UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

UNIDAD ACADÉMICA PROFESIONAL TIANGUISTENCO

"Generación automática de resúmenes de múltiples documentos utilizando secuencias frecuentes maximales y método de grafos"

TESIS

QUE PARA LA OBTENER EL TÍTULO DE
MAESTRA EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:
SELENE ARAI VILCHIS SEPÚLVEDA

TUTORA ACADÉMICA:
DRA. YULIA NIKOLAEVNA LEDENEVA

TUTORES ADJUNTOS:
DR. RENÉ ARNULFO GARCÍA HERNÁNDEZ
M. EN C.C. JOSÉ RAFAEL CRUZ REYES

TIANGUISTENCO, ESTADO DE MÉXICO ENERO 2022

Resumen

El crecimiento exponencial de internet ha provocado un bombardeo de información que se produce día a día aumentando de manera exponencial.

La información masiva se ha vuelto un problema de sobrecarga de información al momento de realizar una búsqueda de información específica, lo cual ha provocado que las ciencias computacionales se vean involucradas en la búsqueda de una solución.

La Generación Automática de Resúmenes de Texto (GART) es una tarea del Procesamiento del Lenguaje Natural (PLN) que busca contrarrestar los efectos negativos de la sobrecarga de información.

Actualmente existen diferentes métodos del estado del arte para la GART basados en una arquitectura de tres etapas:

1. Identificación de Tópicos.
2. Transformación o interpretación.
3. Síntesis o generación del resumen.

Entre los métodos del estado del arte se encontró un método que a diferencia de los otros propone una cuarta etapa. La cuarta etapa busca darle un valor a cada término de las oraciones. El método propuesto por (Ledeneva y García-Hernández, 2017) demostró dar buenos resultados para la tarea Generación Automática de Resúmenes de Texto de Un solo documento (GART-1).

Con referencia a los resultados obtenidos del método de (Ledeneva y García-Hernández, 2017) en este trabajo se propone ajustar los parámetros en las diferentes etapas y adaptar el método para la tarea de Generación Automática de Resúmenes de Texto de Múltiples documentos (GART-M).

En el método propuesto se optó por la extracción de las Secuencias Frecuentes Maximales (SFM's) para ser empleadas como modelo de texto y la utilización de un método basado en grafos para realizar el pesado de las oraciones.

El corpus empleado fue DUC-02, el cual está conformado por 59 colecciones de documentos de noticias.

La evaluación de los resúmenes se hizo con el sistema ROUGE-N, el cual permite comprar los resúmenes generados a partir del método con los resúmenes generados por un humano.

Los resultados obtenidos de los experimentos realizados se dividieron en tres etapas. En la primera etapa se buscó la mejor configuración del método. En la segunda etapa se buscó

probar la importancia de la longitud de las SFM's. En la tercera etapa de busco emplear una nueva configuración para la selección de oraciones.

Los resultados obtenidos por el método propuesto se compararon con otros métodos del estado del arte y las heurísticas. Los resultados obtenidos con el método propuesto logran superar las heurísticas y métodos del estado del arte.

Contenido

Figuras

Tablas

Ecuaciones

CAPÍTULO 1

Introducción

A lo largo del tiempo el ser humano ha evolucionado y a su paso ha evolucionado la gestión de la información/conocimiento (González Suárez, 2004). Hace algunos años gran parte del conocimiento se encontraba en documentos escritos (Sánchez Arteche et al., 2012) pero con la evolución de las tecnologías y la llegada del internet en la actualidad contamos con grandes volúmenes de información en todo el mundo en diferentes idiomas (Vilchis Sepúlveda & Ledeneva, 2019). Durante las últimas décadas se ha presentado un crecimiento acelerado de manera exponencial respecto a la información publicada en todo el mundo. Según el Banco Mundial de Datos, hasta el 2018 reportó 26 mil millones de artículos publicados de ciencia y tecnología. La sobrecarga de información ha traído consigo la dificultad de encontrar información específica de un tema. En la búsqueda de información de algún tema se cuenta con una gran cantidad de documentos encontrados lo que complica revisar cada documento ya que esto requiere una gran cantidad de tiempo, desgaste y esfuerzo humano (Alvarado Bolaños, 2017; Hernández Maya, 2018; Matias Mendoza, 2016; Rojas Sánchez, 2016).

Afortunadamente las ciencias computaciones han abordado el problema de sobrecarga de información a través del Procesamiento del Lenguaje Natural (PLN).

El PLN es una disciplina de la Inteligencia Artificial (IA) que se encarga de la investigación de mecanismos computacionales para la comunicación entre humanos y computadoras mediante el uso de lenguajes naturales (Cortez Vásquez et al., 209 C.E.; Martín Mateos & Ruiz Reina, 2013; Pardo et al., 2012).

El PLN ha abordado la dificultad de sobrecarga de información a través de la tarea de Generación Automática de Resúmenes de Texto (GART) creando; técnicas, herramientas y sistemas para el manejo de la información (Vázquez et al., 2019).

La GART consiste en la extracción de la información más relevante contenida en uno o varios documentos fuente para ser reescrita en una versión más corta sin perder el contexto del documento (Gillick et al., 2009; Ledeneva & García-Hernández, 2017; Neri Mendoza, 2019; Simón et al., 2018).

Existen dos tipos de resúmenes: de tipo extractivo y abstractivo (Ledeneva, García-Hernández, Gelbukh, 2014; Hovy 2009). Los resúmenes abstractivos son textos que describen el contenido y el sentido de un documento original. Consiste en entender el contexto de un documento y después reescribirlo sin perder el sentido del texto utilizando nuevos conceptos en un número menor de palabras, sin perder el contexto del documento original (Ledeneva, García Hernández, 2014; Hovy, 1998). Los resúmenes extractivos consisten en reducir el contenido textual de un documento o un grupo de documentos mediante la selección de un conjunto de frases u oraciones del texto original (Rojas, Ledeneva, García-Hernández, 2018; Hovy, 1998). Generalmente, las personas realizan resúmenes de tipo abstractivo, mientras que la mayoría de las herramientas comerciales y los métodos del estado del arte generan resúmenes de forma extractiva.

La GART se divide en dos tareas: la Generación Automática de Resúmenes de Texto de un Solo Documento (GART-1) y la Generación Automática de Resúmenes de Texto de Múltiples Documentos (GART-M).

Hoy en día, la GART-M ha ganado interés con el desarrollo de talleres de evaluación, como: la Conferencia de Comprensión de Documentos, por sus siglas en inglés DUC y la Conferencia de Análisis de Texto por sus siglas en inglés TAC.

Los resúmenes generados a partir de múltiples documentos presentan un mayor grado de dificultad a diferencia de los generaos de un solo documento. Los resúmenes de múltiples documentos presentan problemas como: - Mayor grado de redundancia, los documentos que se emplean para estos resúmenes pueden estar compuestos de diferentes subtemas. - Dimensión temporal, esto aplica cuando los documentos para generar el resumen son noticias ya que siguen una línea de tiempo sobre la que se desarrollan. - Relación de comprensión, se refiere al tamaño del resumen respecto al número de documentos. Cuando el conjunto de documentos es muy grande la dificultad de realizar el resumen aumenta debido a que se cuenta con un mayor número de oraciones y cada

oración representa una idea diferente (Mcdonald, 2007; Neri Mendoza, 2019; Villatoro-Tello et al., 2009).

A lo largo de los años se han realizado múltiples investigaciones para el desarrollo de herramientas comerciales, heurísticas, algoritmos de optimización, algoritmos evolutivos y diversos métodos para GART-M.

Las heurísticas son procedimientos de búsqueda que no garantizan la obtención de buenos resultados, se trata de una búsqueda de técnicas inspiradas en la naturaleza para generar conocimiento como:

- *baseline:* selecciona las primeras oraciones de un documento de texto, se tiene la idea que en las primeras oraciones se encuentra la información más importante de un documento.
- *baseline-aleatorio:* está inspirada en tomar oraciones aleatoriamente del documento original y reescribirlas para formar un resumen (Ledeneva & García-Hernández, 2017; Marino Cuéllar Chacón et al., 2018; Neri Mendoza, 2019).

Por otra parte, se han desarrollado métodos más sofisticados. Algunos de los métodos se encuentran basados en mecanismos biológicos de la naturaleza, la evolución de las especies y modelos matemáticos, sin embargo, todos estos métodos se encuentran basados en una arquitectura tradicional conformada por diferentes etapas.

La arquitectura para la GART desde el punto de vista clásico se encuentra conformada por tres etapas:

1. Identificación de tópicos, se seleccionan las unidades que contaran como entrada.
2. Transformación o interpretación, se selecciona la información más destacada de diferentes partes de la fuente.
3. Síntesis o generación del resumen, combina fragmentos extraídos de la fuente y descarta el material que no es importante (Hovy, 2003; Cardoso & Pérez-Abelleira, 2013).

Ledeneva y García Hernández en su trabajo (Ledeneva & García-Hernández, 2017) describen un nuevo método por etapas aplicado para la tarea GART-1 en el que proponen una cuarta etapa. Las etapas que proponen son:

- Selección de términos, optaron por utilizar las Secuencias Frecuentes Maximales (SFM's) propuestas por (Ahonen-Myka, 1999) para ser utilizadas como un modelo de representación de texto.
- Pesado de términos, optaron por darle un valor a las SFM's extraídas de los documentos. Los pesos se dieron de acuerdo con el tamaño de la SFM y a la cantidad de ocasiones en que apareció la SFM en el texto.
- Pesado de oraciones, propusieron el uso del algoritmo TextRank para darle un valor a cada oración de acuerdo con las SFM's que contenga cada oración asignándole un valor de utilidad a cada oración.

- Selección de oraciones, seleccionaron las oraciones con mayor valor hasta completar el tamaño deseado del resumen.

El pesado de términos fue propuesto con el fin de asignarle un valor a los términos obtenidos y después darles un valor a las oraciones para seleccionar las mejores y generar el resumen. A diferencia del punto de vista clásico que solo identifica los tópicos que considera importantes.

En la propuesta realizada por Ledeneva y García Hernández se observó que la implementación de la cuarta etapa permitió obtener mejores resultados en comparación de otros métodos del estado del arte para la tarea GART-1. De acuerdo con los resultados reportados, concluyeron que su método generaba buenos resúmenes. Las SFM's obtenían la información más importante del documento y en la etapa de pesado de oraciones el algoritmo TextRank permitió seleccionar las oraciones que contenían los tópicos más importantes.

1.1 Problema

Se desconoce si ¿es posible que al aplicar el método de cuatro etapas podría mejorar la calidad de los resúmenes de múltiples documentos?

1.2 Hipótesis

Se tiene la idea que las SFM's obtienen la información que expresa las ideas más importantes en un texto ya que aparecen en repetidas ocasiones y el algoritmo de TextRank permite ordenar las oraciones a modo de recomendación. Con base en lo anterior se formuló la siguiente hipótesis:

Es posible adaptar las cuatro etapas de la GART para generar resúmenes de múltiples documentos. Si se extraen las SFM's por colección se obtendrán los tópicos más importantes en la etapa de selección de términos. En la etapa de ponderación de términos se pueden utilizar las características de las SFM's para darle un valor a cada término. En el pesado de oraciones se puede utilizar el algoritmo de TextRank para darle un valor a las oraciones más importantes de cada colección y en la etapa de selección de oraciones elegir las oraciones que contengan los tópicos más importantes de la colección de documentos, descartando las oraciones menos relevantes.

1.3 Objetivo general

Generar los resúmenes de texto que contengan la información más relevante extrayendo las SFM's. Usar el algoritmo de ranqueo para obtener las oraciones que contengan la información más relevante de los documentos y generar resúmenes de texto de múltiples documentos.

1.3.1 Objetivos específicos

- Analizar el método a implementar.
- Analizar el conjunto de datos.
- Extraer las SFM's de la colección de documentos
- Configurar el algoritmo de grafos para usarlo en la etapa de pesado de términos para múltiples documentos.
- Realizar pruebas con diferentes configuraciones de selección de términos, pesado de términos y selección de oraciones.
- Realizar la evaluación de los resúmenes generados por el método propuesto.
- Realizar una comparación de los resultados obtenidos contra otros métodos del estado del arte para la GART-M.

1.4 Estructura de la tesis

En este capítulo se abordan algunos elementos que guían al presente trabajo de investigación, lo que permite plantear el problema que se aborda en este trabajo.

En el capítulo 2 se definen conceptos fundamentales que describen el proceso de la generación automática de resuenes de texto, cuáles son las disciplinas que abordan esta tarea, la clasificación de los resúmenes de texto y algunos conceptos fundamentales para la comprensión del presente trabajo.

En el capítulo 3 se presentan algunos de los trabajos que los investigadores han propuesto para la generación automática de resúmenes de texto de múltiples documentos y el método en el que se basó para proponer esta investigación.

En el capítulo 4 se describe a detalle la metodología empleada para llevar a cabo este trabajo, así como la descripción del desarrollo de cada una de las etapas que la conforman.

En el capítulo 5 se describen todas las pruebas que se realizaron con el método propuesto, así como los diversos ajustes de parámetros que se fueron realizando a lo largo de la investigación. Así mismo se describen los resultados obtenidos por las pruebas. Los resultados se describen en tablas que detallan las pruebas realizadas y los resultados obtenidos, del mismo modo se realiza una comparación del método propuesto con otros métodos encontrados en el estado del arte.

CAPÍTULO 2
Marco teórico

En este capítulo se aborda la conceptualización de términos empleados para esta investigación, a continuación, se presentan algunos de los conceptos más importantes:

2.1 Inteligencia Artificial

Según autores de los años 80's y 90's como (Charniak, 1985; Kurzweil, 1990) mencionaban que la Inteligencia Artificial (IA) es el arte de desarrollar máquinas con la capacidad de realizar funciones propias de un ser humano, (Haugeland, 1998) lo definía como máquinas con mentes, mientras que (Knight, 1991) mencionaba que las máquinas tendrían la capacidad de realizar tareas que hasta ese momento los humanos hacían mejor, sin embargo, hoy en día (INCyTu 2018) reafirma la idea de los autores al mencionar que el desarrollo de la IA se ha acelerado gracias a una mayor disponibilidad de datos, recursos tecnológicos y financieros, mencionando que actualmente las aplicaciones de la IA van desde el reconocimiento imágenes o videos de objetos y personas hasta el habla y el tratamiento automático de textos, mencionando también el diagnóstico y tratamiento de enfermedades así como la toma de decisiones.

La IA ha servido para facilitar las tareas o acciones complejas dentro de un sistema, mismas que para un ser humano resultan complejas o que requieren de una gran cantidad de tiempo, sin mencionar el desgaste físico que estas tareas provocan, a lo que (Horgan, 2004) dice que, la IA construye procesos que al ser ejecutados sobre una arquitectura física producen acciones o resultados que maximizan una medida de rendimiento determinada.

2.2 Lingüística computacional

La lingüística computacional es considerada un campo científico interdisciplinar generada a partir de la lingüística y la computación. Esta área lleva más de 50 años de investigación y desarrollo, cuyo objetivo es la elaboración de modelos computacionales que provienen de la inteligencia artificial. El objetivo de los modelos es reproducir uno o más aspectos el lenguaje humano como lo son el procesamiento del lenguaje natural y el reconocimiento de voz (Domínguez Burgos, 2002; Guinovart, 1998; Sidorov, 2013; Sierra Martínez & Cuétara Priede, 2015).

2.3 Procesamiento del lenguaje natural

El PLN es una disciplina encargada del diseño e implementación de los elementos software necesarios para el tratamiento computacional del lenguaje natural. (Alonso et al., 2012). El PLN se aborda desde la lingüística aplicada a la inteligencia artificial cuyo objetivo principal es la realización de estudios informáticos que simulen la capacidad humana de hablar y entender (Cortez Vásquez et al., 209 C.E.). Para un computador el conocimiento humano no pasa de ser un simple archivo o una dirección de memoria física, de esta manera lo que es conocimiento para los seres humanos para las máquinas es una secuencia de señales digitales(Torres & Medina, 2013).

