

**Universidad Nacional Autónoma de México**





**Facultad de Ingeniería**



División de Ingeniería Eléctrica

Asignatura: Análisis y Procesamiento Inteligente de Textos

C:\Users\yareli\analisisInteligenteDeTextos\AnalisisYProcesamientoDeTextosProyectoFinal

Profesor (a):M.P. Octavio Augusto Sánchez Velázquez

PROYECTO

**“RESUMEN DE TEXTOS AUTOMÁTICO”**

Alumna(s): Buendía Valdivia María Angélica

Isabel Gómez Yareli Elizabeth

Zagoya Mellado Roberto Uriel

Grupo 1

Semestre 2020-1

Fecha de entrega: 2020

Índice

[**I.Introducción**](#_30j0zll) **3**

[Motivación](#_ecilo6iy6jw8) 3

[Justificación](#_14l0qks1afp1) 3

[Problema](#_vegg65aaymt2) 3

[Solución Propuesta](#_vegg65aaymt2) 3

[**II.Marco teórico**](#_myrxt2j8zwhc) **3**

[Antecedentes](#_8zxiqsrw9dn7) 3

[Trabajos relacionados](#_foktee8t9uic) 3

[Descripción](#_lvtpvn3jxxos) 3

[**III.Método experimental**](#_17tn2qdhaq0x) **3**

[Datos ¿Cómo se obtuvieron?](#_cjcl4r9sr9fz) 3

[Descripción de su método](#_myohg9kd7osg) 3

[Descripción del experimento](#_flv771hsngry) 3

[Presentación de resultados](#_1ttobnke38lp) 3

[**IV.Conclusión**](#_tajh1x6g08d) **3**

[Discusión de resultados](#_jjcxji5veuf) 3

[Descripción de observaciones](#_g91yfzfw2xgv) 3

[Trabajo futuro](#_l1d4xpkv055f) 3

[**V.Referencias**](#_m4xtkaq771sx) **4**

# **Introducción**

Durante la existencia de la humanidad el conocimiento se ha convertido esencial para su desarrollo y crecimiento, por lo cual la lectura de información ha sido un punto primordial para desempeñar y llegar a cumplir cualquiera de sus objetivos.

Para comprender la mayoría de la información se tiene que leer un texto sumamente largo, lleno de conceptos y con una letra muy pequeña que en algunas ocasiones no se alcanza a ver. Para algunos, la lectura no es su hobby, además consume mucho tiempo y energía para comprender una idea, así que surge la necesidad de acortar la información más relevante. Dado esto, nace el resumen, este es un escrito que logra captar todas las ideas principales de un texto y sintetizarlas en un escrito mucho más pequeño. La extensión de un resumen varia, ya que no todos los textos son de la misma extensión, pero la mayoría tiene el 25% de la extensión original.

El objetivo del resumen es lograr reducir información a la más importante de cualquier tema, demostrar que se entiende lo resumido del texto original y que sea de utilidad para las personas. Los resúmenes han sido de gran ayuda ya que han servido para la realización de presentaciones, para estudiar algún tema, para realizar alguna exposición y para un trabajo académico.

Existen tipos de resúmenes como el indicativo y el informativo el indicativo es aquel donde se incluyen los temas del documento resumido y la metodología y no supera la extensión de las 100 o 200 palabras. El resumen informativo es aquel en donde se profundiza sobre el texto original siendo un método científico (objetivos, metodología, resultados y conclusiones), en el proyecto se implementa más el resumen indicativo.

En décadas anteriores aún se veía que se elaboraban con más frecuencia los resúmenes a mano, es decir, la persona leía todo el texto y al final o durante la lectura hacia anotaciones y subrayaba los puntos que a esa persona se le hacían relevantes e importantes en la actualidad aún se sigue utilizando pero conforme paso el tiempo se ha hecho se suma importancia el elaborar resúmenes en gran volumen así que gracias a que la tecnología ha ido evolucionando hoy en día existen diferentes softwares que nos permiten realizar estos resúmenes de una manera más práctica, automática. Estas herramientas ahora permiten ingresar el texto y resumirlo en cuestión de milisegundos, por lo que así se puede acortar mucho más el tiempo que inviertes en realizar esta actividad al lector y obtener el mismo resultado de haber captado la misma información.

Algunas tecnologías actuales pueden hacer un resumen coherente de un texto al tener en cuenta variables como la longitud, la escritura y la sintaxis.

