados en cada iteración. Seguidamente se procede a entrenar el modelo. Se imprimen los resultados de la búsqueda de parámetros óptimos:

```
[ ] gs.best_params_

    {'classification__class_weight': {0: 1, 1: 1},
     'classification__max_depth': 30,
     'classification__min_samples_leaf': 3,
     'classification__min_samples_split': 3,
     'classification__n_estimators': 120,
     'feature_extraction__norm': 'l2',
     'feature_extraction__smooth_idf': False,
     'feature_extraction__sublinear_tf': True,
     'feature_extraction__use_idf': True}

[ ] gs.score(X_train, y_train)

    0.9339622641509434

[ ] gs.score(X_test, y_test)

    0.5107719928186715
```

Figura 75: mejores parámetros para el clasificador 8

Con los parámetros previamente encontrados con la clase GridSearchCV se procede construir el octavo clasificador y se realiza su entrenamiento. Seguidamente se encontrarán los resultados del entrenamiento del octavo modelo:

```
[19] y_pred8 = clf8.predict(X_test)
     print(metrics.classification_report(y_test, y_pred8))
     print(confusion_matrix(y_test, y_pred8))
     print(precision_score(y_test, y_pred8))

                  precision    recall  f1-score   support

               0       0.54      0.56      0.55       575
               1       0.51      0.49      0.50       539

        accuracy                           0.53      1114
       macro avg       0.53      0.53      0.53      1114
    weighted avg       0.53      0.53      0.53      1114

    [[324 251]
     [276 263]]
    0.5116731517509727

[20] print(clf8.score(X_train, y_train))
     print(clf8.score(X_test, y_test))

    0.9409254267744834
    0.5269299820466786
```

Figura 76: resultados del clasificador 8

4.6 Resumen de resultados

En la siguiente tabla se puede encontrar un resumen de los resultados de los distintos clasificadores que se probaron para analizar la relación entre el IPSA y las noticias de la categoría "economía":

Tipo Clasificador	Score entrenamiento	Score prueba
Clasificador 1: RandomForestClassifier	1	0,5233
Clasificador 2: LogisticRegression	0,8603	0,5215
Clasificador 3: LinearSVC	0,9773	0,5206
Clasificador 4: MultinomialNB	0,9602	0,5
Clasificador 5: LogisticRegression con solve = 'lbfgs'	0,8603	0,5215
Clasificador 6: SVC	0,9935	0,4919
Clasificador 7: MLPClassifier	1	0,5197
Clasificador 8: RandomForestClassifier con GridSearchCV	0,9409	0,5269

Tabla 1: comparación de resultados de los distintos clasificadores

5 Análisis de Resultados

Los algoritmos de aprendizaje automático lineal ajustan un modelo en el que la predicción (variable de salida) es la suma ponderada de los valores de las variables de entrada. Estos algoritmos incluyen regresión lineal, regresión logística y extensiones que agregan regularización. Todos estos algoritmos encuentran un conjunto de coeficientes que ponderan a las variables de entrada para calcular la variable de salida como la suma ponderada. Estos coeficientes se pueden utilizar directamente como puntaje para indicar la importancia de la variable de entrada o característica.

Para determinar cómo influyen las variables de entrada, es decir, las palabras que componen las noticias de la categoría "economía", utilizaremos el clasificador regresión logística del cual podemos obtener un conjunto de coeficientes mediante el atributo coeff_, utilizando la implementación de la librería scikit learn, cada coeficiente se puede asociar con una variable de entrada.

Los coeficientes se pueden utilizar como una puntuación de importancia de cada variable o característica de entrada. Esto supone que las variables de entrada tienen la misma escala, lo cual es efectivo al utilizar las clases CountVectorizer y TfidfTransformer al transformar las palabras a números.

Se debe tener presente que el problema de clasificación tratado en este trabajo tiene dos clases 0 y 1, adicionalmente los coeficientes que entrega el atributo coeff_ del clasificador regresión logística son positivos y negativos. Dado lo anterior los coeficientes o puntuaciones positivas indican que la variable o característica de entrada predice la clase 1, mientras que las puntuaciones negativas indican la variable o característica de entrada predice la clase 0.

Noticias con efecto positivo

En la siguiente figura se observan las palabras o características de entrada que tienen una puntación positiva y por lo tanto influyen en la clase 1 o variación positiva del IPSA al ajustar el clasificador:

