Extracción de Tópicos y Sumarización utilizando Redes Neuronales Recurrentes Aplicado a las Normas Legales Peruanas

Lennin Quiso Córdova
Milton Palacin Grijalva
Carlos Tello Tapia

Resumen

La sumarización de textos es una labor necesaria en muchas actividades y con la ampliación de fuentes de información y de textos el uso de herramientas que ayuden en este proceso se hace cada vez más necesario. Asimismo para muchos individuos esta labor de recopilación de información y lectura puede ser muy tediosa o en todo caso tomar mucho tiempo para que de manera manual se pueda extraer la información de diversas fuentes. Asimismo el resumen o sumarización de textos aumenta la eficiencia para seleccionar los textos adecuados y acordes con la investigación o labor realizada, además con la popularidad en el uso de smartphones, permite simplificar textos en documentos reducidos que serían más fáciles de leer en dispositivos pequeños así como ayudar al proceso de traducción al acortar los textos.

Palabras Clave— Extracción de tópicos, Sumarización, NLP, Redes Neuronales, *Deep Learnig*

1. Introducción

La recuperación de información es una labor necesaria en muchas actividades y con la ampliación de fuentes de información y de textos el uso de herramientas que ayuden en este proceso se hace cada vez más necesario. Asimismo para muchos individuos esta labor de recopilación de información y lectura puede ser muy tediosa o en todo caso tomar mucho tiempo para que de manera manual se pueda extraer la información de diversas fuentes. Asimismo el resumen o sumarización de textos aumenta la eficiencia para seleccionar los textos adecuados y acordes con la investigación o labor realizada, además con la popularidad en el uso de smartphones, permite simplificar textos en documentos reducidos que serían más fáciles de leer en dispositivos pequeños así como ayudar al proceso de traducción al acortar los textos.

Para el proceso de sumarización y extracción de tópicos se utilizan diversos métodos que utilizan técnicas de clasificación, redes neuronales, gráficos semánticos, árboles de decisión, modelos de regresión, etc.

2. Definición del Problema

Cada día se publican decenas de normas en el país lo cual demanda un gran esfuerzo de revisión y análisis por parte de los agentes económicos, dado que la normatividad aprobada y publicada es de estricto cumplimiento en un estado de derecho. En este sentido cobra vital importancia un adecuado seguimiento de las mismas así como su correcta clasificación y resumen. Por otro lado, la clasificación y seguimiento de las mismas también puede servir como un instrumento que ayude al propio estado, congreso y gobiernos regionales a llevar a cabo un adecuado seguimiento de la carga regulatoria que se impone sobre los ciudadanos, con la finalidad de llevar a cabo proyectos de simplificación normativa así como la mejorar del marco normativo actual.

Sin embargo el trabajo señalado en los párrafos previos consume mucho tiempo y recursos destinados a esta labor porque la producción normativa diaria es bastante extensa. Asimismo la labor de revisión normativa también resulta bastante tediosa y es uno de los trabajos menos preferidos por los analistas. Para este trabajo, se va a recolectar y procesar todas las normas publicadas en el diario oficial El Peruano para el periodo comprendido entre 23 de mayo de 2021 al 21 de agosto de 2021 comprendiendo un periodo de 3 meses. Asimismo se recoge toda la normativa, lo cual incluye la normativa publicada tanto por el poder ejecutivo, poder legislativo y gobiernos locales y regionales.

Más específicamente el trabajo realizado consiste en extraer toda la normativa de dicho periodo y procesar el texto con la finalidad de poder trabajar la extracción de tópicos así como entrenar un modelo de redes neuronales para realizar la tarea de sumarización o generación automática de resúmenes.

3. Marco Conceptual

3.1. *Natural Language Processing*

El procesamiento del lenguaje natural (NLP) es un campo de la informática, ciencias de la computación, la inteligencia artificial y la lingüística. Abarca las interacciones entre máquinas y lenguajes humanos con la finalidad de realizar tareas de análisis, comprensión y generación del lenguaje, que el ser humano utiliza de forma natural para interactuar con computadoras tanto de forma oral como escrita. Es un área interdisciplinaria basada en un campo de estudio versátil, que proporciona métodos para el modelado y diseño de algoritmos. Asimismo también emplea las matemáticas así como herramientas lingüísticas, modelos y teorías de comportamiento humano.https://www.overleaf.com/project/611e7752780bf203f7426be1

3.2. Sumarización de Textos

El proceso de sumarización o resumen de texto automático es la tarea de producir un resumen conciso y fluido de un texto mayor conservando el contenido de la información clave y el significado general del mismo. Existen dos enfoques diferentes que se utilizan para el resumen de texto: Resumen extractivo y Resumen abstractivo.

El resumen extractivo identifica las oraciones o frases importantes del texto original y solo las extrae. Esas frases extraídas serían nuestro resumen. La Figura 1 ilustra el resumen extractivo.

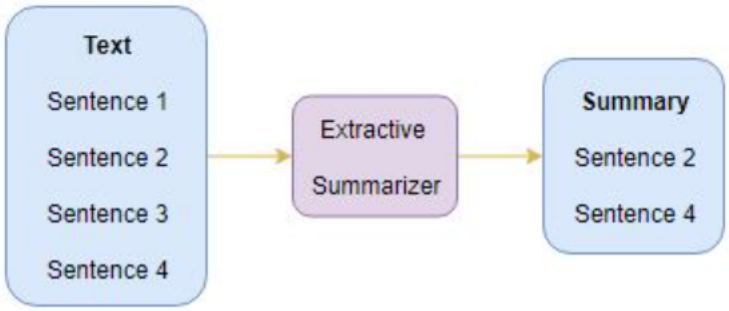

Figura 1: Sumarización Extractiva.

El resumen abstractivo genera nuevas oraciones a partir del texto original. Esto es diferente al enfoque extractivo en donde se utiliza únicamente oraciones que están presentes. De esta manera las oraciones generadas a través del resumen abstractivo pueden no estar presentes en el texto original. La Figura 2 ilustra la sumarización abstractiva.