El resumen automático es parte del área del aprendizaje automático y de la mimería de datos. Existen dos enfoques del resumen automático: la extracción y la abstracción, la extracción funciona seleccionando un subconjunto de palabras, frases u oraciones existentes en el texto original para formar el resumen, en la abstracción es un método que construye una representación semántica interna, a la que le aplican técnicas de generación de lenguaje natural para crear un resumen más similar al que realizaría un humano que más adelante se explicarán a más detalle.

Para este proyecto realizaremos una aplicación la cual podrá resumir un texto conservando las ideas más importantes que hay en este, se podrá ingresar (copiar y pegar) el texto, directamente en esta aplicación o, con la URL de la página que contiene un texto del cual tú quieres obtener un resumen. Este no será un sistema sin fallas y puede ser que, en los resúmenes, contenga equivocaciones porque funcionará con base en los algoritmos propios de las librerías de Python, las cuales se describirán en secciones posteriores, y no usa un pensamiento lógico y crítico del ser humano, por lo tanto, deberá estar atento al resumen obtenido y se deberá revisar los párrafos finales, para verificar la coherencia y verificar si es el resultado.

### Motivación (interna o Externa)

La motivación para realizar este proyecto fue ver la necesidad principalmente de los estudiantes para realizar los resúmenes, muchas veces se junta la carga de trabajo y no da tiempo de leer todos los documentos completos porque algunos tienen gran extensión de contenido y la mayoría de los profesores dejan muchos artículos y algunos libros para leer, pero para eso también sirven los resúmenes para reducir el tiempo y poner captar las ideas esenciales del documento sin tener que haber leído todo. Al elaborar este proyecto lo que se pretende es que los alumnos tengan una herramienta útil para y así facilitar y analizar su contenido desde artículos a textos de internet, de igual manera el ayudar a tener una mejor comprensión lectora.

### Justificación (que es lo que justifica la existencia de ese trabajo)

El proyecto fue elegido entre tres diferentes proyectos, las opciones que existían eran “El análisis de sentimientos”, “Chatbot” y “El resumen de textos”. Estos tres fueron elegidos por diferentes motivos, lo que nos llamó la atención en el análisis de sentimientos es que este es muy utilizado en redes sociales para verificar la opinión que se tiene de las comunidades sobre ciertos temas, también es utilizada en organizaciones mundiales, un ejemplo es Obama, él lo utilizo para sondear la opinión en las elecciones. Lo que encontramos de este proyecto es que el entender los sentimientos no es una tarea fácil ya que se requiere de conocer culturas, interpretar palabras dentro de un contexto, asumir faltas de ortografía y a partir de ello determinar diferentes reglas que es a lo que el usuario se quiere referir, etc., en pocas palabras es un proyecto que requería de mucho tiempo y de algunos conceptos extra clase. Analizando el proyecto de chatbot, este nos pareció interesante ya que este es un programa informático con el que es posible mantener una conversación, tanto si queremos pedirle algún tipo de información o que lleve a cabo una acción. Para la elaboración de este proyecto el profesor dijo que teníamos poco tiempo para su elaboración ya que en un semestre no lo íbamos a poder terminar a menos que obtuviéramos ciertos programas ya elaborados, pero ese no era el objetivo, ya que el profesor quería que nosotros empezáramos el proyecto desde cero, considerando la opinión del profesor opinamos que la dificultad era muy alta para elaborar este proyecto en un semestre teniendo más materias por las cuales se debe de poner atención al igual que en la que estamos elaborando el proyecto, así que se llegó a la conclusión que los tres proyectos desde nuestro punto de vista valían la pena realizarlos, los consideramos muy interesantes para implementarlo en la materia.

Dado que nuestras opciones fueron aprobadas por el profesor, pero tomando en cuenta el factor de tiempo, el profesor nos sugirió realizar el más óptimo para realizarlo durante el semestre y el que se eligió fue el de “Resumen de textos”.

### Problema/Pregunta (¿Qué es lo que pretende responder con este trabajo?)

Hoy en día existen muchas actividades que hacer, la sociedad cada vez trata de realizar todo de una forma más práctica. Conforme pasa el tiempo el internet de las cosas (IoT) y todo lo que se refiere a TIC ha ido entrando más en nuestros hogares, esto ha provocado que la mayoría de las personas quieran realizar sus actividades de una manera más fácil y automática por esta razón ha surgido este proyecto el “El resumidor automático”. Lo que se pretende es precisamente eso, es permitirles a las personas realizar sus resúmenes de una manera más fácil, práctica y en poco tiempo, sin tener que escribir a mano y sin invertir mucho tiempo.

