PLN para análisis de periódicos Mexicanos: Complejidad, Ideología Política y Temáticas.

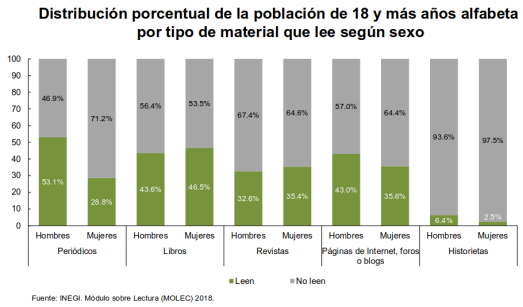
Jorge Ramos Kenny Méndez Adrián Rodríguez
Centro de Investigación en Matemáticas. Unidad Monterrey.

***Resumen*—En la actualidad la estructura que presentan los datos generados por dispositivos, personas, organizaciones viene presentada en "formas" que quedan fuera de la manera tradicional de hacer análisis estadístico sobre ellos. Es por ello, que surge la necesidad de recurrir a técnicas no convencionales, como lo son las abarcadas dentro del 'Procesamiento del Lenguage Natural". Es así que en este trabajo se analiza un conjunto de noticias de diversos periódicos a lo largo del país dentro de la sección local, policiaca y de opinión, del cual se exploran características de las noticias tales como complejidad de lectura, ideología política de columnas de opinión, similaridad de noticias y clasificación de noticias por tópico.**

I. Introducción

Desde hace muchos años se ha identificado a México como uno de los países más alejados de la lectura. Hace poco más de 100 años se registraba que un 82 % de la sociedad era analfabeta, a la fecha esta cifra ha disminuido únicamente 6.9 %.

- Casi cuatro de cada diez personas son cercanas al hábito de la lectura.
- 48 % de los jóvenes nunca han estado en una biblioteca.
- Del total de la población, el 42 % dedica su tiempo libre a ver televisión, tan sólo el 12 % se va por la lectura [1].



Uno de los problemas generados de la falta de lectura se puede ver en la desinformación que presenta una gran parte de la población del pais. La presencia de varios periodicos en medios digitales puede permitir que las generaciones nuevas que dejan de leer libros puedan comenzar a generar un habito de lectura a través de dichos medios.

En el presente trabajo se presentan varios esquemas de análisis de periódicos de México para explorar características de las noticias: Complejidad de lectura, ideología política de columnas de opinión y clasificación de noticias por tópico. Para realizar el trabajo se genero una base de datos con noticias de 38 periodicos de varios estados del país que conforman una colección hasta el día de hoy de 10,937 noticias.

II. Complejidad

En el español la complejidad se refiere a la legibilidad que viene condicionada por el léxico y las construcciones gramaticales utilizadas, y no por el tamaño, forma o diseño de la fuente y sus caracteres. En pocas palabras se puede decir que lo que se busca con la complejidad es medir la facilidad para comprender un texto.

II-A. índice Flesch-Szigriszt

Una metodología usual en las investigaciones de ciencias sociales para clasificar la complejidad de la lectura de un texto esta basada en una metodología abordada en la tesis doctoral de Inés Mª Barrio Cantalejo (Barrio, 2017), en la cual se adecúa el uso índice *Flesch-Szigriszt* para medir la complejidad de los textos.

En su tesis Inés hace una renovación de los pesos del índice F-S y los incorpora en un software llamado **Inflesz** en el cual de acuerdo a algunos umbrales del índice se categoriza un texto como: muy fácil, fácil, normal, díficil, muy díficil.