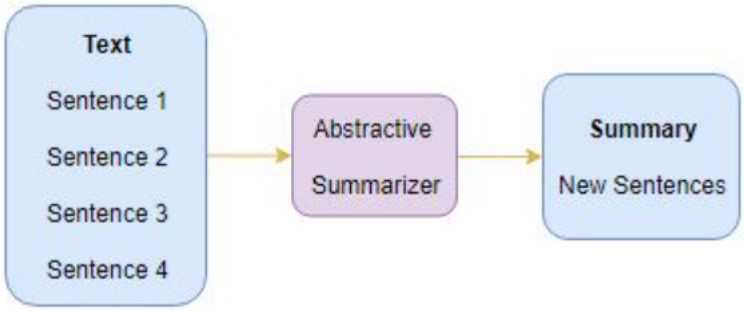

Figura 2: Sumarización Abstractiva.

3.3. Redes Neuronales

Los métodos de sumarización y extracción de tópicos emplean redes neuronales. Una red neuronal artificial es un modelo que busca imitar el comportamiento de las neuronas humanas. La figura siguiente muestra cómo funciona el método. Partimos de un conjunto de entradas, en este caso serían nuestros atributos, la red neuronal asigna pesos a cada uno de estos y luego se conecta con el cuerpo de la red neuronal que está conformado por una o más capas ocultas que se encargan de transformar las variables con la finalidad de ajustarse a los datos, luego esto sale y se aplica una función de activación que lo convierte en datos de salida. Para la estimación de estos modelos se suele utilizar algoritmos de búsqueda de gradiente conocidos

como back propagation o retroalimentación que en términos sencillos lo que hace es partir de la capa final para ir hacia atrás calculado de manera recursiva el vector gradiente.

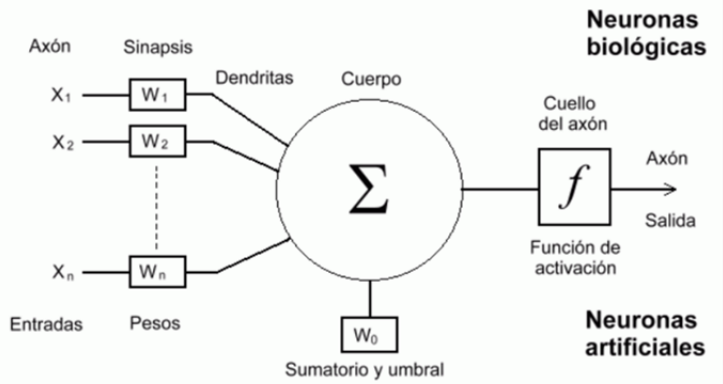

Figura 3: Estructura de una red neuronal artificial y su similitud con una biológica.

3.4. Redes Neuronales Recurrentes

Una red neuronal recurrente (RNN) es un tipo de red neuronal artificial en donde las conexiones entre nodos siguen una estructura temporal. Esto permite a las redes neuronales capturar patrones dinámicos en los datos. Esto implica que las redes neuronales puedan seguir una estructura autoregresiva para que se tome en consideración información de palabras o letras rezagadas. Por esta razón las RNN son ampliamente utilizadas en reconocimiento de escritura y de voz así como la generación de lenguajes.

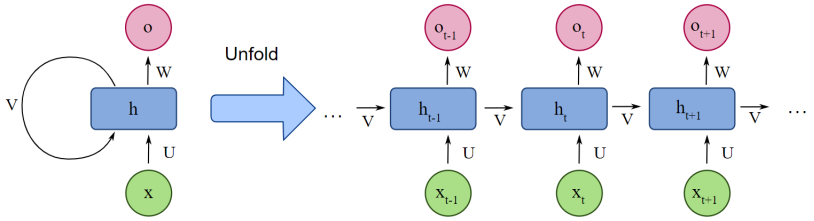

Figura 4: Estructura de una red neuronal recurrente (RNN).

3.5. Arquitectura Encoder-Decoder

Las arquitecturas Encoder-Decoder utilizan RNN para construir un modelo que tiene una parte que se encarga de recoger los datos de entrada y convertirla en formato vectorial (Encoder), y otra parte que se encarga de tomar la salida del

Encoder y transformarla nuevamente al formato de entrada (Decoder). Esta arquitectura al construirse sobre Redes Neuronales Recurrentes toma en consideración una estructura dinámica que permite que la red tome en consideración aspectos dinámicos de tal forma que la red entienda el contexto de las palabras. Este tipo de modelos son también utilizados para labores de traducción en donde el encoder va tomando bloques de texto que posteriormente envía al decoder.

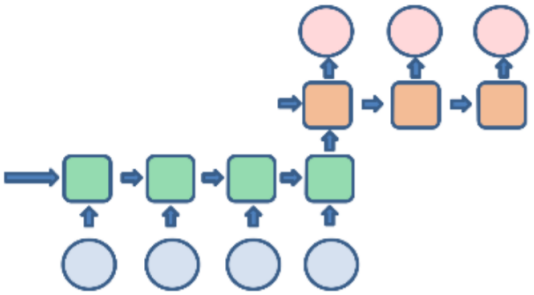

Figura 5: Arquitectura Encoder Decoder.

3.6. Topic Modeling

El modelado de temas es una técnica de aprendizaje automático no supervisada, que es capaz de escanear un conjunto de documentos, detectar patrones de palabras y frases dentro de ellos y agrupar automáticamente grupos de palabras y expresiones similares que caracterizan mejor un conjunto de documentos.

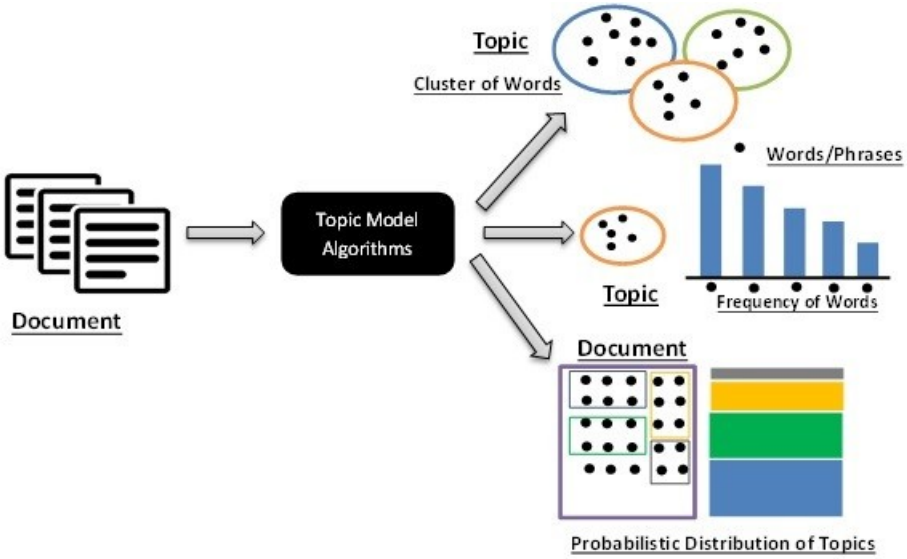

Figura 6: Topic Modeling.