### Solución Propuesta

Para la solución a nuestro proyecto estuvimos investigando en varias fuentes por lo que encontramos varias opciones para la solución. La opción que nos pareció como la mejor es la siguiente.

En lo que primero que se pensó fue en el lenguaje de programación, estuvimos investigando y “Python” fue el elegido ya que cuenta con Bibliotecas útiles como NLTK, esta sirve para el análisis de procesamiento de lenguaje natural (PLN) por lo cual fue de suma importancia para realizar el proyecto.

Otro punto que se tomó en cuenta fue el proceso para la elaboración del resumen paso a paso, el siguiente procedimiento es el que se tomó en cuenta para la solución y llegar al resumen del texto.

1.-Elegir el texto

Este paso consiste en elegir dos opciones un texto que el usuario quiera o que tenga escrito en computadora, y la segunda es elegir un texto que se encuentre en una página web.

Para la explicación de la solución propuesta pondremos un ejemplo de un párrafo de un discurso del famoso Denzel Washington en los premios de Imagen NNAACP.

**“Entonces, sigue trabajando. Sigue esforzándote. Nunca te rindas. Cáete siete veces, levántate ocho. La facilidad es una mayor amenaza para el progreso que las dificultades. La facilidad es una mayor amenaza para el progreso que las dificultades. Entonces, sigue moviéndote, sigue creciendo, sigue aprendiendo. Nos vemos en el trabajo”.**

En el párrafo anterior básicamente lo que está diciendo el famoso es que está motivando a otros para que trabajen duro y nunca se rindan. Para resumir el párrafo anterior se utilizan las técnicas basadas en NLP mencionadas anteriormente, por las que para hacer el resumen se deben de realizar los siguientes pasos.

2.-Convertir párrafos a oraciones

Primero necesitamos convertir todo el párrafo en oraciones es dividir el párrafo cada vez que se encuentra un período. Entonces, si dividimos el párrafo en discusión en oraciones, obtenemos las siguientes oraciones:

* Así que sigue trabajando
* Sigue esforzándote
* Nunca te rindas
* Cáete siete veces, levántate ocho
* La facilidad es una mayor amenaza para el progreso que las dificultades
* La facilidad es una mayor amenaza para el progreso que las dificultades
* Así que sigue moviéndote, sigue creciendo, sigue aprendiendo.
* Nos vemos en el trabajo

3.-Preprocesamiento de texto

Después de convertir el párrafo en oraciones, debemos eliminar todos los caracteres especiales, detener palabras y números de todas las oraciones. Después del preprocesamiento, obtenemos las siguientes frases:

sigue trabajando

sigue esforzándote

nunca dar

caer siete veces conseguir ocho

aliviar la amenaza mayor progreso dificultades

aliviar la amenaza mayor progreso dificultades

sigue moviendo sigue creciendo sigue aprendiendo

ver trabajo

4.-Tokenizado las oraciones

Necesitamos tokenizar todas las oraciones para obtener todas las palabras que existen en las oraciones. Después de tokenizing las oraciones, obtenemos una lista de las siguientes palabras:

['mantener',

 'trabajando',

 'mantener',

 'esforzarse',

 'Nunca',

 'dar',

 'otoño',

 'Siete',

 'hora',

 'obtener',

 'ocho',

 'facilitar',

 'mayor',

 'amenaza',

 'Progreso',

 'privación',

 'facilitar',

 'mayor',

 'amenaza',

 'Progreso',

 'privación',

 'mantener',

 'Moviente',

 'mantener',

 'creciente',

 'mantener',

 'aprendizaje',

 'ver',

 'trabajo']

5.-Encontrar la frecuencia ponderada de ocurrencia

A continuación, necesitamos encontrar la frecuencia ponderada de las ocurrencias de todas las palabras. Podemos encontrar la frecuencia ponderada de cada palabra al dividir su frecuencia por la frecuencia de la palabra que aparece con mayor frecuencia. La siguiente tabla contiene las frecuencias ponderadas para cada palabra:

| Palabra | Frecuencia | Frecuencia ponderada |
| --- | --- | --- |
| facilitar | 2 | 0.40 |
| ocho | 1 | 0,20 |
| otoño | 1 | 0,20 |
| obtener | 1 | 0,20 |
| dar | 1 | 0,20 |
| mayor | 2 | 0.40 |
| creciente | 1 | 0,20 |
| privación | 2 | 0.40 |
| mantener | 5 | 1.00 |
| aprendizaje | 1 | 0,20 |
| emocionante | 1 | 0,20 |
| Nunca | 1 | 0,20 |
| Progreso | 2 | 0.40 |
| ver | 1 | 0,20 |
| Siete | 1 | 0,20 |
| esforzarse | 1 | 0,20 |
| amenaza | 2 | 0.40 |
| hora | 1 | 0,20 |
| trabajo | 1 | 0,20 |
| trabajando | 1 | 0,20 |

Dado que la palabra "mantener" tiene la frecuencia más alta de 5, por lo tanto, la frecuencia ponderada de todas las palabras se ha calculado al dividir su número de ocurrencias por 5.