2.4 Resumen

Un resumen es un texto que transmite la información de otro documento de manera abreviada. Hacer un resumen es una técnica de estudio fundamental: exige una lectura atenta y comprensiva para identificar la información más importante incluida en el documento original. Para la creación de un resumen elaborado por un humano, existen múltiples técnicas, por lo cual no existe un método establecido para poder realizarlos (Rojas Sánchez, 2016), sin embargo, el principal objetivo de un resumen es extraer las características más importantes de un documento y plasmarlo en un nuevo texto más pequeño (Mendoza, 2013), este puede ir situado al inicio o al final del documento original que indica al lector las ideas más relevantes de un texto de origen (Rojas Simón, 2017).

2.5 Generación automática de resúmenes de texto

La generación automática de resúmenes de texto surge a partir de la necesidad de utilizar un método que otorgue a los usuarios el acceso a la información que se considere más importante de un texto, sin tener que leer el documento completo (Mendoza, 2013). La generación automática de resúmenes de texto es una tarea del área de procesamiento de lenguaje natural, que tiene por objetivo resumir el contenido de un documento conservando la información más importante de un texto (Mendoza, 2015), el texto del documento está dividido en fragmentos (oraciones, párrafos, etc.), los fragmento elegidos no sufren alguna modificación respecto del documento original y son colocados en el nuevo documento, en el mismo orden de su selección, formando así el resumen (Rojas Sánchez, 2016).

2.6 Clasificación de resúmenes

Existen diferentes tipos de resúmenes, estos se clasifican de acuerdo con su entrada, según su función y de acuerdo con su salida, a su vez estos contienen otras subclasificaciones como se muestra en la fig. 1.

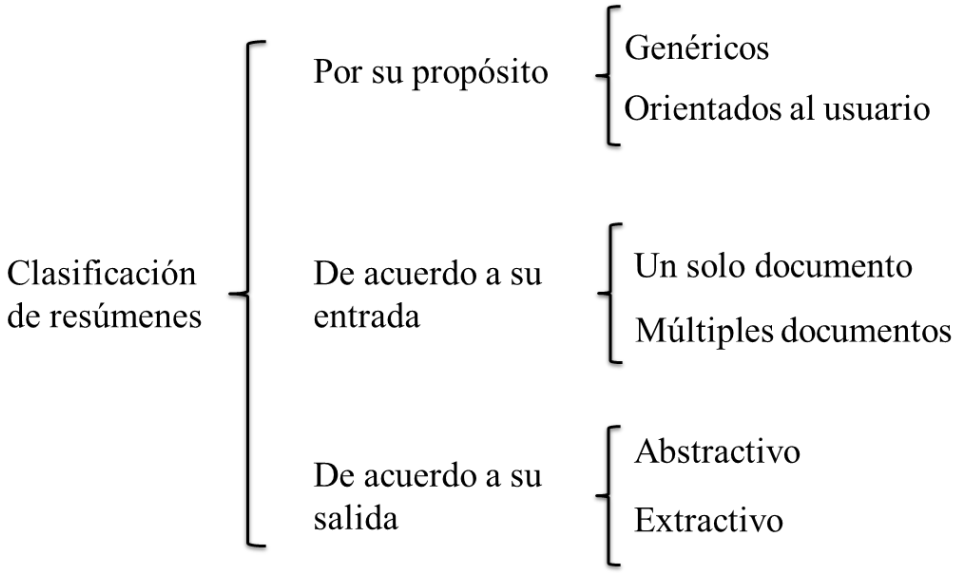

Fig. 1 Clasificación de los resúmenes (Matías, 2013; Rojas Simón, 2017).

2.6.1 Por su propósito

Este tipo se refiere a la audiencia a la que va dirigido el resumen (Acero Et Al., 2001; Anacona, 2015; Cardoso & Pérez-Abelleira, 2013; Elian & Becerra, 2012; Rojas Simón, 2017):

- **Genéricos:** toman los temas principales del documento ya que están destinados a una comunidad amplia de lectores.

- **Orientados al usuario:** pretenden generar un resumen de acuerdo con los intereses particulares de un usuario (temas de interés o necesidades de información), está enmarcado en un paradigma más ajustado a la recuperación de información.

2.6.2 De acuerdo con su entrada

Este tipo se refiere a la cantidad de documentos que se procesan para la generación del resumen(Anacona, 2015; Elian & Becerra, 2012; Neri Mendoza, 2019; Rojas Simón, 2017):

- **Un solo documento:** se realizan a partir de un solo documento como se muestra en la Fig. 2.

Fig. 2 Generación Automática de Resúmenes de un Solo Documento, (GART-1).

- **Múltiples documentos:** considera como entrada un conjunto de documentos para generar un resumen como se muestra en la Fig. 3, que no pierda las ideas principales de cada documento de entrada, tomando en cuenta el multilenguaje y el tipo de documento sobre el cual se hace el resumen (artículo científico, noticia, blog, etc.), este tipo de resumen tiene por objetivo evitar la redundancia que pueda surgir entre cada documento:

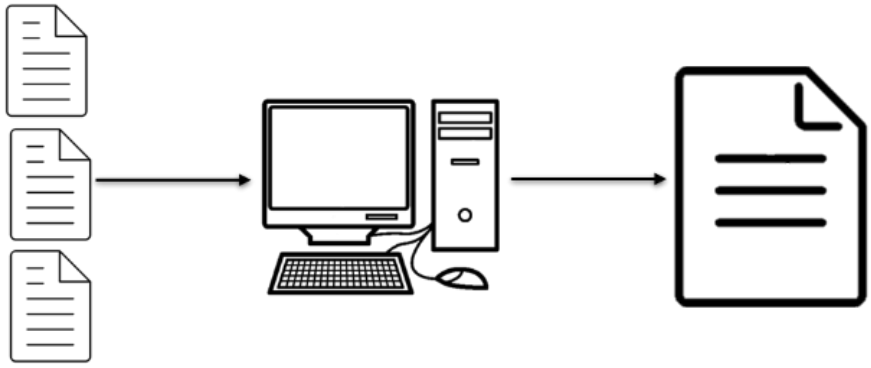

Fig. 3 Generación Automática de Resúmenes de Múltiples Documentos (GART-M).

2.6.3 De acuerdo con su salida

Basado de acuerdo a su estrategia de condensación (Anacona, 2015; Ledeneva & García-Hernández, 2017; Neri Mendoza, 2019; Rojas Simón et al., 2018; Valderrama Vilca, 2017; Vilchis Sepúlveda & Ledeneva, 2019):

- **Extractivo:** consiste en tomar fragmentos del documento original, secuencias de palabras, oraciones o párrafos de los documentos fuente y sin ser modificados, en el mismo orden son colocadas en el nuevo documento.
- **Abstractivo:** estos resúmenes son de naturaleza arbitraria que utilizan métodos lingüísticos y semánticos para comprender el texto fuente, esto consiste en entender el contexto de un documento y después reescribirlo sin perder el sentido del texto utilizando nuevos conceptos en un número menor de palabras.

2.7 Modelo de GART por etapas

A lo largo de los años se han realizado múltiples investigaciones para la generación de resúmenes automáticos, uno de los métodos que ha demostrado dar buenos resultados, es método por etapas propuesto por (Ledeneva & García-Hernández, 2017), su método consta de cuatro etapas: Selección de términos, pesado de términos, pesado de oraciones y selección de oraciones, como se muestra en la fig. 4.

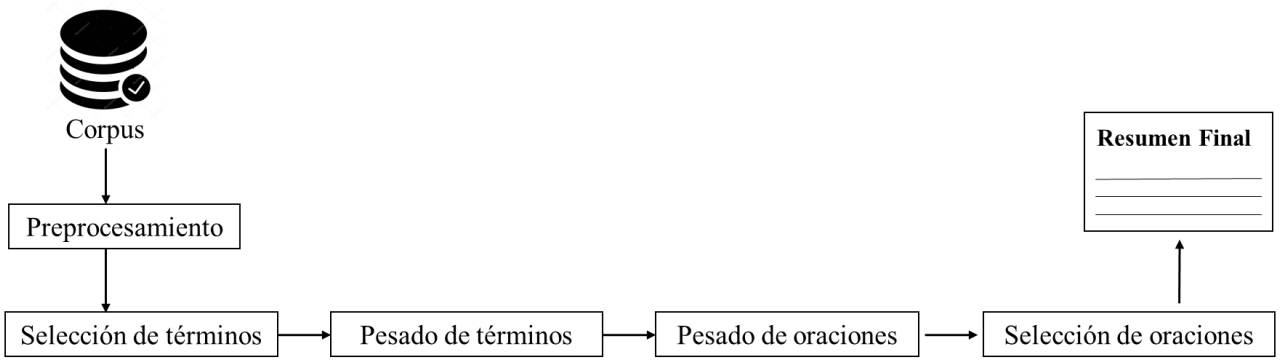

Fig. 4 Método por etapas propuesto por (Ledeneva y García Hernández 2017).

2.7.1 Selección de términos

En esta etapa se debe decidir qué unidades contaran como términos, por ejemplo, palabras, n gramas, oraciones o SFM, a continuación, se describen algunos de los modelos de texto:

- **N-grama:** Un n-grama es una subsecuencia de n elementos consecutivos en una secuencia dada. Los n-gramas se construyen con base en distintos tipos de elementos como fonemas, silabas, letras o palabras, el maño de los n-gramas se pueden definir de diferentes tamaños, cuando son de tamaño uno se llaman n-gramas, de tamaño dos se llaman bígamas, de tamaño tres trigramas, etc., el tamaño de los n-gramas se da dependiendo del problema que

se quiera resolver (Alguliev et al., 2013; Rojas Simón, 2017; Ledeneva & García Hernández, 2017; Matías Mendoza, 2013; Neri Mendoza, 2019).

- **Syntactic n-gram:** Este modelo de texto es similar a los n-gramas, sin embargo, los sn-gramas se basan en relaciones sintácticas de palabras, por lo que cada palabra está vinculada a sus vecinos "reales" (Sidorov, 2019; Stamatatos, Gelbuckh & Chanona Hernández, 2013; Neri Mendoza, 2019).

- **Frases:** Son una secuencia de palabras que representan las ideas fundamentales expresadas por el autor, las técnicas de selección y extracción de frases tienen independencia del dominio y del idioma (Hernández Casimiro, 2016; Ortiz et al., 2010; Neri Mendoza, 2019).

- **Secuencias Frecuencias Maximales (SFM):** Es un conjunto de secuencias frecuentes que no están contenidas en otras secuencias frecuentes, una secuencia frecuente es una secuencia de palabras que aparecen en el mismo orden secuencial y de manera repetida. Algunas de las ventajas que presentan las SFM con respecto a los n-gramas, es que estas no requieren una configuración humana, es decir su tamaño lo determina el contenido del texto, además de ser independientes del lenguaje, a diferencia de las frases (García Hernández 2007, Ledeneva & García Hernández, 2017, Matías Mendoza 2013; Neri Mendoza 2019; Rojas Simón, 2017; Villatoro-Tello, 2007).

2.7.2 Pesado de términos

Se trata de un proceso de ponderación de los términos individuales con respecto al contenido del documento, a continuación, se describen algunos tipos de pesados:

- **Frecuencia de términos:** Las palabras con mayor frecuencia en un documento indican cuales son los tópicos más importantes y asigna un peso a cada oración del documento en función de su relevancia (Cardoso & Pérez, 2013; Conroy & Davis, 2018; Ledeneva & García-Hernández, 2017, Lin & Hovy, 2007; Loret et al., 2008; Neri-Mendoza, 2019).

- **Frecuencia inversa:** La frecuencia inversa se define como el número de documentos en los que aparece el término entre el número de documentos que tiene la colección. Esto significa que cuando los términos aparecen en menos documentos son más importantes, lo que permite caracterizar de mejor manera a cada documento, este tipo de pesado evalúa la distribución de los términos en el documento (Ledeneva & García Hernández, 2017; Neri Mendoza, 2019).

- **Entropía:** Es la relación que existe entre el número de veces que aparece un término en una oración, el número de veces que aparece un término en todos los documentos y su distribución (Coroy & Davis, 2018, Neri Mendoza, 2019).

- **Frecuencia de término-Frecuencia inversa de documento (tf-idf**): Es común que la frecuencia de los términos (tf) y la frecuencia inversa del documento (idf) se utilicen juntas, el fin es determinar la relevancia de cada término, considerando tanto la importancia que tiene el término en la colección de documentos como su importancia en ese documento (Salton & Buckley, 1988; Ledeneva & García Hernández, 2017; Neri Mendoza, 2019)

- **Pesado interpolado:** Consiste en otorgar mayor peso a las primeras 100 palabras (Takamura & Okumura, 2009)

- **Pesado booleano:** Es la forma más fácil de pesar un término. Se asigna un valor de 1 si el término aparece en la oración y un valor de 0 si no aparece (Ledeneva & García Hernández, 2017; Takamura & Okumura, 2009).

2.7.3 Pesado de oraciones

Es el proceso de asignación de una medida numérica de utilidad de la oración decir la suma de los pesos de utilidad de los términos de los que se compone una oración, a continuación, se describen algunos tipos de pesado de oraciones:

- **Suma de los pesos de utilidad:** Se realiza una suma de todos los términos que conforman la oración y la sumatoria total es el valor que se le asigna a la oración (Ledeneva & García Hernández 2017).

- **Ranqueo:** Se genera un grafo con todas las oraciones de la colección en donde, las oraciones se consideran nodos y las aristas son la relación que existe entre las oraciones, después se itera el algoritmo hasta la convergencia y a cada nodo se le asigna un peso realizando una sumatoria de todos los pesos, de acuerdo con el peso que contengan las aristas adyacentes al nodo (Ledeneva & García Hernández, 2017; Mihalcea, 2006).

2.7.4 Selección de oraciones

En esta etapa se asigna un puntaje a cada oración de los documentos, teniendo en cuenta modelos que determinan la relevancia de las oraciones, las oraciones con altos puntajes son elegidas para formar parte del resumen, a continuación, se describen algunos de selección de oraciones:

- **Similitud con el título:** En este modelo se plantea, que las palabras del título son un factor importante en el proceso del resumen, asignando mayor importancia a las oraciones que incluyen términos que aparecen en el título (Hirao et al., 2003; Vázquez et al., 2018).

- **Similitud con otras oraciones:** (O centralidad de oración). A partir de una oración denominada oración central, se concede importancia a las demás oraciones del documento que contengan superposición de palabras con esta (Fattah & Ren, 2008).

- **Longitud de las oraciones:** En este modelo se tiene la suposición de que la longitud de una oración puede indicar si es relevante para el resumen final o no. En el estado del arte se han propuesto diversas formas de calcularlo. Por ejemplo, este modelo se puede emplear para penalizar oraciones que son demasiado cortas, ya que no se espera que estas oraciones pertenezcan al resumen. Sin embargo, la relevancia obtenida a través de este modelo puede ser engañosa, debido a que generaliza si vale la pena incluir una oración en un resumen simplemente en función de si es corta o larga (Cao, Ziqiang; Wei, 2015; Fattah & Ren, 2008; Mutlu et al., 2019; Vázquez et al., 2018)

- **Reducción de redundancia:** Se espera que la información redundante o duplicada contenida en el resumen generado se minimice (Alguliev, Aliguliyev, & Hajirahimova, 2012; Alguliev et al., 2013; Alguliyev et al., 2018b; Erkan & Radev, 2004; Jung, Datta, & Segev, 2017; Mcdonald, 2007; Saleh, Kadhim, & Attea, 2015; Sanchez-Gomez, Vega-Rodríguez, & Pérez, 2018).

- **Posición de las oraciones:** La hipótesis aquí es que las frases importantes están situadas en ubicaciones particulares, que dependen del género del texto. Las primeras oraciones se consideran un indicativo de una oración relevante, por ejemplo, en el dominio de noticias, las primeras oraciones se consideran indicativo de una oración relevante, comúnmente el primer párrafo contiene la información principal acerca de la noticia, mientras que el resto del texto posee detalles como información general sobre el evento. Por lo tanto, la selección de oraciones desde el inicio del texto es una estrategia razonable (Conroy et al., 2002; Kato, Matsushita, & Kando, 2007; C.-Y. Lin & Hovy, 2002, 2007; Matías, 2013; Moens, Angheluta, & Dumortier, 2005; Vázquez et al., 2018).