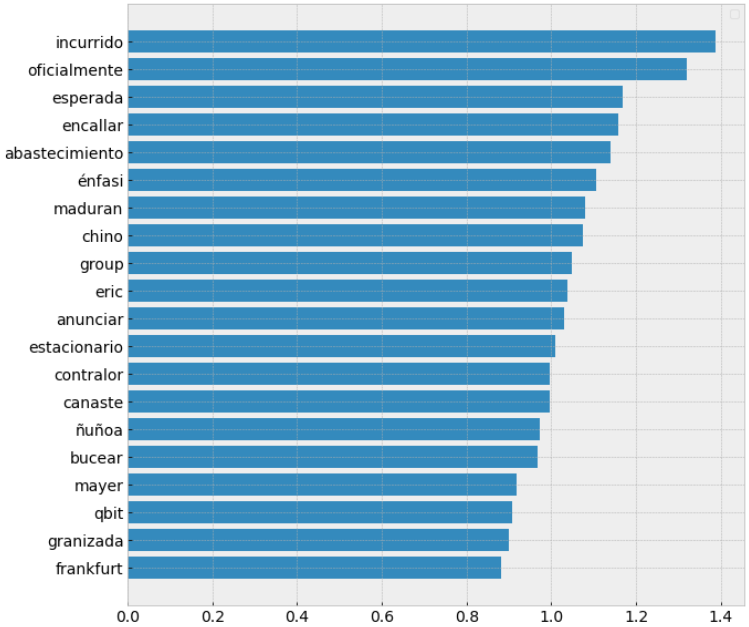


Figura 77: palabras que afectan al clasificador para la categoría 1 o variación positiva del IPSA

Se selecciona un subconjunto de palabras de las que se muestran en el grafico con influencia en la clase 1 o variación positiva del IPSA:

- oficialmente
- abastecimiento
- énfasi
- chino
- group
- anunciar

Con este listado se revisan noticias aleatorias (en la siguiente tabla) con el objeto de verificar que dichas noticias pudieran tener un efecto positivo en la variación del IPSA:

Palabra	Fecha y Hora	Titulo	Texto noticia
oficialmente	16.03.2018 / 08:58	En 2017 Codelco alcanzaría excedentes seis veces más altos que en el año anterior	A días de que la compañía estatal informe oficialmente los resultados, el . Cifra seis veces mayor que la lograda en 2016 donde se reportaron US$435 millones. De lograr esos números, además, la cifra se convertiría en , lo que Landerretche valoró señalando que , según consigna Así, enfatizó en que uno de sus logros era, a poco más de un mes de abandonar el directorio, estabilizar la deuda de Codelco. “, añadió.

Palabra	Fecha y Hora	Titulo	Texto noticia
abastecimiento	07.06.2018 / 16:53	El cobre alcanza su precio más alto en cuatro años y medio	El precio del cobre pasa por una muy buena racha y continúa al alza. El metal rojo , situación que permite relajar un poco las arcas fiscales. Según información de Cochilco, al cierre de las operaciones en Londres el cobre se cotizó en , 1,38% superior con respecto a la jornada anterior. Se trata del mayor valor del metal rojo desde el , cuando estuvo en US$ 3,28174 la libra. Uno de los motivos que impulsa esta alza es el , donde el vocero del Sindicato N°1 de Minera Escondida, Carlos Allendes, advirtió que "estamos preparados para una nueva huelga". El vicepresidente ejecutivo de Cochilco, Sergio Hernández, dijo a Cooperativa que "lo de Escondida ciertamente que tiene influencia porque es la mina individual más grande del mundo del cobre. Cualquier paralización de una mina importante como ésta genera lógicamente expectativas de mayor apetito por abastecimiento de cobre y alza de precios".
énfasi	30.09.2013 / 08:20	Elogios y críticas recibió el presupuesto 2014	Educación, salud y empleo, son algunos de los énfasis del . Además el próximo gobierno contará con US$600 millones de libre disposición. El Presidente Sebastián Piñera se comprometió asimismo a reducir el "déficit fiscal estructural" heredado, y recuperar los ahorros externos, los que llegarán –según el mandatario– a más de US$22 mil millones el 2013. Para el ex director del Banco Mundial y miembro del equipo económico de Marco Enríquez-Ominami, Andrés Solimano, el presupuesto para el 2014 "es serio porque es preparado con proyecciones razonables. Lo interesante es que vaya disminuyendo el déficit fiscal". Sin embargo, el problema que ve Solimano son las prioridades, y "esto levanta el tema que en Chile se deberían preparar presupuestos multianuales en que pudiera reflejar el cambio de una administración a otra". Dentro de las propuestas económicas del candidato del PRO está el impuesto a la riqueza privada alta, mayor tributación de las mineras privadas de cobre, reducción del gasto militar y terminar con la distinción de entre utilidades devengadas y redistribuidas. "Queremos financiar con más impuestos y reducir gastos militares", aclaró. El economista y líder del Movimiento Evolución Política (Evópoli), Felipe Kast, recordó que el gobierno de Piñera recibió el país "con un terremoto y un déficit estructural, es decir, estábamos gastando más allá de lo que ganamos en promedio". Pese a lo anterior, "logró que la renconstrucción se financiara, logró hacer una agenda social muy fuerte y la economía creció", destacó. Kast también valoró el aumento del fondo de libre disposición. Vea el debate completo en el video adjunto.
chino	10.03.2021 / 00:37	OCDE mejora proyección económica mundial para 2021, pero advierte sobre factores de riesgo	Por fin tenemos una buena noticia: las La (OCDE) actualizó este martes sus previsiones de crecimiento mundial y el pronóstico tiene una mejora significativa. El organismo indicó que "las y apuntó al desarrollo de las vacunas contra el coronavirus y los anuncios de estímulos adicionales. La agencia con sede en París espera que la Esta cifra representa una en comparación con su estimación de diciembre. Se prevé que la , más de tres puntos porcentuales de lo que se estimaba en diciembre. La agencia señaló los efectos del en el paquete de estímulo del presidente Biden con un valor de US$ 1,9 billones. Sin embargo, la OCDE también hizo énfasis en que y advirtió que muchos factores podrían poner en riesgo la recuperación. Un ejemplo de esto: los inversores están cada vez más preocupados de que un aumento en la actividad pueda