6.-Reemplazar palabras por frecuencia ponderada en oraciones originales

El paso final es conectar la frecuencia ponderada en lugar de las palabras correspondientes en las oraciones originales y encontrar su suma. Es importante mencionar que la frecuencia ponderada de las palabras eliminadas durante el preprocesamiento (palabras de detención, puntuación, dígitos, etc.) será cero y, por lo tanto, no es necesario agregarlas, como se menciona a continuación:

| Frase | Suma de frecuencias ponderadas |
| --- | --- |
| Así que sigue trabajando | 1 + 0.20 = 1.20 |
| Sigue esforzándote | 1 + 0.20 = 1.20 |
| Nunca te rindas | 0.20 + 0.20 = 0.40 |
| Cáete siete veces, levántate ocho | 0.20 + 0.20 + 0.20 + 0.20 + 0.20 = 1.0 |
| La facilidad es una mayor amenaza para el progreso que las dificultades | 0.40 + 0.40 + 0.40 + 0.40 + 0.40 = 2.0 |
| La facilidad es una mayor amenaza para el progreso que las dificultades | 0.40 + 0.40 + 0.40 + 0.40 + 0.40 = 2.0 |
| Así que sigue moviéndote, sigue creciendo, sigue aprendiendo. | 1 + 0.20 + 1 + 0.20 + 1 + 0.20 = 3.60 |
| Nos vemos en el trabajo | 0.20 + 0.20 = 0.40 |

7.-Ordenar oraciones en orden descendente de la suma

El paso final es ordenar las oraciones en orden inverso de su suma. Las oraciones con las frecuencias más altas resumen el texto. Por ejemplo, mire la oración con la suma más alta de frecuencias ponderadas:

*Así que sigue moviéndote, sigue creciendo, sigue aprendiendo.*

Puedes juzgar fácilmente de qué se trata el párrafo. De manera similar, puede agregar la oración con la segunda suma más alta de frecuencias ponderadas para tener un resumen más informativo. Echa un vistazo a las siguientes oraciones:

*Entonces, sigue moviéndote, sigue creciendo, sigue aprendiendo. La facilidad es una mayor amenaza para el progreso que las dificultades.*

Estas dos oraciones dan un resumen bastante bueno de lo que se dijo en el párrafo.

# **II. Marco teórico**

### Antecedentes

Para cualquier problema planteado con el análisis y procesamiento de textos existe un campo amplio de información en el cual nos basamos para realizar nuestro trabajo y esperamos que con lo realizado los resultados obtenidos contribuyan y aporten a dicho tema. En el trabajo presenta se mostrarán diversos trabajos que nos permitirán o aportarán en mayor o menor medida recursos necesarios para el desarrollo de nuestra aplicación.

Uno de los principales documentos que consultamos fue la tesis titulada *Sistema Resumidor-Traductor Automático* del Ing. Jouse Antonio Careaga Moya publicada en el año 2012 por la UNAM, la cual nos permite reforzar la parte que el resumidor que construyamos debe contener las ideas principales y dar una idea completa del texto. Nos guiamos en este trabajo ya que emplea la librería NLTK la cual brinda la posibilidad de segmentar el texto en enunciados, separados por comas, puntos o saltos de línea, por lo tanto, podemos indicar que el trabajo citado nos orienta en cómo llevar a cabo el proceso de elaboración ya mencionado. [[1]](#footnote-1)

A lo largo de que el lector conozca más sobre el trabajo presentado se dará cuenta que el trabajo citado en el párrafo anterior es de gran ayuda para la solución de nuestro planteamiento, sin embargo, se analizaron otros trabajos como el de Pedro Luis Mateo, José Carlos González, Julio Villena y José Luis Martínez titulado *Un sistema para resumen automático de textos en castellano* en el cual se presenta las técnicas que deben ser aplicadas para detectar palabras clave, dar ponderaciones con el fin de no solo crear extracciones simples y sin coherencia ya que el fin de un resumen es mostrar la idea del texto basado en el original y no sólo copiar las frases idénticamente. El trabajo nos resulta útil e interesante ya que consideramos que nos permite evaluar si el resumen creado está cumpliendo con el concepto, si este está siendo analizado correctamente para crear finalmente un resumen de la mejor calidad, los cuales son puntos que hemos considerados importantes y necesarios en nuestro trabajo. [[2]](#footnote-2)

Consideramos que los dos trabajos citados son los más relevantes ya que contienen una amplia descripción y brindan una o más soluciones al problema planteado, además que coinciden con la solución propuesta.

