ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

MODELOS ESTADÍSTICOS PARA LA DETECCIÓN DE PATRONES EN MEDIO AMBIENTE Y ECONOMÍA

MEDICIÓN DE INCERTIDUMBRE POLÍTICA ECONÓMICA DEL ECUADOR A PARTIR DEL USO DE TÉCNICAS DE MACHINE LEARNING

TRABAJO DE INTEGRACIÓN CURRICULAR PRESENTADO COMO REQUISITO PARA LA OBTENCIÓN DEL TÍTULO DE INGENIERO MATEMÁTICO

GABRIEL SEBASTIÁN AMAGUA SANDOBALIN

gs.amagua@gmail.com

DIRECTOR: PHD. MIGUEL ALFONSO FLORES SÁNCHEZ

miguel.flores@epn.edu.ec

FEBRERO 2022

CERTIFICACIONES

Yo, GABRIEL SEBASTIÁN AMAGUA SANDOBALIN, declaro que el trabajo de integración curricular aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

GABRIEL SEBASTIÁN AMAGUA SANDOBALIN

Certifico que el presente trabajo de integración curricular fue desarrollado por GABRIEL SEBASTIÁN AMAGUA SANDOBALIN, bajo mi supervisión.

PHD. MIGUEL ALFONSO FLORES SÁNCHEZ
DIRECTOR

DECLARACIÓN DE AUTORÍA

A través de la presente declaración, afirmamos que el trabajo de integración curricular aquí descrito, así como el(los) producto(s) resultante(s) del mismo, es(son) público(s) y estará(n) a disposición de la comunidad a través del repositorio institucional de la Escuela Politécnica Nacional; sin embargo, la titularidad de los derechos patrimoniales nos corresponde a los autores que hemos contribuido en el desarrollo del presente trabajo; observando para el efecto las disposiciones establecidas por el órgano competente en propiedad intelectual, la normativa interna y demás normas.

GABRIEL SEBASTIÁN AMAGUA SANDOBALIN

PHD. MIGUEL ALFONSO FLORES SÁNCHEZ

DEDICATORIA

A cada uno de los familiares y amigos que durante el transcurso de mi formación académica estuvieron presentes y con su apoyo me alentaron a seguir adelante.

RESUMEN

En promedio cada año la información disponible en la red se duplica, y como tal, blogs, redes sociales, páginas web, resultan en fuentes de grandes cantidades de información a las cuales se puede acceder por medio de técnicas de Web-Scraping.

En los últimos años se han propuesto diversas metodologías para la medición de la incertidumbre a partir del conteo de noticias relacionadas con un conjunto de palabras clave, el presente trabajo sigue de cerca dichas metodologías. Debido a la variedad de opiniones y la cantidad de datos que se pueden obtener, se hace uso del modelado de tópicos que son modelos probabilísticos, que se basan en dos suposiciones: 1) En una gran colección de documentos existen varios grupos o fuentes de texto diferentes, 2) Los textos de diferentes fuentes tienden a usar un vocabulario diferente.

En particular se hará uso del algoritmo de Asignación Latente de Dirichlet (LDA), abordando desde la problemática de escoger el número óptimo de tópicos hasta finalmente clasificar los diferentes documentos relacionados con incertidumbre, en diferentes tópicos que permitan descomponer a la Incertidumbre en sus posibles causas.

Palabras clave: Incertidumbre, Web-Scraping, LDA.

ABSTRACT

On average each year the information available on the network is duplicated and sources such as blogs, social networks, and web pages generate huge amounts of information that can be accessed through Web-Scraping techniques.

In recent years, various methodologies have been proposed for measuring uncertainty based on news counts related to a set of keywords. The present work closely follows these methodologies. I use Topics modeling due to the variety of opinions and the amount of data obtained, which are probabilistic models. These models are based on two assumptions: 1) There are several different groups or text sources in an extensive collection of documents. 2) Texts from different sources tend to use different vocabulary.

In particular, the Latent Dirichlet Allocation (LDA) algorithm addresses this work. From the problem of choosing the optimal number of topics to classifying the different documents related to uncertainty in different topics. Finally, the uncertainty to be broken down into its possible causes.

Keywords: uncertainty, Web-Scraping, LDA.

Índice general

Índice de figuras

Capítulo 1

Descripción del componente desarrollado

Muchas de las opiniones acerca de la incertidumbre se basan en impresiones y no en datos que las respalden, por lo cual la falta de una medida o aproximación de la incertidumbre no permite que se pueda estudiar su trayectoria, así como tampoco estimar su impacto en el desempeño de una economía [22].

En los últimos años, diversos autores han estudiado el impacto de la incertidumbre en diferentes ambientes, por ejemplo Blei en[6], menciona el impacto de la misma al momento de tomar decisiones dentro del hogar, de las empresas y por parte de los políticos y financieros, por otro lado Bloom en [7], expresa que la incertidumbre afecta indirectamente al crecimiento y calidad del empleo.

El aumento en el interés de la investigación por la incertidumbre ha sido impulsado por varios factores. 1) El salto en la incertidumbre en 2008 y su probable papel en la configuración de la Gran Recesión. 2) La mayor disponibilidad de indicadores empíricos para la incertidumbre, como paneles de resultados a nivel de empresa, bases de datos de noticias en línea y encuestas. 3) El salto en la capacidad computacional [4].

Recientemente, nuevas metodologías han permitido a los investigadores medir mejor la incertidumbre macroeconómica [18] y la política económica [4], lo cual ha permitido enriquecer las metodologías para la medición de la incertidumbre económica tradicionales basadas en la va-

riabilidad de los precios o el tipo de cambio. Si bien ha habido un progreso sustancial, una serie de preguntas siguen abiertas en torno a la medición, la causa y el efecto de la incertidumbre, lo que hace que esta sea un área fértil para la investigación continua [8].

Baker en [4] propone una metodología para construir un indicador empírico de incertidumbre llamado índice de incertidumbre de política económica (EPU), dicha metodología consiste en el conteo de noticias que contienen palabras relacionadas con tres categorías: Economía, Política e Incertidumbre.

Gran cantidad de indicadores de incertidumbre de política económica se han construido a partir de la metodología propuesta por Baker, entro ellos a nivel latinoamericano resaltan los construidos por Cerda [9], Perico [24] y Padilla [22] para Chile, Colombia y Ecuador respectivamente. Uno de los principales problemas que se presenta al generar la serie de noticias de incertidumbre de política económica, es la necesidad de construir un conjunto discreto de categorías formadas por un conjunto de palabras, a partir de las cuales se clasificará una noticia como perteneciente a la serie, cuando esta contiene palabras dentro de las mismas.

La creación del conjunto de categorías presenta un gran problema, ya que se debería contar con el conocimiento y el juicio para poder formar los conjuntos de palabras, y que estos comprendan los contenidos semánticos requeridos de manera correcta, Azqueta en [2] propone un enfoque menos costoso y más flexible a través del uso del modelo de tópicos Latent Dirichlet Allocation (LDA).

Por lo cual el presente trabajo de integración curricular abordará el problema de crear el conjunto de categorías para la construcción de la serie de incertidumbre de política económica, para ello se ocupará el enfoque propuesto por Azqueta, a partir del cual se buscará inferir las diferentes componentes que forman a la incertidumbre presente en el país a través de la construcción de conjuntos de palabras con contenido semántico coherente que permitan clasificar un documento.

1.1. Objetivo general

Para el desarrollo del Trabajo de Integración Curricular, se ha considerado el siguiente objetivo general:

Construir un índice de incertidumbre política económica para el Ecuador.

1.2. Objetivos específicos

Con respecto, a los objetivos específicos se han reformulado de tal forma que el alcance del plan de titulación original tenga una reducción en su elaboración, estos son:

1. Construir un corpus de noticias relacionadas con la Incertidumbre extrayendo la información a través de técnicas de web scraping.
2. Procesar el corpus de noticias, a través de técnicas de procesamiento de lenguaje natural.
3. Desarrollar un modelo de tópicos mediante el uso del algoritmo LDA.
4. Inferir campos semánticos de los resultados del algoritmo LDA.

1.3. Alcance

En este trabajo de integración curricular se propone identificar las diferentes causas que componen a la incertidumbre presente en Ecuador, medida a través del número de noticias publicadas en el diario El Comercio.

Para el desarrollo del trabajo, se implementara un modelo de tópicos en el software estadístico R, con la finalidad de relacionar los diferentes tópicos resultado del modelo con campos semánticos coherentes.

1.4. Marco teórico

1.4.1. Minería de Texto

La minería de texto consiste en descubrir nueva información, previamente desconocida de diferentes recursos escritos [16].

En [28], Witten considera que la *minería de texto* se usa para referirse a cualquier sistema que analice grandes cantidades de texto en lenguaje natural y detecta patrones de uso léxico o lingüístico, en un intento de extraer información probablemente útil.

La minería de texto es un campo multidisciplinario que involucra la recuperación de información, análisis de texto, extracción de información, agrupamiento, categorización, visualización, base de datos, tecnología, aprendizaje automático y minería de datos.

En la figura 1.1 se muestran diferentes actividades asociadas a la minería de texto.

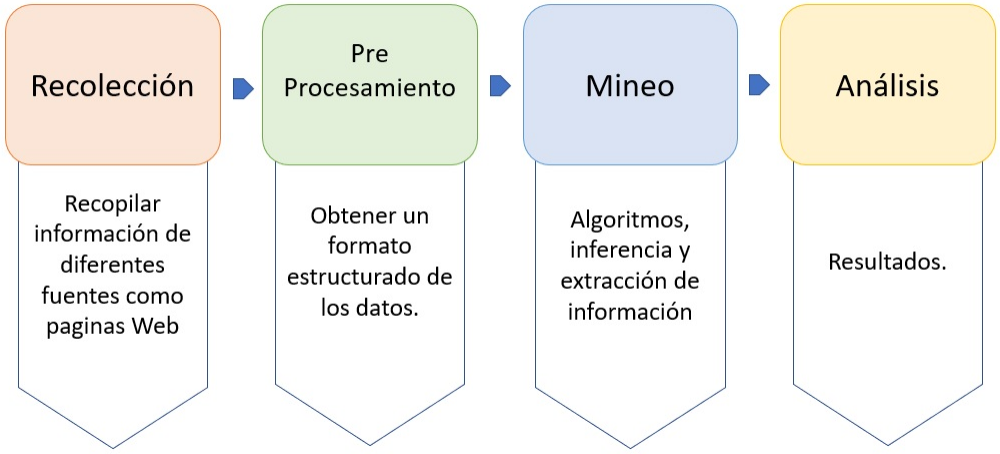


Figura 1.1: Actividades de Minería de Texto

Generalmente la forma más natural de guardar información es en texto, se cree que la minería de texto tiene un potencial comercial superior al de la minería de datos.

De hecho, un estudio reciente indicó que el 80 % de la información de una empresa está contenida en documentos de texto [1]. Sin embargo, la minería de texto también es mucho más compleja, ya que implica tratar

con datos no estructurados y confusos.

1.4.2. Procesamiento estadístico del lenguaje natural

El procesamiento del lenguaje natural se encuentra dentro de la inteligencia artificial y la lingüística, cuyo objetivo es hacer que las computadoras entiendan las declaraciones o palabras escritas en un lenguaje humano [10].

El procesamiento estadístico del lenguaje natural se basa en que todo documento se encuentra definido como un conjunto de palabras claves a las cuales se las denomina términos índice [25], además cada índice tiene un peso según su importancia, es decir, la frecuencia con la que el índice aparece en el documento, de tal forma que no se toma en consideración, la estructura, o el significado de las palabras.

El procesamiento estadístico del lenguaje natural se conforma de 2 etapas:

1. **Preprocesado de los documentos:** Durante esta etapa se busca limpiar los documentos, eliminando de ellos términos que sean considerados irrelevantes. Dentro de esta etapa se encuentran los procesos de tokenización, lematización y reducción del número de palabras.

2. **Parametrización:** Durante esta etapa, dado que se han identificado los términos clave, se procede a cuantificarlos haciendo uso de su frecuencia.

1.4.3. Modelado de Tópicos

Dentro del análisis de datos, uno de los objetivos más importantes consiste en determinar las características que los datos poseen.

En el análisis de texto, esto a menudo significa determinar qué eventos o conceptos se abordan dentro de un documento, como tal esta información es clara para un ser humano que lee un documento, pero a un programa solo se le proporciona el texto tal como está escrito y no el tema de cada documento [17].

Los modelos de tópicos son algoritmos de aprendizaje automático que se caracterizan por la posibilidad de descubrir y extraer temas latentes, o tópicos, de grandes y no estructuradas colecciones de documentos.