Palabra	Fecha y Hora	Titulo	Texto noticia
			provocar un suba de los precios en los próximos meses. Esto podría obligar a los bancos centrales a Según la OCDE, un , especialmente de China, está haciendo subir los precios de los alimentos y los metales. Los , por su parte, han experimentado una fuerte recuperación. El grupo dijo que va a ser fundamental que los responsables políticos mantengan el , incluso si la inflación supera algunos objetivos. Sin embargo, la posibilidad de que haya una subida pronunciada de los precios no es el único temor. La agencia señaló que las avanzan en diferentes velocidades en el mundo y que todavía podrían surgir variantes del coronavirus que resistan las vacunas. "El avance lento en el despliegue de las vacunas y la aparición de nuevas mutaciones del virus resistentes a las vacunas existentes darían lugar a una , una y un ", señaló en su informe. La OCDE también dijo que incluso cuando la situación empieza a mejorar. La presidenta del Banco Central Europeo, Christine Lagarde, lanzó una advertencia similar al decir que los "Hay que evitar un endurecimiento prematuro de la política fiscal", dijo el grupo. Otra preocupación son los La OCDE se centró en la carga de la deuda de las empresas en particular. Las cargas del servicio de la deuda están en niveles iguales o superiores a los de la crisis financiera de 2018 incluso cuando las rasas de interés están en mínimos históricos. "Aunque algunas empresas han recurrido al endeudamiento para acumular reservas de efectivo considerables desde el inicio de la pandemia, un apalancamiento alto podría moderar las nuevas inversiones", señaló. Si la recuperación es más lenta de lo que se espera, o si los programas de apoyo gubernamentales terminan demasiado pronto, esto podría
chino	07.03.2017 / 18:04	¿Por qué aumentó el Imacec minero?	Entre las cifras que se encuentra la de la actividad minera. Ello pese a que, según el INE, la producción cayera en casi un 2% en ese mes. El vicepresidente ejecutivo de Cochilco, , explicó que , y eso tiene que ver con el margen de diferencia entre los costos y el valor de la producción". Otra explicación radica en el alza del valor del metal rojo debido a la alta demanda china y la huelga de Minera Escondida, que elevó el precio hasta US$ 2,70 la libra. En esa línea, los expertos esperan que tras casi un mes de movilizaciones en aquella empresa, el precio del cobre gire en torno a US$ 2,50 en 2017.
group	07.11.2013 / 06:38	Twitter informó que sus acciones tendrán un valor de 26 dólares en la bolsa de Nueva York	Setenta millones de acciones estarán a la venta de la conocida red social Twitter, lo que reportará a la empresa una ganancia de 1.820 millones de dólares y estarán disponibles desde este jueves bajo las siglas TWTR. Las entidades colocadoras serán los bancos Goldman Sachs, Morgan Stanley, JP Morgan, Merrill Lynch y el Deutsche Bank. Los pequeños inversores, deberán esperar a que abran la bolsa alrededor del mediodía de nuestro país. El lanzamiento de Twitter en la bolsa coincide con los descensos en la cotización de redes sociales. Marcas como Groupon bajó un 2,7%, Facebook disminuyó un 2,0 % y LinkedIn 1,7 %.

Palabra	Fecha y Hora	Titulo	Texto noticia
anunciar	28.02.2013 / 07:49	Expertos anuncian signos de recuperación en el mercado inmobiliario estadounidense	Según explicó Daniel Soto, porfolio manager de ForexChile, se han conocido datos positivos en la economía de Estados Unidos, pues la venta de casas llegó a niveles que no se observaban desde el 2008. La situación, según expertos es positiva, ya que el Banco Central anunció que "la fiesta continúa" y que los estímulos, actualmente inyectados en la economía, seguirán realizando su trabajo sin modificación alguna. "El mercado inmobiliario era uno de los que más impedía el crecimiento en EE.UU. desde la crisis subprime", explicó Soto. Además, agregó que existe una mayor confianza en el consumidor, lo que es un indicio favorable, pues el consumidor representa alrededor del 70% del Producto Interno Bruto del mercado estadounidense.