### Trabajos relacionados. (Comparación con los trabajos y el nuestro)

Algunos de los trabajos relacionados son los mencionados en la parte de antecedentes, para el primero de ellos se implementa Python y algunas de las librerías que son útiles para la creación de un resumidor de textos, el segundo es un trabajo internacional que de igual manera nos permite tener una guía para dar solución a nuestro problema. En general como ya se mencionó en un inicio contamos con una amplia variedad de artículos y trabajos que nos proporcionan ideas y soluciones que debemos implementar en nuestro propio proyecto.

Además, tenemos como guía las ventajas que son definidas en todas las aplicaciones disponibles online o en aplicaciones, las cuales nos permiten a través de una comparativa identificar cuáles son los principales puntos con los que debe contar nuestro programa.

**Descripción.**

**-Resumen de textos automático.**

Un resumen de textos automático es aquel que se ejecutará por el usuario pero que no requerirá la intervención de este en la generación del documento final, dicho documento debe contener una idea concisa del documento original el cual será procesado por diferentes funciones del programa clasificando las ideas más relevantes y ordenándolas con el fin de generar un documento que sea lo más coherente posible. El hecho de generar un resumen de textos es una tarea que corresponde a la disciplina del procesamiento de lenguaje natural.

**-Procesamiento de Lenguaje Natural.**

Corresponde a una disciplina que emplea diversas técnicas con el fin de que el lenguaje natural, es decir el lenguaje hablado por los humanos sea procesado por las computadoras o una máquina, dicho problema no tiene una solución fácil ya que se requiere que el programador conozca tanto el lenguaje natural como el de la máquina con el fin de hacer que una computadora genere los procesos necesarios con el fin de obtener resultados que los humanos realizamos como el comparar textos, resumir textos o encontrar una idea principal.

**-Recuperación de la información.**

Indica el problema que se tiene con poder recuperar la información extraída por el documento en un inicio, con proporcionar la información solicitada y mostrarla de manera exitosa.

**-Técnicas para generar un resumen de textos.**

Podemos encontrarnos con diversas técnicas que van desde generarlo a partir de temas específicos, otro ejemplo es introducir un aprendizaje para que compare posteriormente los textos solicitados o bien los que se basan en análisis discursivos a partir de árboles, por ejemplo, en el trabajo presenta se pretende generar un análisis a partir de los tokens identificados a través de NLTK de Python. En general en una aplicación de resumen de textos veremos como punto clave el uso de la lingüística y sus conceptos la cual nos permitirá identificar los elementos necesarios en un texto y a partir de esto generar un resumen lo más adecuado y completo posible.

# **III. Método experimental**

### Datos ¿Cómo se obtuvieron?

### Descripción de su método

### Descripción del experimento (Responder a la pregunta)

### Presentación de resultados

# **IV. Conclusión**

### Discusión de resultados

### Descripción de observaciones

### Trabajo futuro

**Buendía Valdivia María Angélica**

**Isabel Gómez Yareli Elizabeth**

**Zagoya Medallo Roberto Uriel**

# **V. Referencias**

# <https://redhistoria.com/historia-y-evolucion-del-resumen-de-textos/>

<https://www.instintoprogramador.com.mx/2019/07/resumen-de-texto-con-nltk-en-python.html?m=1>

1. Careaga, J.2012. *Sistema Resumidor-Traductor Automático. México, CDMX. UNAM. Consultado 27/03/2020 en:* [*http://www.ptolomeo.unam.mx:8080/xmlui/bitstream/handle/132.248.52.100/2640/Tesis.pdf?sequence=1*](http://www.ptolomeo.unam.mx:8080/xmlui/bitstream/handle/132.248.52.100/2640/Tesis.pdf?sequence=1) [↑](#footnote-ref-1)
2. Mateo, P. González, J.Villena, J. Martínez, J. *Un sistema para resumen automático de textos en castellano. Centro de Empresas la Arboleda. Madrid. Consultado el 20/03/2020 en:* [*http://rua.ua.es/dspace/handle/10045/1486*](http://rua.ua.es/dspace/handle/10045/1486) [↑](#footnote-ref-2)