Los algoritmos se apoyan en las relaciones estadísticas existentes entre las palabras de los documentos para agruparlas en tópicos [11]. En particular hacen uso del enfoque llamado bolsa de palabras, es decir, se ignora el orden de las palabras dentro de cada documento de forma que para captar la estructura temática de un documento, basta con describir su distribución de palabras [20], los modelos de tópicos se basan en dos suposiciones.

1. Cada documento es una mezcla de tópicos
2. Cada tópico es una mezcla de palabras.

El resultado de un modelo de tópicos puede usarse para organizar la colección de documentos de acuerdo con los tópicos descubiertos.

Los algoritmos de modelado de tópicos se pueden adaptar a muchos tipos de datos. Entre otras aplicaciones, se han utilizado para encontrar patrones en datos genéticos, imágenes y redes sociales [5].

1.4.4. Latent Dirichlet Allocation (LDA)

Uno de los primeros y hasta hoy día más utilizado método de modelado de tópicos se conoce como Latent Dirichlet Allocation (LDA) [6], el cual es un algoritmo de aprendizaje automático no supervisado que aprende los tópicos subyacentes de un conjunto de documentos. Se basa en un enfoque probabilístico generativo para inferir la distribución de palabras que define un tópico, al mismo tiempo que anota documentos con una distribución de tópicos [2].

Si bien el algoritmo lleva varios años, su popularidad ha aumentado recientemente. Esta popularidad reciente se debe a que la potencia del algoritmo necesita un volumen de documentos relativamente grande, algo que sólo ha estado disponible en los últimos años [13].

Blei en [5], parte de intentar recrear de manera inversa el modelo teórico por el cual los textos sos generados.

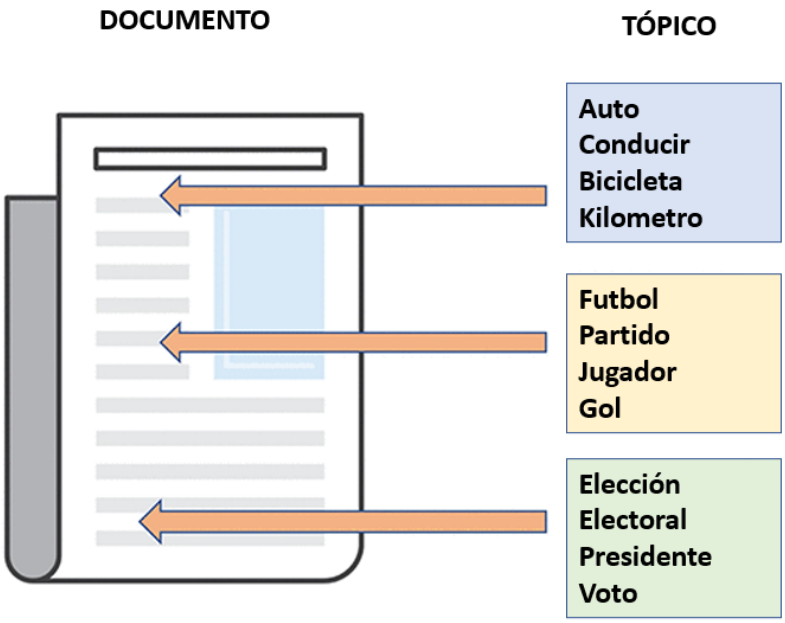


Figura 1.2: Generación teórica de textos

Según él, los autores solo disponen de un conjunto delimitado y cerrado de tópicos, cada uno de esos tópicos contiene palabras, y el autor va sacando palabras de los diferentes tópicos para escribir su texto, como se observa en la figura 1.2.

El método depende de la suposición de que cada documento exhibe una combinación aleatoria de dichos tópicos y que todo el documento se generó mediante el siguiente proceso de dos pasos que se basan en las suposiciones de un modelo de tópicos vistos en el apartado anterior:

1. Para cada documento d en una colección D, existe una distribución aleatoria θ_d sobre K tópicos donde cada valor de $\theta_{d,k}$ representa la proporción del tópico k en el documento d.

2. Para cada palabra w en el documento d, se selecciona un tópico z de θ_d y se observa su distribución en un vocabulario fijo dado por β_z [11].

Un modelo LDA que se construye en base a los dos puntos anteriores tiene la siguiente estructura gráfica.

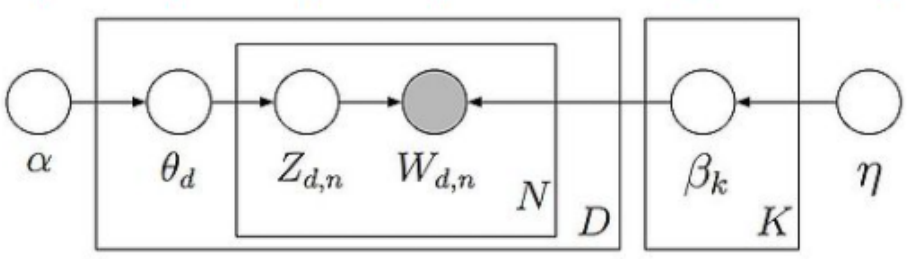


Figura 1.3: Representación gráfica LDA

Notación:

α = Hiper parámetro para la proporción de tópicos por documento.

θ_d = Distribución de los tópicos en el documento d

$Z_{d,n}$ = Tópico asignado a la n-ésima palabra en el documento d.

$W_{d,n}$ = n-ésima palabra en el documento d.

η = Hiper parámetro para la proporción de palabras por tópico.

βk = Distribución de palabras del tópico k.

K = Número de tópicos.

N = Número de palabras en el documento.

D = Número de documentos dentro del corpus.

LDA pertenece al grupo de modelos bayesianos, un modelo bayesiano cuenta con distribuciones de probabilidad *a priori* y *a posteriori*.

LDA consta de dos distribuciones a priori diferentes: distribución de probabilidad de tema - documento y distribución de probabilidad de palabras - tópico las cuales corresponden a una distribución de Dirichlet [5].

El modelo parte de aquellas dos distribuciones obteniendo los parámetros del modelo que maximizan la probabilidad de que cada palabra aparezca en cada artículo dado el número total de temas K. La probabilidad de que aparezca una palabra w_i en un artículo viene dada por:

$$P(w_i) = \sum_{j=1}^{K} P(w_j|z_j)P(z_j = j) \tag{1.1}$$

donde z_j es una variable latente que indica el tópico en el cual la i-ésima palabra se encuentra, $P(w_j|z_j)$ es la probabilidad de que la palabra w se extraiga del tópico j, y $P(z_j = j)$ es la probabilidad de extraer una palabra del tema j en el artículo i.

$P(w|z)$ indica que palabras son importantes para un tópico, mientras que $P(z)$ indica cuáles de esos tópicos son importantes para un artículo. Por lo tanto, el objetivo es maximizar $P(w_j|z_j)$ y $P(z_j = j)$ de la ecuación 1.1 [2].

1.4.5. Validación de modelos de tópicos

Coherencia semántica

Este indicador mide si las palabras de un tópico tienden a coexistir, es decir, se puede lograr un valor alto de coherencia semántica si las palabras con una alta probabilidad para cada tema a menudo aparecen juntas dentro de los documentos.

$$\sum_{i} \sum_{j<i} log \frac{D(w_j, w_i)}{D(w_i)} \rho \tag{1.2}$$

Donde, $D(w)$ es el número de documentos que contienen al menos un token de tipo w, y $D(w_1, w_2)$ es el número de documentos que contienen al menos un w_1 y un w_2.

Para evitar errores de registro cero, se agrega el parámetro de suavizado ρ. Al ser probabilidades logarítmicas, son negativas y los valores negativos grandes indican palabras que no se repiten con frecuencia; los valores más cercanos a cero indican que las palabras tienden a coexistir con más frecuencia.

Exclusividad de tópicos

Este indicador mide hasta que punto las palabras principales de un tópico no aparecen como palabras principales en otros tópico, es decir, la medida en que sus palabras principales son exclusivas. El valor es el promedio, sobre cada palabra principal, de la probabilidad de esa palabra en el tópico dividida por la suma de las probabilidades de esa palabra en todos los tópicos.

Capítulo 2

Metodología

2.1. Creación del Corpus de Noticias

En esta sección se detallará el procedimiento realizado para la construcción del corpus de noticias.

2.1.1. Fuente de Información

Diversos medios de comunicación a nivel nacional disponen de páginas web, de las cuales mediante el uso de técnicas de web scraping, se puede extraer un gran número de diversa información.

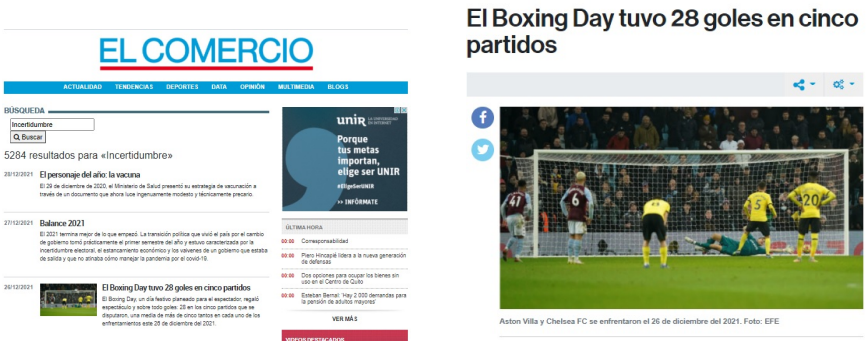
El Comercio es considerado como el diario de mayor circulación a nivel nacional, por cual se ha decidido usar al sitio web del diario como la fuente de información para el presente trabajo, es de considerar además que el sitio web ofrece diversas ventajas en su estructura lo que facilitará el proceso de extracción de noticias.

2.1.2. Extracción de Noticias

El sitio web de El Comercio, ofrece un sistema de búsqueda de artículos basado en etiquetas, con lo cual en primer lugar se buscó dentro de la pagina web los artículos relacionados con las palabras {*Incertidumbre,*

Incierto, Incerteza}, como resultado de la búsqueda se despliegan hasta diez artículos por página web, si existen más, el sitio web los divide entre n páginas.

A través de cada uno de los artículos mostrados como resultado de la búsqueda, se puede acceder al sitio web del mismo, donde se encontrará el texto de la noticia.



(a) Búsqueda de Artículos (b) Artículo

Figura 2.1: Noticias relacionadas con Incertidumbre

Para la extracción de las noticias se hará uso del paquete **rvest** [26], el cual ofrece una variedad de funciones para extraer el contenido de una página web.

En la figura 2.2 se presenta la estructura que se siguió para la extracción de noticias.

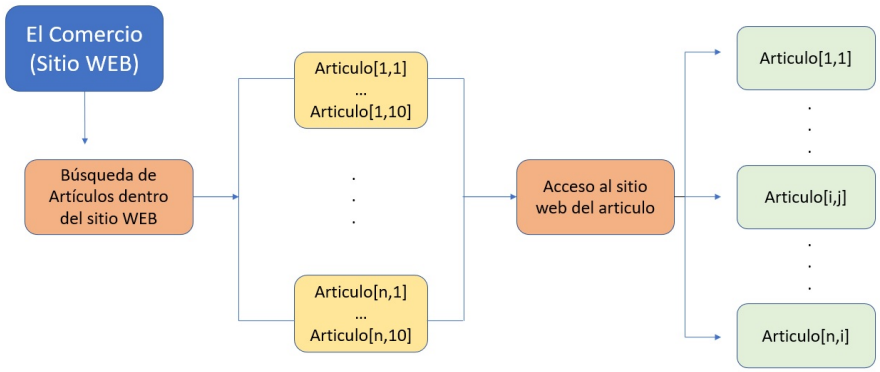


Figura 2.2: Estructura del acceso a las noticias.

.

Durante la extracción de información es de importancia identificar correctamente los elementos que se desea extraer, en el presente trabajo se

ha hecho uso de la herramienta **SelectorGadget** de Google Chrome.

Figura 2.3: Identificación de elementos.

.

En la figura 2.3, se muestra a la herramienta **SelectorGadget** siendo utilizada para identificar la fecha de una noticia.

Los elementos extraídos fueron los siguientes:

Elemento	**Descripción**
titulo	Título de la noticia.
fecha	Fecha de publicación de la noticia.
link	Dirección web de la noticia.
texto	Noticia completa.

El corpus de noticias resultante consta de 5277 noticias en un intervalo de tiempo desde Abril del 2009 hasta diciembre del 2021, el corpus de noticias fue almacenado dentro de un archivo de valores separado por comas *(csv)*, con codificación utf-8[1] para no perder la estructura del lenguaje español.