En las noticias seleccionadas asociadas al listado de palabras previamente indicado, se observan noticias que pudieran tener un efecto positivo en los índices de la bolsa de comercio, noticias que tratan temas como:

- Excedentes en las ganancias de Codelco
- Alza en los precios del cobre
- Proyección de mejora económica de la OCDE
- Aumento del IMACEC, indicador creado por el Banco Central que permite medir de forma constante y más detallada la actividad económica nacional
- Aumento en el valor de las acciones de Twitter
- Reactivación del mercado inmobiliario de EEUU

Noticias con efecto negativo

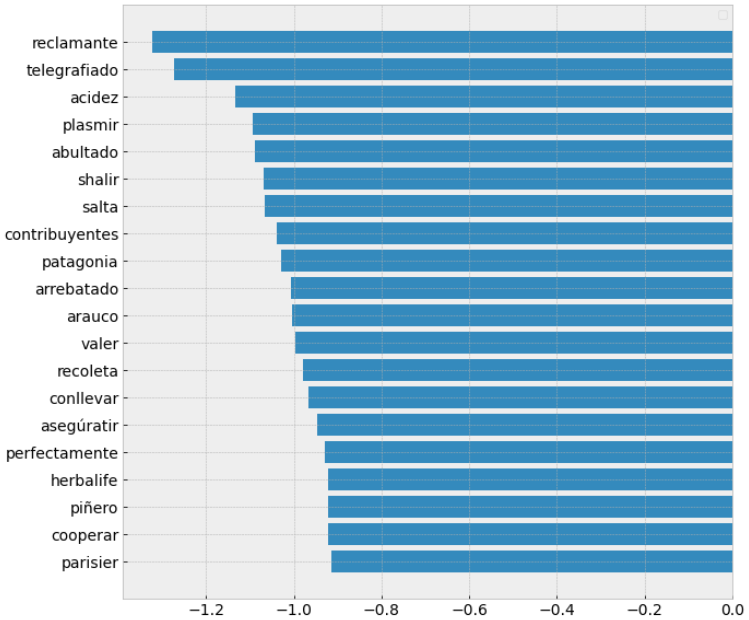


Figura 78: palabras que afectan al clasificador para la categoría 0 o variación negativa o cero del IPSA

Se selecciona un subconjunto de palabras de las que se muestran en el grafico anterior con influencia en la clase 0 o variación negativa del IPSA:

- salta
- valer
- conllevar
- piñero
- perfectamente

Con este listado se revisan noticias aleatorias (en la siguiente tabla) con el objeto de verificar que dichas noticias pudieran tener un efecto negativo en la variación del IPSA:

Palabra	Fecha y Hora	Titulo	Texto noticia
salta	24.12.2013 / 18:57	Eduardo Aninat analizó los escenarios económicos de Chile de cara al 2014	Eduardo Aninat, ex ministro de Hacienda, conversó con Agenda Económica para analizar el futuro escenario de la economía chilena el 2014, con los efectos de la desaceleración, llegada de un nuevo gobierno y una posible reforma tributaria. El economista sostuvo que Chile "seguirá creciendo el 2014", aunque lo hará a una velocidad menor que en años anteriores, debido a los efectos de la desaceleración económica, y también a la disminución de los precios de los comodities. Aninat también destacó el impulso que recibirá el sector exportador ante el aumento del tipo de cambio, indicando que el dólar puede llegar a un valor cercano a los $560 pesos. Sobre la llegada de un nuevo gobierno, el ex ministro resaltó la importancia de Hacienda para el ordenamiento y equilibrio macro del Estado. "Es importante que el futuro ministro tenga la mayor estabilidad posible", agregó, resaltando que también debe tener la capacidad para dialogar.
salta	15.03.2022 / 15:40	Ministra de Minería: "Las empresas se han ido dando cuenta de que no hay ánimo expropiatorio por parte del pdte. Boric"	La creación de una son algunos de los desafíos que tendrá el del gobierno del presidente Gabriel Boric. Retos que abordó la titular de la cartera, , en entrevista con de . La secretaria de Estado destacó que ambas iniciativas " Pensamos que el royalty es uno de los pilares de la forma en que se van a recaudar recursos para financiar todas estas cosas". Consultada sobre cómo abordará el Ejecutivo el proyecto de royalty, Hernando explicó que deberán hacer un balance, ya que "en el Parlamento no existen grandes mayorías y por lo tanto uno no puede contar con todos los votos. ". Existen dos caminos a ser evaluados por el Gobierno: "Uno es que vaya . Me imagino que el ministro de Hacienda tiene algunas observaciones respecto del proyecto de royalty tal como está", comentó la ministra. El otro es "incorporar alguno de estos aspectos o completamente ". La jefa de la cartera de Minería aseguró que ha conversado con las grandes empresas mineras privadas en Chile y "entienden que a nivel internacional existen alzas de impuestos, que los gobiernos están en el derecho de aumentar los impuestos, y ". En esa línea, destacó que "lo que hay que establecer es una escala y que sea razonable este aumento de impuestos, que quedemos todos contentos, tanto el Estado que va a recaudar, como ellos". "Para la industria, es importante que nosotros demos certezas. ; eso es lo que me han transmitido y yo se los agradezco", añadió. Por otro lado, la ministra aseguró que "la incerteza que producía el tema de las expropiaciones se ha ido calmando, en el sentido que ". En ese sentido, resaltó que " porque sigue siendo un país seguro". Sobre la propuesta de crear una , la secretaria de Estado señaló que "al interior del ministerio estamos volviendo a , y eso va a tener la responsabilidad de la creación de la empresa del litio y ver cómo se va a institucionalizar aquello".