La fórmula utilizada en el índice es:

$$IFSZ = 206.84 - \left(62.3 \times \frac{Silabas}{Palabras} - \frac{Palabras}{Frases}\right)$$

[1] Módulo sobre lectura MOLEC, Febrero de 2018

ESCALA INFLESZ		
PUNTOS	GRADO	TIPO DE PUBLICACIÓN
< 40	MUY DIFICIL	UNIVERSITARIO, CIENTÍFICO
40-55	ALGO DIFÍCIL	BACHILLERATO, DIVULGACIÓN CIENTIFICA, PRENSA ESPECIALIZADA
55-65	NORMAL	E.S.O., PRENSA GENERAL, PRENSA DEPORTIVA
65-80	BASTANTE FÁCIL	EDUCACIÓN PRIMARIA, PRENSA DEL CORAZÓN, NOVELAS DE ÉXITO
> 80	MUY FÁCIL	EDUCACIÓN PRIMARIA, TEBEOS, CÓMIC

Figura 1: Categorías del índice Flesch-Szigriszt

Cabe señalar que la herramienta *Inflesz*, además de las clasificaciones proporciona la cantidad de silabas, palabras y frases.

Problemas del índice

- Número de oraciones en voz pasiva. (Abuso → complejidad)
- **Número y porcentaje de tipos de palabras (verbos, adjetivos, conjunciones, etc.)**
- Número y porcentaje de oraciones en afirmativa o negativa.
- Número y porcentaje de abreviaturas y acrónimos sin explicar.
- Errores ortográficos y gramaticales.

II-B. Metodología

Se utilizo un modelo entrenado para establecer el POS-TAG en los textos para extraer el número de verbos, adjetivos y sustantivos. Además se calcularon el numero de sílabas, numero de palabras y numero de frases en el texto y el etiquetado se hizo con base a **Inflez**.

Debido a que la clasificación de noticias estaba muy desigual en los extremos los incorporamos en las categorias mas cercanas, por lo que nos quedamos con tres categorías: **Facil, Normal** y **Dificil**.

II-C. Resultados

Con la base de datos establecida se generaron modelos SVM con un kernel lineal para establecer la clasificación de los textos etiquetados. Además para evitar el sesgo del uso de una semilla, por lo que se usaron 100 semillas distintas para la elección de los conjuntos de entrenamiento (80 %) y prueba (20 %).

Con lo anterior se obtuvo una precisión promedio de entrenamiento de 96.91 % y de prueba de 91.05 %. Implicando que la complejidad de los textos puede ser separado de manera lineal de una buena manera.

El modelo entrenado será utilizado para cuando se tenga una nueva noticia, una vez que se obtengan las características del texto respectivas, se logre clasificar al texto en alguna categoría de complejidad.

III. Ideología política

La ideología es una de las formas que pueden revestir los diversos modelos integradores de las creencias morales y cognitivas sobre el hombre, la sociedad y el universo… que florecen en las sociedades humanas.

Particularmente el tema de la ideologia política es una cuestión de múltiple controversia en varios circulos intelectuales, ya que no existe una definición única sobre como determinar la posición política de un texto.

Todo sistema político está caracterizado por un cierto número de conflictos: sobre la distribución del ingreso, sobre la intervención del estado en la economía, sobre las relaciones estado-iglesia, etc. Por lo que establecer la ideología de las personas a cargo de la toma de decisiones y de los periodistas de opinión nos permiten establecer un panorama mas amplio sobre lo que sucede en el Estado.

Existen diversos enfoques que se basan en establecer la clasificación por autor y por círculo social, otros en crítica y conformismo. Otros autores han establecido categorias mucho más amplias agregando mas categorias, y en otros se generan una categorización mucho más extrema como lo es el cuadrante de Nolan pero visualmente mas agradable.

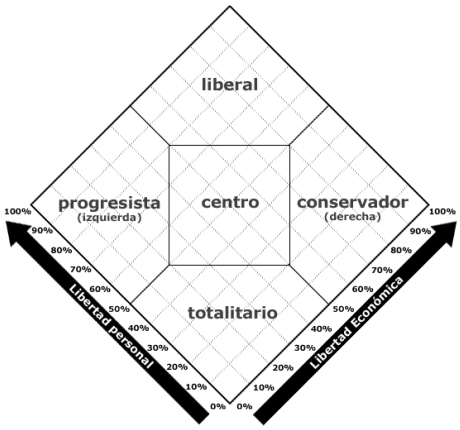


Figura 2: Cuadrante de Nolan

En “Political Ideology Detection Using Recursive Neural Networks“ varios autores presentan un marco de trabajo usando redes neuronales profundas para lograr la clasificación de frases con respecto a la ideología política de izquierda-derecha, para ello hacen uso de textos bien identificados como de izquierda o derecha, pero aún con ese esquema bien

armado solo se logra obtener una precisión de cerca del 70 %.