Un extracto de las noticias recolectadas, se muestra a continuación:

[1]utf-8 - Tipo de Codificación de Caracteres

titulo	fecha	link	texto
Bélgica cierra cines...	22/12/2021	https://www.elcom...	Bélgica ha endureci...
Debemos cumplir l...	22/12/2021	https://www.elcom...	En noviembre fue u...
Cartas al director / ...	22/12/2021	https://www.elcom...	Estado vs. narco del...
Ómicron: cómo act...	20/12/2021	https://www.elcom...	Ómicron suena a ap...
Las amenazas de Ó...	20/12/2021	https://www.elcom...	Ecuador se suma a l...
Europa afronta la ...	19/12/2021	https://www.elcom...	La imparable expan...

Figura 2.4: Extracto de noticias recolectadas

2.1.3. Análisis Exploratorio

Una de las expectativas durante este proyecto, es la búsqueda de un histórico de noticias lo bastante antiguo como sea posible, con lo cual, durante el proceso de levantamiento de noticias nos podemos encontrar con algunos errores como; el *error 404*[2], y la existencia de noticias duplicadas.

(a) Error 404

(b) Noticias Duplicadas

Figura 2.5: Errores al recolectar noticias.

En la figura 2.5a se observa que algunos textos de las noticias ya no se encuentren disponibles, los cuales al momento de levantar la información, fueron considerados como datos faltantes.

[2]Error 404 - El recurso no está disponible en el servidor

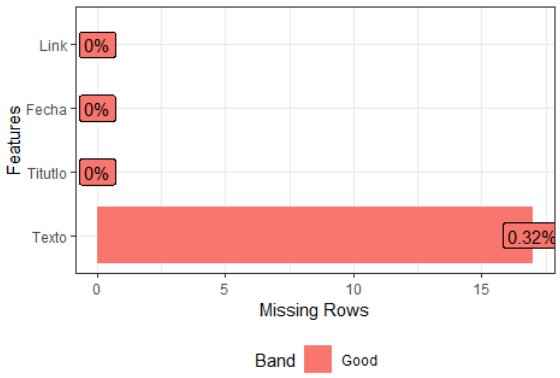


Figura 2.6: Casos Completos

En la figura 2.6, se observa que solo el $0{,}32\,\%$ de los casos están incompletos, al no presentar información en el apartado del Texto de la noticia, aunque representan un porcentaje pequeño dentro del corpus de noticias, se los eliminará pues uno de los objetivos del presente trabajo requiere del manejo de los textos de las noticias, se eliminarán además aquellos datos que estén duplicados.

Como resultado a partir de este punto se trabajará con el corpus de noticias conformado de 5218 observaciones, desde el 6 de Abril del 2009 hasta el 22 de Diciembre del 2021.

2.2. Pre-procesamiento de texto

La incertidumbre se encuentra presente en diferentes escenarios, como la Salud, la Política e incluso el Deporte, por ejemplo el siguiente fragmento de noticia posee dentro de su *wording*[3] la palabra incertidumbre y como tal la noticia esta relacionada con el deporte.

"Piero Hincapié se fogueó en el exterior para llegar a la Tricolor"
"Los defensores cotizados en el fútbol internacional, además de sus condiciones técnicas, superan el 1,80 metros. Por ello, había incertidumbre sobre si el esmeraldeño alcanzaría esa estatura para rendir en la alta competencia."
Fuente: El Comercio

Al ser un solo fragmento corto, su interpretación puede resultar sencilla, pero es una historia completamente diferente si se habla de grandes cantidades de documentos extensos.

En esta sección se detallaran todos los pasos previos para que la cantidad de noticias que se tiene sean interpretables por una máquina que no habla ni entiende nuestro idioma.

Para el presente trabajo se hará uso de la librería **UdPipe** [27] la cual consta de herramientas de procesamiento de lenguaje natural como 'tokenización', 'etiquetado de partes del habla', 'lematización' y 'análisis de dependencia' de texto sin formato e independiente del idioma.

2.2.1. Tokenización y Lematización

El proceso de tokenización, consiste en tener como unidad mínima del análisis a cada palabra dentro del documento, cada una de estas será conocida como *token*, mientras que la lematización es el proceso en el cual dadas todas las diferentes formas flexionadas de una palabra se halla su forma base o lema.

Para realizar lo anteriormente mencionado, se hará uso de las siguientes funciones dentro de la librería.

[3]Wording - Conjunto de palabras usadas

Cuadro 2.1: Funciones UDPipe

UDPipe	
Función	**Descripción**
udpipe_download_model	Descarga el modelo de idioma a usar
udpipe_load_model	Almacena el modelo de idioma en memoria
udpipe_annotate	Tokenización, lematización

La salida de **udpipe_annotate**, consta de varias variables, de las cuales solo se hará uso de las siguientes.

Cuadro 2.2: Variables pre-procesamiento de texto.

Variable	**Descripción**
doc_id	Identificación del documento
token	Descomposición por palabras
lemma	Lema o raíz de la forma de la palabra
upos	Etiqueta universal de partes del habla.

La variable *upos*, marca las categorías principales de la parte gramatical, para distinguir propiedades léxicas y gramaticales adicionales de las palabras, usando características universales, un resumen de las categorías, se muestra a continuación:

Cuadro 2.3: Categorías gramaticales

Categoría	**Representación**
ADJ	Adjetivos
NOUN	Sustantivos
PROPN	Nombres Propios
VERB	Verbo
ADV	Adverbio.
...	...

Como resultado del proceso de tokenizado y lematizado se obtuvo un data-frame que consta de 3,605,825 de observaciones, de las variables mencionadas en 2.2.

A continuación se presenta un extracto del resultado.

doc_id	token	lemma	upos
1	Bélgica	bélgica	PROPN
1	ha	haber	AUX
1	endurecido	endurecer	VERB
1	las	el	DET
1	restricciones	restricción	NOUN
1	contra	contra	ADP

Figura 2.7: Resultado tokenización y lematización

En la figura 2.7, se observa como dentro de cada documento cada palabra ha pasado a ser considerada como unidad en el análisis (tokenización), también se tiene que cada palabra fue transformada a su forma raíz o lema.

Por ejemplo se considera la siguiente frase, extraída del primer documento de nuestro corpus de noticias:

Bélgica ha endurecido las restricciones contra la pandemia

Al tomar, las raíces de las palabras anteriores, se tiene.

bélgica haber endurecer el restricción contra el pandemia

Además cada palabra se encuentra categorizada por la variable **upos**, la cual como se mencionó antes representa las partes gramaticales, por ejemplo se observa que la palabra *endurecido*, es categorizado como verbo.

La gran cantidad de observaciones que se tiene corresponde a cada una de las palabras dentro del wording del corpus de noticias, pero no todas las palabras como tal proveen de información relevante, por lo cual el próximo paso es abordar la reducción del número de palabras

2.2.2. Reducción del número de palabras.

Tipo de palabras

En la sección anterior, se categorizó cada palabra respecto a su parte gramatical, en esta sección se seleccionaron para trabajar las palabras dentro de las siguientes categorías: adjetivos (ADJ), verbos (VERB), sustantivos (NOUN) por ser las partes más importantes dentro de una oración y nombres propios (PROPN) debido a la naturaleza de las noticias, ya que nombres de países o mandatarios resultan de interés en el análisis.

Stopwords

Conocidas como *palabras vacías*, las *stopwords* son palabras frecuentemente usadas en un idioma, y como tal no proveen de un valor de información para la comprensión del contexto de un documento, ejemplos de estas palabras en el idioma español, son artículos tales como: el, la, los, las entre otras.

Palabras típicas

Debido a la temática general del corpus de noticias, palabras como incertidumbre, incierto e incerteza estarán presentes frecuentemente, otras palabras que se consideraron como frecuentes son aquellas relacionadas a periodos de tiempo, día, mes y año.

Como resultado se obtuvo un data-frame que consta de 1,334,616 observaciones. A continuación se muestra un extracto del resultado del proceso de reducción del número de palabras.

doc_id	lemma
1	bélgica
1	endurecer
1	restricción
1	pandemia
1	víspera
1	navideño

Figura 2.8: Reducción de palabras

Como se observa en la figura 2.8, a comparación de las figura 2.7, palabras como *el* y *haber*, ya no se encuentran. Además a partir de este punto se trabajará con los lemas de las palabras.

2.2.3. Vectorización

Al principio de esta sección se planteo como objetivo conseguir que los documentos presentes en el corpus de noticias se conviertan en estructuras que sean interpretables por una máquina, en este apartado se desarrollará los pasos finales a dicho objetivo.

Generalmente, los modelos de machine learning son entrenados con datos estructurados en forma de tablas, cuando se trabaja con texto se debe construir estas tablas a partir de las palabras del documento con el que se este trabajando. Esto se consigue con el vectorizado.

Para este proceso se considerará la matriz de términos o TDM por sus siglas en inglés, misma que es una forma de representar las palabras en el texto como una tabla (o matriz) de números, la cual describe la frecuencia de las palabras que ocurren en una colección de documentos. En una TDM, las filas corresponden a los documentos de la colección y las columnas corresponden a las palabras.

Para la construcción de matriz de términos del corpus de noticias, se hará uso de la lematización antes realizada, partiendo del resultado ob-

tenido en el paso de reducción del número de palabras, basta con contar cuantas veces aparece una palabra.

Como resultado se obtuvo una matriz de 5218 filas por 60606 columnas. En este punto se realizó una nueva reducción el número de palabras, usando la estructura de la matriz, se eliminaron aquellas palabras únicas, es decir, que aparecen en muy pocos documentos, para este caso, se eliminaron todas aquellas palabras que solo aparezcan en menos de 5 noticias, por otro lado también se eliminaron aquellas palabras que aparezcan en más del 90 % de las noticias, tras lo cual se obtuvo una matriz de 5218 filas por 17474 columnas, a continuación se presenta un extracto de la TDM del corpus de noticias obtenida.

Cuadro 2.4: Extracto TDM

Docs	económico	ecuador	gobierno	llegar	millón	parte	pasado
2338	5	14	13	2	6	8	1
2478	2	3	22	4	10	6	5
2509	4	16	10	3	2	1	1
2750	3	9	11	3	3	5	4
3330	4	17	17	5	18	1	4
3456	2	8	11	2	6	4	0
3468	0	14	13	3	8	5	1
3551	8	6	13	3	7	0	0
422	39	120	89	21	63	13	6
431	16	58	41	11	21	6	3

La interpretación de la tabla interior, se sigue de la siguiente manera: El documento 2338 contiene dentro de su wording la palabra *económico* un total de 5 veces, mientras que dentro del documento 2478 aparece 2 veces por otro lado la palabra *ecuador* aparece un total de 14 veces y 3 veces en los respectivos documentos.

2.3. Modelado de Tópicos

En este capítulo se detalla la metodología empleada para entrenar uno de los algoritmos del modelado de tópicos.

Para construir el modelo de tópicos se hará uso del algoritmo Latent Dirichlet Allocation, a través de la librería **topicmodels** [14].

Posteriormente, en el momento del análisis de los resultados del modelo, se buscará inferir una temática a partir de las palabras que más contribuyen a cada tópico. Según Blei en [5], la interpretabilidad de la mayoría de los tópicos es el resultado de "la estructura estadística del lenguaje y cómo interactúa con los supuestos probabilísticos específicos de LDA".

Al ser LDA un modelo parametrizado, se debe determinar previamente el número de tópicos k; dado que por el momento solo existen lineamiento generales para determinar el valor de k se comenzará abordando dicho problema.

2.3.1. Número de Tópicos

En [21] Nikita comenta que el mejor camino para estimar el número de tópicos k, es un proceso en el cual se debe considerar el número de iteraciones y la estimación de tópicos, ya que si se toma un número muy grande o pequeño esto infiere en la credibilidad del valor de k.

Si k es muy pequeño, los tópicos obtenidos corresponderán a pocos campos semánticos muy generales, mientras que para números elevados de k, no se suelen obtener tópicos interpretables.

Algunos estudios como, [20] y [15] incluso consideran que para un valor de k mayor a 50, los resultados no serán humanamente interpretables, debido a la dificultad de leerlos sin confundir u olvidar los tópicos, por lo cual en el presente proyecto se realizó la búsqueda del número de tópicos entre 2 y 60.

Deveaud en [12], propone una metodología para determinar el número de tópicos basada en la maximización del índice de coherencia, en diversos trabajos como [3], [23] han empleado dicha metodología junto al

índice de exclusividad.

En una fase inicial, se presentan los valores del índice Deveaud2014 entre 3 y 60 tópicos en un intervalo de 3, 2.9.