4. Procesamiento de los Datos y Topic Modelling

4.1. *Web Scrapping*

Se hace uso de la librería Selenium para realizar la busqueda de las normas legales que se hayan publicado desde el 23 de mayo al 21 de agosto de 2021 en la web "https://diariooficial.elperuano.pe/Normas".

Luego usamos BeautifulSoup para generar un arreglo con cada resultado de la búsqueda y posteriormente iterar sobre cada uno para extraer: **La categoría, Título, URL de la página de contenido, Fecha, Resumen/Abstract**. Tomando descansos de 10 segundos para prevenir posibles bloqueos por sospechas de ataques DDoS.

Con ello, obtenemos un primer dataset como se muestra en la siguiente imágen.

category	title	date	abstract	content
PRODUCE	RESOLUCION MINISTERIAL N° 00146-2021-PRODUCE	26/05/2021	Disponen la publicación en el portal instituci...	Lima, 24 de mayo de 2021 VISTOS: El Informe Nº...
DEFENSA	RESOLUCION SUPREMA N° 026-2021-DE	28/07/2021	Dan por concluido el nombramiento de Jefe del ...	Lima, 27 de julio de 2021 CONSIDERANDO: Que, e...
TRABAJO Y PROMOCION DEL EMPLEO	RESOLUCION SUPREMA N° 018-2021-TR	15/07/2021	Renuevan designación de representante del Esta...	Lima, 14 de julio de 2021 VISTO: El Oficio N° ...
ECONOMIA Y FINANZAS	DECRETO SUPREMO N° 129-2021-EF	30/05/2021	Autorizan Transferencia de Partidas en el Pres...	EL PRESIDENTE DE LA REPÚBLICA CONSIDERANDO: Qu...
CONGRESO DE LA REPUBLICA	LEY N° 31230	23/06/2021	Ley que declara el distrito de Sacsamarca pueb...	LA PRESIDENTA A. I. DEL CONGRESO DE LA REPÚBLI...

Figura 7: Dataset inicial

4.2. *Limpieza del texto*

Se elabora una lista de palabras y oraciones que por su gran número de repeticiones es conveniente eliminar, así como la lista de stopwords y el diccionario español para la lematización.

Definimos funciones para realizar distintos niveles de limpieza: Para tópicos con lematización, Para tópicos sin lematización y para Sumarización.

Luego, utilizando estas funciones de limpieza, ampliamos nuestro dataset, incluyendo nuevas columnas como se aprecia en la siguiente imagen.

category	title	date	abstract	content	content_for_summary	abstract_for_summary	content_for_topic	content_for_topic_lemma
JUSTICIA Y DERECHOS HUMANOS	RESOLUCION MINISTERIAL N° 0107-2021-JUS	11/06/2021	Aprueban el Trigésimo Primer Listado de Benefi...	Lima, 7 de junio de 2021 VISTOS, el Oficio N° ...	lima de junio de vistos el oficio jus se cman ...	aprueban el trigésimo primer listado de benefi...	lima junio vistos oficio cman secretaría ejecu...	lima junio visto oficio cmar secretaría ejecut...
ERGIA Y MINAS	RESOLUCION MINISTERIAL N° 252-2021-MINEM/DM	23/07/2021	Constituyen el derecho de servidumbre de ocupa...	Lima, 22 de julio de 2021 VISTOS: El escrito c...	lima de julio de vistos el escrito con registr...	constituyen el derecho de servidumbre de ocupa...	lima julio vistos escrito registro presentado ...	lima julio visto escrito registro presentado e...
RINTENDENCIA NACIONAL DE EDUCACION SUPERIO...	RESOLUCION N° 059-2021-SUNEDU/CD	10/06/2021	Deniegan la modificación de licencia instituci...	Lima, 8 de junio de 2021 VISTOS: La Solicitud ...	lima de junio de vistos la solicitud de modifi...	deniegan la modificación de licencia instituci...	lima junio vistos solicitud modificación licen...	lima junio visto solicitud modificación licenc...

Figura 8: Dataset aumentado

Posteriormente se realiza una limpieza adicional, donde se busca eliminar aquellas filas que procedan de una corrección de una publicación anterior. Dado que frecuentemente se emplea la palabra "errata"para estos casos, empleamos esta palabra para ayudarnos en la búsqueda y posterior eliminación de estas entradas.

Observamos que en nuestro Corpus, al usar el tokenizador sin lematización, obtenemos 2,898,532 tokens y un vocabulario de 40,616 palabras, pero al usar lematización obtenemos 2,888,275 tokens y un vocabulario de 30,380 palabras.

	words	counts
0	resolución	24242
1	decreto	23373
2	nacional	20258
3	general	18819
4	mediante	16886
5	presente	16840
6	supremo	14321
7	reglamento	13487
8	ministerio	13381
9	informe	12389

	words	counts
0	resolución	25667
1	decreto	24044
2	nacional	21463
3	general	20639
4	público	17865
5	presente	16953
6	mediante	16886
7	supremo	15885
8	reglamento	13728
9	informe	13657

Figura 9: Tokenización sin (Izq.) y con lematización (Der.)

4.3. Procedimiento Utilizado para el Topic Modeling

Los datos utilizados para el Topic Modeling comprenden todas las normas publicadas en el diarios oficial El Peruano entre el 23 de mayo de 2021 al 21 de agosto de 2021 comprendiendo un periodo de 3 meses.
En total el dataset está compuesto por 4037 filas, donde cada fila representa a una norma. Se realizó el preprocesado como se indicó en el punto anterior con y sin lematización.