- **Cobertura:** La cobertura de contenido significa que el resumen generado debe cubrir todos los subtemas tanto como sea posible. Se refiere a la medida en que la información proporcionada en los documentos originales se incluye en el resumen generado (Alguliev et al., 2013; Angheluta, De Busser, & Moens, 2002; John & Wilscy, 2015; Jung et al., 2017; C.-Y. Lin, 2004b; Matías, 2013; Mendoza et al., 2014; Saleh et al., 2015; Sanchez-Gomez et al., 2018; Vázquez et al., 2018).

- **Identificación de entidades nombradas:** Este modelo se basa en la premisa de que la oración que contiene más sustantivos, o nombres propios tiene mayor probabilidad de ser una oración relevante (Cardoso & Pérez-Abelleira, 2013; Fattah & Ren, 2008; Hirao et al., 2003).

- **Identificación de números:** En este modelo, se considera relevante a la oración que contiene datos numéricos. Por consiguiente, es probable que esta oración aparezca en el resumen final (Fattah & Ren, 2008).

2.8 Secuencias Frecuentes Maximales

El proceso de descubrimiento de información en grandes volúmenes de datos resulta una tarea complicada para el ser humano, por lo que Ahonen en su trabajo (Ahonen-Myka, 1999) propone el uso de las Secuencia Frecuentes Maximales (SFM's), con el objetivo de extraer datos que se presentan frecuentemente, perseverando su orden secuencial dentro de la información.

Un documento de texto se encuentra descrito de manera secuencial mediante palabras, por lo que es posible descubrir conocimiento a partir del orden secuencial de las palabras. El texto puede ser representado en n-gramas. Un n-grama es una secuencia de *n* palabras. Se dice que un n-grama ocurre en un texto si las palabras aparecen en el texto en el mismo orden inmediatamente una tras otra. El tamaño del n-grama es dado según la tarea, cuando el n-grama es de tamaño uno se le conoce como grama, cuando es de tamaño dos se conoce como bigrama, cuando es de tamaño tres se le llama trigrama, de tamaño cuatro cuatrigrama y, así sucesivamente. Cuando un n-grama aparece con frecuencia en un texto también es llamado Secuencia Frecuente (SF), sin embargo, la cantidad de SF extraídas de un texto puede llegar a ser muy grande, por lo que únicamente se consideran las SFM's.

De acuerdo con (García Hernández, 2007) una SFM es un conjunto de SFs que no se encuentran contenidas en otras SFs. Una SFM se considera frecuente si ésta se encuentra en al menos un cierto número de documentos β (umbral de frecuencia), que se refiere al número mínimo de ocasiones en que aparece una SFM en una colección de documentos.

Cuando menor sea el umbral es mayor la cantidad de SFM's extraídas. A diferencia de los n-gramas las SFM's cuentan con una restricción de separación máxima en el texto, a esta restricción se le llama GAP. La restricción de GAP se refiere al número de palabras que pueden aparecen entre las palabras que son consideradas SFM's, por ejemplo, en las oraciones que se muestran en la figura 5 se observan dos SFM's con restricción de GAP.

A: he National Hurricane Center said a hurricane watch was on the Texas coast .
B: The government issued a hurricane watch for on the coast of northeast Mexico from Tampico north.

Fig. 5 Ejemplo de una colección de dos oraciones

En la fig. 5 se muestra un documento con dos oraciones en donde la extracción de las SFM's para las oraciones A y B se utilizó una restricción de GAP =1 donde se encontraron las SFM's *"a hurracane watch on the"* y *"the coast"*, en la oración A se encuentra la frase *a hurracane watchwas on the* donde la palabra *was* no pertenece a la SFM y en la oración B se encuentra la frase *a hurracane watchs for on the* donde la palabra *for* no pertenece a la SFM.

Una de las ventajas que presentan las SFM's es que a diferencia de los n-gramas estas no requieren de una configuración humana para determinar el número de gramas. El tamaño de las SFM's es determinado por el contenido del texto, permitiendo obtener la longitud y la frecuencia de las SFM's. Otra característica importante es que su extracción es independiente del lenguaje.
El uso de las SFM se ha dado en tareas como: determinación del autor de un texto, detección de tópicos, extracción de frases clave, recuperación de información, resúmenes de documentos, análisis de secuencias de ADN, entre otras.

2.9 Grafo

Un grafo es una estructura matemática representada simbólicamente de los elementos constituidos de un sistema en el que exista una relación binaria entre ciertos objetos (Villalpando Becerra, García Sandoval, 2014) (Johnsonbaugh, 2005) (Jara, 2005).
Un grafo G= (V, E) es una estructura combinatoria constituida por un par de conjuntos.
Un conjunto V de vértices o nodos, donde $V \neq \{\emptyset\}$
Un conjunto E de aristas o arcos, parejas de elementos de V
V= {A, B, C, D}
E={(AB), (BD), (DC), (CA)}

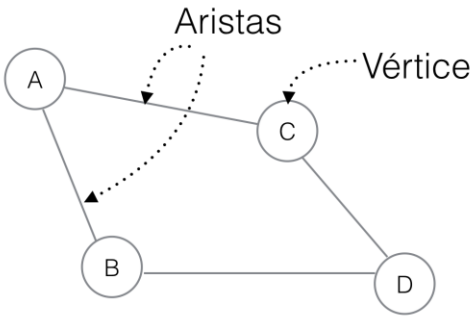

Fig. 6 Ejemplo de un grafo no dirigido.

Existen diversos tipos de grafos, los cuales se presentan a continuación.

2.9.1 Topología de grafos

Grafo no dirigido: En un grafo no dirigido el par de vértices que representa un arco no está ordenado. Por lo tanto, los pares (A, B) y (B, A) representan el mismo arco, como se muestra en la fig. 6.

Grafo dirigido: En un grafo dirigido cada arco está representado por un par ordenado de vértices, de forma que (A, B) y (B, A) representan dos arcos diferentes, como se muestra en la fig. 7.

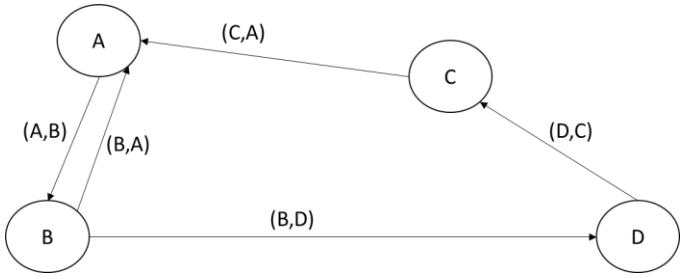

Fig. 7 Ejemplo de un grafo dirigido.

Grafo etiquetado: Un grafo cuyos nodos tienen asociada información. Un ejemplo común es que los nodos contengan los nombres de ciudades en un mapa, como se muestra en la fig. 8.

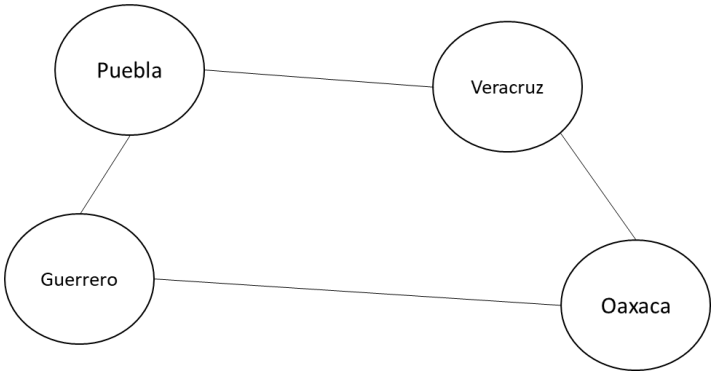

Fig. 8 Ejemplo de un grafo etiquetado.

Grafo ponderado: Un grafo donde cada arco tiene asociado un valor numérico o peso.
Ejemplo: Conexiones entre algunas ciudades de la república mexicana, donde el peso de cada arista significa la distancia en kilómetros desde la ciudad capital de un estado al otro, como el ejemplo de la fig. 9.

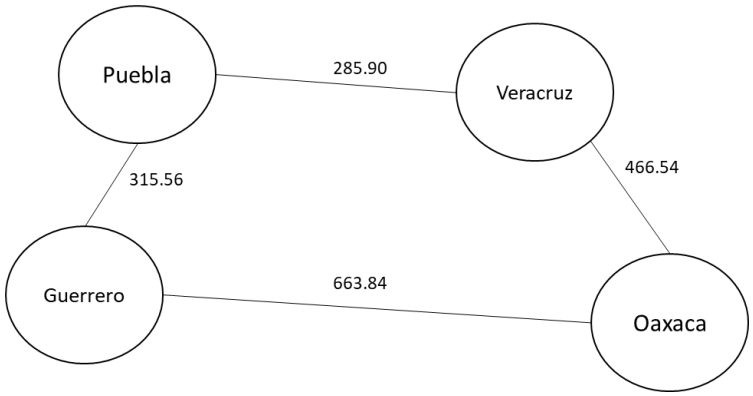

Fig. 9 Ejemplo de un grafo etiquetado y ponderado.

2.10 TextRank

Es un método que genera un grafo a partir de un documento de texto, consta de un par de conjuntos G= (V, E), los elementos de V se llaman vértices (oraciones) y los elementos de E se llaman aristas (éstas contienen un peso que se calcula utilizando la similitud del coseno). (Mihalcea, 2006) presentó el modelo TextRank, una variación del algoritmo de clasificación PageRank basado en grafos, TextRank es un método no supervisado e independiente del lenguaje. Este método fue utilizado para las tareas de extracción de palabras clave y GART, el método consta de cuatro pasos; 1: identifica las unidades de texto (palabras, oraciones) y los agrega como vértices en el grafo, 2:identifica las relaciones que conectan las unidades de texto y dibuja las aristas, estas pueden ser dirigidas o no dirigidas, 3: itera el algoritmo de clasificación hasta la convergencia, 4: ordena los

vértices según su puntuación final y selecciona las oraciones con mayor peso que formaran parte del resumen.

TextRank, es un algoritmo de ranqueó propuesto por (Mihalcea, 2006) basado en grafos, es un método no supervisado e independiente del lenguaje. Este método fue utilizado para las tareas de extracción de palabras clave y GART.

El algoritmo trabaja en la construcción de un grafo que representa el texto e interconecta palabras u otras entidades de texto con las relaciones significativas.

Los nodos o vértices del grafo se definen en función de la aplicación, los cuales contienen segmentos de texto, estos pueden ser de diferentes tamaños y características, por ejemplo, palabras, sentidos de palabras, oraciones completas, documentos completos, etc.

Las aristas o arcos que conectan a los nodos del grafo se determinan al tipo de relación que existe entre los nodos, por ejemplo, las relaciones léxicas o semánticas, medidas de cohesión del texto, superposición contextual, la pertenecía de una palabra en una oración, etc.

El algoritmo se compone de los siguientes pasos:

1: identifica los segmentos de texto y los agrega como vértices, como se muestra en la figura 10.

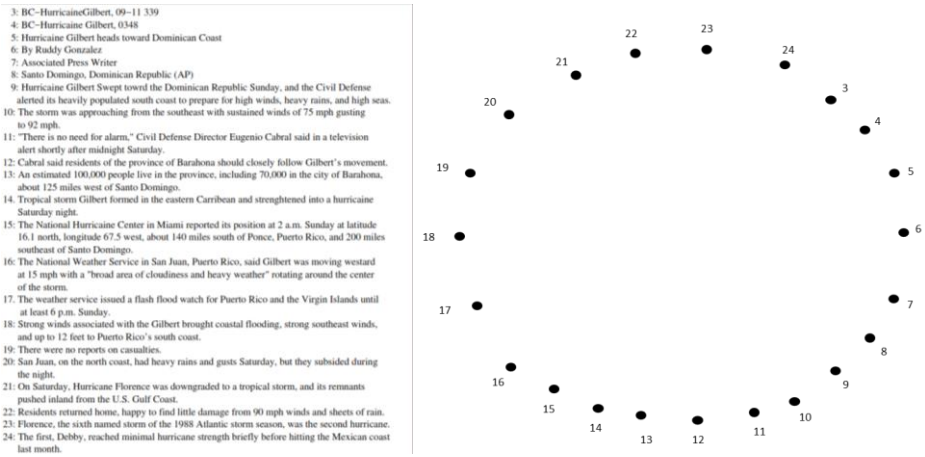

Fig. 10 Representación de las oraciones en un grafo (Mihalcea, 2006).

2. En la figura 11 se muestra el ejemplo de cómo se dibujan las aristas entre los vértices, se define una relación de similitud (coseno), la relación entre dos oraciones puede verse como una recomendación para referirse a otras oraciones que abordan el mismo o conceptos similares.

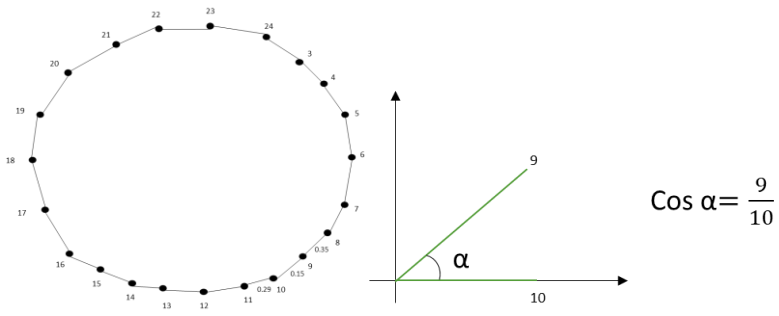

Fig. 11 Ejemplo de la relación coseno entre las oraciones de un grafo (Mihalcea, 2006).

3. Itera el algoritmo de clasificación basada en grafos hasta la convergencia.

El algoritmo de clasificación de grafos se basa en un modelo de caminata aleatoria, donde un caminante da pasos aleatorios en el grafo G como se muestra en la figura 12.

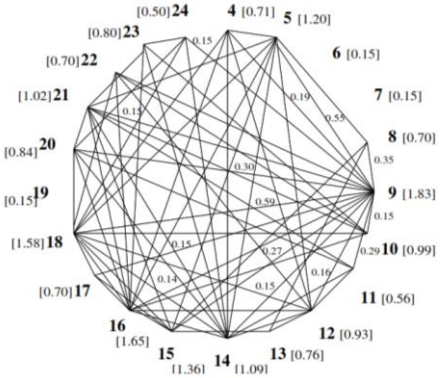

Fig. 12 Ejemplo del ranqueo de un texto (Mihalcea, 2006).

4: Selecciona las oraciones con mayor peso que formarán parte del resumen como se muestra en la figura 13.

9: Hurricaine Gilbert Swept towrd the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains, and high seas.

15: The National Hurricaine Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo.

16: The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westard at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the storm.

18: Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds, and up to 12 feet to Puerto Rico's south coast.

Fig. 13 Selección de las 4 mejores oraciones por el algoritmo TextRank (Mihalcea, 2006).

CAPÍTULO 3
Estado del arte

La GART tiene una trayectoria de más de 60 años de investigación, sin embargo, gano interés en la década de los 90´s, su desarrollo comenzó en las competencias de DUC (Document Understanding Conferences) y TAC (Text Analysis Conferences), donde se han presentado diversas investigaciones compitiendo por obtener los resúmenes automáticos más parecidos a los resúmenes generados por los humanos (Das & Martins, 2007; Galanis & Malakasiotis, 2008).

En este capítulo se presentan algunos de los trabajos que a lo largo del tiempo los investigadores han ido proponiendo, como el uso de las heurísticas, algoritmos evolutivos, algoritmos genéticos, métodos basados en modelos matemáticos y método por etapas.

3.1 Metodologías para la GART-M

En la tarea de GART-M se han empleado las siguientes metodologías:

Metodología que no considera todas las oraciones de la colección

Este método consiste en crear resúmenes individuales de los documentos y posteriormente colocar todos los resúmenes generados en un solo documento al que se le conoce como “meta documento”, después de genera un resumen de este documento usando el mismo o diferente método al del paso anterior (Lloret et al., 2008; Villatoro-Tello et al., 2009).