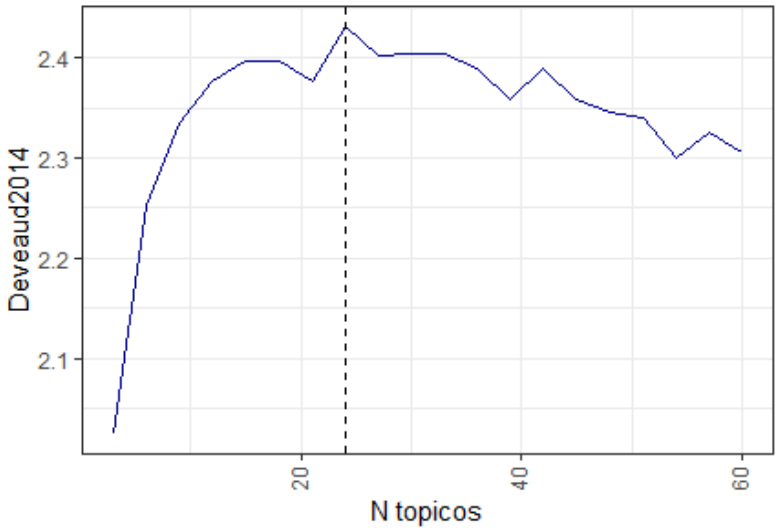


Figura 2.9: Variación del índice Deveaud2014

Una primera inspección sugiere que el número de tópicos apropiado se encuentra entre 20 y 30.

En un siguiente paso, se prueban diferentes modelos entre 20 y 30. En la figura 2.10 se muestran los valores de coherencia semántica para cada modelo mientras que en la figura 2.11, se muestran los valores de coherencia semántica y de exclusividad para cada modelo.

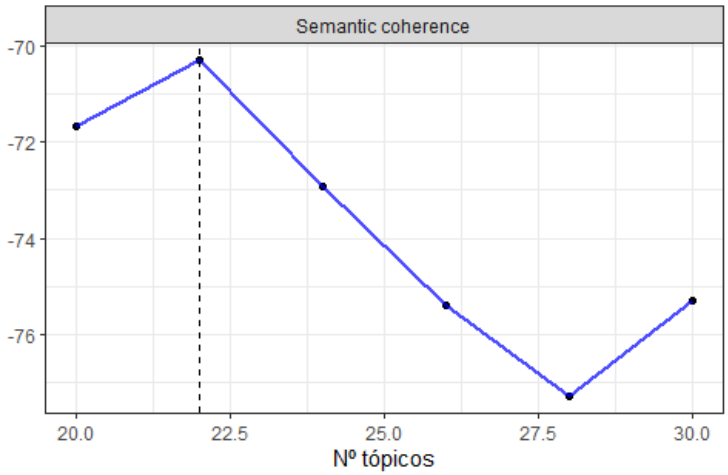


Figura 2.10: Coherencia semántica

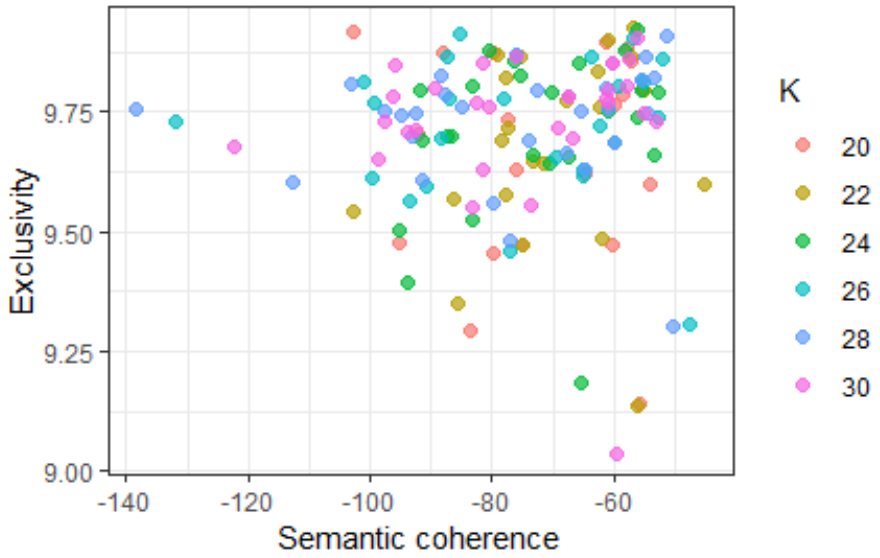


Figura 2.11: Exclusividad - Coherencia

Un valor más alto de coherencia semántica indica una mayor interpretabilidad del tópico, mientras que un tópico puede considerarse exclusivo si las palabras con alta probabilidad de ocurrencia condicional a un tópico en particular tienen baja probabilidad condicional a otros tópicos [19]. Por lo cual las figuras 2.10 y 2.11 sugieren que se utilice un modelo con 22 tópicos.

2.3.2. Selección del modelo

En el apartado anterior se llego a la conclusión de usar un modelo con 22 tópicos, una inspección manual, sugirió que el modelo de 22 tópicos tiene el mejor poder interpretativo entre los diferentes modelos que se evaluaron.

A continuación se muestra un resumen de los principales términos (palabras) de cada tópico, además del título de la noticia más representativo dentro del tópico.

Tópico	**Palabras principales**	**Título más representativo**
1	presidente, nacional, comisión, proyecto, ley, asamblea	*"Oposición y oficialismo miden fuerzas"*
2	familia, saber, hijo, casa, padre	*"Bertha: 'Nunca podremos decirle adiós a mi hermana Inés y Filadelfio, nunca sabremos si son sus cuerpos'"*
3	mejor, primero, obra, historia, mundo	*"Los Globos de Oro premian a Michael Douglas y 'The Americans'"*
4	primero, euro, españa, europeo, español,	*"La carrera en Reino Unido para suceder a Theresa May toca a su fin con la amenaza de un Brexit brutal."*
5	impuesto, sector, empresa, comercio, producto,	*"La venta de etanol para elaboración de la gasolina ecopaís se redujo."*
6	ecuatoriano, ecuador, querer, creer, correa,	*"Siete candidatos presentaron sus propuestas en el primer día del Debate"*

7	marzo, llegar, medida, lunes, abril	*"Ministerio de Turismo aclara que la salida de vuelos internacionales no está prohibida en el país"*
8	policía, seguridad, militar, paz, fuerza,	*"Condenan a 28 años de prisión a secuestrador del equipo periodístico de El Comercio"*
9	ecuador, venezuela, colombia, méxico, venezolano	*"Es necesario dejar puertas abiertas para ciudadanos venezolanos, según responsable OEA"*
10	salud, pandemia, persona, caso, coronavirus	*"Así funciona el bamlanivimab, el primer medicamento específico contra la covid-19"*
11	primero vez, cambio, gran, último	*"Actividad volcánica bajo el mayor y más inestable glaciar antártico"*
12	social, medio, diario, red, información	*"Facebook, Instagram y WhatsApp, con errores de conexión a escala mundial"*
13	millón, pagar, dinero, deuda, crédito	*"La reserva, en su punto más bajo desde el 2012"*
14	persona, centro, casa, ciudad, zona	*"Cierre de la intersección de la av. Mariscal Sucre y Rodrigo Chávez genera congestión vehicular en el sur de Quito"*
15	educación, trabajo, público, universidad, estudiante	*"Déficit de 6 000 docentes en Ecuador"*
16	derecho, norma, poder, jurídico, constitución	*"La certeza: Virtud de la legalidad"*

17	tiempo, social, humano, vez, vida	*"Todo vale"*
18	candidato, presidente, política, partido, político	*"Sondeos de boca de urna dicen que Luis Arce ganó las elecciones en Bolivia en primera vuelta"*
19	primero, jugar, partido, final, jugador	*"12 clubes vuelven a la disputa del Campeonato ecuatoriano"*
20	economía, económico, dólar, precio, fmi	*"La economía ecuatoriana crecerá 0,2% este año, según el FMI"*
21	presidente, acuerdo, ee.uu, Trump, obama	*"Obama da un primer paso para encauzar las relaciones con el Golfo Pérsico"*
22	empleo, económico, público, obrero, política	*"Inquietudes nacionales"*

Cuadro 2.5: Modelo con 22 tópicos

En el cuadro 2.5, las palabras principales son extraídas de la distribución de probabilidad palabras - tópico, y corresponden a aquellas palabras con mayor probabilidad dentro del tópico. Mientras que el título más representativo corresponde a la distribución de probabilidad tópico - documento y representa al documento cuya probabilidad de pertenecer a un tópico es mayor.

La distribución de probabilidad tópico - documento, permite etiquetar a cada documento con aquellos tópicos cuya proporción dentro del documento es significativa, mientras que la distribución de probabilidad palabras - tópico permite realizar un proceso cualitativo para inferir campos semánticos dentro de los tópicos, estos puntos se abordaran en apartados posteriores.

2.3.3. Etiquetado de los documentos

En este apartado se detalla la metodología que se siguió para etiquetar las diferentes noticias dentro del corpus.

Se comienza mostrando las probabilidades tópico - documento para las primeras cinco noticias dentro del corpus.

Tópico	Noticia 1	Noticia 2	Noticia 3	Noticia 4	Noticia 5
1	0.019	0.010	0.011	0.006	0.033
2	0.025	0.044	0.135	0.014	0.012
3	0.061	0.010	0.029	0.005	0.022
4	0.058	0.044	0.004	0.004	0.022
5	0.016	0.018	0.013	0.011	0.028
6	0.028	0.023	0.064	0.017	0.012
7	0.203	0.023	0.010	0.014	0.086
8	0.013	0.044	0.109	0.006	0.022
9	0.028	0.018	0.029	0.006	0.038
10	0.279	0.023	0.057	0.693	0.359
11	0.010	0.239	0.010	0.048	0.012
12	0.013	0.014	0.013	0.033	0.022
13	0.016	0.036	0.015	0.006	0.028
14	0.046	0.010	0.006	0.013	0.028
15	0.019	0.027	0.066	0.005	0.028
16	0.022	0.023	0.124	0.007	0.070
17	0.013	0.049	0.183	0.027	0.022
18	0.028	0.014	0.008	0.006	0.028
19	0.043	0.014	0.010	0.005	0.017
20	0.043	0.062	0.008	0.007	0.028
21	0.010	0.161	0.011	0.013	0.012
22	0.007	0.092	0.086	0.056	0.070
Suma	1	1	1	1	1

Cuadro 2.6: Probabilidades Tópico - Documento

En el cuadro 2.6, se observa que cada noticia posee una probabilidad de pertenecer a cada tópico, dicha probabilidad es más significativa para ciertos tópicos por documento, por ejemplo la primera noticia, tiene una mayor probabilidad de pertenecer al tópico 10 y una probabilidad menor de pertenecer al tópico 22.

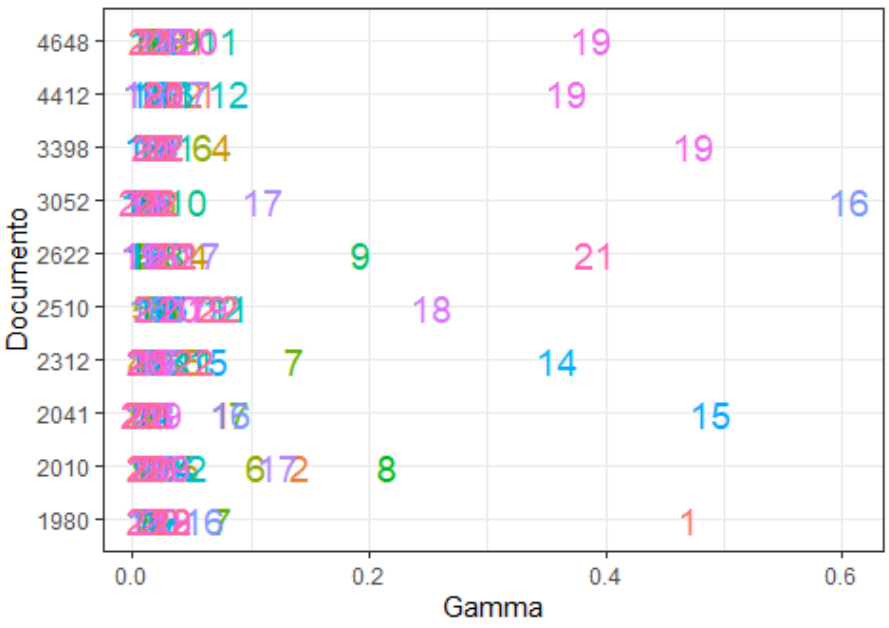


Figura 2.12: Probabilidad Tópico-Documento

En la figura 2.12, se presentan las probabilidades tópico - documento para una muestra aleatoria de documentos, visualmente cada uno de ellos presenta una mayor probabilidad en ciertos tópicos y por otro lado una gran cantidad de tópicos se agrupan en probabilidades de bajo valor.

Etiquetar un documento con tópicos de baja probabilidad puede convertirse en un error, debido a que dentro de cada documento dichos tópicos no se encuentran bien representados, con el objetivo de asignar a cada documento con tópicos que lo puedan explicar, se hará uso del criterio de Pareto para lo cual se propone que los tópicos que tengan una probabilidad de al menos el 20 % describirán al menos el 80 % de lo que trata el documento.