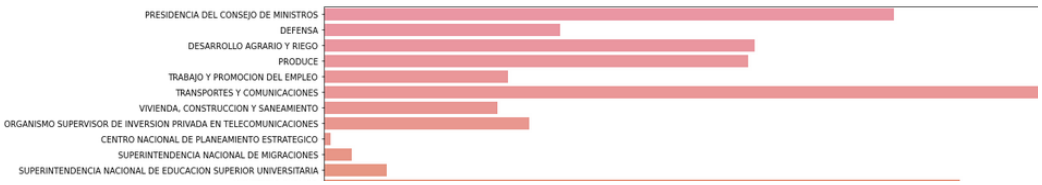

Figura 10: Categorías del dataset (brindadas por la pagina)

Al explorar el dataset, se observa que las palabras mas representativas son las que se muestran en la figura 10.

Se cuenta en total con un vocabulario de 30380 palabras y tenemos 2888275 tokens.

Se realiza la vectorización del texto con TF-IDF, usando una frecuencia minima para la palabra de 0.01 y una máxima de 0.99, con un número maximo de 20000 palabras. Con este procedimiento se obtiene una matríz de 4037 filas por 3226 columnas.

```
[('resolución', 25667),
 ('decreto', 24044),
 ('nacional', 21463),
 ('general', 20639),
 ('público', 17865),
 ('presente', 16953),
 ('mediante', 16886),
 ('supremo', 15885),
 ('reglamento', 13728),
 ('informe', 13657)]
```

Figura 11: Palabras mas repetidas

4.4. Ejecución del modelo

Se ejecuta el modelo Latent Dirichlet Allocation (LDA) durante 10 iteraciones con los siguientes hiperparámetros *'learning_decay':* 0.7 y un *'n_components':* 5

	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
Word 0	municipalidad	designar	decreto	fiscal	judicial
Word 1	ordenanza	cargo	regional	electoral	justicia
Word 2	municipal	ministerio	salud	fiscalía	corte
Word 3	gerencia	designación	nacional	elección	superior
Word 4	tributario	ministro	ministerio	provincial	osinergmin
Word 5	distrital	ejecutivo	supremo	acta	resolución
Word 6	concejo	señor	general	distrito	juez
Word 7	alcalde	general	presupuesto	penal	juzgado
Word 8	administrativo	confianza	público	junta	poder
Word 9	alcaldía	concesión	resolución	voto	cultural

Figura 12: Cuadro que muestra las 10 palabras asociadas a cada tópico (0,1,2,3,4)

Posteriormente hacemos un análisis de tópicos para cada norma publicada. Los resultados para una muestra de normas se puede ver en la Figura 13. Este análisis puede permitir clasificar las normas publicadas en función a determinados tópicos para luego generar reportes o hacer un seguimiento de las mismas. Sin embargo en la siguiente sección se desarrollará un análisis más completo que buscará realizar el trabajo de sumarización utilizando el método abstractivo.

	Topic0	Topic1	Topic2	Topic3	Topic4	TOPIC
Norma 3209	0.017	0.017	0.933	0.017	0.017	2
Norma 1302	0.025	0.025	0.902	0.024	0.024	2
Norma 646	0.018	0.018	0.928	0.018	0.018	2
Norma 1200	0.923	0.019	0.019	0.019	0.019	0
Norma 980	0.018	0.017	0.439	0.017	0.509	4
Norma 2551	0.021	0.021	0.918	0.020	0.020	2
Norma 434	0.023	0.909	0.023	0.023	0.023	1
Norma 3674	0.920	0.019	0.019	0.022	0.019	0
Norma 1116	0.020	0.016	0.017	0.017	0.930	4
Norma 1185	0.919	0.020	0.021	0.020	0.020	0

Figura 13: Análisis de topicos en las normas publicadas

5. Arquitectura del Modelo para la Sumarización

El modelo utilizado para la sumarización es un modelo secuencia a secuencia (SeqToSeq) que toma una secuencia de objetos (palabras, letras, series de tiempo, etc.) y genera otra secuencia de objetos. Todo el procesado de la información lo realiza una estructura de redes neuronales recurrentes (RNN) que primero transfiere una cadena de entrada de longitud variable en una representación vectorial de longitud fija, que luego se utiliza para generar una cadena de salida.

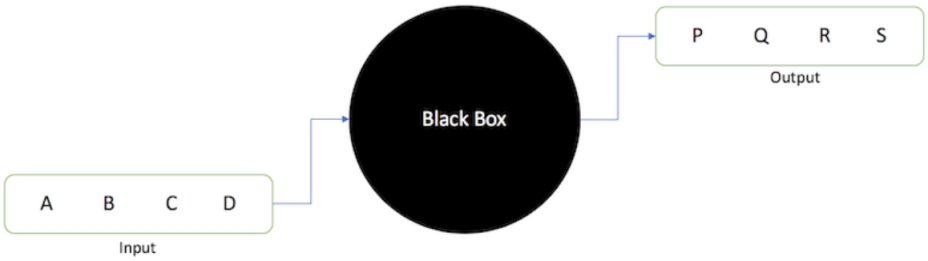

Figura 14: Modelo secuencia a secuencia

Hay varios tipos de modelos SeqToSeq y, según el tipo de problema, se debe seleccionar el correcto. En este caso de resumen de texto, la aplicación más apropiada es un modelo many to many en donde tanto la entrada como la salida consisten en varias palabras.

5.1. Encoder-Decoder

Los modelos de secuencia a secuencia se basan en lo que se conoce como una arquitectura de codificador-decodificador, es decir una combinación de redes neuronales recurrentes en capas que están organizadas de manera que les permite realizar las tareas de codificar una secuencia de palabras y luego pasar esa secuencia codificada a una red de decodificadores para producir una salida. En el caso de la sumarización, la secuencia de entrada primero se tokeniza y luego se introduce palabra por palabra en el codificador. El codificador transforma la secuencia reduciendo su dimensionalidad para que después de pasar al decodificador, se convierta en la base para producir una secuencia de salida, que en nuestro caso sería la versión resumida del texto original.

Sin embargo el proceso de sumarización plantea una serie de problemas e inconvenientes y está relacionado a que la arquitectura del codificador-decodificador es adecuada para considerar el contexto, esta arquitectura plantea dificultades al momento de que las series de entrada se agrandan. Con secuencias de entrada largas, el vector de estado final que genera el codificador puede perder información contextual importante de puntos anteriores de la secuencia y esto debido a que todo el contexto del texto de entrada se comprime en un solo vector de estado.

5.2. *Attention*

El *attention* sirve para ayudar al modelo codificador-decodificador a enfocarse en ciertas secciones o palabras relevantes en el texto de entrada al predecir el siguiente *token* de salida. Esto ayuda a mitigar el problema de la pérdida de contexto de fragmentos anteriores de una secuencia de entrada, de esta manera, en lugar de un solo vector de contexto basado en el último estado oculto del codificador, el vector de contexto se construye utilizando todos los estados ocultos del codificador.