Palabra	Fecha y Hora	Titulo	Texto noticia
salta	08.06.2022 / 10:01	IPoM: Banco Central eleva proyección de crecimiento entre 1,5% a 2,25% para 2022	El dio a conocer el correspondiente al mes de junio, El principal factor tras el alza continúa siendo el significativo aumento de la demanda en 2021. Asimismo, el informe destaca que en los últimos meses, se ha profundizado el impacto de las altas presiones de costos globales, consecuencia de los mayores precios de las materias primas, la energía y los alimentos. Además, el informe también menciona que la inflación mundial ha aumentado significativamente en varias economías. En esa línea, resalta que sino también a un alza de sus perspectivas. Respecto al plano de actividad interna, el informe precisó que los datos de inicios de 2022 muestras que la economía ya entró en una fase de ajuste, reflejando un leve descenso desde los elevados niveles que alcanzó en 2021. En ese contexto, se prevé que el crecimiento de 2022 estará entre 1,5% y 2,25%. Esto considera una reducción de la actividad que irá adquiriendo fuerza a medida que avanza año. Mientras que Para 2024, se sigue proyectando una expansión del Producto Interno Bruto (PIB) entre 2,25% y 3,25%, valores que están en línea con su crecimiento potencial. Respecto a la inflación, , durante el tercer trimestre. A partir de ahí, comenzará a descender en torno al 10%. Asimismo, hacia 2023 y 2024, el escenario central sigue contemplando que la inflación total tendrá un sostenido descenso, ubicándose en torno al 3% para mediados de 2024.
valer	04.08.2022 / 14:40	Tras viaje de Pelosi: Cómo los ejercicios militares de China en Taiwán amenazan con impactar aún más el comercio mundial	– Los amenazan con interrumpir el comercio y los viajes en el este de Asia, al obligar a las embarcaciones a desviarse de una de las rutas fluviales más transitadas del mundo y al presionar aún más las ya tensas cadenas de suministro a nivel global. Este jueves, China comenzó sus entrenamientos con la participación de la , la y otros cuerpos militares en el mar y el espacio aéreo que rodean Taiwán. Estos ejercicios –sin precedentes en número– son una demostración directa de , presidenta de la Cámara de Representantes de Estados Unidos, a la isla autónoma. Un viaje contra el cual Beijing advirtió en repetidas ocasiones. El Ministerio de Defensa de China publicó este martes un mapa de alrededor de la isla, donde informó que realizaría ejercicios aéreos y marítimos, además de entrenamientos con fuego real de largo alcance que se extenderán hasta el domingo. En ese sentido, se advirtió a los barcos y aviones que deben mantenerse fuera de esas áreas durante los simulacros militares. Por su parte, Taiwán señaló que estos ejercicios equivalen a un " ", además de vulnerar "las aguas territoriales de Taiwán y su zona contigua". También amenazan con alterar los flujos comerciales en una de las rutas de navegación más concurridas del mundo. El Estrecho de Taiwán, una arteria de 177 kilómetros de ancho que separa la isla de Taiwán y Asia continental, es una para las embarcaciones que transportan mercancías entre las principales economías del noreste de Asia, como China, Japón y Corea del Sur, y el resto del mundo. VesselsValue, una consultora de transporte marítimo con sede en Londres, dijo que actualmente hay y otras embarcaciones en las aguas territoriales de Taiwán. Y añadió que se estima que otros 60 lleguen entre este jueves y el domingo, justo cuando se realizarán los ejercicios militares. "Existe la posibilidad de una interrupción sustancial en el comercio de la región", señaló , analista de flujo comercial de VesselsValue. Cerrar las rutas comerciales alrededor de Taiwán, aunque sea de manera temporal, "despierta preocupaciones sobre si China podría volver a hacerlo con éxito. Y también sobre lo que esto podría significar, no solo para el comercio, los viajes y los , sino para los ", apuntó , analista líder de comercio global en The Economist Intelligence Unit. Aún no está claro cuál será el impacto a largo plazo. Sin embargo, los transportistas ya anticipan debido a cambios de ruta, posibles y que deben hacer más horas. , que interrumpieron el flujo