A partir de este punto, cada noticia fue etiquetada con todos los tópicos que tengan una probabilidad superior al 20 % de representar dicha noticia, así pues de ser el caso algunas noticias podrán pertenecer a más de un tópico a la vez o por lo contrario no pertenecer a ninguno.

2.3.4. Etiquetado de los tópicos

Partiendo de la idea de un proceso de codificación cualitativo en este apartado se formaliza el hecho de que los tópicos se encuentran relacionados con alguna temática en particular.

Para realizar este proceso cualitativo, se uso como fuente de información, las palabras principales dentro de cada tópico, se evaluó si existe un campo coherente entre las palabras y finalmente se asigna una etiqueta que describa al tópico.

A continuación se muestra el análisis realizado para cada tópico.

1 Palabras como asamblea, asambleísta, ley y proyecto son nociones relacionas con **política asamblearia**.

2 Familia, padre, madre, hijo y casa son claramente conceptos relacionados con un entorno **familiar**.

3 Obra, serie, artista, película, libro, corresponden a **entretenimiento**.

4 Para este tópico la presencia de palabras como Europa, España y Reino Unido lleva a considerar que esta relaciona con un aspecto **internacional (Europa)**.

5 Dentro de este tópico palabras como empresa y comercio son las de mayor importancia y en adición se encuentran palabras como impuesto, venta, mercado, negocio, así se considera que se relaciona con **empresa y comercio**.

6 Público, ciudadano, presidente, ecuatoriano, se encuentran dentro de este tópico como tal hace referencia a política, pero se recalca la presencia de palabras como Correa, Rafael y Moreno con lo que se tiene referencia al movimiento político **correista**.

7 En este caso palabras como medida, lunes, abril, marzo, diciembre no parecen tener un campo coherente detrás, una investigación más profunda llevo a identificar que las noticias dentro de este tópico, hablan sobre medidas tomadas ante diferentes situaciones de

impacto en movilidad, cierre de rutas aéreas, y cierre de vías, con lo cual se considera que este tópico se encuentra relacionado con **restricciones de movilidad**.

8 Policía y militar son las palabras más representativas junto a otros como, seguridad, violencia, fuerza, víctima lo cual sugiere que el tópico tiene relación con **Las fuerzas armadas y policía nacional**.

9 La presencia de la palabra migrante entre el grupo de palabras más representativas junto a otras palabras como Chile, Venezuela, Colombia, crisis, frontera son referencias a la **Migración**.

10 Este tópico resulta bastante sencillo de analizar pues dentro de sus palabras principales se encuentran, salud, enfermedad, pandemia, coronavirus, términos referentes a **Salud**.

11 La presencia de palabras como tierra, agua, clima, cambio, tiempo, llevaron a considerar su relación con el **medio ambiente**.

12 Medios, comunicación, periodista, diario, prensa, noticia entre otras son referentes a **medios de comunicación**.

13 Millón, deuda, pago, dinero, fondo e interés son algunas de las palabras más representativas del tópico, pero junto a banco, sugiere que el tópico se encuentra asociado a la actividad **bancaria**.

14 Metro, calle, vivienda, agua, zona, ciudad la idea que proveen estas palabras es estar hablando sobre los eventos dentro de los **política municipal**.

15 Dadas las palabras principales, educación, colegio, universidad la conclusión es que el tópico esta relacionado con la **educación**.

16 Juez, derecho, ley, constitución, son algunas de las palabras que guardan relación con la **función judicial**.

17 Vivir, personal, mismo, son conceptos relacionados con cada persona, dentro de este tópico se encontró una relación con noticias que tratan sobre la percepción de una persona de lo que vendrá en el futuro, al parecer un tema recurrente y por lo cual se lo ha relacionado con aquello que sera nombrado como incertidumbre **personal**.

18 Elección, electoral, voto, partido son claras referencias a la **política electoral**.

19 Como se mencionó en apartados pasados, este tópico se encuentra relacionado con el ambiente **deportivo**.

20 Economía, económico son la base para considerar que el tópico trata sobre economía, profundizando la investigación, diferentes palabras como fmi, petróleo, precio y dólar llevaron a considerar que el tópico se relaciona con la **economía ecuatoriana**, pues como bien se sabe, el petróleo y el fondo monetario internacional son pilares en la economía del país.

21 Estados Unidos, Trump, Obama, acuerdo son referentes a un aspecto **internacional (USA)**.

22 Dentro de este tópico la palabra economía, llevo a considerar que tiene que ver al igual que el tópico 20 con la economía ecuatoriana, al examinar a profundidad y en conjunto a otras palabras como social, empleo, cambio, políticas, se determinó que se encuentra más relacionado con **políticas laborales** y su impacto en la economía.

Los nombres de cada tópico no son un resultado del modelo, sino una descripción hecha para el presente proyecto del tema que trata cada agrupación de palabras.

Capítulo 3

Resultados, conclusiones y recomendaciones

3.1. Resultados

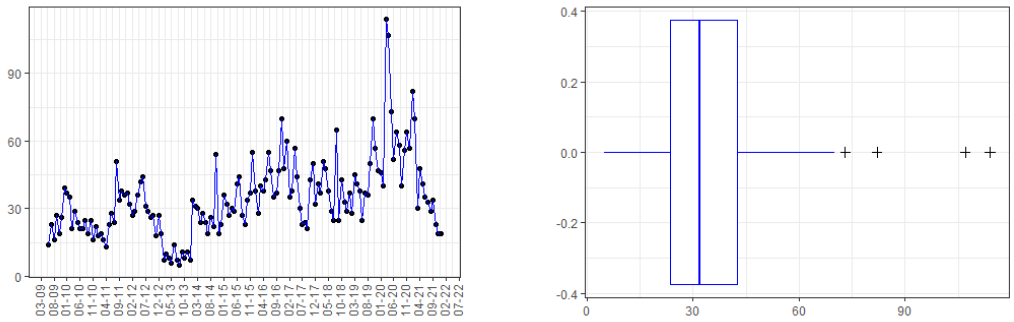
3.1.1. Niveles de incertidumbre en Ecuador

Sea **N(t)**, el número de noticias relacionadas con incertidumbre publicadas durante un periodo de tiempo **t**, a continuación se muestra un resumen de las principales estadísticas de la variable para un periodo de tiempo anual, mensual y diario.

	Anual	**Mensual**	**Diaria**
Nº	13	151	2869
Mínimo	113	5	1
Media	401.4	34	1.82
Máximo	771	114	12
Sd	167.28	17.41	1.18

Cuadro 3.1: Número de noticias

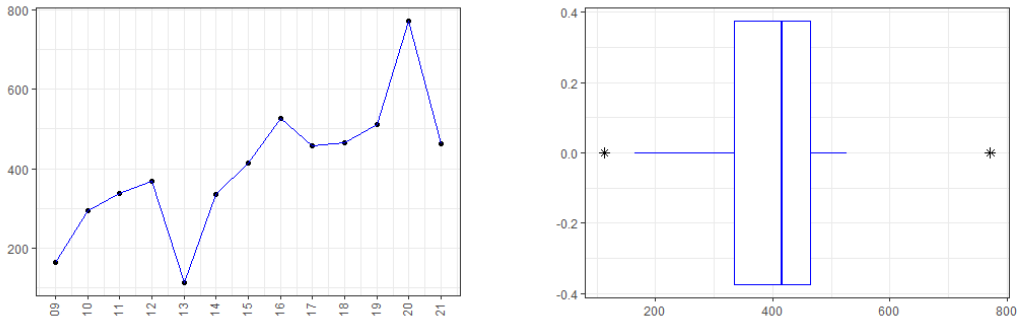
Del cuadro 3.1, se nota que, el número máximo de noticias relacionadas con incertidumbre publicadas al día es de 12, por mes 114 y por año 771, mientras que en media se publican 1.82, 34 y 401.4 respectivamente de lo cual se concluye que existen instantes de tiempo en los cuales los niveles de incertidumbre se elevan.



(a) Variación (b) Diagrama de Caja

Figura 3.1: Número de noticias mensuales

En la figura 3.1, se muestran representaciones gráficas de interés respecto al número de noticias en frecuencia mensual, se puede observar en el diagrama de caja 3.1b, la existencia de valores inusuales, correspondientes a meses de alta incertidumbre.



(a) Variación (b) Diagrama de Caja

Figura 3.2: Número de noticias anuales

En la figura 3.2, se presentan representaciones gráficas de interés respecto al número de noticias en frecuencia anual, se puede observar en el diagrama de caja 3.2b, la existencia de valores inusuales, correspondientes a años de alta incertidumbre, pero también de muy poca incertidumbre, lo cual se puede observar de mejor manera en la figura 3.2a, donde se observa como ha variado la incertidumbre medida por el número de noticias a través de los años.

En la figura 3.2a, se observa una tendencia creciente durante los primeros años, pero abruptamente el valor cae en el año 2013, lo cual coincide con el año de aprobación de la Ley Orgánica de Comunicación (LOC),

por otro lado el valor más alto se encuentra en el año 2020 donde a destacar la pandemia del covid-19, la corrupción en los hospitales y la falta de un plan de vacunación nublaron el panorama de los ecuatorianos.

La metodología del conteo de noticias, permite identificar instantes de tiempo en donde la cantidad de noticias relacionadas con incertidumbre es alta, a estos valores altos se los conoce como *shocks de incertidumbre*, y por lo general se encuentran relacionados con eventos de impacto en el entorno de estudio.

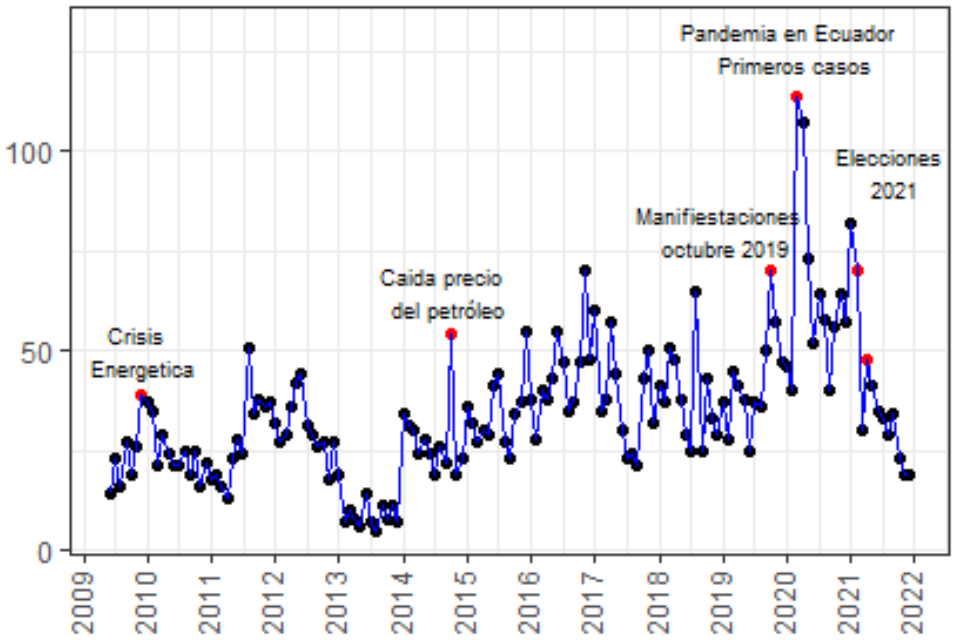


Figura 3.3: Incertidumbre en Ecuador

En la figura 3.3, se resaltan algunos niveles altos de incertidumbre y se los ha asociado con eventos que a criterio personal resultan importantes, comenzando por septiembre del 2009, una época marcada por una crisis económica fruto de una fuerte crisis energética que sumió al país durante periodos de tiempo en total oscuridad, en medio eventos como la caída del precio del petroleo y las manifestaciones de octubre del 2019 provocaron en el país una duda sobre el futuro generalizada, más reciente nos ubicamos en febrero del 2020, el mundo comenzaba a entrar de lleno a la pandemia del covid-19, y en Ecuador los primeros casos se reportaban.