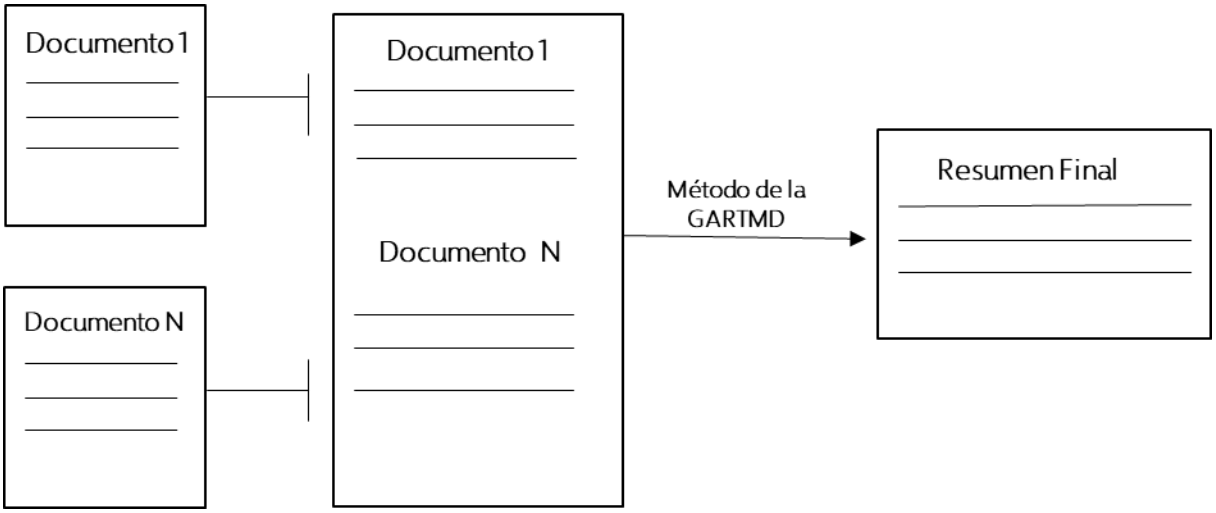

Fig. 14 Metodología que si considera todas las oraciones (Nery Mendoza, 2019).

Metodología que considera todas las oraciones de la colección

Este método consiste en crear un solo documento con todos los documentos de la colección y posteriormente se realiza un resumen del documento creado (Lloret et al., 2008; Mcdonald, 2007; Neri Mendoza, 2019).

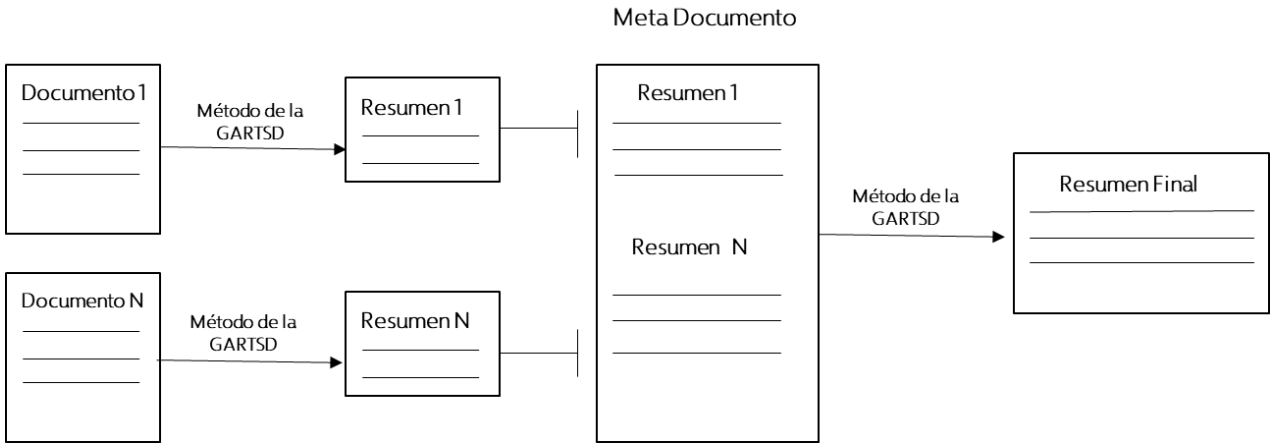

Fig. 15 Metodología que no considera todas las oraciones (Nery Mendoza, 2019).

3.2 Algoritmo evolutivo

Los algoritmos evolutivos están inspirados en la capacidad de las especies de evolucionar y adaptarse a su entorno, basados en una población de individuos que emplean mecanismos biológicos como la mutación, recombinación, selección natural y supervivencia de los más aptos para ir ajustando o refinando un conjunto de soluciones, esta idea corresponde con el principio de la evolución natural de Darwin.

3.2.1 Colonia de abejas

Sánchez Gómez en su trabajo (Sánchez Gómez et al., 2018) propone el algoritmo la colonia de abejas artificiales, en el que se consideran tres tipos de abejas; abejas obreras, observadoras y explotadoras. En primer lugar se inicia un archivo de almacenamiento (documento de texto) que contiene las soluciones no dominadas, la colonia inicial se genera aleatoriamente (resumen), las oraciones se seleccionan de forma aleatoria, en el siguiente paso las abejas obreras aplican una mutación para mejorar la solución, después se envían las abejas observadoras, cuya tarea es seleccionar su abeja obrera asignada, teniendo en cuenta las probabilidades de selección y el tamaño de la colonia se duplica, las abejas explotadoras se encargan de verificar las soluciones agotadas, que son aquellas que no han mejorado después de un determinado número de intentos, estas abejas agotadas son reemplazadas por abejas explotadoras, para finalizar el ciclo el tamaño de la colonia se reduce al tamaño original.

En su trabajo (Sánchez Gómez et al., 2018) considera todas las oraciones, este método considera los modelos de selección de oraciones de cobertura y reducción de redundancias, el método fue probado con el corpus de datos DUC 2002 en la que únicamente genero resúmenes de 3 colecciones, generando resúmenes de 200 palabras, los sistemas de evaluación que usaron fueron ROUGE-2 y ROUGE-1.

3.2.2 Algoritmo luciérnaga

En el trabajo de Tomer & Kumar (Tomer, Kumar, 2021) proponen el uso del algoritmo luciérnaga para la generación de resúmenes de múltiples documentos, como función de aptitud utiliza la relación de tema (TRF), cohesión (CF) y factor de legibilidad (RF). El algoritmo se basa en el comportamiento de apareamiento de las luciérnagas, está basado en el concepto de que la luciérnaga se siente atrida por la luciérnaga más brillante representa la solución óptima, mientras que las luciérnagas restantes no son la solución óptima. Las luciérnagas se mueven hacia luciérnagas más brillantes, de manera similar, las soluciones no óptimas se mueven hacia la solución óptima.

Cada luciérnaga está representada por un vector de tamaño N, donde N es el número total de oraciones de todos los documentos combinados, se emplea el sistema binario para decidir si una oración se va a incluir en el resumen, donde 1 representa a la oración incluida y 0 que no será incluida. Hay k luciérnagas inicializadas al azar, la intensidad de la luz de cada luciérnaga se calcula en base a la función de aptitud en donde se emplean las características TRF, CF y RF, el algoritmo termina hasta cierto número de iteraciones.

El método fue empleado para las colecciones DUC 2002, DUC 2003, DUC 2004 con resúmenes de 100 palabras, en donde el método dio buenos resultados en comparación con los métodos enjambre de partículas y un algoritmo genético.

3.3 Algoritmo genético

Los algoritmos genéticos (AG) implementan métodos de búsqueda a partir de modelos de fenómenos de la naturaleza. Los AG no buscan modelar la evolución biológica sino derivar estrategias de optimización, se basan en la generación de poblaciones de individuos (oraciones) mediante la reproducción de los padres y a diferencia de los algoritmos evolutivos estos permiten que individuos con menor aptitud puedan aparearse con individuos de mejor aptitud. Dando un mayor número se soluciones, el AG trabaja sobre una población de soluciones (resúmenes) y enfatiza la importancia de la cruza de los genes (palabras), sobre el de la mutación y usa la selección probabilística basada en la aptitud de soluciones a diferencia de otros paradigmas evolutivos (Ledeneva & García-Hernández, 2017; Mendoza, 2013; Neri Mendoza, 2019; Vázquez Et Al., 2018).

Nery Mendoza en su trabajo (Neri Mendoza, 2019) propuso un AG que, si considera todas las oraciones de la colección, en el operador de selección utilizó ruleta, con un modelo de representación de bigramas, el método propuesto fue empleado para resúmenes de 10, 50,100 y 200 palabras, se generaron resúmenes con las colecciones DUC 2001 y DUC 2002 y se evaluaron con los sistemas ROUGE-1 y ROUGE.2, su método se probó para resúmenes extractivos y abstractivos.

3.4 Método basado en modelos matemáticos

También se han realizado método basados en modelos matemáticos como los grafos, se han empleado grafos semánticos y grafos basados en caminatas aleatorias como TextRank.

3.4.1 Grafos semánticos

En este trabajo (Del camino Valle et al., 2019), proponen la implementación de un grafo como modelo de representación de texto, a diferencia de (Mihalcea, 2006), ellos generar un grafo para cada documento de texto, los documentos se representan mediante grafos semánticos generados automáticamente, a partir de la identificación de los conceptos y relaciones semánticas entre ellos desde WordNet. La construcción de los grafos parte de extraer de WordNet los sentidos (synset) de los conceptos incluidos en cada una de las oraciones, así como las relaciones semánticas de

hiperonimia, hiponimia, meronimia y holonimia existentes entre ellos. Estos elementos capturados son usados para obtener el grafo que representa a cada oración, donde los synset constituyen los nodos y las relaciones se establecen según los vínculos identificados en WordNet. Luego, los grafos generados son integrados para obtener un único grafo del documento, cuyo proceso se realiza unificando los synset comunes en los diferentes grafos de las oraciones. Este método se probó para los idiomas inglés y español con las colecciones EnCol y SPCol, de acuerdo con los resultados obtenidos, en comparación con otros métodos, los grafos semánticos obtuvieron los mejores resultados para ambas colecciones.

3.5 Método por etapas

Adicionalmente se han desarrollado métodos compuestos por diferentes etapas, de acuerdo con el punto de vista clásico la GART consta de tres etapas. (Hovi, 2003) propone la primera etapa como la identificación de tópicos, la segunda etapa corresponde a la interpretación y la última etapa es la generación del resumen. (Cardoso & Pérez-Abelleira, 2013) coinciden al proponer tres etapas: análisis, transformación y síntesis. Sin embargo, (Ledeneva & García-Hernández, 2017) proponen una epata más. Las etapas que ellos proponen son: selección de términos, pesado o ponderación de términos, pesado o ponderación de oraciones y selección de oraciones. La cuarta etapa se propuso con el fin de asignarle un valor a las palabras obtenidas y después darles un valor a las oraciones para poder seleccionar las mejores que pasaran a formar parte del resumen, a diferencia del punto de vista clásico que solo identifica los tópicos que considera importantes.

3.5.1 Método por etapas con SFM

En el trabajo realizado en (Ledeneva, García-Hernández, Gelbuk, 2010), realizaron una primera aproximación en la generación de resúmenes de texto de múltiples documentos empleando a las SFM como modelo de representación de texto. En esta primera aproximación consideraron el método por etapas, que consta de cuatro etapas, en la primera etapa extrajeron las SFM de cada colección de documentos, en la segunda etapa de pesado de los términos consideraron las características de las SFM (longitud y frecuencia), en la tercer etapa de selección de oraciones optaron por realizar la suma de los términos individuales de cada oración y en la última etapa se extrajeron las oraciones que tenían los pesos más altos, para su trabajo utilizaron el corpus DUC2002, sin embargo, sus resultados no fueron tan buenos, ya que el puntaje que obtuvieron de acuerdo a la medida F del sistema ROUGE-2 fue 0.312.

CAPÍTULO 4
Método propuesto

En el estado de arte, el resumen de un solo documento se han probado algunos métodos como el método por etapas propuesto por (Ledeneva & García Hernández, 2017) y el algoritmo de TextRank propuesto por (Mihalcea, 2006) por lo que en este trabajo de investigación se propone implementar el método por etapas y en la etapa de pesado de oraciones utilizar el algoritmo de TextRank para pesar las oraciones de acuerdo a las características dadas por las SFM's encontradas en la colección de documentos, en la Figura 16 se muestra el método propuesto.

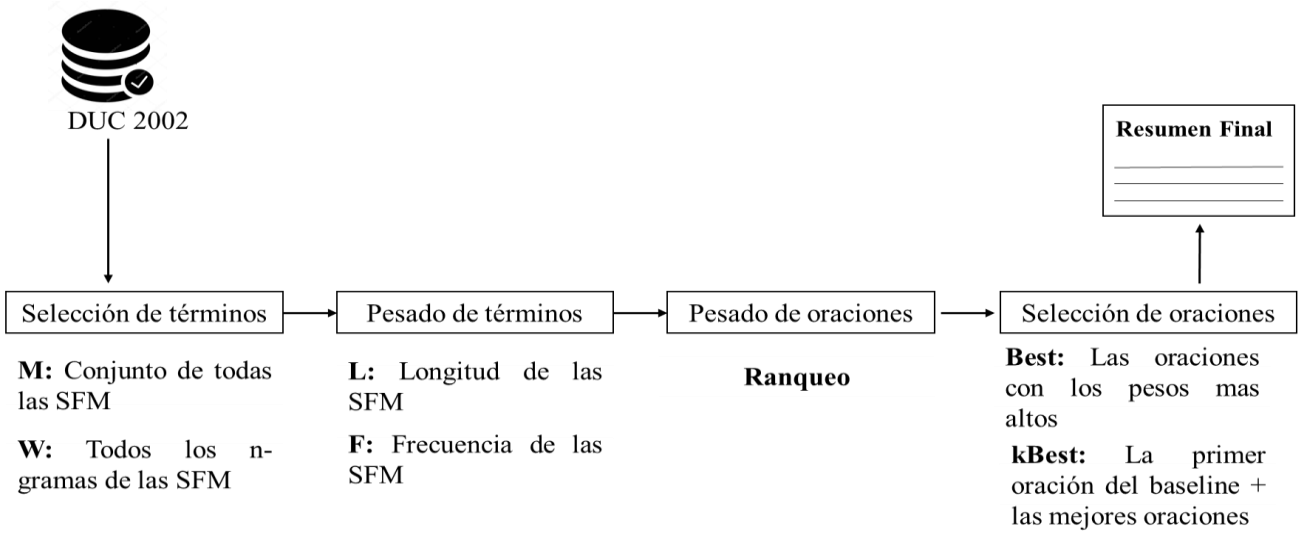

Fig. 16 Método propuesto.

4.1 Selección de Términos

En esta etapa se debe definir el término que será extraído del texto para poder representar y cuantificar la importancia de las oraciones. En la propuesta de investigación se optó por la utilización de las SFM's como modelo de texto, debido a que las SFM's no limitan su longitud, como lo hacen los n-gramas.

Nuestra hipótesis es que las SFM's expresan las ideas más importantes de una colección de documentos. Esto se puede ver como tf-idf (frecuencia de término-frecuencia inversa del documento). Por un lado, la idea expresada por una SF es importante para el documento si aparece en repetidas ocasiones (alta frecuencia del término). Por otro lado, la idea correspondiente debe ser específica para este documento, de lo contrario no existiría en el lenguaje una sola palabra o por lo menos una abreviatura para expresarlo (alta frecuencia inversa de documento).

Es importante tener en cuenta que las SFM's representan al conjunto de las SFs, puesto que es posible reproducir al conjunto de SFs a partir de las SFM's, si cada SFM se descompone en todas las subsecuencias. Esta propiedad debe tenerse en cuenta ya que si una oración contiene algunas subsecuencias de una SFM también representa un grado de dependencia. Es por lo que se prueban varios esquemas de sub-términos derivados de las SFM's.

Una de las principales ventajas que presentan las SFM es que a diferencia de los n-gramas, éstas no requieren una configuración humana para determinar el número de gramas y su extracción es independiente del lenguaje.

Las variantes que proponemos para la selección de términos son las siguientes:
M: Conjunto de todas las SFM's, es decir, un n-grama m ϵ M si es una SFM con algún umbral β (se consideran SFM's de dos o más palabras y $\beta \geq 2$).
W: Palabras derivadas (unigramas) de elementos de M. Es decir, una palabra w ϵ W si existe una SFM m ϵ M tal que w ϵ m.