Palabra	Fecha y Hora	Titulo	Texto noticia
			de bienes y dispararon la inflación en muchas partes del mundo. Cualquier conflicto en , que , podría exacerbar la escasez mundial de chips de computadoras. Justamente, componentes vitales para prácticamente todos los dispositivos electrónicos modernos. . El puerto de Kaohsiung, ubicado en la costa suroeste, es el más grande de Taiwán y el decimoquinto más grande del mundo, según el Consejo Mundial de Transporte Marítimo. La Oficina Marítima y Portuaria de Taiwán emitió tres avisos este miércoles, en los que que utilicen rutas alternativas para los puertos en las ciudades de Keelung, Taipei, Kaohsiung y otras más. Además, luego de negociaciones con Japón y Filipinas. Aproximadamente debido a los cambios de ruta, informó este miércoles el ministro de Transporte de Taiwán, Wang Kwo-tsai. “Las repercusiones no han terminado: apenas comienzan”, advirtió Clifford Bennett, economista en jefe de ACY Securities, una firma de corretaje australiana. “Será mucho peor en la relación entre Taiwán y China como resultado de la visita de Pelosi”, añadió. China ya ha golpeado a Taiwán con desde el miércoles. Entre ellas, la suspensión de algunas importaciones de frutas y pescado de Taiwán y las exportaciones de arena natural a la isla. Todo el evento puede “continuar resonando y causar más daño durante meses, incluso años, tanto a las relaciones de Taiwán como de Estados Unidos con China continental”, dijo Bennett.
conllevar	09.04.2015 / 14:39	Tomás Flores explicó las razones tras la caída del Ipec	En entrevista con CNN Chile, , ahondó en la caída de la confianza en los consumidores. En este contexto aseguró que “en la medida que uno tiene temor hacia el futuro, la conducta se modifica y se hace más cauta”, enfatizando en que eso conlleva a una caída en las ventas. Respecto a la catástrofe en el norte del país, destacó que el hecho “genera un temor natural”, el que se traduce en que “nada es muy seguro, por lo tanto, lo que pensaba comprar mejor lo dejo para después”. Además, Flores también se refirió a la preocupación del Banco Central en torno a la inflación, la última cifra del Imacec y la tasa de desempleo en nuestro país, entre otros temas. Para más detalles sobre el análisis del mercado laboral, revisa la entrevista completa en el video adjunto.
piñero	05.08.2014 / 19:05	Imacec de junio marca peor registro desde marzo de 2010	Todo el mercado coincidía en que la actividad económica durante junio seguiría desacelerada y las estimaciones más pesimistas estaban en torno a 1,5% de crecimiento. Sin embargo, el Imacec sorprendió a todos por lo bajo de la cifra: solo fue de un 0,8%. Un resultado que refleja el peor Imacec desde marzo 2010, justo después del terremoto en Chile. Según lo informó el Banco Central, esta cifra se dio a pesar de que junio de este año tuvo un día hábil más que el mismo mes de 2013. En el resultado incidió, principalmente, la caída de la industria manufacturera y del comercio mayorista y automotor. De esta forma, el país habría cerrado el primer semestre con una

Palabra	Fecha y Hora	Titulo	Texto noticia
			expansión de sólo 2,2%. “A mediados de nuestro gobierno esperamos estar con nuestra casa bien ordenada”, explicó la presidenta Bachelet en este escenario de desaceleración. Sin embargo, Felipe Larraín recalcó que las economías “no crecen por encanto” y que las cifras de la gestión Piñera no fueron una casualidad. Conoce más detalles en el informe adjunto.
piñero	05.08.2019 / 09:05	Imacec confirma merma en el crecimiento y economía sólo crece un 1,3% durante junio	Este lunes 5 se conocieron las cifras del correspondiente a junio, números que vienen a confirmar una desaceleración en el crecimiento económico del país. El reporte del Banco Central da cuenta que el Imacec Minero “Este último se vio favorecido por el desempeño de las actividades de servicios y de construcción, efecto que fue en parte compensado por la caída de la industria manufacturera”, se aprecia. “De acuerdo con la información preliminar, el IMACEC de junio 2019 creció 1,3% en comparación con igual mes del año anterior. respecto del mes precedente y aumentó 1,7% en doce meses. El mes registró un día hábil menos que junio de 2018”, explican desde el organismo. Con el resultado de junio, , la economía chilena acumuló una expansión de 1,73% en lo que va de 2019, lo que supone que el cálculo oficial actualizado de la administración Piñera (entre 3% y 3,5%) se ve muy complejo de alcanzar. Para lograrlo, la economía debe acelerar notablemente en los cinco meses que quedan y promediar un crecimiento superior al 4,5%. Este, además, es condicionado por lo que suceda en la guerra comercial que llevan a cabo China y Estados Unidos. El enfrentamiento, que ha recrudecido en los últimos días, hace aún más cuesta arriba que la salud de la economía nacional tenga un mejor prospecto para los meses restantes.
perfectamente	05.09.2022 / 09:13	Experto dice que dólar bajaría a $800 tras triunfo del Rechazo, pero “la incertidumbre debería volver a generar alzas”	¿Cómo se ven las tras el triunfo del Rechazo en el plebiscito de salida? Una son parte de las repercusiones que esperan los expertos para las próximas horas y días. Sin embargo, la caída de la divisa estadounidense sería un fenómeno pasajero. , jefe de Estudios de Mercados de Capitaria, explicó en conversación con que “no sería extraño ver y perfectamente se podría acercar a los , pero pensando en los próximos días”. “Hay que considerar que no se finalizó un camino, ahora comienza un nuevo camino, ¿vendrá una nueva Convención Constitucional? y la incertidumbre debería “. “Pronto deberíamos ver alzas del dólar considerando un y un escenario de incertidumbre en lo político”, añadió el jefe de Estudios de Mercado de Capitaria. En ese sentido, sostuvo que “las dudas que van a existir con respecto en la situación chilena, tanto en lo económico como en lo político, deberían por parte de los inversionistas extranjeros”. Finalmente, indicó que “ , o sea un decrecimiento económico, y los inversionistas no van a querer invertir en un país que puede tener una importante. Sumado a una incertidumbre del proceso constituyente, sería un mix perfecto para que, lamentablemente, a fin de año volvamos a ver caídas en la venta variable local”.