Para este trabajo se considero hacer el etiquetado de noticias en izquierda y derecha con base en las desviaciones del discurso con respecto al estado de derecho e institucional. Una desviación con respecto a este sistema representa un texto de la izquierda y un apoyo al sistema implica un texto de derecha. Y se separaron los textos de insulto con respecto a este mismo sistema.

Por ejemplo la siguiente frase puede considerarse de derecha

"El rostro de la Cuarta Transformación exhibe una moral bifurcada a conveniencia: el perdón a funcionarios corruptos colaboradores, y el azote a las instituciones incómodas."

esto debido a que denota una opción de conformismo con respecto al sistema actual. Mientras que la frase

"La prohibición de la mariguana produce más daños, más infelicidad y a la larga más crímenes y más criminales que el consumo de la sustancia."

se etiqueto como de izquierda dado a que indica que se debe hacer un cambio en el estado de derecho actual.

III-A. Resultados

Para este trabajo se clasificaron 124 noticias de caracter político: 60 de derecha y 64 de izquierda. Para trabajar con ellas se realizo un proceso de vectorización, usando el método Doc2Vec, con el fin de obtener una representación vectorial de cada documento y con ello obtener una estructura manejable para usar métodos de clasificación mas tradicionales como SVM o Adaboost

Se probaron dos modelos: SVM y Adaboost. Se usaron vectores embebidos de tamaño 400 usando "negative sampling" de orden 10.

También para eliminar el efecto que puede producirse por hacer uso de una semilla conveniente se corrieron los modelos usando varios parámetros y con 100 semillas distintas.

Haciendo lo anterior y calculando la media y la mediana de todos los niveles de precisión se obtuvieron los siguientes resultados:

- SVM 53.6 % (Media)
- Adaboost 58 % (Media), 57.9 % (Mediana)

Que aunque son muy bajos, Al compararlos con el trabajo previo y considerando que se trabajo con textos mucho mas amplios que solo frases puede considerarse un buen resultado en general.

IV. Similaridad entre noticias

IV-A. Problema

Uno de los problemas a los que se enfrentan investigadores de ciencias sociales y cualquier investigador que deba consultar una hemeroteca es el de encontrar todas las noticias relevantes sobre un tema en particular.

Sin embargo, esto puede convertirse en una labor titánica por lo que encontrar un método automatizado que nos permita hacer esto se vuelve de suma importancia.

IV-B. Posibles metodologías

- Búsqueda de palabras específicas en el documento.
- Extracción y búsqueda por tópico.
- **Doc2Vec y distancia coseno.**[2]

IV-C. Resultados

En general este esquema mostró muy buenos resultados, así como también nos permitió encontrar que algunos periódicos reciclan noticias de otros periódicos para rellenar sus secciones. Un ejemplo de los resultados obtenidos se muestra a continuación

Titulo de la noticia a buscar similaridades:

"Abandonan cadáver a orilla de la carretera"

Noticias similares:

- Tiran a ejecutado en carretera
- Abandonan a mujer ejecutada, le dejan mensaje
- Hallan ejecutado en interior de camioneta

V. Tópicos

Una necesidad que por lo general se presenta es la categorización de textos de manera automatizada, en este caso la categorización corresponde a asignar un tópico, definiendo un tópico como un conjunto de palabras que caracteriza a todo un texto. Para ello existen varias tecnicas en el contexto del procesamiento de lenguaje natural.