4.2 Pesado de Términos

En esta etapa se propone un esquema de pesado de las SFM's teniendo en cuenta la longitud y la frecuencia de los términos.

F: frecuencia del término en una SFM, es decir, el número de veces que el término aparece en el texto, f es el número de veces que el término aparece en la colección de documentos de texto.
En esta forma de pesado se emplea una matriz en la que se le asigna un peso a cada SFM dependiendo de en número de ocasiones en que aparezca la SFM en un documento, posterior a ello se realiza un sumatorio total del número de veces que aparece la SFM en la colección como se muestra en la tabla 1. Nuestra hipótesis es que cuando mayor es la frecuencia de una SFM, la oración que contenga dicha SFM tiene mayor relevancia en la generación del resumen.

Tabla 1 Pesado de términos por frecuencia.

SFM\ Documento	AP880911-0016	AP880912-0095	AP880912-0137	WSJ880912-0064	AP880915-0003	AP880916-0060	**Frecuencia**
In Puerto Rico	0	0	0	2	0	0	2
The Hurricane Center Said	0	0	0	1	0	1	2
In The Dominican Republic	0	2	0	1	2	0	5
The Brunt Of The	0	0	2	0	0	0	2
The National Hurricane Center Said	0	0	0	0	2	1	3
Gilbert Was Expected	0	1	1	0	0	0	2
South Of The	0	1	0	1	0	0	2
Gilbert swept toward	1	0	0	1	0	0	2
The storm was	1	0	0	0	1	0	2
the north coast	1	1	0	0	1	0	3

L: Se refiere a la longitud de las SFM's, este pesado se refiere al número de palabras por las que está conformada una SFM, en la tabla 2 se muestra el pesado de las SFM por longitud.

Tabla 2 Pesado de términos por longitud.

SFM	Longitud
In Puerto Rico	3
The Hurricane Center Said	4
In The Dominican Republic	4
The Brunt Of The	4
The National Hurricane Center Said	5
Gilbert Was Expected	3
South Of The	3

Gilbert swept toward	3
The storm was	3
the north coast	3

4.3 Pesado de Oraciones

En esta etapa se configura el algoritmo de TextRank para realizar una clasificación de los nodos en el grafo.

Vértices: Los vértices del grafo son cada una de las oraciones, representadas mediante los términos derivados de las SFM propuestas en el apartado 2.2 de selección de términos.

Aristas: Las relaciones que conectan las SFM's son relaciones de ponderación de términos como la frecuencia de un SFM's en un texto y la longitud de las SFM's.

Configuración del algoritmo TextRank: El objetivo es clasificar las SFM's, por lo tanto, se agrega un vértice al grafo para cada SFM's en el texto, se dibujan las aristas entre los vértices para definir una relación de ponderación de términos. Se puede ver como un proceso de recomendación en el que una oración que aborda ciertos conceptos en un texto le da al lector una recomendación para referirse a otras oraciones que aborden el mismo o conceptos similares. El grafo resultante contiene un peso asociado a cada vértice.

Como pesado de oraciones se utiliza la variante del peso $(t_f) = f(t_k)^{L(t_k)}$ que significa que se toma como relevancia de término t_k a la frecuencia de la SFM elevada a la longitud de dicha SFM. El peso inicial de cada arista entre el vértice V_i y el vertex V_j se calcula como:

$$\left(Peso\ arista(V_i, V_j\right) = \Pi t_{k=\left\{t\middle|t\epsilon v_j \cap t\ \in\ v_j\right\}} peso\left(t_j\right)$$

Ecuación 1 Cálculo del peso inicial de cada arista.

Para calcular la relevancia de cada vértice se suman todas las aristas que tiene ese vértice como se muestra en la fórmula 2:

$$Peso\ inicial\left(V_j\right) = \Sigma_{i=j} peso\ arista\ \left(V_i, V_j\right)$$

Ecuación 2 Cálculo de la relevancia de cada vértice.

Por ejemplo, considerando las SFM's como términos de las oraciones de la figura 17 se puede construir el grafo de la figura 18, donde para el peso de la arista entre dos oraciones se puede calcular de acuerdo con la frecuencia.

A: *The National Hurricane Center said a hurricane watch was in effect on the Texas coast from Brownsville to Port Arthur and along the coast of northeast Mexico from Tampico north.*
B:*The National Hurricane Center said Gilbert was the most intense storm on record in terms of barometric pressure.*
C: *In Jamaica, the government issued a hurricane watch for the entire island.*

Fig. 17 Ejemplo de tres oraciones de un texto arbitrario.

La oración A tiene las SFM's {The National Hurricane Center said, a hurricane watch}, la oración B tiene la SFM's{The National Hurricane Center said} y la oración C tiene la SFM{a hurricane watch}, por lo que la intersección entre las oraciones (A, B) sería { The National Hurricane Center said } y la intersección entre las oraciones (A, C) sería { a hurricane watch}, por lo que el peso de la arista (A, B) = 3 y el peso de (A, C) = 2.

Para calcular la relevancia inicial de cada vértice se suman todas las aristas que tiene ese vértice como se muestra en la ecuación 2.

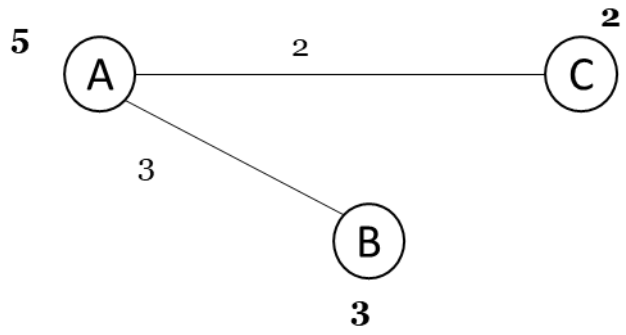

Fig. 18 Representación del grafo usando SFM como términos de las oraciones de la fig. 17.

Considerando el ejemplo anterior, el vértice A = 2 + 3 = 5, como se ve en la figura 15, los vértices pueden ordenarse de mayor a menor importancia en A, B, C.

4.4 Selección de Oraciones

En esta etapa se considerarán diversas configuraciones las cuales se enlistan a continuación:

best: Selecciona las oraciones con mayor peso dado por el algoritmo de TextRank hasta obtener el tamaño deseado del resumen (100 o 200 palabras) como se muestra en la figura 19.

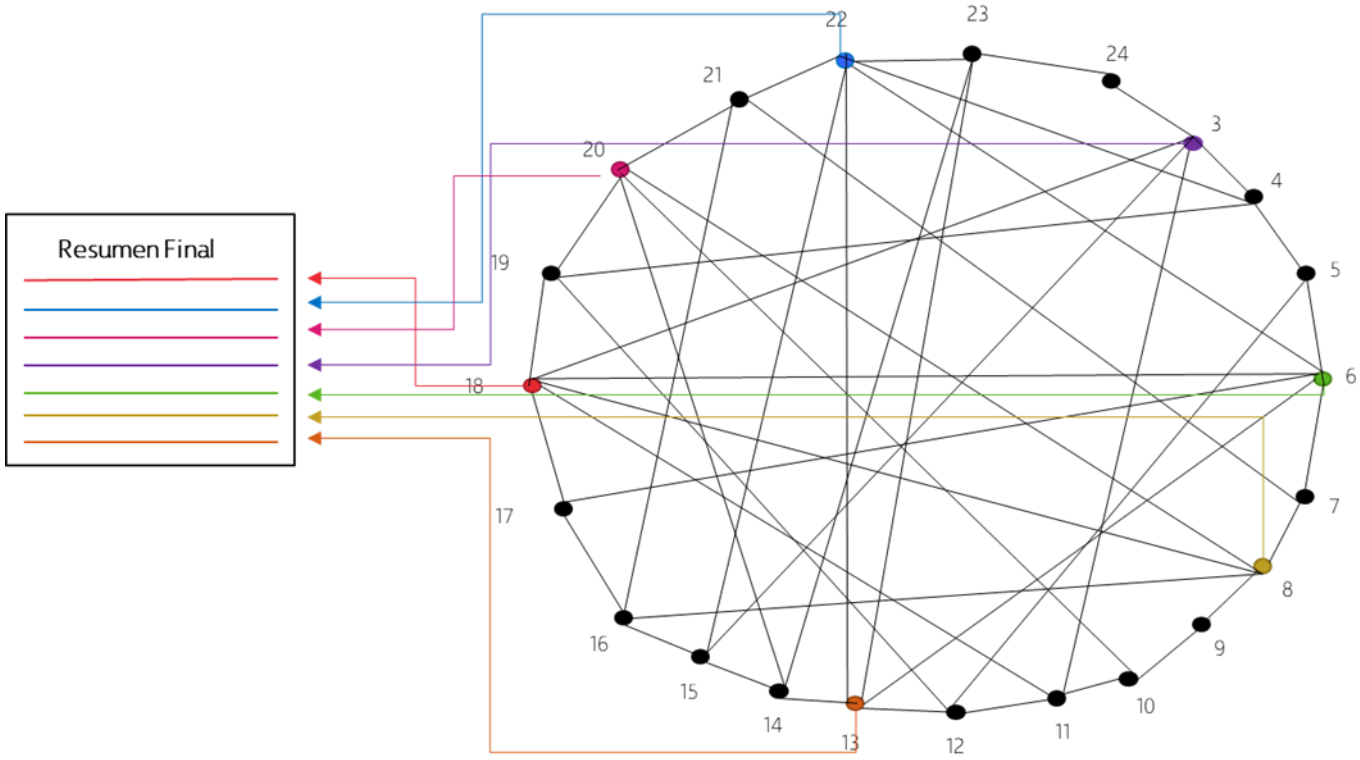

Fig. 19 Selección de oraciones *best*.

kbest + first: Selecciona las *k* mejores oraciones dadas por TextRank y después toma las primeras oraciones del primer documento de texto hasta obtener el tamaño deseado del resumen, en la figura 20 se muestra la forma en que se realiza este tipo de selección. Esto fue motivado por la heurística *Baseline-first-document.*

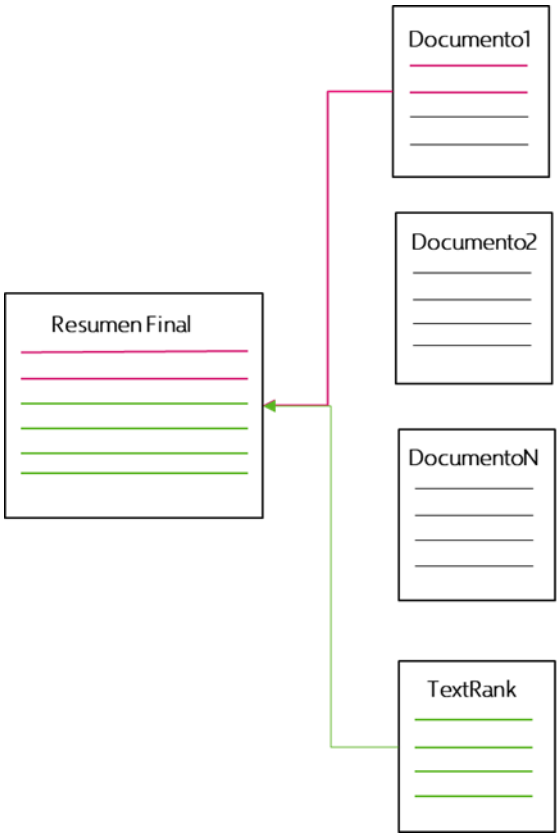

Fig. 20 Selección de oraciones *kbest + first*.

kbaseline + best: Se extrae la primera oración de los *k* primeros documentos originales ordenados cronológicamente y luego las mejores oraciones dadas por el algoritmo de TextRank hasta obtener el tamaño deseado para el resumen. Esta selección fue motivada por la heurística *Baseline*.

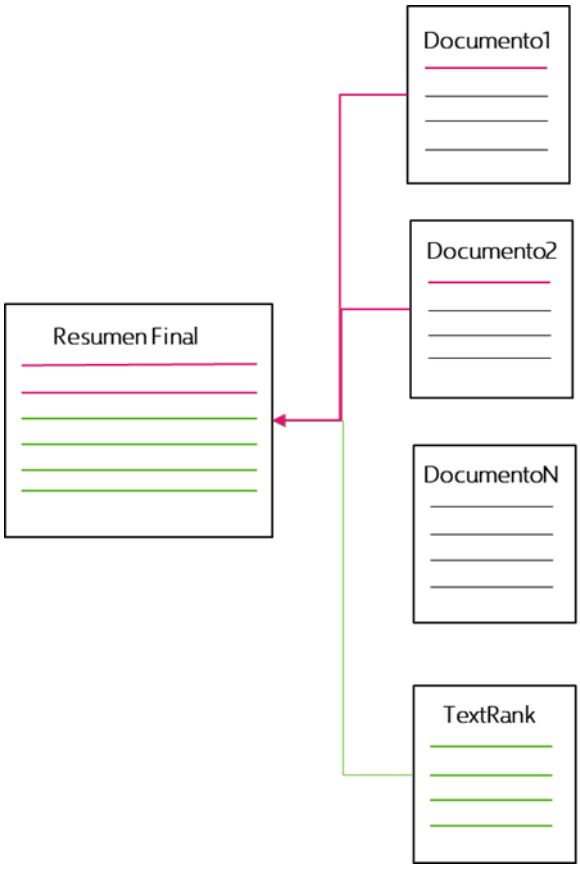

Fig. 21 Selección de oraciones *kbaseline + best*.

CAPÍTULO 5
Experimentos y resultados

En este capítulo describimos algunos de los experimentos que se probaron para esta investigación, así como la colección de datos que se utilizó para las pruebas y el sistema de evaluación empleado.

5.1 Corpus DUC-02

El Instituto Nacional de Estándares y Tecnología (NIST) realiza competencias para la tarea de resúmenes de texto, llamada Conferencia de Comprensión de Documentos (DUC), para esta competencia pone a disposición de los investigadores diferentes colecciones de datos, entre las que destacan las colecciones de datos DUC-01 y DUC-02, esto debido a que son las más utilizadas por los investigadores.

Para este trabajo se utilizó la colección de datos DUC-02, la cual consta de 567 documentos en idioma inglés, agrupados en 59 colecciones. Cada colección contiene en promedio de 5 a 8 documentos de noticias periodísticas, que abordan temas sobre; desastres naturales, información biográfica de un individuo, temas políticos, etc. Esta colección de datos también contiene los resúmenes modelo humano. Éstos están hechos de 50, 100, 200 y 400 palabras.

5.2 Evaluación

En el presente trabajo de investigación, se optó por utilizar el sistema de evaluación ROUGE-N, el cual se encarga de medir la similitud que existe entre un resumen generado por un modelo y un resumen generado con juicios humanos. Para la comparación, utiliza estadísticas de n gramas, en los que se mide la precisión, el recuerdo y la exhaustividad, se calcula de la siguiente manera:

$$ROUGE - N \ = \frac{\sum_{S \in \{ReferenceSummaries \ \ \}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries \ \ \}} \sum_{gram_n \in S} Count(gram_n)}$$

Ecuación 3 Cálculo de ROUGE-N (Lin, 2004).

donde *n* es la longitud del n-grama y $Count_{match}(gram_n)$ el número máximo de n-gramas que coocurren en el resumen candidato y el conjunto de resúmenes de referencia.

Para cada una de las medidas de ROUGE se calculan las medidas de Recuerdo, Precisión, F-measure, las cuales se enlistan a continuación:

- Precisión (*P*): Refleja la cantidad de buenas oraciones extraídas por el sistema y se calcula con la ecuación 4.

$$P = \frac{\#(oraciones\ correctas)}{\#(oraciones\ correctas + oraciones\ incorrectas)}$$

Ecuación 4 Cálculo de la precisión (Lin, 2004).

- Recuerdo (*R*): Refleja la cantidad de oraciones correctas que olvido el sistema se calcula con la ecuación 5.

$$R = \frac{\#(oraciones\ correctas)}{\#(oraciones\ correctas + oraciones\ no\ extraidas)}$$

Ecuación 5 Cálculo del recuerdo (Lin, 2004).