En las noticias seleccionadas asociadas al listado de palabras previamente indicado, se observan noticias que pudieran tener un efecto negativo en los índices de la bolsa de comercio, noticias que tratan temas como:

- Eduardo Aninat analizó futuros escenarios económicos para el 2014, indicando que la economía chilena seguirá creciendo, pero más lento que otros años

- Ministra de la minería del gobierno de Boric habla sobre el royalty para el sector minero
- Informe del banco central menciona que la inflación mundial ha aumentado significativamente en varias economías
- IMACEC de junio de 2014 marca peor registro desde marzo de 2010
- Durante agosto de 2019 IMACEC confirma merma en el crecimiento del país
- Incertidumbres en el aspecto económico debido al rechazo a la propuesta de la nueva constitución

En general se puede observar que en esta pequeña muestra de noticias si existe una correlación entre los temas tratados y las clases que debe clasificar el modelo utilizado, en este caso el clasificador de regresión logística.

6 Conclusiones

Seguidamente se analiza cómo se desarrollaron cada uno de los objetivos particulares o específicos planteados al comienzo de este trabajo:

- Mediante la librería o framework Scrapy de Python, se implementó una metodología con la cual se rasparon datos de la web de CNN Chile, descargando todas las noticias publicadas entre noviembre de 2012 y septiembre de 2022. Para cada noticia se extrajeron los siguientes datos: categoría, título, fecha y hora de publicación, resumen y el texto completo, con estos datos se pudo construir el dataset que se utilizo en el desarrollo de este trabajo.

- Para realizar la limpieza de datos se desarrollaron las siguientes etapas:

    - Etapa 1: Mediante la librería SQLite3 se implementó una función de limpieza la cual se aplicó directamente a la base de datos mediante sentencias del lenguaje SQL, lo cual permite optimizar el trabajo con gran cantidad de datos. En esta etapa se transformo el texto a minúsculas, se eliminaron los tags de HTML, se eliminaron los espacios y dígitos.
    - Etapa 2: Mediante la librería pandas se analizaron los datos resultantes de la etapa anterior, implementando funciones para quitar caracteres como puntos suspensivos, quitar stop words o palabras vacías y eliminar emoticones. También en esta etapa se desarrollo el proceso de lematización mediante la librería spacy de Python.

- Para la exploración de datos se utilizaron las librerías pandas, matplotlib y wordcloud de Python. Se implementaron gráficos de barras horizontales para medir la frecuencia de aparición de las palabras, bigramas y trigramas, para el dataset completo como para cada categoría dentro de las noticias. Además, se construyeron nube de palabras para el dataset completo y para cada categoría de noticias. También se realizó una exploración de los datos históricos del IPSA mediante graficas de series de tiempo con la librería matplotlib.

- Como análisis complementario se realizó una separación o clasificación de los datos mediante clusterización, en particular se utilizó el algoritmo de clasificación no supervisado KMeas de la librería scikit-learn, con el objetivo de determinar si existe una mejor clasificación de los datos que la que se encontraba en el sitio web de CNN Chile. Una vez realizada la clusterización se utilizaron las librerías pandas, matplotlib y wordcloud para explorar las nuevas categorías, sin embargo, se determino que en esta nueva clasificación de los datos no existe una categoría

asociada al concepto economía. Dado lo anterior se concluyó que para encontrar una relación entre la variación del índice de la bolsa de comercio IPSA con las noticias publicadas en el sitio CNN Chile se utilizarían las noticias que en dicho sitio se encuentran clasificadas bajo el concepto de economía.