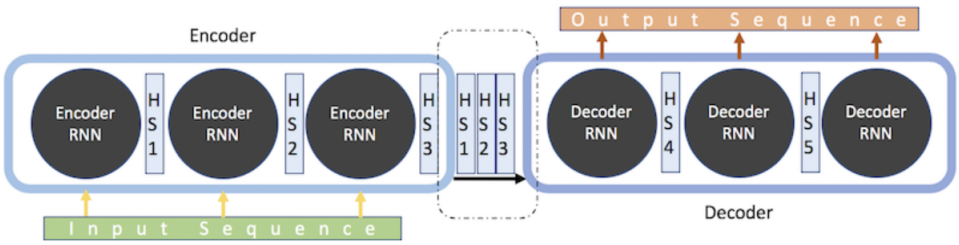

Figura 15: Arquitectura *Encoder-Decoder* con *Attention*

Cuando se combinan los estados ocultos en la salida final del codificador, cada vector o estado oculto obtiene su propio peso aleatorio. Estos pesos son recalculados mientras que otra red neuronal es entrenada en paralelo al decodificador que verifica qué tan bien se ajusta la última salida del decodificador a los diferentes estados transmitidos desde el codificador. Dependiendo de las puntuaciones de ajuste respectivas, los pesos del modelo de alineación se optimizan mediante propagación hacia atrás. A través de esta ponderación dinámica, la importancia de los diferentes estados ocultos varía entre las instancias de entrada y salida, lo que permite que el modelo preste más atención (peso / importancia) a los diferentes estados del codificador en función de la entrada.

6. Procedimiento Utilizado para la Sumarización

Los datos utilizados para la sumarización comprenden todas las normas publicadas en el diarios oficial El Peruano entre el 23 de mayo de 2021 al 21 de agosto de 2021 comprendiendo un periodo de 3 meses. Este conjunto de normas comprende tanto normas del poder ejecutivo, poder legislativo y normas de caracter local y regional. La información con la que se realizó el trabajo de sumarización utiliza el mismo proceso de captura y recopilación de la información mostrado anteriormente para el análisis de tópicos, sin embargo dado que el trabajo de sumarización no se requiere el mismo nivel de pre procesado dado que aquí se busca mantener las características esenciales del texto a analizar, en ese sentido el preprocesado es muy parecido al que se realiza para las labores de traducción automática utilizando redes neuronales.

Al preparar los datos para el modelado, se realiza una limpieza en los títulos así como en los textos que se repiten más y tal como se mencionó en el párrafo anterior a diferencia de la extracción de tópicos los textos son pre-procesados sin considerar stemmer o lemmatizer. Es necesario mantener la palabras originales para ayudar a la red a mejorar la exactitud en la generación del resumen.

	category	title	date	abstract	content	content_for_summary	abstract_for_summary
1481	MINISTERIO PUBLICO	RESOLUCION N° 1016-2021-MP-FN	17/07/2021	Aceptan renuncia de Fiscal Adjunto Provincial ...	Lima, 16 de julio de 2021 VISTOS Y CONSIDERAND...	lima de julio de vistos el oficio mp fn fsnced...	aceptan renuncia de fiscal adjunto provincial ...
2369	MINISTERIO PUBLICO	RESOLUCION N° 915-2021-MP-FN	28/06/2021	Aceptan renuncia de fiscal del Distrito Fiscal...	Lima, 28 de junio de 2021 VISTOS Y CONSIDERAND...	lima de junio de vistos los oficios nros mp fn...	aceptan renuncia de fiscal del distrito fiscal...
117	PRODUCE	RESOLUCION MINISTERIAL N° 00256-2021-PRODUCE	18/08/2021	Designan Asesor II del Despacho Ministerial	Lima, 17 de agosto de 2021 CONSIDERANDO: Que, ...	lima de agosto de que se encuentra vacante el ...	designan asesor ii del despacho ministerial
1671	AGENCIA DE PROMOCION DE LA INVERSION PRIVADA	ACUERDO N° 109-3-2021-CD	14/07/2021	Aprueban la incorporación de los Proyectos Enl...	Sesión Nº 109 del 09 de julio de 2021 Acuerdo ...	sesión del de julio de acuerdo proinversión cd...	aprueban la incorporación de los proyectos enl...
3591	ORGANISMO SUPERVISOR DE LA INVERSION EN ENERGI...	RESOLUCION N° 042-2021-OS/GRT	02/06/2021	Aprueban costos administrativos y operativos d...	Lima, 31 de mayo de 2021 CONSIDERANDO: Que, co...	lima de mayo de que con la ley en adelante la ...	aprueban costos administrativos operativos del...

Figura 16: Muestra del Dataset utilizado

Un primer paso consiste en eliminar las secuencias de palabras con extensión grande tanto del texto original como del texto sumarizado. Viendo el histograma del tamaño de las secuencias que se muestra en la Figura 17 se utiliza una extensión máxima para el texto original de $3,000$ palabras, mientras que la extensión máxima para las palabras del resumen es de 35.

El filtrado de normas con muchas palabras nos muestra que el número de normas con palabras entre 1 y 3000 representa el $90.46\,\%$ del total de normas, mientras que el número de resúmenes con palabras menores a 35 representa el $92.02\,\%$ del total de normas.

El siguiente paso consiste en agregar al texto resumen la cadena START al inicio y END al final, esto con la finalidad de encontrar los límites del resumen en la fase de inferencia. Los resultados de este procesamiento se muestran en la Figura 18.