- F-measure (*F*): Donde se toma a oraciones correctas como el número de oraciones extraídas por el sistema y por los humanos, a oraciones incorrectas como el número de oraciones extraídas por el sistema, pero no por el ser humano, y a oraciones no extraídas como el número de oraciones extraídas por el ser humano, pero no por el ser humano, se calcula con la ecuación 6.

$$F = \frac{2PR}{P + R}$$

Ecuación 6 Cálculo de la medida F (Lin, 2004).

5.3 Experimentos de 200 palabras

En este apartado se explican las diversas etapas de experimentos que se realizaron para este trabajo de investigación, en la primera etapa se describen las principales configuraciones propuestas en el método, en la segunda etapa se seleccionan las mejores configuraciones de la etapa 1 y se prueban con diferentes longitudes de las SFM's y en la tercera etapa se pone a prueba una nueva técnica se selección de oraciones.

5.3.1 Primera etapa de experimentos y resultados

Es esta etapa se muestran las principales combinaciones en las cuatro etapas de método propuesto, en esta configuración no se le realizó algún tipo de preprocesamiento al corpus, se extrajeron las SFM's por colección, con un umbral de frecuencia mayor igual a 2 y un GAP = 3.

Tabla 3 Tabla de resultados de la primera etapa de experimentos para resúmenes de 200 palabras.

N. E	Selección de términos	Pesado de términos	Pesado de oraciones	Selección de oraciones	ROUGE-1.5.5			SFM (tamaño)
					R	P	F	
1	**M**	**F**	**TextRank**	***best***	**0.43268**	**0.44323**	**0.43778**	**1-9**
2	M	L	TextRank	*k+best*	0.38622	0.42504	0.40346	1-9
3	M	F	TextRank	*best*	0.41533	0.43099	0.42274	1-9
4	M	L	TextRank	*k+best*	0.37685	0.42015	0.39594	1-9
5	W	F	TextRank	*best*	0.40800	0.42850	0.41752	1-9
6	W	L	TextRank	*k+best*	0.34163	0.42003	0.37430	1-9
7	W	F	TextRank	*best*	0.32919	0.41046	0.36275	1-9
8	W	L	TextRank	*k+best*	0.39114	0.40555	0.39812	1-9

En la Tabla 4 se muestran las 8 diferentes configuraciones que se realizaron en las cuatro etapas, en esta primera etapa de experimentos, se probó con SFM's de tamaño 1-9 extrayendo únicamente las 8 mejores oraciones dadas por el algoritmo de ranque.

En esta primera etapa de experimentación se descubrió que la configuración que daba mejores resultados fue usando la configuración del experimento 1 en donde: se utilizaron el conjunto de todas las SFM's como selección de términos, y en la etapa de pesado de términos el pesado de acuerdo con la frecuencia dio mejores resultados. Esto se debe a que cuando mayor número de ocasiones aparece un término en toda la colección más relevante es en el resumen y en la selección de oraciones al seleccionar las mejores oraciones resultó ser mejor que seleccionar la primera oración del primer documento más las mejores oraciones dadas por el algoritmo de ranqueo.

5.3.2 Segunda etapa de experimentos y resultados

En esta etapa de experimentos se utilizó la misma configuración para la extracción de las SFM's como en el experimento 1. El objetivo fue darle prioridad a la longitud de las SFM, por lo que se probó utilizando diferentes longitudes, la SFM con mayor tamaño tuvo una longitud de 27 palabras. Este experimento nos permitió observar mejores resultados con respecto a la primera etapa de experimentos como se muestra en la tabla 5.

Tabla 4 Resultados de experimentos con SFM's de diferentes longitudes para resúmenes de 200 palabras.

N. E	Selección de términos	Pesado de términos	Pesado de oraciones	Selección de oraciones	ROUGE-1.5.5			SFM (tamaño)
					R	P	F	
1	M	F	TextRank	*best*	0.43556	0.45185	0.44331	**1-15**
2	M	F	TextRank	*best*	0.41884	0.42977	0.42414	1-9
3	M	L	TextRank	*best*	0.42449	0.43442	0.42931	1-27
4	**M**	**F**	**TextRank**	***best***	**0.43856**	**0.45154**	**0.44486**	**1-27**
5	M	F	TextRank	*k+best*	0.43274	0.42891	0.43073	1-27
6	W	F	TextRank	*best*	0.43235	0.44224	0.43714	1-27

5.3.3 Tercera etapa de experimentos y resultados

Para esta tercera etapa de experimentos optamos por probar una nueva técnica en la selección de oraciones, en donde pusimos a prueba la selección *kbaseline+best* descrita en la sección de método propuesto, para este experimento se tomaron en cuenta las mejores configuraciones descubiertas en los experimentos anteriores, en donde se mostró una mejoría en los resultados obtenidos, como se muestra en la tabla 6.

Tabla 5 Resultados de experimentos con selección *kbaseline + best* para resúmenes de 200 palabras.

N. E	Selección de términos	Pesado de términos	Pesado de oraciones	Selección de oraciones	ROUGE-1.5.5			SFM (tamaño)
					R	P	F	
1	M	F	TextRank	*1bestline+best*	0.45776	0.46895	0.46320	1-27
2	M	F	TextRank	*2bestline+best*	0.46569	0.47438	0.46990	1-27
3	M	F	TextRank	*3bestline+best*	0.46986	0.47548	0.47256	1-27
4	M	F	TextRank	*4bestline+best*	0.48072	0.48377	0.48214	1-27
5	**M**	**F**	**TextRank**	***5bestline+best***	**0.48798**	**0.48847**	**0.48813**	**1-27**

5.4 Experimentos de 100 palabras

En este apartado se realizaron los experimentos para resúmenes de 100 palabras, los resúmenes de generaron en tres diferentes etapas de experimentación. En la primera etapa se probó con todas las configuraciones posibles respecto al método propuesto. En la segunda etapa se probo con las mejores configuraciones dadas por la esta anterior y se probo para diferentes longitudes de las SFM's.

5.4.1 Primera etapa de experimentos y resultados

Se extrajeron las SFM's por colección de documentos. Las configuraciones que se ocuparon para la extracción de las SFM's fueron: umbral de frecuencia mayor igual a 2 y una restricción de GAP=3.

Para esta primera etapa de experimentos únicamente se usaron las SFM's con una longitud de 1 a 9 palabras.

Tabla 6 Todas las configuraciones para resúmenes de 100 palabras

N. E	Selección de términos	Pesado de términos	Pesado de oraciones	Selección de oraciones	ROUGE-1.5.5			SFM (tamaño)
					R	P	F	
1	**M**	**F**	**TextRank**	***best***	**0.34533**	**0.36017**	**0.35251**	**1-9**
2	M	L	TextRank	*k+best*	0.10619	0.39056	0.16340	1-9
3	M	F	TextRank	*best*	0.3096	0.31876	0.31404	1-9
4	M	L	TextRank	*k+best*	0.10540	0.38373	0.16177	1-9
5	W	F	TextRank	*best*	0.31672	0.32561	0.32103	1-9

6	W	L	TextRank	*k+best*	0.31073	0.32746	0.31883	1-9
7	W	F	TextRank	*best*	0.28862	0.30599	0.29698	1-9
8	W	L	TextRank	*k+best*	0.31672	0.32561	0.32103	1-9

En la tabla 6 se muestran los resultados obtenidos para los 8 experimentos realizando en esta etapa de experimentación, donde se demostró que la mejor configuración es utilizando en la etapa de -Selección de términos: el conjunto de todas las SFM's. -Pesado de términos: por longitud. -Pesado de oraciones: TextRank. -Selección de oraciones: *best*.

5.4.2 Segunda etapa de experimentos y resultados

En la tabla 7 se muestran tres experimentos con las mismas configuraciones en las 4 etapas del método, la diferencia de estos experimentos es la longitud de las SFM's empleadas, se optó por probar con tamaños de 1-9, 1-15, 1-27, donde se demostró que a mayor longitud de las SFM's el método genera mejores resúmenes.

Tabla 7 Mejores configuraciones con SFM's de diferente longitud para resúmenes de 100 palabras.

N. E	Selección de términos	Pesado de términos	Pesado de oraciones	Selección de oraciones	ROUGE-1.5.5			SFM (tamaño)
					R	P	F	
1	M	F	TextRank	*best*	0.35213	0.36599	0.35886	**1-15**
2	M	F	TextRank	*best*	0.33912	0.35362	0.34617	1-9
3	M	L	TextRank	*best*	0.32354	0.34524	0.33413	1-27
4	**M**	**F**	**TextRank**	***best***	**0.36493**	**0.38197**	**0.37320**	**1-27**
5	W	F	TextRank	*k+best*	0.31242	0.32041	0.31629	1-27
6	W	F	TextRank	*best*	0.31672	0.32561	0.32103	1-27

5.4.3 Tercera etapa de experimentos y resultados

En la tabla 8 se muestran cinco experimentos con las mismas configuraciones en las etapas de selección de términos, pesado de términos y pesado de oraciones, sin embargo, en este apartado se muestran diferentes tipos de selección de oraciones.

Tabla 8 Experimentos con selección *kbaseline + best* para 100 palabras.

N. E	Selección de términos	Pesado de términos	Pesado de oraciones	Selección de oraciones	ROUGE-1.5.5			SFM (tamaño)
					R	P	F	
1	M	F	TextRank	*1bestline+best*	0.41629	0.43110	0.42350	1-27
2	M	F	TextRank	*2bestline+best*	0.42906	0.44041	0.43460	1-27
3	M	F	TextRank	*3bestline+best*	0.44682	0.45260	0.44963	1-27
4	M	F	TextRank	*4bestline+best*	0.45606	0.46055	0.45824	1-27
5	**M**	**F**	**TextRank**	***5bestline+best***	**0.45673**	**0.46108**	**0.45885**	**1-27**

CAPÍTULO 6
Conclusiones

- En este trabajo pudimos comprobar nuestra hipótesis, donde planteábamos que cuando más veces se repitiera una SFM, la oración que contenga dicha SFM contiene información relevante, lo que permite generar buenos resúmenes.

- En la tarea GART-M con SFM y grafos se descubrió que la configuración que da mejores resultados es empleando el conjunto de todas las SFM's en la etapa de selección de términos, en la etapa de pesado de términos resulto dar mejores resultados el pesado con frecuencia, en la etapa de pesado de oraciones fue de gran impacto la implementación del algoritmo de TextRank y en la etapa de selección de oraciones la mejor opción fue *kbaseline+best* propuesta en este trabajo.

- Los resultados obtenidos en esta investigación lograron competir con los resultados de un método más sofisticado como lo es el algoritmo genético, en la figura 22 se muestra una gráfica de comparación del método propuesto con otros métodos encontrados en el estado del arte.

- El método propuesto en este trabajo logró obtener mejores resultados a los observados en métodos propuestos en el estado del arte y en algunas heurísticas.

- Una de las ventajas que presenta el método propuesto es que, a diferencia de los métodos supervisados, este tiene la capacidad de adaptarse fácilmente a nuevos lenguajes y dominios sin requerir datos de entrenamiento para cada nuevo tipo de datos.

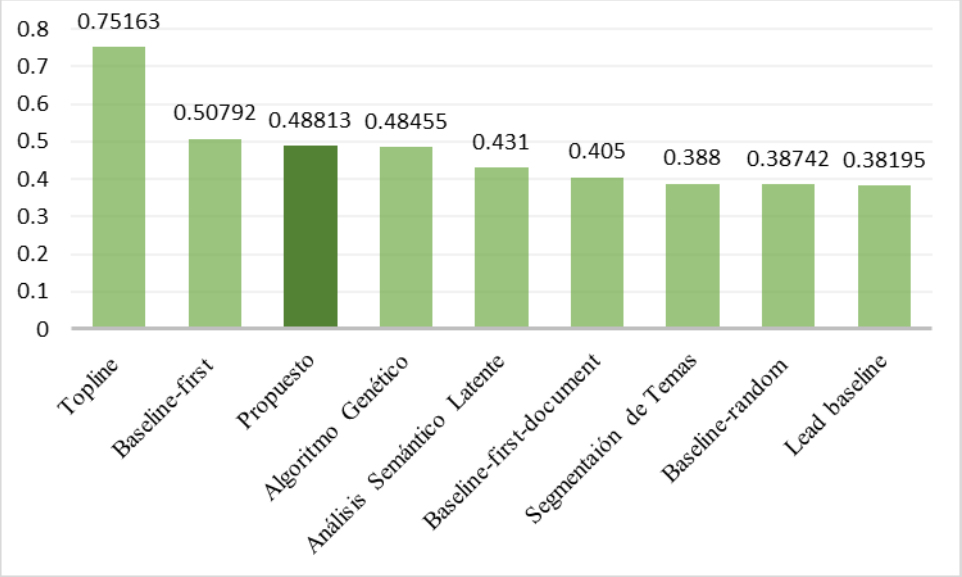

Fig. 22 Gráfica de comparación con otros métodos.

6.1 Aportaciones

Derivado de la investigación realizada se realizaron las siguientes aportaciones:

- Fue propuesta la selección *kbaseline + best*.
- Se escribió un artículo derivado de esta investigación el cual lleva por título *Multi-document Extractive Text Summaries using a Ranking Algorithm and Maximal Frequent Sequences*

6.2 Trabajo futuro

Dentro de las áreas de oportunidad que presenta este trabajo se consideran las siguientes:

- Probar el método con otra colección de datos como la colección DUC-05 y DUC-07.
- Generar resúmenes con la metodología que no considera todas las oraciones, es decir generar resúmenes de manera individual y posteriormente generar un resumen general.
- Generar resúmenes con las SFM's empleando método basados en la naturaleza como el algoritmo luciérnaga.

Referencias

Acero, I., Alcojor, M., Díaz Esteban, A., Gómez Hidalgo, J. M., & Maña López, M. J. (2001). Generación Automática De Resúmenes Personalizados. Procesamiento Del Lenguaje Natural, 27(34), 281–290.

Ahonen-Myka, H. (1999). Finding All Maximal Frequent Sequences In Text. Proc. Icml Workshop On Machine Learning In Text Data Analysis, 11–17.

Alonso, M. P., Alonso, M. R., Rodríguez, C. G., Gil, J. G., & Ferro, J. V. (2012). Language Technologies For Opinion Analysis In Social Networks: From Text To Microtext View Project Word Order Theory View Project.

Alvarado Bolaños, A. (2017). Evaluación De La Calidad De Las Herramientas Comerciales Y Métodos Del Estado Del Arte Para La Generación De Resúmenes Del Corpus Duc-2001. Universidad Autónoma Del Estado De México.

Anacona, F. A. A. (2015). Algoritmo Para Generación Automática De Resúmenes Extractivos Genéricos De Múltiples Documentos Basado En Consensos [Universidad Del Cauca Facultad De Ingeniería Electrónica Y Telecomunicaciones]. Https://Doi.Org/10.13140/Rg.2.2.25122.22721

Cardoso, A., & Pérez-Abelleira, M. A. (2013). Generación Automática De Resúmenes. 1er Congreso Nacional De Ingeniería Informática/Sistemas De Información, Conaiisi, 1(1), 10. Http://Conaiisi.Frc.Utn.Edu.Ar/Pdfsparapublicar/1/Schedconfs/1/11-495-1-Dr.Pdf

Cortez Vásquez, M. A., Vega Huerta, M. H., & Pariona Quispe, J. (209 C.E.). Procesamiento De Lenguaje Natural. Revista De Ingeniería De Sistemas E Informática, 6, 45–54.

Das, D., & Martins, A. F. . (2007). A Survey On Automatic Text Summarization. Journal Of Ai And Data Mining, 0(0). Https://Doi.Org/10.22044/Jadm.2018.6139.1726

Del Camino Valle, O., Simón-Cuevas, A., Valladares-Valdés, E., Olivas, J., Romero, F., Olivas Varela, J., Simón Cuevas, A., & Valladares Valdés, E. (2019). Multi-Document Extractive Summarization Using Semantic Graph. Procesamiento Del Lenguaje Natural, 63, 103–110. Https://Doi.Org/10.26342/2019-63-11

Domínguez Burgos, A. (2002). Lingüística Computacional: Un Esbozo. Boletín De Lingüística, 18, 104–119.

Elian, M., & Becerra, A. M. (2012). Una Revisión De La Generación Automática De Resúmenes Extractivos A Review Of The Extractive Text Summarization. L, 7–27.