Entre estas tecnicas se encuentran

- Latent Semantic Analysis
- Non-Negative Matrix Factorization
- Latent Dirichlet Allocation

Para este trabajo se considero utilizar Latent Dirichlet Allocation (LDA), que es un modelo probabilístico generativo de un corpus.

La idea basica es que los documentos son representados como mezclas aleatorias sobre tópicos latentes, donde cada tópico esta caracterizado por una distribución sobre las palabras.

[2]Incluso pueden considerarse noticias disimilares para filtrar.

V-A. *Metodología*

- Para la obtención de las noticias se diseñó todo un proceso automático de "*Scrapping*" soportado por el lenguaje *Python*.

Para realizar el análisis de tópicos de la noticias se generó un modelo por cada sección (definida para los objetivos de este trabajo), las cuáles son las siguientes a su vez son categorizadas por el tipo de importancia de difusión:

Sección	Nivel de difusión
Opinión	Nacional
Opinión	Local
Local	Local
Policiaca	Local

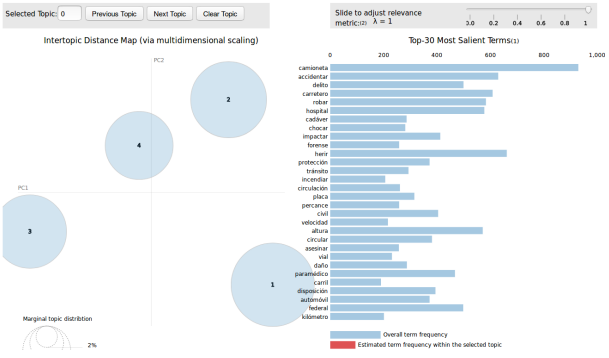
Tabla I: Categorías de las sección y nivel de importancia.

Los modelos se generaron acorde al siguiente protocolo:

- Se extrae la información de la BD
- Se procesa los cuerpos de las noticias (lematización, tokenización, etc..)
- Se crea un corpus creando una lista de Bolsa de Palabras (BOW)
- Se procesa el corpus (parametros de esparcidad)
- Tunear parametros LDA (corpus, passes, iterations, eta, alpha)
- Validar visualmente topicos y distribución (PCA)

Una vez teniendo entrenado cada modelo es posible asignarle a una nueva noticia su clasificación de tópico mas probable, solo que debe considerarse que al actualizar el modelo cada determinado tiempo se pueden captar tópicos que puedan ser de temporada y de corta duración, por lo que pueden clasificarse noticias nuevas en una categoria nada significativa solo porque contiene algunos elementos parecidos.

LDAVis



Análisis Exploratorio Tópicos

V-B. *Resultados*

Se obtuvieron 4 modelos de topicos:

- Modelos de Importancia Local

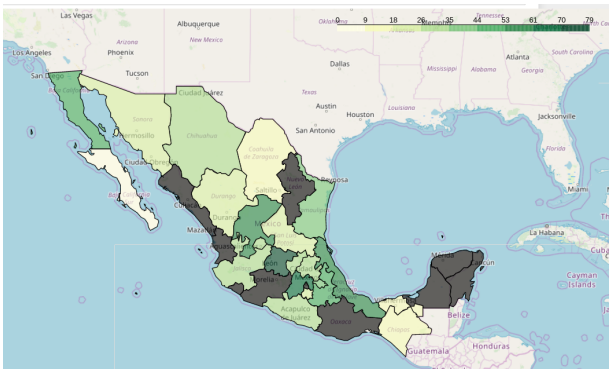


Figura 3: Noticias policíacas (homicidios) en todo el periodo.

  - Local : se encontró 8 tópicos entre los cuales se lista (Reformas y Legislaciones), (Lopez Obrador y Politica), (Economia y Finanzas), (Delitos y feminicidios), (topico navideño y familiares), (Educación y Maestros), (Programas sociales), (Migrante, asuntos internacionales)
  - Opinion Se encontro 5 topicos entre ellos se puede listar (Educacion y maestros), (relativo a la 4ta transformacion) etc,.
  - Policia: Se encontro 4 topicos (Matanzas y balacera), (delitos y detenciones), (Choques y accidentes viales), (Emergencias medicas e incendios (accidentes))
- Modelo Opinion Nacional : se encontro 15 tópicos, enre los tópicos encontrados se puede ver topicos referentes al SNTE y sindicato, tópico de futbol, tópico de economia, tópico referente a AMLO, topicos de cine y música (eventos), etc...