- En base a metodologías encontradas durante el análisis bibliográfico se asoció la variación del IPSA con las noticias de la categoría de economía del sitio web de CNN Chile. Dado que se cuenta con los datos históricos diarios del IPSA se asignó un valor de 1 a las noticias publicadas durante un día con variación positiva del IPSA, en cambio, cuando el IPSA registraba una variación negativa o cero se asignó el valor 0 a las noticias publicadas durante ese día. Esta subclasificación de las noticias de la categoría de economía se implementó mediante la librería pandas.

- A partir de la etapa anterior se procedió a entrenar distintos tipos de clasificadores, como Logistic Regression, Linear SVC, Multinomial NB, MLP Classifier, Random Forest, etc. Para las variables de entrada (noticias) se aplicaron metodologías de vectorización de tal manera de transformar las palabras en una representación numérica con las clases CountVectorizer y TfidfTransformer de la librería scikit-learn. El objetivo era poder encontrar una relación entre las palabras que componen a los distintos documentos o noticias con las variaciones positivas o negativas del IPSA (variable de salida) mediante la implementación de distintos clasificadores.

- Para poder evidenciar si existe alguna relación entre las variaciones negativas y positivas del IPSA con las noticias de la categoría economía se utilizó el clasificador entrenado del tipo Logistic Regression que se indicó en el párrafo anterior. Mediante el atributo coeff_ de la clase LogisticRegression de la librería scikit-learn se pueden determinar las características o variables de entrada que más influyen en cada categoría de la variable de salida (1 y 0, variación positiva o negativa del IPSA, correspondientemente), con lo anterior se asociaron los puntajes o ponderaciones del atribute coeff_ con las variables de entrada o las palabras que componen a las noticias, mostrando las más influyentes para cada categoría de resultado mediante gráficos de barras.

7 Pasos futuros

- Se podría realizar un análisis temporal de cuando aparecen términos claves, como los encontrados en el apartado anterior, con el objeto de determinar si existe algún suceso en el tiempo que se pueda relacionar con la mayor o menor frecuencia de aparición de un término.
- Se podría realizar el análisis incluyendo todas las categorías de noticias y revisar si existen otros términos claves que pudieran afectar a la clasificación.
- Se podrían generar un clasificador con redes neuronales artificiales LSTM, las cuales tienen en consideración la secuencialidad de las palabras dentro de un documento, para ver si se obtienen mejores resultados.
- Se podría implementar alguna metodología para optimizar los parámetros de los clasificadores que se probaron, para obtener mejores resultados.
- Se podría explorar la utilización de otras librerías para la lematización. En este caso se utilizó la librería Spacy dado que dispone de diccionarios de lemas en español, sin embargo, se evidencia en el desarrollo de este trabajo que el proceso de lematización no es suficientemente optimo dado que no se redujo significativamente el vocabulario al aplicar dicho proceso. También existen

formas de complementar el diccionario de lemas incluyo algunos casos de lematización que la librería Spacy no incluye.

- Otra posible mejora es durante la etapa de vectorización, en este trabajo se utilizo las funciones que provee scikit-learn, tales como CountVectorizer y TfidfTransformer, las cuales utilizan la frecuencia de aparición de las distintas palabras en los documentos o noticias y en el dataset completo, sin embargo, no extraen la posible relación que puede existir entre las palabras en base a su secuencia de aparición, para este fin existen métodos como TextVectorization de la librería TensorFlow o Word2Vec de la librería Gensim. En particular el método TextVectorization de la librería TensorFlow es una capa que se entrena con el conjunto de datos y determina una representación numérica de las palabras que rescata la relación entre palabras dada su secuencia de aparición en los documentos.

8 Bibliografía

[1] G. Costa, «Tesis: Diseño y desarrollo de un módulo de clasificación de páginas web en base a las características de su contenido utilizando técnicas de minería de datos,» Universidad de Chile, 2016.

[2] S. Patni, Introduction - XML, JSON, Apres, 2017.

[3] J. Purushothaman, RESTful Java Web Services, Packt, 2015.

[4] U. Rodríguez, «Tesis: Investigación y Desarrollo de Técnicas de Scraping,» Universidad de Alcalá Escuela Politécnica Superior, 2019.

[5] T. Alice, Sistema de Bases de Datos Administracion y Uso, Prentice Hall, 1990.

[6] M. Mannino., Administración de Bases de Datos Diseño y Desarrollo de aplicaciones, 3ra Edición, McGraw Hill, 2007.

[7] D. K. A. K. y. S. S. Kiran Khatter, «Natural Language Processing: State of the Art, Current Trends and Challenges,» *Multimedia Tools and Applications,* 2022.

[8] N. Chomsky, Aspects of the Theory of Syntax, The MIT Press, 2014.

[9] Z. C. C.-W. T. Q. H. y. H. L. Yue Kang, «Natural Language Processing (NLP) in Management Research: A Literature Review,» *Journal of Management Analytics ,* vol. 7, nº 2, pp. 139-172, 2020.

[10] S. Mirjalili, Python Machine Learning, Packt Publishing, 2017.