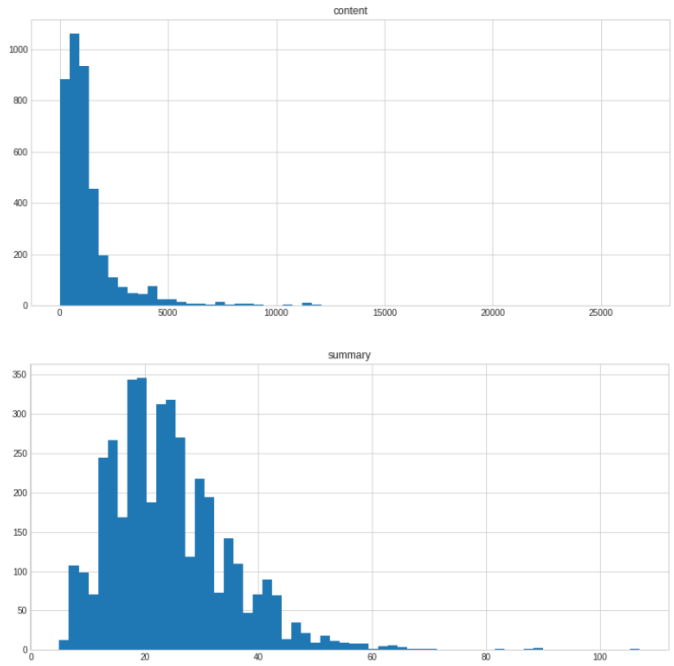

Figura 17: Distribución del tamaño de secuencia de palabras

	content_for_summary	abstract_for_summary
3896	lima de mayo de visto el informe técnico grt e...	_START_ fijan margen comercial banda de precio...
2488	lima de junio de vistos el memorando minam vmg...	_START_ modifican la primera actualización del...
2297	lima de junio del visto el expediente que cont...	_START_ aprueban el presupuesto analítico de p...
1110	lima de julio de visto el acta de la octogésim...	_START_ formalizan el acuerdo de la octogésima...
3220	san bartolo de mayo de el concejo municipal de...	_START_ ordenanza que regula la expedición de ...

Figura 18: Muestra del Dataset procesado

6.1. Determinación de los hiperparámetros del modelo

Para el entrenamiento del modelo hacemos una partición para tener una muestra de entrenamiento correspondiente al 90 % de los datos y una muestra de validación o test correspondiente al 10 % de los datos. Esto hace que la muestra final utilizada en el entrenamiento sea de $2,901$ normas legales y un tamaño de vocabulario de $1,333$.

Para el encoder utilizamos una dimensión del embeding de 200 y 300 neuronas en la capa LSTM. Utilizamos 3 capas LSTM. Para el decoder utilizamos una capa LSTM con un dropout de 0,4 al igual que en el encoder. También utilizamos una capa correspondiente al Attention y una capa de concatenación.

En la Figura 19 puede verse el modelo final en donde se observa que la arquitectura utilizada requiere un total de $6,548,633$ de parámetros entrenables.

```
Model: "model"
__________________________________________________________________________________________________
Layer (type)                    Output Shape         Param #     Connected to
==================================================================================================
input_1 (InputLayer)            [(None, 3000)]       0
__________________________________________________________________________________________________
embedding (Embedding)           (None, 3000, 200)    2655800     input_1[0][0]
__________________________________________________________________________________________________
lstm (LSTM)                     [(None, 3000, 300),  601200      embedding[0][0]
__________________________________________________________________________________________________
input_2 (InputLayer)            [(None, None)]       0
__________________________________________________________________________________________________
lstm_1 (LSTM)                   [(None, 3000, 300),  721200      lstm[0][0]
__________________________________________________________________________________________________
embedding_1 (Embedding)         (None, None, 200)    266600      input_2[0][0]
__________________________________________________________________________________________________
lstm_2 (LSTM)                   [(None, 3000, 300),  721200      lstm_1[0][0]
__________________________________________________________________________________________________
lstm_3 (LSTM)                   [(None, None, 300),  601200      embedding_1[0][0]
                                                                 lstm_2[0][1]
                                                                 lstm_2[0][2]
__________________________________________________________________________________________________
attention_layer (AttentionLayer ((None, None, 300),  180300      lstm_2[0][0]
                                                                 lstm_3[0][0]
__________________________________________________________________________________________________
concat_layer (Concatenate)      (None, None, 600)    0           lstm_3[0][0]
                                                                 attention_layer[0][0]
__________________________________________________________________________________________________
time_distributed (TimeDistribut (None, None, 1333)   801133      concat_layer[0][0]
==================================================================================================
Total params: 6,548,633
Trainable params: 6,548,633
Non-trainable params: 0
__________________________________________________________________________________________________
```

Figura 19: Modelo utilizado

6.2. Entrenamiento del modelo

Para el entrenamiento del modelo se utilizó como optimizador un algoritmo RMSprop, este algoritmo mantiene un promedio móvil del cuadrado de gradientes para posteriormente dividir la gradiente en por la raíz cuadrada de dicho promedio móvil.El learning rate por default en este algoritmo es de 0.001. Se utilizaron 30 épocas y un batch size de 8.

En la Figura 20 se puede apreciar que la función de pérdida a lo largo de las épocas/iteraciones es decreciente lo cual indica que el modelo tiene un buen comportamiento en esta etapa. Además podemos inferir que la pérdida de validación es mayor entre la época 1 y 4 de forma sucesivas. Por tanto, el entrenamiento se detiene en épocas mayores a 20.

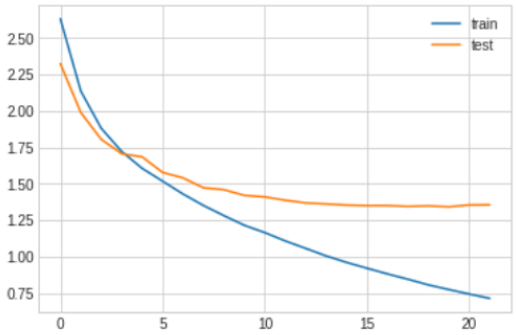

Figura 20: Entrenamiento del Modelo

7. Resultados de la Sumarización

Luego del proceso de modelamiento y entrenamiento detallados en la sección anterior, los resultados generados con el modelo son bastante satisfactorios. En la siguiente tabla se puede apreciar que el desempeño del proceso de sumarización es bastante bueno y los resúmenes muestran coherencia y capturan la esencia del documento original.