VI. Conclusiones

- Al incorporar información del número de verbos, sustantivos y adjetivos de un texto podemos ver que se puede generar una separación lineal para la complejidad de los mismos.
- El problema de clasificación de la ideología política es un problema de una complejidad muy alta, tanto que existen muchas posiciones distintas para su categorización. El uso de tecnicas de "machine learning" para categorizarlos se ha vuelto un tópico de investigación para solventar el problema de la clasificación subjetiva. Sin embargo, hasta el día de hoy los resultados muestran que aún se requiere mas investigación en el tema.
- Según la librería utilizada para el preproceso se pueden obtener distintos resultados, por ejemplo, la manera de lematizar o bien la definición de *Stop words*.
- A través de LDA pudimos reconocer temas reelevantes sobre los distintos estados estudiados. Esto también da una gran posibilidad a analizar el comportamiento de lo que pasa a través del tiempo.

- El enfoque de Doc2Vec mostró muy buenos resultados para establecer la búsqueda de noticias similares, lo que puede ayudar a resolver un gran problema de tiempo de búsqueda manual.

VII. Trabajo futuro

El trabajo se pretende aumentar de manera que se integren en una aplicación web los resultados obtenidos. La idea de establecer esta aplicación web es la de generar una base de datos de usuarios para generar un sistema de recomendación con base en los resultados obtenidos y con ello permitirle al usuario leer solo las noticias que considere relevantes, así como fomentar el hábito de la lectura de dicho usuario mostrando noticias de mayor complejidad de lectura con respecto al uso de la plataforma.

Otro trabajo que se planea establecer es la creación de una plataforma que permita el etiquetado interactivo para modelos de POS-TAG y NER, con el fin de generar modelos mucho más específicos para poder extraer características del texto y encontrar localizaciones dentro del contexto de las noticias.

También se planea realizar una aplicación web incorporando la información obtenida por la extracción de tópicos con el fin de mostrar información espacio-temporal sobre los tópicos importantes a nivel estado con el fin de identificar cuestiones importantes como establecer los tipos de crimen más frecuentes por estado.

Referencias

[1] Inés Barrio. Legibilidad y salud. los métodos de medicón de legibilidad y su aplicación alos folletos educativos sobre salud, 2017.

[2] Terry Eagleton. *Ideology*. Routledge, 2014.

[3] Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1113–1122, 2014.

[4] Dan Jurafsky and James H Martin. *Speech and language processing*, volume 3. Pearson London, 2014.

[5] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196, 2014.

VIII. Anexo

VIII-A. Estructura de archivos en el servidor

Para mantener organizada la estructura para la ingesta automatizada de noticias y el análsis de los textos, en el árbol de archivos del usuario 1 en el servidor CIMAT se tiene:

```
nlp_news_2018/

    analysis/
            *.R
            *.ipynb
    scrapps/
            logs/
            *.py
```

VIII-B. Tecnologías utilizadas

Server Desarrollo

- Ubuntu Server 16.04.5
- 4 GB RAM
- Procesador Corei5 (4 hilos)

Server CIMAT

- Linux Mint Sonya
- 16 Gigas de Ram
- 8 hilos (Procesadores Xeon)

Dentro del servidor de desarrollo se instaló la versión de docker actual, y para este proyecto dado que se requiere realizar escrapeo, almacenamiento (DB) y análisis se ocuparon 4 contenedores (instancias docker).