3.1.2. Componentes de la incertidumbre

Anteriormente se mencionó que la incertidumbre como tal es un tema que abarca muchas cosas, durante el apartado de etiquetado de tópicos se asignó con un campo semántico coherente a cada tópico, a continuación se muestra un resumen de los resultados de dicho proceso:

Incertidumbre			
Nombre	**Nº**	**Media**	**Máximo**
1. Política asamblearia	187	2.25	11
2. Familiar	229	2.22	12
3. Entretenimiento	253	2.53	10
4. Política Internacional (Europa)	200	2.86	18
5. Empresa y Comercio	289	2.58	10
6. Correismo	43	1.39	5
7. Movilidad	85	2.18	20
8. Fuerzas del orden	217	2.41	20
9. Migración	126	2.52	28
10. Salud	230	4.34	17
11. Ambiente	161	1.85	11
12. Comunicación	75	1.39	4
13. Banca	212	2.02	7
14. Política Municipal	250	2.43	11
15. Educación	149	1.71	5
16. Jurídica	146	1.64	6
17. Personal	290	2.48	12
18. Política Electoral	276	3.21	12
19. Deportes	551	3.99	18
20. Economía Ecuatoriana	480	4	27
21. Política Internacional (USA)	227	2.99	28
22. Políticas Laborales	194	1.81	6

Cuadro 3.2: Causas aproximadas de incertidumbre en Ecuador

En la tabla 3.2, se muestra las etiquetas que se les asigno a cada tópico además se muestra el número de noticias clasificadas dentro de cada tópico, el número medio de noticias publicadas por mes y el número máximo de noticias publicadas por mes relacionadas con el respectivo tópico.

Cada uno de los tópicos que se han encontrado en el transcurso de este trabajo forman parte de la incertidumbre que esta presente en el país,

si bien se han encontrado diversas componentes, cada una de ellas aporta en cierta medida a construir el nivel de incertidumbre que se percibe.

El número de noticias dentro de cada tópico, permite identificar cuales de ellos son más significativos, es decir, aquellos con mayor frecuencia.

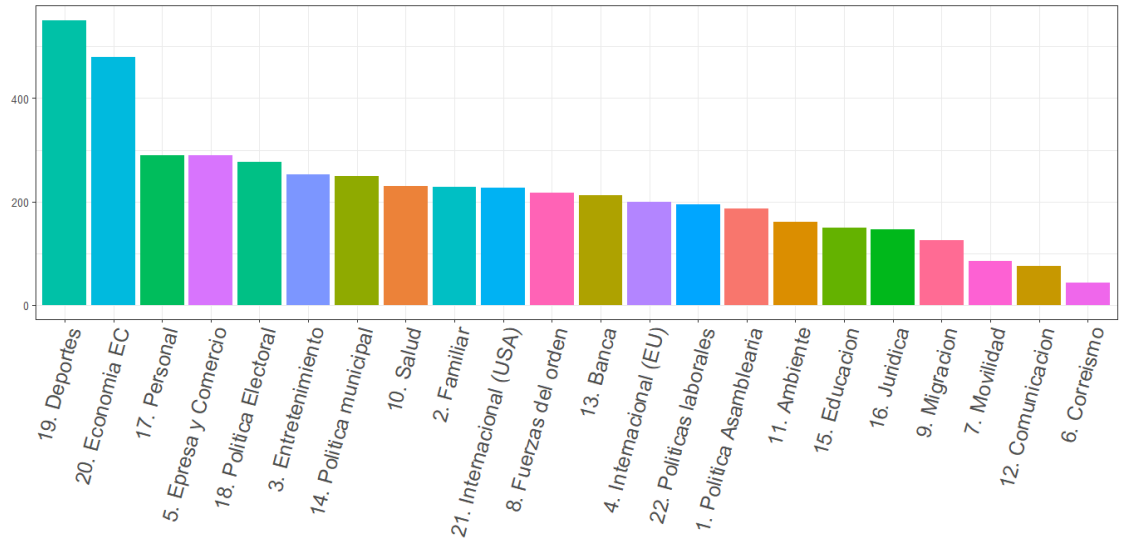


Figura 3.4: Frecuencia de los tópicos

En la figura 3.4, se observan como los tópicos 19. Deportes y 20. Economía ecuatoriana resaltan entre todos al contener la mayor cantidad de noticias, por otro lado los tópicos 6. Correismo, 7. Movilidad y 12. Comunicación presentan la menor cantidad de noticias.

Es claro que la incertidumbre en el deporte, y la incertidumbre en la economía ecuatoriana constituyen en gran parte a la incertidumbre presente en el país, en la figura 3.5 se muestra el aporte de las primeras cinco componentes principales que se han encontrado.

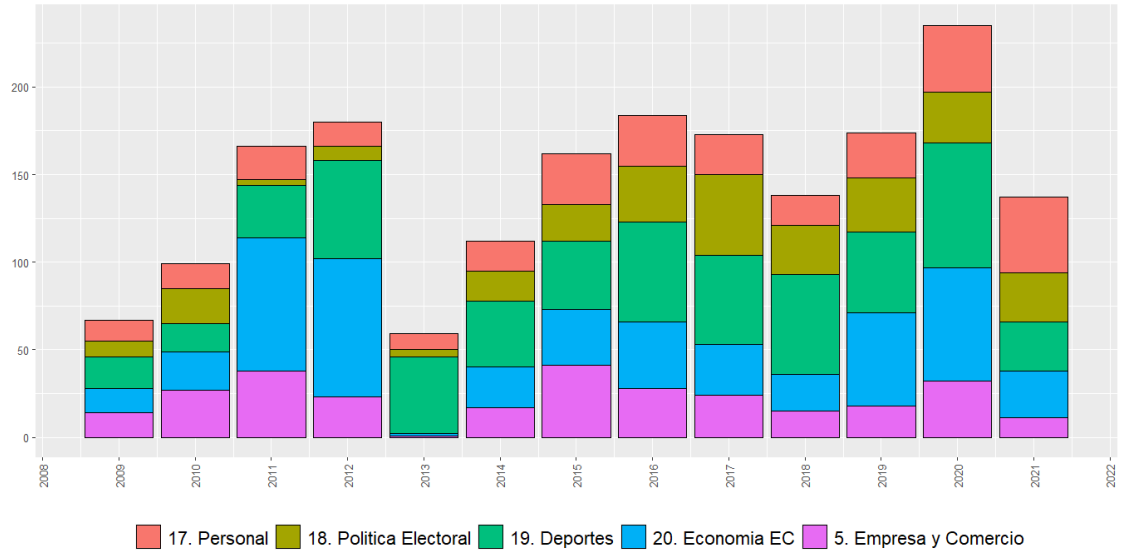


Figura 3.5: Contribución de las componentes más significativas

En la figura 3.5, se observa que durante el transcurso de los años, las principales componentes de incertidumbre en el país están presentes en diferente medida, la incertidumbre en los deportes, representa en gran medida a la incertidumbre percibida durante cada año, la incertidumbre en la economía ecuatoriana presenta una contribución variable en el transcurso del tiempo, enfatizando en los años 2011, 2012 y 2020.

Por otra parte, la incertidumbre personal esta presente en diferentes niveles a lo largo de cada instante de tiempo, lo cual sugiere que existe una permanente sensación de duda y desconcierto en el país.

Acerca de la incertidumbre relacionada con la política electoral, presenta contribuciones significativas en los años finales y posteriores al mandato del ex-presidente Rafael Correa, alcanzando un máximo valor en el año 2017, durante el cual Moreno asumió la presidencia, y desde entonces se observa como la incertidumbre relacionado con los candidatos presidenciales y elecciones mantiene contribuciones significativas.

Si bien las anteriores componentes resultan ser las más significativas al presentar mayor frecuencia, ciertamente componentes con menos frecuencia también forman parte de la incertidumbre percibida, por lo cual a continuación se muestra las contribuciones de las últimas cinco componentes.

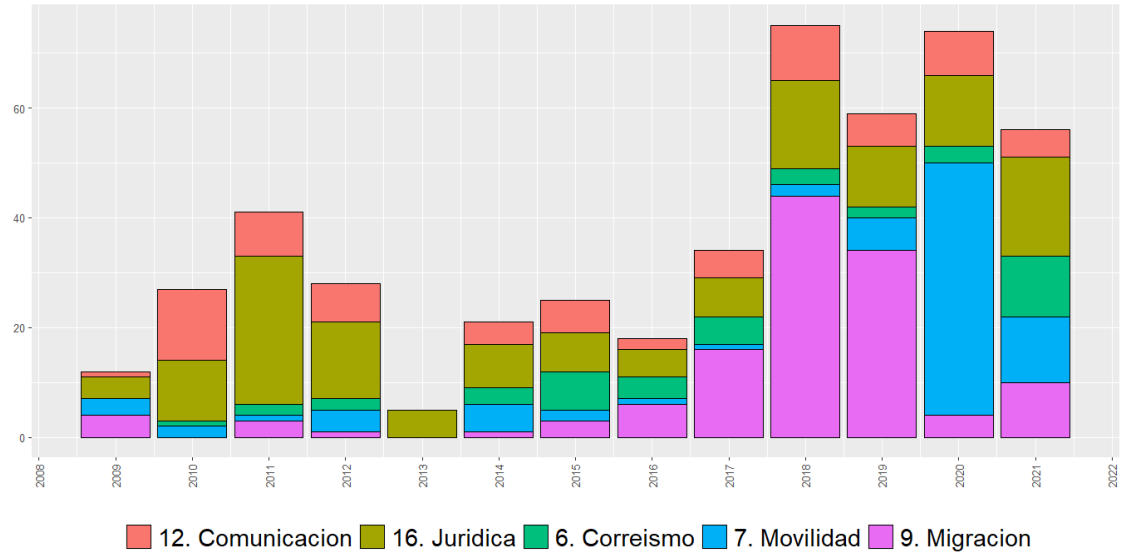
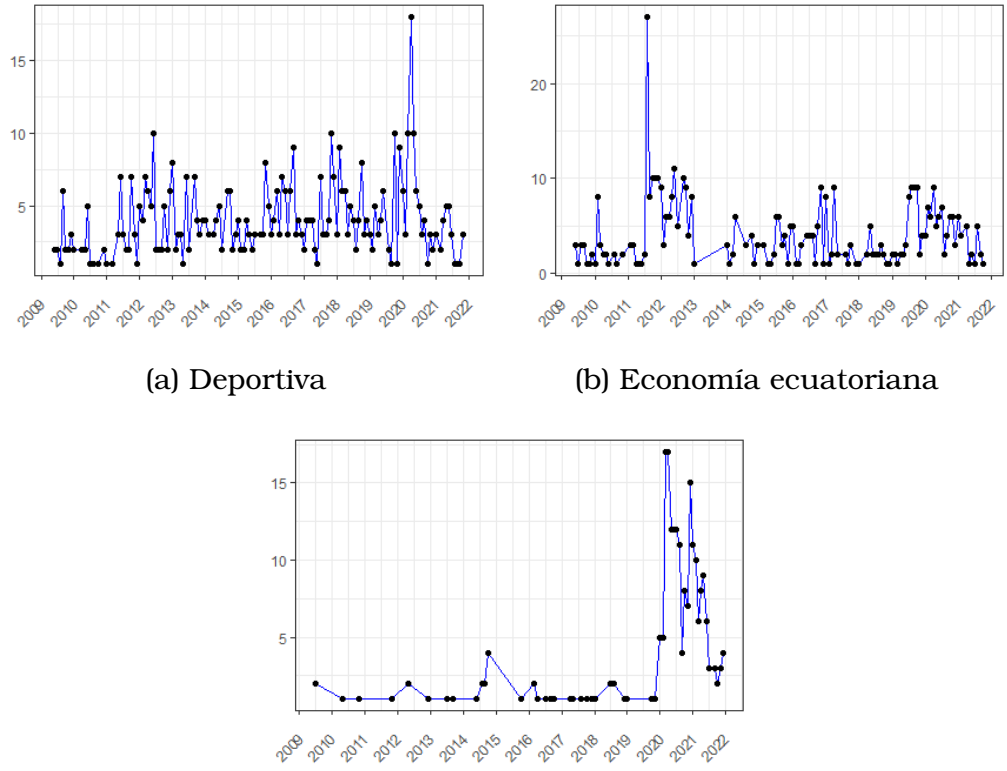


Figura 3.6: Contribución de las componentes menos significativas

En la figura 3.6, lo primero a resaltar se relaciona con la incertidumbre en movilidad, la cual muestra una gran contribución durante el año 2020, año marcado por la pandemia del covid-19 y el subsecuente confinamiento.

Por otra parte la incertidumbre relacionada con migración, muestra una contribución creciente desde el 2014, alcanzando el máximo valor durante 2018, como tal, la migración ha sido un tema constante en el país, pero el éxodo venezolano se ha convertido en causa de desborde en los gobiernos de América latina, entre ellos, el gobierno ecuatoriano se vio en medio de panoramas oscuros sobre dicha situación. Resulta de interés además la ausencia de incertidumbre sobre comunicación durante el año 2013, mientras que el resto del tiempo dicha componente contribuye en la incertidumbre percibida.

Cada componente por separado es un mundo de posibilidades para estudiar, si bien, un análisis a profundidad de cada componente excede el alcance del trabajo de integración curricular, se presenta un análisis de algunas componentes que resaltan entre las demás.