Tabla del Proceso de Inferencia		
Contenido de la Norma	**Resumen Original**	**Resumen Predicido**
lima de agosto de vistos el informe mimp omi de la oficina de modernización institucional el memorándum mimp ogpp de la oficina general de planeamiento presupuesto que mediante decreto legislativo modificatoria se aprueba la ley de organización funciones del ministerio de la mujer poblaciones vulnerables que determina su ámbito de competencia funciones estructura orgánica básica como organismo rector en las políticas nacionales sectoriales sobre mujer promoción protección de las poblaciones vulnerables que los numerales del de la ley ley orgánica del poder ejecutivo modificatorias establecen que los ministerios son organismos del poder ejecutivo que comprenden uno varios sectores considerando su homogeneidad finalidad que estos diseñan establecen ejecutan supervisan políticas nacionales sectoriales asumiendo la rectoría respecto de ellas que el de la ley ley marco de modernización de la gestión del estado modificatorias establece que el proceso de modernización de la gestión del estado...	aprueban el texto integrado del reglamento de organización funciones del ministerio de la mujer poblaciones vulnerables	aprueban el plan operativo institucional poi poi del ministerio de comercio exterior turismo
lima diez de febrero de dos mil veintiuno vista la investigación número seiscientos dieciocho guion dos mil diecisiete guion del santa que contiene la propuesta de destitución de la señora soledad rojas en su actuación como técnico judicial del juzgado de paz letrado investigación preparatoria de pallasca cabana corte superior de justicia del santa remitida por la jefatura de la oficina de control de la magistratura del poder judicial en merito la resolución número quince del diecinueve de diciembre de dos mil diecinueve de fojas doscientos noventa dos trescientos dos primero que mediante resolución número uno del once de setiembre de dos mil diecisiete de fojas trece al dieciséis el jefe de la unidad de quejas investigaciones visitas defensoría del usuario judicial de la oficina desconcentrada de control de la magistratura de la corte superior de justicia del santa abrió procedimiento administrativo disciplinario contra la servidora judicial soledad rojas en su actuación como técnico ...	imponen la medida disciplinaria de destitución técnico judicial del juzgado de paz letrado investigación preparatoria de corte superior de justicia del santa	imponen medida disciplinaria de destitución juez de paz de segunda nominación del distrito de provincia de departamento distrito judicial de la libertad

Tabla del Proceso de Inferencia		
Contenido de la Norma	**Resumen Original**	**Resumen Predicido**
lima de mayo de vistos el oficio extra fap secre fap de la secretaría general de la comandancia general de la fuerza aérea del perú los oficios mindef vpd digrin mindef vpd digrin de la dirección general de relaciones internacionales el informe legal mindef sg ogaj de la oficina general de asesoría jurídica que mediante resolución suprema de se designa al comandante fap carlos enrique para ocupar el cargo de delegado alterno de la delegación de perú ante la junta interamericana de defensa jid participar en la maestría acreditada en ciencias de defensa seguridad interamericana clase lxi desempeñarse como asistente en el colegio interamericano de defensa en la ciudad de washington de los estados unidos de américa órdenes del ministerio de relaciones exteriores del de julio de al de junio de que con oficio fap fap la dirección general de personal de la fuerza aérea del perú remite la jefatura del estado mayor general de la fap la documentación que sustenta la presente autorización de viaj...	autorizan viaje de oficial de la fuerza aérea del perú ee uu en comisión especial	autorizan viaje de personal militar de la marina de guerra del perú ee uu en comisión especial en el exterior
lima de junio de visto la nota de elevación mtc de la dirección ejecutiva del proyecto especial de infraestructura de transporte nacional provias nacional que la quinta disposición complementaria final de la ley ley que facilita la adquisición expropiación posesión de bienes inmuebles para obras de infraestructura declara de necesidad pública la adquisición expropiación de bienes inmuebles afectados para la ejecución de diversas obras de infraestructura entre otros declara de necesidad pública la ejecución de la obra red vial tramo pativilca santa trujillo puerto salaverry empalme autoriza la expropiación de los bienes inmuebles que resulten necesarios para tal fin que el del texto único ordenado del decreto legislativo decreto legislativo que aprueba la ley marco de adquisición expropiación de inmuebles transferencia de inmuebles de propiedad del estado liberación de interferencias dicta otras medidas para la ejecución de obras de infraestructura aprobado por decreto supremo vivienda ...	aprueban ejecución de expropiación de área de inmueble afectado por la ejecución de la obra red vial tramo pativilca santa trujillo puerto salaverry empalme el valor de su tasación	aprueban ejecución de expropiación de inmueble afectado por la obra red vial tramo pativilca santa trujillo puerto salaverry empalme el valor de su tasación
que los del decreto legislativo decreto legislativo del sistema nacional de presupuesto público establecen que las leyes de presupuesto del sector público consideran una reserva de contingencia que constituye un crédito presupuestario global dentro del presupuesto del ministerio de economía finanzas destinada financiar los gastos que por su naturaleza coyuntura no pueden ser previstos en los presupuestos de los pliegos disponiendo que las transferencias habilitaciones que se efectúen con cargo la reserva de contingencia se autorizan mediante decreto supremo refrendado por el ministro de economía finanzas que mediante los oficios ns grl gr goremad gr gru gr los gobiernos regionales de los departamentos de loreto madre de dios ucayali respectivamente solicitan una demanda adicional de recursos para financiar las actividades relacionadas al programa presupuestal competitividad aprovechamiento sostenible de los recursos forestales de la fauna silvestre con el fin de cumplir las metas físic...	autorizan transferencia de partidas en el presupuesto del sector público para el año fiscal favor de diversos gobiernos regionales	autorizan transferencia de partidas favor del ministerio de salud en el presupuesto del sector público para el año fiscal

8. Conclusiones y Recomendaciones

En este documento se desarrolla el análisis de normas legales peruanas publicadas en el diario oficial El Peruano utilizando dos técnicas de procesamiento de lenguaje natural (NLP) que son la extracción de tópicos y la sumarización. Existen dos enfoques diferentes que se utilizan para el resumen de texto que son el resumen extractivo y resumen abstractivo. El resumen extractivo identifica las oraciones o frases importantes del texto original y solo las extrae. Esas frases extraídas serían nuestro resumen. En este trabajo hemos realizado un proceso de sumarización abstractiva. Para la sumarización hemos utilizado redes neuronales recurrentes empleando métodos de atención (*attention*) dado que estos ayudan en el entrenamiento del modelo así como mejoran mucho la calidad predictiva de los resúmenes.

Referencias

[1] Md. Majharul Haque, Suraiya Pervin, y Zerina Begum.(2013) Literature Review of Automatic Single Document Text Summarization Using NLP . Innovative Space of Scientific Research Journals, Vol. 3 pp 857-865.

[2] Goodfellow, I., Bengio, Y., y Courville, A. (2016). Deep Learning. The MIT Press.