Contenedores Docker

- Contenedor (anacondajup_new)
  - Configuracion:
    - Este contenedor se basa en la imagen oficial de anaconda, la cual ya incluye librerías como matplotlib, numpy, scipy, entre otras, y ademas trae preinstalado jupyter-notebook
    - Esta imagen se extendio usando un Dockerfile en el cual se mando instalar otras librerias necesarios pára realizar el proyecto como gensim, pyldavis, plotly, folium, etc
  - Uso: Este contenedo es el que mas se usa para el proyecto dado que todos los integrantes pueden contectarse a el remotamento para acceder al jupyter-notebook y crear analisis con la inforamcion de la base de datos contenida en la instancia ”nlpdb”,
- Contenedor (seleniumdocker)
  - Configuracion: Esta imagen se extiende de una imagen encontrada en dockerhub, se dejo intacta dado que cumple con el proposito
  - Uso: Esta imagen solo sirve para realizar escrapeo automatizado en páginas las cuales tienen contenido dinámico (Contenido que se genera al vuelo con javascript y AJAX); el cual no puede ser captado por los escrapeadores comunes como urllib o BeautifulSoap, por lo cual se usa “Selenium” el cual usa un navegador y emula un usuario al escrapear
- Contenedor (nlpdb) Este contenedor es una instancia de la imagen oficial de “postgress”, en este contenedor reside nuestra base de datos, para esta imagen se expone el puerto 5433 y se redirige al puerto 5432 el cual es el puerto oficial de postgress, Se realizó la configuración de un usuario para la base y se creó una tabla, la cual es la que tiene la estructura que necesitamos para almacenar los datos
- Contenedor (rockergisnew)
  - Este contenedor es una instancia de la imagen oficial “rocker GIS”, el cual es una versión optimi-

zada de R con la paqueteria necesaria de GIS (rgdal,maptools,etc.....), ademas usa una versión optimizada de BLAS OpenBLAS lo cual acelera los cálculos matriciales en comparación con una instalación de R normal, y otra ventaja es que se puede acceder a un Rstudio remotamente desde un navergador web usando el puerto 8787,

- Uso: el objetivo de esta imagen es realizar prototipado de análisis, y llevar a cabo análisis que nos son algo más fáciles en este lenguaje, ademas tambien se usa para realizar mapas animados combinando las librerias "maptools" y "animation"

VIII-C. Infraestructura

VIII-D. Proceso de extracción de textos ("scrapping")

Para realizar el escrapeo de las noticias, se usa 3 de las 4 instancias (anacondajup_new,nlpdb,seleniumdocker) , y se usa una serie de programas para poder realizar el escrapeo, dado que son mas de 40 sitios web los que se consultan diariamente, para el caso de noticias locales se usa la cadena de noticias OEM, la ventaja de obtenerlos de una misma cadena de noticias es que casi todos los subsitios de las diferentes partes de méxico que se incluyen ahí, tienen una estructura muy similar html, por lo cual para escrapear 32 sitios se usa el mismo programa, al cual se le manda de parámetro del sitio, para los demas sitios no se tuvo tanta suerte y fue necesario adaptar el código para poder lidiar con la estructura.

En el caso de las noticias de OEM tiene la peculiaridad que para obtener las ligas o enlaces a escrapear es necesario usar "selenium" dado que en el sitio se presenta como contenido dinámico y es la única manera de obtenerlo en forma consistente.

Una vez que ya se tienen programadas las rutinas de código necesarias para escrapear e insertar en la base de datos, dado que es un proceso que se requiere hacer diariamente, el paso obvio es crear un "job" en cron, el cual se ejecute diarimaente, solo que hay que tomar ciertas consideraciones, lo mejor es no traslapar tareas de sitios sobre todo si no se cuenta con una buena conexión y para evitar fastidiar a los servidores de noticias y por lo mismo ocasionar el ser bloqueados por ip.

Se pleaneo una programacion en "CRON" para que contemplara todos los sitios y que se diese tiempo entre petición y petición.