Galanis, D., & Malakasiotis, P. (2008). Aueb At Tac 2008. Theory And Applications Of Categories.

García-Hernández, R. A. (2007). Desarrollo De Algoritmos Para El Descubrimiento De Patrones Secuenciales Máximales. Https://Inaoe.Repositorioinstitucional.Mx/Jspui/Bitstream/1009/600/1/Garciahra.Pdf

Gillick, D., Favre, B., Hakkani-Tür, D., Bohnet, B., Liu, Y., & Xie, S. (2009). The Icsi Summarization System At Tac 2008. 801–815.

González Suárez, E. (2004). Conocimiento Y Evolución De La Humanidad. Acimed, 12(2).

Guinovart, J. G. (1998). Fundamentos De Lingüística Computacional : Bases Teóricas , Líneas De Investigación Y Aplicaciones. January 1998.

Hernández Maya, P. T. (2018). Desempeño De Los Métodos Del Estado Del Arte Para La Generación Automática De Resúmenes Extractivos Para El Corpus Textruss. Universidad Autónoma Del Estado De México Unidad.

Ledeneva, Y., & García-Hernández, R. A. (2017). Automatic Generation Of Text Summaries: Challenges, Proposals And Experiments.

Lloret, E., Ferránes, Ó., Muños, R., & Palomar, M. (2008). Integración Del Reconocimiento De La Implicación Textual En Tareas Automáticas De Resúmenes De Texto. Procesamiento Del Lenguaje Natural, 41, 183–190. Https://Www.Redalyc.Org/Pdf/5157/515751741020.Pdf

Marino Cuéllar Chacón, C., Eliana Mendoza Becerra Co-Director, M., & Alberto Cobos, C. (2018). Generación Automática De Resúmenes De Múltiples Documentos Mediante La Hibridación De La Metaheurística De La Mejor Búsqueda Armónica Global Y El Algoritmo Basado En Grafos Lexrank.

Martín Mateos, F. J., & Ruiz Reina, J. L. (2013). Procesamiento Del Lenguaje Natural Contenidos.

Matias Mendoza, G. A. (2016). Generación Automática De Resúmenes Independientes Del Lenguaje. Universidad Autónoma Del Estado De México Unidad.

Mcdonald, R. (2007). A Study Of Global Inference Algorithms In Multi-Document Summarization. Advances In Information Retrieval, 4425. Https://Doi.Org/10.1007/978-3-540-71496-5_51

Mendoza, G. A. M. (2013). Generación Automática De Resúmenes Usando Algoritmos Genéticos. Universidad Autónoma Del Estado De México.

Mihalcea, R. (2006). Random Walks On Text Structures. Springer, Berlín, Heidelberg, 978-3-540-32205–4, 249–262.

Neri Mendoza, V. (2019). Modelado De Posición De Oraciones Y Cobertura Mediante Un Algoritmo Genético Para La Generación Automáticade Resúmenes De Múltiples Documentos. 76.

Pardo, M. A. A., Ramos, M. A., Rodríguez, C. G., Gil, J. G., & Ferro, J. V. (2012). La Enseñanza Del Procesamiento Del Lenguaje Natural En Facultades De Informática Y Filología. May 2014.

Rojas Sánchez, J. M. (2016). Evaluación De Herramientas Comerciales Y Métodos Del Es Tado Del Arte Para La Generación De Resúmenes En Idioma Ruso. Universidad Autónoma Del Estado De México.

Rojas Simón, J. (2017). Cálculo De Topline Para La Generación Automática De Resúmenes Usando Algoritmos Genéticos. Universidad Autónoma Del Estado De México.

Rojas Simón, J., Ledeneva, Y., & García-Hernández, R. A. (2018). Calculating The Upper Bounds For Multi-Document Summarization Using Genetic Algorithms. Computacion Y Sistemas, 22(1), 11–26. Https://Doi.Org/10.13053/Cys-22-1-2903

Sánchez Arteche, A., Suárez, F., & Chávez Maya, M. A. (2012). Cronología De La Escritura, La Lectura Y El Libro (Consejo Editorial De La Administración Públicaestatal (Ed.)). Http://Servicio.Bc.Uc.Edu.Ve/Educacion/Eduweb/V11n1/Art04.Pdf

Sidorov, G. (2013). Construcción No Lineal De N-Gramas En La Lingüística Computacional (Sociedad Mexicana De Inteligenciaartificial (Ed.); Primera Ed).

Sierra Martínez, G., & Cuétara Priede, J. (2015). Lingüística Computacional En México (F. De F. Y L. Instituto De Ingeniería, Unam (Ed.); Primera Ed).

Simón, J. R., Ledeneva, Y., & García-Hernández, R. A. (2018). Calculating The Significance Of Automatic Extractive Text Summarization Using A Genetic Algorithm. Journal Of Intelligent And Fuzzy Systems, 35(1), 293–304. Https://Doi.Org/10.3233/Jifs-169588

Torres, C., & Medina, Z. (2013). Procesamiento Del Lenguaje Natural , Un Reto De La Inteligencia Artificial Natural Language Processing.

Valderrama Vilca, G. C. (2017). Generación Automática De Resúmenes Abstractivos Mono Documento Utilizando Análisis Semántico Y Del Discurso. Pontificia Universidad Católica Del Perú.

Vázquez, E., Arnulfo García-Hernández, R., & Ledeneva, Y. (2018). Sentence Features Relevance For Extractive Text Summarization Using Genetic Algorithms. Journal Of Intelligent And Fuzzy Systems, 35(1), 353–365. Https://Doi.Org/10.3233/Jifs-169594

Vázquez, E., Ledeneva, Y., & García-Hernández, R. A. (2019). Learning Relevant Models Using Symbolic Regression For Automatic Text Summarization. Computacion Y Sistemas, 23(1), 127–141. Https://Doi.Org/10.13053/Cys-23-1-2921

Vilchis Sepúlveda, S. A., & Ledeneva, Y. (2019). Evaluación De Resúmenes Automáticos Con Y Sin Resúmenes De Referencia Para El Idioma Inglés. Research In Computing Science, 148(7), 241–252. Https://Doi.Org/10.13053/Rcs-148-7-18

Villatoro-Tello, E., Villaseñor-Pineda, L., Montes-Y-Gómez, M., & Pinto-Avendaífot, D. (2009). Multi-Document Summarization Based On Locally Relevant Sentences. 8th Mexican International Conference On Artificial Intelligence - Proceedings Of The Special Session, Micai 2009, November, 87–91. Https://Doi.Org/10.1109/Micai.2009.10

Takamura, H., & Okumura, M. (2009). Text summarization model based on maximum coverage problem and its variant. Retrieved from https://dl.acm.org/citation.cfm?id=1609154

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), 513- 523.

Ortiz, R., Pinto, D., Tovar, M., & Jiménez-Salazar, H. (2010). BUAP: An unsupervised approach to automatic keyphrase extraction from scientific articles. Proceedings of the 5th International Workshop on Semantic Evaluation.

Minakshi Tomer, Manoj Kumar.(2021). Multi-document Extractive Text Summarization Based on Firefly Algorithm.

Anexos

En este apartado se muestra un conjunto de noticias que corresponden al corpus DUC2002, las secuencias frecuentes maximales extraídas de la misma colección asimismo como algunos de los resúmenes extraídos con diferentes configuraciones y los resultados dados por el sistema ROUGE-N.

1. Documentos de la colección D061J

1.1 Documento AP880911-0016

Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas.
The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph.
``There is no need for alarm," Civil Defense Director Eugenio Cabral said in a television alert shortly before midnight Saturday.
Cabral said residents of the province of Barahona should closely follow Gilbert's movement.
An estimated 100,000 people live in the province, including 70,000 in the city of Barahona, about 125 miles west of Santo Domingo.
Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night.
The National Hurricane Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo.
The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westward at 15 mph with a ``broad area of cloudiness and heavy weather" rotating around the center of the storm.
The weather service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday.
Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds and up to 12 feet feet to Puerto Rico's south coast.
There were no reports of casualties.
San Juan, on the north coast, had heavy rains and gusts Saturday, but they subsided during the night.
On Saturday, Hurricane Florence was downgraded to a tropical storm and its remnants pushed inland from the U.S. Gulf Coast.
Residents returned home, happy to find little damage from 80 mph winds and sheets of rain.
Florence, the sixth named storm of the 1988 Atlantic storm season, was the second hurricane.
The first, Debby, reached minimal hurricane strength briefly before hitting the Mexican coast last month.

1.2 Documento AP880912-0095

Hurricane Gilbert, packing 110 mph winds and torrential rain, moved over this capital city today after skirting Puerto Rico, Haiti and the Dominican Republic.
There were no immediate reports of casualties.
Telephone communications were affected.
``Right now it's actually moving over Jamaica," said Bob Sheets, director of the National Hurricane Center in Miami.
``We've already had reports of 110 mph winds on the eastern tip.
``It looks like the eye is going to move lengthwise across that island, and they're going to bear the full brunt of this powerful hurricane," Sheets said.
Forecasters say Gilbert was expected to lash Jamaica throughout the day and was on track to later strike the Cayman Islands, a small British dependency northwest of Jamaica.
Meanwhile, Havana Radio reported today that 25,000 people were evacuated from Guantanamo Province on Cuba's southeastern coast as strong winds fanning out from Gilbert began brushing the island.

All Jamaica-bound flights were canceled at Miami International Airport, while flights from Grand Cayman, the main island of the three-island chain, arrived packed with frightened travelers.
``People were running around in the main lobby of our hotel (on Grand Cayman) like chickens with their heads cut off,'' said one vacationer who was returning home to California through Miami.
Hurricane warnings were posted for the Cayman Islands, Cuba and Haiti.
Warnings were discontinued for the Dominican Republic.
``All interests in the Western Caribbean should continue to monitor the progress of this dangerous hurricane,'' the service said, adding, ``Little change in strength is expected for the next several hours as the hurricane moves westward over Jamaica''.
The Associated Press' Caribbean headquarters in San Juan, Puerto Rico, was unable to get phone calls through to Kingston, where high winds and heavy rain preceding the storm drenched the capital overnight, toppling trees, causing local flooding and littering streets with branches.
Most Jamaicans stayed home, boarding up windows in preparation for the hurricane.
Some companies broadcast appeals for technicians and electricians to report to work.
The weather bureau predicted Gilbert's center, 140 miles southeast of Kingston before dawn, would pass south of Kingston and hit the southern parish of Clarendon.
Flash flood warnings were issued for the parishes of Portland on the northeast and St. Mary on the north.
The north coast tourist region from Montego Bay on the west and Ocho Rios on the east, far from the southern impact zone and separated by mountains, was expected only to receive heavy rain.
Officials urged residents in the higher risk areas along the south coast to seek higher ground.
``It's certainly one of the larger systems we've seen in the Caribbean for a long time,'' said Hal Gerrish, forecaster at the National Hurricane Center.
Forecasters at the center said the eye of Gilbert was 140 miles southeast of Kingston at dawn today.
Maximum sustained winds were near 110 mph, with tropical-storm force winds extending up to 250 miles to the north and 100 miles to the south.
Prime Minister Edward Seaga of Jamaica alerted all government agencies, saying Sunday night: ``Hurricane Gilbert appears to be a real threat and everyone should follow the instructions and hurricane precautions issued by the Office of Disaster Preparedness in order to minimize the danger''.
Forecasters said the hurricane had been gaining strength as it passed over the ocean after it dumped 5 to 10 inches of rain on the Dominican Republic and Haiti, which share the island of Hispaniola.
``We should know within about 72 hours whether it's going to be a major threat to the United States,'' said Martin Nelson, another meteorologist at the center.
``It's moving at about 17 mph to the west and normally hurricanes take a northward turn after they pass central Cuba''.
Cuba's official Prensa Latina news agency said a state of alert was declared at midday in the Cuban provinces of Guantanamo, Holguin, Santiago de Cuba and Granma.
In the report from Havana received in Mexico City, Prensa Latina said civil defense officials were broadcasting bulletins on national radio and television recommending emergency measures and providing information on the storm.
Heavy rain and stiff winds downed power lines and caused flooding in the Dominican Republic on Sunday night as the hurricane's center passed just south of the Barahona peninsula, then less than 100 miles from neighboring Haiti.
The storm ripped the roofs off houses and flooded coastal areas of southwestern Puerto Rico after reaching hurricane strength off the island's southeast Saturday night.
Flights were canceled Sunday in the Dominican Republic, where civil defense director Eugenio Cabral reported some flooding in parts of the capital of Santo Domingo and power outages there and in other southern areas .

1.3 Documento AP880912-0137

Hurricane Gilbert slammed into Kingston on Monday with torrential rains and 115 mph winds that ripped roofs off homes and buildings, uprooted trees and downed power lines.
No serious injuries were immediately reported in the city of 750,000 people, which was hit by the full force of the hurricane around noon.
For half an hour, the hurricane lashed the city, tearing branches from trees, blowing down fences and whipping paper through the air.
The National Weather Service reported heavy damage to Kingston's airport and aircraft parked on its fields.
The first shock let up as the eye of the storm moved across the city.

Skies brightened, the winds died down and people waited for an hour before the second blow of the hurricane arrived.
All Jamaica-bound flights were canceled at Miami International Airport.
Flights from the Cayman Islands, reportedly next in the path of the hurricane, arrived in Miami packed with travelers cutting short their vacations.
``People were running around in the main lobby of our hotel (on Grand Cayman Island) like chickens with their heads cut off," said one man.
A National Weather Service report said the hurricane was moving west at 17 mph with maximum sustained winds of 115 mph.
It said Jamaica would receive up to 10 inches of rain that would cause flash floods and mud slides.
``Right now it's actually moving over Jamaica," said Bob Sheets, director of the National Hurricane Center in Miami.
``It looks like the eye is going to move lengthwise across that island, and they're going to bear the full brunt of this powerful hurricane," he said.
Gilbert reached Jamaica after skirting southern Puerto Rico, Haiti and the Dominican Republic.
Hurricane warnings were issued Monday for the south coast of Cuba east of Camaguey, the Cayman Islands, and Haiti, while warnings were discontinued for the Dominican Republic.
High winds and heavy rain preceding the storm drenched Kingston overnight, toppling trees, causing local flooding and littering streets with branches.
Most of Jamaica's 2.3 million people stayed home, boarding up windows in preparation for the hurricane.
The popular north coast resort area, on the other side of the mountains, was expected to receive heavy rain but not as much damage from the hurricane as the south coast, where officials urged residents to seek higher ground.
Havana Radio, meanwhile, reported Monday that 25,000 people were evacuated from coastal areas in Guantanamo Province on the nation's southeastern coast as Gilbert's winds and rain began to brush the island.
In Washington, the Navy reported its bases at Guantanamo Bay, Cuba, and Roosevelt Roads, Puerto Rico, had taken various precautionary steps but appeared to be safe from the brunt of the hurricane.
Lt. Ken Ross, a spokesman, said the Navy station at Guantanamo reported that as of 2:30 p.m. EDT, the brunt of the storm appeared to be passing southeastern Cuba.
``They have reported maximum winds of 25 knots and gusts up to 50 knots," said Ross.
``But there are no reports of injuries or damage".
The spokesman said earlier in the day, Guantanamo had moved to ``Condition Two," meaning electrical power usage was cut back to only essential uses and ``all non-essential personnel sent to their barracks".
The storm also skirted Puerto Rico without causing any damage to military facilities, Ross said.
Sheets said Gilbert was expected next to sweep over the Cayman Islands, on its westward track, and in two to three days veer northwest into the southern Gulf of Mexico.
Residents of the neighboring Caymans, a British dependency to the northwest, were urged to ``rush all preparatory actions".
The National Weather Service warned that the Caymans could expect high waters and large waves ``which may undermine buildings along the beaches".
``All interests in the Western Caribbean should continue to monitor the progress of this dangerous hurricane," the service advised.
Forecaster Hal Gerrish on Sunday described Gilbert ``certainly one of the larger systems we've seen in the Caribbean for a long time."