(a) Deportiva

(b) Economía ecuatoriana

(c) Salud

Figura 3.7: Variación componentes de la incertidumbre

En la figura 3.7, se muestra la variación de los niveles de incertidumbre para las dos componentes mas representativas, siendo estas Deportes y Economía Ecuatoriana, en adición se presenta la variación de la componente Salud.

La incertidumbre en el deporte se ha mostrado como la componente mas significativa, como tal en la figura 3.7a, se observa como varia alrededor de ciertos instantes de tiempo, visualmente se presenta un crecimiento antes del 2012 luego un descenso y nuevamente un crecimiento antes del 2016, un patrón que se repite cada 4 años y es aun mas visible en el 2020, es claro que la incertidumbre en el deporte varía en función de la copa mundial de fútbol, claramente mientras mas cerca se encuentra el evento, mayor es la duda por quien ganara, quien clasificara, que jugadores no podrán participar entre otras preguntas que nos hacemos. En adición esta componente alcanza su máximo en el 2020, que al igual que con la incertidumbre sanitaria 3.7b su valor en dicho año es reflejo de la pandemia del covid-19.

Por su parte, la incertidumbre respecto a la económica ecuatoriana en la figura 3.7c, muestra un valor elevado a finales del 2011, periodo de tiempo alrededor del cual la Ley de Fomento Ambiental entraba en vigor en medio de declaraciones de inconstitucionalidad a varios nuevos impuestos, si bien dicho valor resalta entre los demás, existen otros que muestran diferentes eventos en la economía ecuatoriana tales como 2010 y la crisis económica resultado de la crisis energética, la crisis económico global a raíz de la pandemia del covid-19 a partir del 2020.

3.2. Conclusiones y recomendaciones

Al termino de este proyecto, se menciona aquella idea que lo origino, ya que aquello que empezó como la construcción de un indicador de incertidumbre política económica para el país, se convirtió en una oportunidad de estudiar las posibles causas de la incertidumbre en el país, además de adentrarse en el campo del aprendizaje automático y procesamiento del lenguaje natural de mano del modelamiento de tópicos.

La cantidad de información disponible en estos días podría resultar abrumadora, y más aún al considerar, noticias, textos, comentarios, como información más allá de datos numéricos, el procesar grandes cantidades de textos, resulta en una tarea imposible para una persona, por lo que en este proyecto se mostró el uso de herramientas estadísticas para encontrar información antes desconocida pero que se encuentra implícita y resulta de utilidad. A todo esto ¿En qué punto se cruza la línea de la ética al extraer información?, la cantidad de datos disponibles en la red es inmensa y puede dar origen a muchos y diversos trabajos de investigación, pero el problema está en acceder a ellos, muchos datos no son públicos y muchos otros están restringidos, esto fue una realidad que se golpeó de frente al empezar el proyecto, aun con la existencia de técnicas para extraer información, estas no son todo poderosas y en muchas ocasiones se ven rezagadas.

Es conocido que el éxito de un algoritmo depende de los datos que se usen para entrenarlo, en el algoritmo LDA eso no es diferente, y con más razón al tratar con textos de gran tamaño no todo el texto aporta información relevante, de mano con el coste computacional el cual es alto, el procesamiento del corpus de texto es un paso de vital importancia, pues en términos de eficiencia es completamente absurdo entrenar un algoritmo con grandes cantidades de información no relevante. La parte más desafiante del proyecto recae en encontrar el número óptimo de tópicos y con ello validar el modelo, como tal no parece haber una forma directa de encontrar dicho óptimo, por ahora lo que se hace es proceder por prueba y error, se realizan diferentes modelos con diferente número de tópicos y se evalúa entre ellos buscando maximizar el número de tópicos que tengan sentido.

Maximizar la coherencia semántica y la exclusividad de los tópicos ha demostrado ser le metodología más popular para determinar el número de tópicos, en el presente trabajo se muestra el uso de dicha metodología en un conjunto de documentos relacionados con un mismo tema, como resultado se obtuvieron tópicos en los cuales las palabras principales en la mayor parte de los casos conformaban un campo semántico bastante claro. El lograr construir tópicos coherentes es de bastante importancia pues en muchas ocasiones, los tópicos resultados del algoritmo LDA, corresponden a mixturas de palabras que pueden resultar incoherentes. Por su parte la exclusividad en los tópicos permite diferenciarlos de mejor manera.

LDA es un algoritmo no supervisado, aun así durante el desarrollo del trabajo se volvió altamente supervisado, debido a la cantidad de trabajo requerido al momento de filtrar información relevante, determinar el número de tópicos y finalmente realizar un análisis cualitativo de los resultados obtenidos, dicho análisis cualitativo no solo esta presente al analizar los resultados del algoritmo, mas bien se encuentra presente durante todo el desarrollo, pues al filtrar información relevante al inicio del trabajo se requiere analizar los términos o palabras que son relevantes basados en el contexto en el que se esta trabajando. Por lo cual, un conocimiento previo la temática en las que se esta trabajando garantiza mejores resultados.

Cada una de las componentes encontradas en el desarrollo del proyecto, contribuyen en los niveles de incertidumbre, aunque ciertas componentes dan la impresión de contribuir bastante poco, lo que se ha encontrado en el trabajo, es que existen momentos en los cuales la incertidumbre en ciertos aspectos se dispara por sucesos de importancia. Por otro lado algunas componentes si bien presentan niveles altos en ciertos puntos, su mayor importancia esta de mano con el hecho de que se encuentran más presentes a lo largo del periodo de tiempo, de hecho de que la incertidumbre deportiva sea la componente más presente a lo largo del tiempo lleva a considerar un escenario en el cual no se le da la suficiente importancia a aspectos que podrían considerarse de mayor valor como la economía, la política e incluso la salud, o quizás en el extremo opuesto se ignora dichos aspectos a razón de darle importancia al entretenimiento,

quizás es un intento por ocultar la realidad o simplemente es fanatismo hacia el fútbol.

El número de noticias juega un papel importante en este trabajo y como tal la cantidad de las mismas se ve sujeta a aspectos que escapan del alcancen de este proyecto, como se vio en la 3.2a, en el año 2013 la cantidad de noticias publicadas sufrió una caída enorme y seria apresurado concluir que se debe a épocas de tranquilidad, pero por otro lado seria aventurado decir que se debe a restricciones sobre los medios, es aquí, que nace una oportunidad de profundizar en la naturaleza de la situación. Por otro lado el conjunto de palabras que forma un documento tiene también bastante importancia, pues durante el desarrollo del trabajo, se mostró que palabras como incertidumbre, incerteza o incierto son frecuentes en medios oficiales, pero como tal, los medios oficiales no son la única fuente de información, una extensión al trabajo es considerar fuentes alternativas de información, dicha situación se exploro en el desarrollo pero sin profundizar en ello, y un resultado casi inmediato es que las personas como tal no acostumbran a usar aquellas palabras en su día a día.