[3] Jurafsky, D. y Martin, J. (2020). Speech and Language Processing. Third Edition (Draft).

[4] Nallapati, R., Zhai, F., y Zhou, B. (2016). Summarunner: A recurrent neural network based sequence model for extractive summarization of documents.

[5] Chopra, S., Auli, M., y Rush, A. (2016). Abstractive sentence summarization with attentive recurrent neural networks. In NAACL HLT 2016, junio 12-17, 2016, pp. 93–98.

Apéndice A Lecciones Aprendidas

Código	L001
Fase	Extracción de datos
Descripción	Se tenía planeado extraer todo el texto de las publicaciones en PDF de las normas legales. Sin embargo, luego de revisión más detallada se pudo observar que el contenido de las normas ya estaban clasificadas y expuestas en la página web del "Diario Oficial el Peruano" y no era necesario leer el PDF.
Acciones	Debido a que no tomo mucho tiempo en darnos cuenta que la información se encontraba estructurada; realizamos el análisis de la página web para realizar un "Webscrapping".
Resultados	Nos permitió canalizar el esfuerzo para las siguientes fases del trabajo.
Recomendación	Antes de realizar el plan de extracción revisar con más de detalle la página web del diario.

Código	L002
Fase	Exploración de datos
Descripción	Dado que el plan inicial consistía en realizar sumarización a través de los tópicos encontrados, cuando se generó la información del "Topic Modelling" nos percatamos que solo se podría generar resúmenes de **tipo extractivo**.
Acciones	Se acordó explorar modelos de NLP que realicen resúmenes de tipo abstractivo.
Resultados	Se eligió el modelo "Sequence to Sequence"que permitió generar resúmenes de tipo abstractivo.
Recomendación	Se recomienda. para trabajos similares, realizar una fase de revisión previa al inicio donde se pueda revisar a detalle los modelos y los resultados que arrojen. Considerar que aún cuando se aplique esta técnica, no se estará exento de que vuelva a suceder, debido a que el entendimiento ocurre cuando se va explorando los datos.

Código	L003
Fase	Extracción y Exploración de datos
Descripción	Nuestro dataset orginal consistía en 10 meses, un apróximado de 15,000 registros, el cual demoró en extraer entre 1.5 a 2 horas. Luego en la exploración nos dimos cuenta que no se podría trabajar con un dataset de esas dimensiones debido a las limitaciones de recursos de cómputo.
Acciones	Se consulto al profesor acerca de requisitos con respecto a los datos. Él confirmo que no es necesario y nos enfoquemos a entenderlo y aplicar lo aprendido en clase.
Resultados	Se eligió trabajar con tres (3) meses de información, que representó un aproximado de 4000 registro .
Recomendación	Confirmar los requisito de datos en un etapa temprana del trabajo, esto se puede realizar preguntando al profesor, incluso se debe considerar la limitaciones de recursos de cómputo para encontrar un equilibrio entre lo solicitado y lo factible.

Código	L004
Fase	Modelado/Entrenamiento de SeqToSeq+Attention
Descripción	En la etapa de entrenamiento del modelo (*"modelo.fit()"*) nos enfrentamos a un problema de recursos de cómputo, tanto en el Colab como en las máquinas personales, los modelos no podían entrenarse por falta memoria.
Acciones	Se encontró que la causa de la falta de memoria era la cantidad de "batch" (configurado a $batch\_size = 128$) sumado a la cantidad de parámetros del modelo (6,548,633) y la cantidad registros con contenido/texto en promedio 3000 tokens. Se configuró el $batch\_size = 8$) para superar el problema en nuestros máquinas personales.
Resultados	La reducción de batch permitió la ejecución del entrenamiento del modelo en Colab y máquinas personales.
Recomendación	Revisar experiencias previas para encontrar el equilibrio entre batch, tamaño del dataset, parámetros del modelo y capacidades del GPU y TPU en máquinas locales o el Colab.

Código	L005
Fase	Limpieza del Dataset y Proceso de Inferencia
Descripción	Para el entrenamiento del modelo SeqToSql+Attention se ingreso el contenido/texto sin stop-words y caracteres menores iguales que tres y abstract/resúmenes que no tenían limpieza de ese tipo. En el proceso de inferencia los resultados, al no poder correlacionar textos que no existían en el contenido, las sumarizaciones/resúmenes no eran claras, incluso se repetían lo textos sin correlación, p.e. "la ley la ley la ley...".
Acciones	Se aplicó el mismo tipo de limpieza al contenido/texto y a los resúmenes; sin quitar stop-words, sin Lematización y solo se quitó caracteres de tamaño uno (1).
Resultados	El proceso de inferencia mejoró de manera considerable.
Recomendación	Aplicar el mismo tipo de limpieza a los datos de entrenamiento y prueba, para evitar un mal rendimiento de los modelos de NLP.

Código	L006
Fase	Modelado y Proceso de Inferencia
Descripción	En un inicio el diseño de modelo solo consideró implementar SeqToSeq (Encoder-Decoder). En el proceso de inferencia nos dimos cuenta que los resúmenes siempre era los mismos.
Acciones	Revisando el modelo encontramos que SeqToSeq aprendía el contexto pero se atascaba (bottleneck) en vectores de contexto de mayor peso/frecuencia por lo que los resultado eran siempre lo mismo. Mediante una actividad adicional de investigación y revisión de lo explicado en clase, superamos el problema agregando un capa de Attention que permitió darle el peso adecuado a los tokens en el entrenamiento a cada entrada (contenido/texto + resumen).
Resultados	El proceso de inferencia mejoró de manera considerable.
Recomendación	Tener siempre presente que uno de los principales problema de los modelos NLP SeqTOSeq presentan el inconveniente de quedarse atascados por están en estado bottleneck.

Código	L007
Fase	Extracción de datos
Descripción	El proceso de extracción consiste en ir a la página web del Diario el Peruano de manera constante y por un periodo largo. La invocaciones a las URL empezarán a fallar debido a una denegación a la conexión.
Acciones	Todo empresa tiene implementado mecanismos anti-procesos de bloqueo o saturación, conocidos como DDoS (Distributed Denial-of-Service). Logramos superar el problema estableciendo pausa periódicas de 10 segundos durante la extracción.
Resultados	Extracción completa de los datos sin corte de conexión.
Recomendación	Considerar siempre que los diarios y empresas en general tienen implementados mecanismos de seguridad, como el DDoS, que evitan consultas masivas..