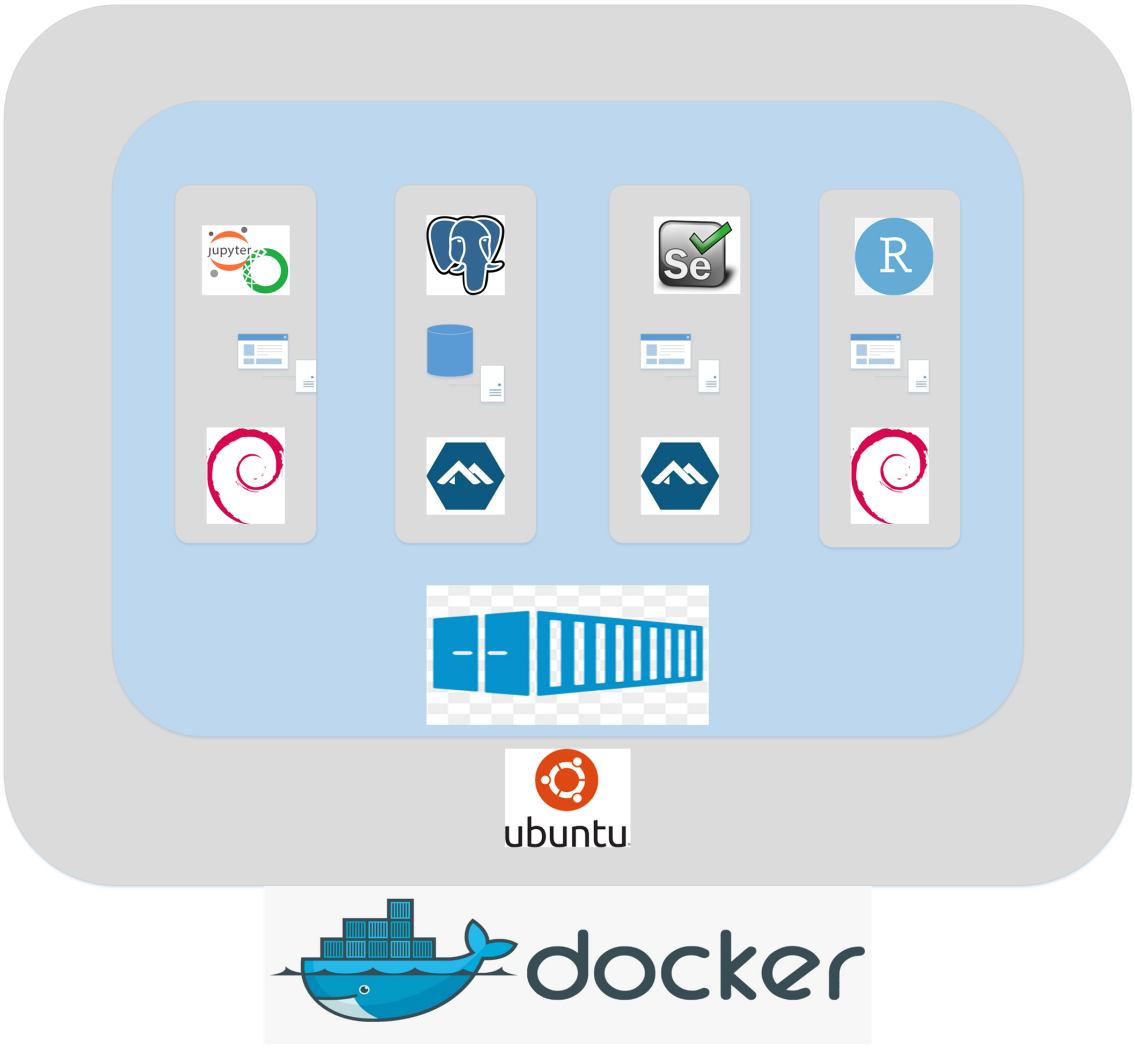


Figura 4: Infraestructura sobre la que implementó la extracción automática de noticas y el análisis de textos. En él se pueden ver las tecnologías empleadas como `Docker`, `PostgreSQL`, `Python`, etc.