Capítulo A

Código fuente

```
########################################
# LIBRERIAS
########################################
# Semilla
set.seed(1727)

# Web scraping
library(rvest)

# Manejo de datos
library(dplyr)
library(lubridate)
library(tm)
library(tidytext)

# Visualizacion
library(ggplot2)
library(ggplotlyExtra)

# Modelado
library(topicmodels)
library(udpipe)
library(stopwords)
library(ldatuning)
```

```r
# Multiprocesos
library(purrr)
library(stm)
library(furrr)
plan(multiprocess)

############################################################
# EXTRACCION DEL CORPUS DE NOTICIAS
############################################################

noticias = data.frame()
for (page_result in 2:529){
  link = paste0("https://www.elcomercio.com/search/page/",
                page_result,"/?s=Incertidumbre")
  page = read_html(link)
  name = page %>%
    html_nodes(".content") %>%
    html_nodes("h3 a") %>%
    html_text()
  fecha = page %>%
    html_nodes(".list-item__date") %>%
    html_text()
  descripcion = page %>%
    html_nodes(".content") %>%
    html_nodes("p") %>%
    html_text()
  enlace = page %>%
    html_nodes(".content") %>%
    html_nodes("h3 a") %>%
    html_attr("href")
  texto = vector(length = length(enlace))
  for (isim in 1:length(enlace)) {
    link.p = paste0(enlace[isim])

  texto[isim] = tryCatch(read_html(link.p) %>%
                           html_nodes(".entry__content")%>%
                           html_nodes("p") %>% html_text()%>%
                           str_c(collapse = " "),
```

```
                          error = function(e){return(NA)})
  }
  noticias = rbind(noticias,
                   data.frame(name,
                              fecha,
                              descripcion,
                              enlace,
                              texto,
                              stringsAsFactors = FALSE))

  print(paste("Page:", page_result))
}

############################################################
# ALMACENAMIENTO DEL CORPUS
# LECTURA DEL CORPUS
# LIMPIEZA DEL CORPUS
############################################################


Informe <- read_csv("Informe.csv")
Incertidumbre_EC <- select(Informe, -...1)
names(Incertidumbre_EC) <- c("titulo",
                             "fecha",
                             "link",
                             "texto")
# Datos faltantes
plot_missing(Incertidumbre_EC)
# Selecion de Variables y eliminar duplicados
incertidumbre <- Incertidumbre_EC %>%
  select(c(fecha,texto)) %>%
  unique()

# Casos completos
incertidumbre <- incertidumbre[complete.cases(incertidumbre),]

incertidumbre <- incertidumbre %>%
  mutate(fecha = gsub("[\r\n]", "", fecha)) %>%
                  #Eliminar saltos de linea
```

```
  mutate(fecha = as.Date(fecha, format =  "%d/%m/%Y"))
glimpse(incertidumbre)

#############################################################
# ESTADISTICA DEL NUMERO DE NOTICIAS
#############################################################

#############################################################
# GRAFICAS DEL NUMERO DE NOTICIAS
#############################################################


# Frecuencia Anual
incertidumbre %>%
  group_by(month = lubridate::floor_date(incertidumbre$fecha,
                                         unit = "year")) %>%
  summarise(n=n()) %>%
  ggplot(aes(x=n))+
  geom_boxplot(col = "blue",
               outlier.colour = "black",
               outlier.shape = 3,
               outlier.fill = "red",
               outlier.size = 2)+
  labs(title = "", x = "",
       y = "") +
  theme_bw()

# Frecuencia mensual

incertidumbre %>%
  group_by(month = lubridate::floor_date(incertidumbre$fecha,
                                         unit = "month")) %>%
  summarise(n=n()) %>%
  ggplot(aes(x = month, y = n)) +
  geom_point() +
  geom_line(col='blue') +
  labs(title = "", x = "",
       y = "") +
  theme_bw() +
```

```
  scale_x_date(date_breaks = "5 month",
               date_labels = "%m-%y")+
  theme(axis.text.x = element_text(angle = 90,
                                   vjust = 0.3,
                                   hjust=1))

##############################################
# MINERIA DE TEXTO
# LEMATIZACION
# TOKENIZACION
# REDUCCION DEL N DE PALABRAS
##############################################
#

# Procesos computacionalmente pesados
# Resultado UdPipe

# descargar el modelo y guardar la referencia
modelo_sp <- udpipe::udpipe_download_model('spanish')
# refrencia al modelo descargado
modelo_sp$file_model
# cargar el modelo en memoria
modelo_sp <- udpipe_load_model(file = modelo_sp$file_model)


noticias_anotadas <- udpipe_annotate(
  object = modelo_sp,
  x = incertidumbre$texto,
  doc_id = c(1:length(incertidumbre$texto)),
  trace = 200
) %>% as.data.frame(.)

# Visualizacion del Resultado
noticias_anotadas$upos %>%
  unique()
noticias_anotadas %>%
  select(doc_id,token,lemma,upos) %>%
  head() %>%
  View()
```

```r
# Procesos menos costosos computacionalmente.

noticias_anotadas2 <- noticias_anotadas %>%
  filter(upos=="ADJ"|upos=="VERB"|upos=="NOUN"|upos=="PROPN") %>%
  select( doc_id, lemma ) %>%
  filter(!lemma %in% stopwords::stopwords(language = "es")) %>%
  filter(!lemma %in% c("ser","dar", "decir", "tener",
                       "haber", "estar", "hacer", "ver",
                       "leer","comentar","ir")) %>%
  filter(!lemma %in% c("semana","mes","anio","dia")) %>%
  filter(!lemma %in% c("foto","windows","nuevo",
                       "pais","incertidumbre"))

# Visualizacion del resultado
glimpse(noticias_anotadas2)
noticias_anotadas2 %>%
    head() %>%
    View()

########################################
# Matriz Documento Termino
########################################

# DTM Y REDUCCION DE PALABRAS COMUNES Y UNICAS
noticias_dtm <- noticias_anotadas2 %>%
  count(doc_id, lemma, sort = TRUE) %>%
  cast_dtm(doc_id, lemma, n)
inspect(noticias_dtm)

palabras_freq<-findFreqTerms(noticias_dtm,
                             lowfreq=5,
                             highfreq=nrow(noticias_dtm)*0.9)

noticias_dtm <- noticias_dtm[ , palabras_freq]
inspect(noticias_dtm)
```

```r
# TOP DE PALABRAS

m <- as.matrix(noticias_dtm)
v <- sort(colSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)

ggplot(head(d,20), aes(x = reorder(word, -freq),
                       y = freq,
                       fill=freq)) +
  scale_fill_gradient(low = "gray",
                      high = "light blue")+
  geom_bar(stat = "identity")+
  labs(title = "", x = "",
       y = "Frecuencia de palabras") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90,
                                   vjust = 0.4,
                                   hjust=1),
        legend.position = "none")

#############################################################
# NUMERO DE TOPICOS
#############################################################

# K entre 5 a 50
result2 <- FindTopicsNumber(
  noticias_dtm,
  topics = seq(from = 5, to = 50, by = 5),
  metrics = c("Griffiths2004",
              "CaoJuan2009",
              "Arun2010",
              "Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 1727),
  mc.cores = 10L,
  verbose = TRUE
)

# Visualizacion
```

```
result2
FindTopicsNumber_plot(result2)

# K entre 15-25
result4 <- FindTopicsNumber(
  noticias_dtm,
  topics = seq(from = 15, to = 25, by = 1),
  metrics = c("Griffiths2004",
              "CaoJuan2009",
              "Arun2010",
              "Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 1234),
  mc.cores = 10L,
  verbose = TRUE
)

# Visualizacion
result4
FindTopicsNumber_plot(result4)
result4 %>%
  ggplot(aes(x = topics, y = Deveaud2014))+
  geom_line(col="blue")+
  geom_point()+
  labs(title = "", x = "N_topicos",
       y = "Deveaud2014") +
  theme_bw()+
  geom_vline(xintercept = 24,
             linetype = 2)+
  theme(axis.text.x=element_text(angle = 90,
                                 vjust = 0.3,
                                 hjust=1),
                                 legend.position="none")

#########################################################
# Coherence Semantica
#########################################################
```

```
# Entrenamiento de modelos para diferentes
# valores de k.

dtm_lda <- Matrix::Matrix(as.matrix(noticias_dtm),
                          sparse = T)
k_list <- seq(20, 30, by = 1)
m_models <- data_frame(K = k_list) %>%
  mutate(topic_model = future_map(K, ~stm(dtm_lda, K = .,
                                          verbose = FALSE)))

# Acceso a los diferentes resultados

heldout <- make.heldout(dtm_lda)
k_result <- m_models %>%
  mutate(exclusivity = map(topic_model,
                           exclusivity),
         semantic_coherence = map(topic_model,
                                  semanticCoherence,
                                  dtm_lda),
         eval_heldout = map(topic_model,
                            eval.heldout,
                            heldout$missing),
         residual = map(topic_model,
                        checkResiduals,
                        dtm_lda),
         bound =  map_dbl(topic_model,
                          function(x) max(x$convergence$bound)),
         lfact = map_dbl(topic_model,
                         function(x)lfactorial(x$settings$dim$K)),
         lbound = bound + lfact,
         iterations=map_dbl(topic_model,
                            function(x)length(x$convergence$bound)))

# Resultado del multiproceso
k_result

# Representacion grafica
# Coherencia semantica
```

```
k_result %>%
  transmute(K,
            `Semantic coherence` = map_dbl(semantic_coherence,
                                           mean)
            ) %>%
  gather(Metric, Value, -K) %>%
  ggplot(aes(K, Value)) +
  geom_point()+
  geom_line(size = 1,
            color = "blue",
            alpha = 0.7, show.legend = FALSE) +
  geom_vline(xintercept = 22,
             linetype = 2)+
  theme_bw()+
  facet_wrap(~Metric, scales = "free_y") +
  labs(x = "N topicos",
       y = NULL,
       title = "",
       subtitle = "")

#########################################################
# Exclusividad de los topicos
#########################################################

k_result %>%
  select(K, exclusivity, semantic_coherence) %>%
  filter(K %in% c(22, 24)) %>%
  unnest() %>%
  mutate(K = as.factor(K)) %>%
  ggplot(aes(semantic_coherence, exclusivity,
             color = K)) +
  geom_point(size = 2, alpha = 0.7) +
  labs(x = "Semantic coherence",
       y = "Exclusivity",
       title = "",
       subtitle = "")+
  theme_bw()
```

```r
#####################################################################
# Modelado de Topicos (22 Topicos)
#####################################################################


k_topics2 <- 22 # numero de topicos
noticias_tm2 <- topicmodels::LDA(
  noticias_dtm,
  k = k_topics2,
  method = "Gibbs",
  control = list(seed = 1:5, nstart=5, verbose=1000))

noticias_tm2


noticias_tm2_beta <- tidy(noticias_tm2, matrix = "beta")
noticias_tm2_gamma <- tidy(noticias_tm2, matrix = "gamma")
glimpse(noticias_tm2_beta)
glimpse(noticias_tm2_gamma)
#View(noticias_tm_beta[1:25,])

# Interpretacion

noticias_tm2_beta %>%
  filter(topic %in% c(1:22)) %>%
  group_by(topic) %>%
  top_n(15) %>%
  ungroup() %>%
  arrange(topic, -beta) %>%
  ggplot(aes(x=reorder(term, beta),y=beta,
             fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~topic, scales = "free") +
  coord_flip()+
  labs(title = "", x = "",
       y = "") +
  theme_bw(base_size = 15)+
  theme(axis.title.x=element_blank(),
```

```
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())

# Encontrar los titulos mas representativos

noticias_tm2_gamma %>%
  group_by(topic) %>%
  filter(gamma == max(gamma))

# Topicos mas frecuentes

noticias_tm2_gamma %>%
  filter(gamma >0.2) %>%
  group_by(topic) %>%
  summarise(n=n()) %>%
  ggplot(aes(x = reorder(topic, -n), y = n, fill=n)) +
  scale_fill_gradient(low = "light␣blue",
                      high = "light␣green")+
  geom_bar(stat = "identity")+
  labs(title = "", x = "Topicos",
       y = "Frecuencia") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90,
                                   vjust = 0.3,
                                   hjust=1),
        legend.position = "none")

################################################
# Etiquetado de topicos (No necesario)
################################################

glimpse(incertidumbre)
incertidumbre$id <- c(1:length(incertidumbre$texto))

# Etiquetas
topicos_nombres <- rbind(
  c(topic = 1 , nombre = "1.␣Politica␣Asamblearia"),
  c(topic = 2 , nombre = "2.␣Familiar"),
  c(topic = 3 , nombre = "3.␣Entretenimiento"),
```

```
  c(topic = 4 , nombre = "4. Internacional (EU)"),
  c(topic = 5 , nombre = "5. Empresa y Comercio"),
  c(topic = 6 , nombre = "6. Correismo"),
  c(topic = 7 , nombre = "7. Movilidad"),
  c(topic = 8 , nombre = "8. Fuerzas del orden"),
  c(topic = 9 , nombre = "9. Migracion"),
  c(topic = 10 , nombre = "10. Salud"),
  c(topic = 11 , nombre = "11. Ambiente"),
  c(topic = 12 , nombre = "12. Comunicacion"),
  c(topic = 13 , nombre = "13. Banca"),
  c(topic = 14 , nombre = "14. Politica municipal"),
  c(topic = 15 , nombre = "15. Educacion"),
  c(topic = 16 , nombre = "16. Juridica"),
  c(topic = 17 , nombre = "17. Personal"),
  c(topic = 18 , nombre = "18. Politica Electoral"),
  c(topic = 19 , nombre = "19. Deportes"),
  c(topic = 20 , nombre = "20. Economia EC"),
  c(topic = 21 , nombre = "21. Internacional (USA)"),
  c(topic = 22 , nombre = "22. Politicas laborales")
) %>% as_tibble() %>% mutate(topic=as.integer(topic))

# Etiquetado
noticias_tm2_gamma %>%
  #filter(topic %in% top5 ) %>%
  filter(
    gamma > 0.2
  ) %>%
  mutate(
    id=as.integer(document),
    topic=as.integer(topic)
  ) %>%
  left_join(x = ., y = incertidumbre, by="id") %>%
  left_join(topicos_nombres, by="topic")

################################################
# Estadisticas de los topicos
################################################
```

```
estadisticas_topicos <- noticias_tm2_gamma %>%
  filter(
    gamma > 0.2
  ) %>%
  mutate(
    id=as.integer(document),
    topic=as.integer(topic)
  ) %>%
  left_join(x = ., y = incertidumbre, by="id") %>%
  left_join(topicos_nombres, by="topic")  %>%
  group_by(nombre,
    month = lubridate::floor_date(fecha,
                                  unit="month")) %>%
  summarise(n=n()) %>%
  summarise(data.frame(sum(n),
                       min(n),
                       mean(n) %>% round(2),
                       max(n))) %>%
  write.csv("est_topicos.csv")
```

Referencias bibliográficas

[1] Ah-Hwee, T. Text mining: The state of the art and the challenges. *Kent Ridge Digital Labs 21 Heng Mui Keng Terrace Singapore 119613*, 2016.

[2] Azqueta-Gavaldón, A. Developing news-based economic policy uncertainty index with unsupervised machine learning. 2017a.

[3] Bai, X., Zhang, X., Li, K. X., Zhou, Y., y Yuen, K. F. Research topics and trends in the maritime transport: A structural topic model. 2021.

[4] Baker, S., Bloom, N., y Davis, S. Mesauring economic policy uncertainty. *The Quarterly Journal of Economics*, págs. 1593–1636, 2016.

[5] Blei, D. M. Probabilistic topic models:surveying a suite of algorithms that offer a solution to managing large document archives. *Communications of the acm*, 55:77–84, 2012.

[6] Blei, D., Ng, A., y Jordan, M. Latent dirichlet allocation. *Journal of machine Learning research, 3(Jan)*, págs. 993–1022, 2003.

[7] Bloom, N. The impact of uncertainty shocks. 2009.

[8] Bloom, N. "fluctuations in uncertainty". *Journal of Economic Perspectives*, 2014.

[9] Cerda, R., Álvaro Silva, y Valente, J. T. Índice de Incertidumbre Económica: Medición e Impacto. *El Faro Economics*, 2016.

[10] Chopra, A., Prashar, A., y Sain, C. Natural language processing. *INTERNATIONAL JOURNAL OF TECHNOLOGY ENHANCEMENTS AND EMERGING ENGINEERING RESEARCH*, 1, 2013.

[11] Consoli, S., Recupero, D. R., y Saisana, M. *Data Science for Economics and Finance Methodologies and Applications*. 2021.

[12] Deveaud, R., SanJuan, E., y Bellot, P. Accurate and effective latent concept modeling for ad hoc information retrieval. 2014.

[13] Echevarría Icaza, V. Desentrañando las causas de la incertidumbre de política económica en españa: una aproximación usando machine learning. 2019.

[14] Grün, B. y Hornik, K. topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, 40(13):1–30, 2011.

[15] Hagen, L. Content analysis of e-petitions with topic modeling: How to train and evaluate lda models? *Information Processing and Management*, 54:1292–1307, 2018.

[16] Hearst, M. What is text mining? *SIMS,UC Berkeley*, 2003.

[17] Ike Vayansky, S. A. K. A review of topic modeling methods. *Information Systems*, 2020.

[18] Jurado, K., Ludvigson, S., y Serena, N. Measuring Uncertainty. *American Economic Review*, 105(3):1177–1216, 2015.

[19] Kuhn, K. D. Using structural topic modeling to identify latent topics and trends in aviation incident reports. 2018.

[20] Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., y Pfetsch, B. Applying lda topic modeling in communication research: Toward a valid and reliable methodology. *ommunication Methods and Measures*, 12:93–118, 2018.

[21] Nikita, M. *ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters*, 2020. R package version 1.0.2.

[22] Padilla, S. Índice de incertidumbre de política económica para Ecuador: Discusiones y una propuesta de cuantificación. *Revista Puce. Issn: 2528-8156*, 108:117–221, 2019.

[23] Park, E., Chae, B., Kwon, J., y Kim, W.-H. The effects of green restaurant attributes on customer satisfaction using the structural topic model on online customer reviews. 2020.

[24] Perico, D. Measuring economic policy uncertainty in colombia: A news based approach. 2018.

[25] Vallez, M. y Pedraza-Jimenez, R. El procesamiento del lenguaje natural en la recuperación de información textual y áreas afines. *Journal of machine Learning research, 3(Jan)*, 2007.

[26] Wickham, H. *rvest: Easily Harvest (Scrape) Web Pages*, 2021. R package version 1.0.1.

[27] Wijffels, J. *udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit*, 2021. R package version 0.8.8.

[28] Witten, I. H. Text mining. *Computer Science, University of Waikato, Hamilton, New Zealand*.