1.4 Documento WSJ880912-0064

Hurricane Gilbert swept toward Jamaica yesterday with 100-mile-an-hour winds, and officials issued warnings to residents on the southern coasts of the Dominican Republic, Haiti and Cuba.
The storm ripped the roofs off houses and caused coastal flooding in Puerto Rico.
In the Dominican Republic, all domestic flights and flights to and from Puerto Rico and Miami were canceled.
Forecasters said the hurricane was gaining strength as it passed over the ocean and would dump heavy rain on the Dominican Republic and Haiti as it moved south of Hispaniola, the Caribbean island they share, and headed west.
"It's still gaining strength.
It's certainly one of the larger systems we've seen in the Caribbean for a long time," said Hal Gerrish, forecaster at the National Hurricane Center in Coral Gables, Fla.

At 3 p.m. EDT, the center of the hurricane was about 100 miles south of the Dominican Republic and 425 miles east of Kingston, Jamaica.
The hurricane was moving west at about 15 mph and was expected to continue this motion for the next 24 hours.
Forecasters said the hurricane's track would take it about 50 miles south of southwestern Haiti.
The hurricane center said small craft in the Virgin Islands and Puerto Rico should remain in port until conditions improve.
The forecasters said the Dominican Republic would get as much as 10 inches of rain yesterday, with similar amounts falling in Haiti last night and tonight.
Hurricane warnings were issued for the south coast of Haiti and Cuba by their respective governments.
In Jamaica, the government issued a hurricane watch for the entire island.
Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night.
In Puerto Rico, besides tearing off several roofs, the storm caused coastal flooding and brought down power lines and trees along roads and highways in the west and southwestern regions.
Three people were injured in Guayanilla, Puerto Rico, when a tree fell on their vehicle as they traveled along Route 97, police reported.
Four policemen stationed on Mona Island, between Puerto Rico and the Dominican Republic, were stranded as a result of the weather .

1.5 Documento AP880915-0003

Hurricane Gilbert, one of the strongest storms ever, slammed into the Yucatan Peninsula Wednesday and leveled thatched homes, tore off roofs, uprooted trees and cut off the Caribbean resorts of Cancun and Cozumel.
Looters roamed the streets of Cancun, stealing from stores whose windows were blown away.
Huge waves battered the beach resorts and thousands were evacuated.
Despite the intensity of the onslaught and the ensuing heavy flooding, officials reported only two minor injuries.
The storm killed 19 people in Jamaica and five in the Dominican Republic before moving west to Mexico.
Prime Minister Edward Seaga of Jamaica said Wednesday the storm destroyed an estimated 100,000 of Jamaica's 500,000 homes when it throttled the island Monday.
The Jamaican Embassy reported earlier that 500,000 of the nation's 2.3 million people were homeless.
In Cancun, amateur radio operators said an oil tanker from the fleet of the government oil monopoly Pemex, the Lazaro Cardenas, had run aground on the beach in the Cancun hotel zone.
Pemex officials however said all their vessels were secure.
Army officials in Mexico City said about 35,000 people were evacuated from Cancun, but Cancun Mayor Jose Sanchez Zapata said about 11,000 fled.
More than 120,000 people on the northeast Yucatan coast were evacuated, the Yucatan state government said.
The eye of the storm passed over Cozumel and Cancun with howling winds clocked at 160 mph at about 8 a.m. EDT.
The winds bent and toppled utility poles and uprooted slum dwellings.
Ham radio operators around Cancun said Gilbert knocked down a radio and television communications tower, uprooted trees and blew the roofs off buildings.
Floods prevented officials from reaching the hotel zone in Cancun and there were no relief efforts under way by late Wednesday.
Shelters had little or no food, water or blankets and power was out.
``We can't do it yet.
The wind would blow them away,'' said an army official at city hall who declined to give his name.
Bands of 25-30 youths roamed the streets of Cancun Wednesday, looting stores with shattered windows, said Alfredo Moro Sanchez, investigative coordinator of the Quintana Roo state judicial police.
He said he asked for army troops to halt the looting but none had arrived by late Wednesday.
About 150 tourists took refuge at the Cancun city hall.
Honeymooner Sheila Fournier of Long Island said she and her husband, Pete, had been evacuated from the Villas Playa Hotel.
``It had an ocean-front view _ which is now washed away,'' she said.
Officials were checking low-lying areas of Cancun for stranded residents.
There was heavy damage visible to the humble wood and thatch homes typical of the Yucatan.
``There were some people who did not want to leave yesterday,'' the army official said.

``It was something new.
They didn't know what a cyclone was".
At midnight EDT Gilbert was centered near latitude 21.5 north, longitude 90.2 west and approaching the north coast of Yucatan, about 60 miles east-northeast of the provincial capital, Merida, the National Hurricane Center in Coral Gables, Fla., said.
The storm was about 550 miles southeast of Brownsville, Texas, the center said in a statement.
Gilbert was moving west-northwest at 15 mph and winds had decreased to 125 mph.
The Mexican National Weather Service reported winds gusting as high as 218 mph earlier Wednesday with sustained winds of 179 mph.
Earlier Wednesday Gilbert was classified as a Category 5 storm, the strongest ad deadliest type of hurricane.
Such storms have maximum sustained winds greater than 155 mph and can cause catastrophic damage.
By Wednesday night the National Hurricane Center downgraded it to a Category 4, but center director Bob Sheets said:``There's no question it'll strengthen again once it comes off the Yucatan Peninsula and gets back in open water".
As Gilbert moved away from the Yucatan Peninsula Wednesday night, the hurricane formed a double eye, two concentric circles of thunderstorms often characteristic of a strong storm that has crossed land and is moving over the water again.
One eye was about eight miles wide, and the second about 25 miles wide, said hurricane center meteorologist Jesse Moore.
``This is one of the features that we expected to see as the hurricane moved back over the water, and we do expect intensification," he said.
Only two Category 5 hurricanes have hit the United States _ the 1935 storm that killed 408 people in Florida and Hurricane Camille that devastated the Mississippi coast in 1969, killing 256 people.
Oil companies evacuated thousands of workers from rigs in the Gulf of Mexico.
The peninsula ports of Campeche, Celestum, Progreso, Sinzal, Ucaltepen, Tel-Chac, Cancun, Puerto Morelos, and Ciudad del Carmen were closed, the government news agency Notimex said.
Airports in the region were closed.
``The sound of the wind outside is horrible," said receptionist Pablo Torres at Cancun's Hotel Carrillos as the storm approached.
The National Hurricane Center said a hurricane watch was in effect on the Texas coast from Brownsville to Port Arthur and along the coast of northeast Mexico from Tampico north.
In Mexico City, the National Civil Defense System said it lost telephone contact with Cancun and Cozumel at about 8 a.m. EDT.
Public buildings in Cancun were used as shelters, said Cecila Lavalle, a spokesman for Quintana Roo state government in Chetumal, 155 miles southeast of Cozumel.
Jennie Valdez, a U.S. consular representative in Cancun, said she did not know how many tourists were in Cancun, but government figures estimate 40,000 to 65,000 monthly visitors.
Hurricane warnings were in effect for the entire Yucatan Peninsula and widespread evacuations were reported.
Winds destroyed at least 100 homes in slums of Valladolid, a town of about 45,000 about 100 miles from Cozumel, Alberto Pol, a judicial police officer, said by telephone.
The National Hurricane Center said Gilbert was the most intense storm on record in terms of barometric pressure.
It was measured at 26.31 inches, breaking the 26.35 inches recorded for the 1935 hurricane that devastated the Florida Keys.
``That's the lowest pressure ever measured in the Western Hemisphere," said forecaster Mark Zimmer.
On Sunday, Monday and Tuesday, Gilbert pounded the Dominican Republic, Jamaica and the Cayman Islands.
Seaga said Jamaica would need $8 billion in aid.
Officials in the Dominican Republic, sideswiped Sunday by the storm, reported five dead.
Gilbert also buffeted the Cayman Islands, but no deaths were reported .

1.6 Documento AP880916-0060

Hurricane Gilbert's growth from a harmless low pressure zone off Africa to a ferocious killer in the Gulf of Mexico was fueled by a combination of heat, moisture and wind that baffles forecasters.
``It's a matter of getting everything together in the right place in the right time," Gil Clark of the National Hurricane Center said Thursday.

``It doesn't happen very often.
How it develops, we don't know".
Gilbert came to the attention of center forecasters Sept. 3 as a dry low pressure trough moving west out of Africa.
``We get 50 or 60 of these off Africa every summer.
About one of six develop," said Clark.
By Sept. 8, the system became a depression.
It reached tropical storm status by Saturday and a hurricane Sunday.
A tropical wave becomes a depression when winds start swirling.
When sustained winds reach 39 mph, the system becomes a named tropical storm.
It reaches hurricane status when sustained winds hit 74 mph.
Why Gilbert organized and strengthened while other systems didn't ``is a mystery more or less," said University of Miami meteorology Professor Rainer Bleck.
``The first part of the summer we were biting our nails, wondering why these (other) disturbances didn't develop," he said Thursday.
``That's something meteorologists would like to know more about".
But the scientists do know what fuels a budding storm once development begins.
And they know that development is sparked when winds converge, and that growth is affected by time and place.
``If that happens in an area where there's plenty of moisture in the lower atmosphere (the bottom 10,000 feet or so), this convergence may lead to upward motion and cloud formation," Bleck said.
``If clouds form, the heat of condensation in the clouds occasionally provides `positive feedback' to the convergence pattern.
That strengthens it," he said, adding that storms can begin budding only off the equator because of the Earth's rotation.
Eventually, a vortex is created.
``Any time you contract an air mass, they will start spinning.
That's what makes the tornadoes, hurricanes and blizzards, those winter storms," Bleck said.
Hurricanes ``are useful to the climate machine.
Their primary role is to transport heat from the lower to the upper atmosphere," he said.
``The sun puts energy into the water, the top of the oceans and lowest part of the atmosphere.
That has to be distributed from the bottom to higher levels of the atmosphere".
When the depression that would become Gilbert neared Barbados, warm Atlantic waters nurtured it.
``This time of year in the northwest Caribbean is best for development," Clark said.
``If you get a storm in this area in September, when the water's warmest, it can just explode.
This is where Camille formed and exploded," referring to the 1969 storm that slammed into the Gulf Coast.
``It is an exciting thing to watch.
If you're on the beach watching the storm surge, it's a different story," he said.
The hurricane center said Gilbert was the most intense storm on record in terms of barometric pressure, measured at 26.13 inches Tuesday night.
That broke the 26.35 inches of the 1935 hurricane that devastated the Florida Keys .

2. Lista de la SFM’s de la colección D061J

2.1 SFM’s con longitud 3

gilbert,swept,toward
south,coast,to
heavy,rains,and
the,storm,was
said,in,a
residents,of,the
of,the,atmosphere
in,the,right
of,santo,domingo
a,hurricane,watch
miles,south,of
south,of,the

gilbert,was,expected
gilbert,was,moving
at,@one@five,mph
@one@five,mph,and
the,center,said
the,virgin,islands
there,were,no
no,reports,of
reports,of,casualties
on,the,north
on,the,northeast
on,the,beach
the,north,coast
mph,winds,and
haiti,and,cuba
caused,coastal,flooding
in,puerto,rico
the,hurricane,s
over,the,water
it,s,a
at,the,center
about,@one@ten@ten,miles
@one@ten@ten,miles,from
was,expected,to
for,the,next
for,the,entire
power,lines,and
the,west,and
slammed,into,the
yucatan,peninsula,and
uprooted,trees,and
cancun,and,cozumel
of,jamaica,s
the,nation,s
in,mexico,city
radio,and,television
by,late,wednesday
quintana,roo,state
he,said,the
mph,earlier,wednesday
maximum,sustained,winds
wednesday,night,the
as,the,hurricane
the,united,states
were,closed,the
measured,at,@two@six
the,cayman,islands
said,jamaica,would
when,sustained,winds
part,of,the
to,be,a
downed,power,lines
mountains,was,expected
officials,urged,residents
guantanamo,province,on
appeared,to,be
@one@one@ten,mph,winds
miles,to,the

2.2 SFM's con longitud 4

the,dominican,republic,all
with,sustained,winds,of
of,the,hurricane,arrived
of,the,dominican,republic
an,estimated,@one@ten@ten,@ten@ten@ten

in,the,dominican,republic
in,the,city,of
the,national,weather,service
puerto,rico,and,the
national,weather,service,reported
the,center,of,the
p,m,edt,the
forecasters,said,the,hurricane
said,the,eye,of
said,the,hurricane,was
the,hurricane,center,said
hurricane,warnings,were,issued
the,yucatan,peninsula,wednesday
the,@two@six,@three@five,inches
to,receive,heavy,rain
to,seek,higher,ground
s,southeastern,coast,as
the,brunt,of,the

2.3 SFM's con longitud 5

civil,defense,director,eugenio,cabral
in,the,gulf,of,mexico
night,the,national,hurricane,center
the,national,hurricane,center,said
@one@four@ten,miles,southeast,of,kingston
in,san,juan,puerto,rico
for,the,south,coast,of
warnings,were,issued,for,the
roamed,the,streets,of,cancun
s,@two,@three,million,people
the,eye,of,the,storm
to,@one@ten,inches,of,rain

2.4 SFM's con longitud 6

the,hurricane,was,moving,west,at
at,about,@eight,a,m,edt
prime,minister,edward,seaga,of,jamaica

2.5 SFM's con longitud 7

puerto,rico,haiti,and,the,dominican,republic
rain,on,the,dominican,republic,and,haiti
warnings,were,discontinued,for,the,dominican,republic
the,@one@nine@three@five,hurricane,that,devastated,the,florida

that,@two@five,@ten@ten@ten,people,were,evacuated,from

2.6 SFM's con longitud 8

the,storm,ripped,the,roofs,off,houses,and
the,national,hurricane,center,in,coral,gables,fla
gaining,strength,as,it,passed,over,the,ocean

2.7 SFM's con longitud 9

high,winds,and,heavy,rain,preceding,the,storm,drenched
like,chickens,with,their,heads,cut,off,said,one

2.8 SFM's con longitud mayor igual a 10

all,jamaica,bound,flights,were,canceled,at,miami,international,airport
stayed,home,boarding,up,windows,in,preparation,for,the,hurricane
1 sfm's of size=[12]----------
overnight,toppling,trees,causing,local,flooding,and,littering,streets,with,branches,most
1 sfm's of size=[14]----------
people,were,running,around,in,the,main,lobby,of,our,hotel,on,grand,cayman
1 sfm's of size=[15]----------
tropical,storm,gilbert,formed,in,the,eastern,caribbean,and,strengthened,into,a,hurricane,saturday,night
1 sfm's of size=[16]----------
hurricane,center,said,gilbert,was,the,most,intense,storm,on,record,in,terms,of,barometric,pressure
1 sfm's of size=[18]----------
all,interests,in,the,western,caribbean,should,continue,to,monitor,the,progress,of,this,dangerous,hurricane,the,service
1 sfm's of size=[19]----------
right,now,it,s,actually,moving,over,jamaica,said,bob,sheets,director,of,the,national,hurricane,center,in,miami
1 sfm's of size=[26]----------
it,looks,like,the,eye,is,going,to,move,lengthwise,across,that,island,and,they,re,going,to,bear,the,full,brunt,of,this,powerful,hurricane
1 sfm's of size=[27]----------

it,s,certainly,one,of,the,larger,systems,we,ve,seen,in,the,caribbean,for,a,long,time,said,hal,gerrish,forecaster,at,the,national,hurricane,center

3 Resúmenes del método de la colección D062J

3.1 Configuración (M, F, TextRank, 5baseline+best)

hurricane gilbert swept toward the dominican republic sunday, and the civil defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas.
hurricane gilbert, packing 110 mph winds and torrential rain, moved over this capital city today after skirting puerto rico, haiti and the dominican republic.
hurricane gilbert slammed into kingston on monday with torrential rains and 115 mph winds that ripped roofs off homes and buildings, uprooted trees and downed power lines.
hurricane gilbert, one of the strongest storms ever, slammed into the yucatan peninsula wednesday and leveled thatched homes, tore off roofs, uprooted trees and cut off the caribbean resorts of cancun and cozumel.
hurricane gilbert's growth from a harmless low pressure zone off africa to a ferocious killer in the gulf of mexico was fueled by a combination of heat, moisture and wind that baffles forecasters.
hurricane gilbert, packing 110 mph winds and torrential rain, moved over this capital city today after skirting puerto rico, haiti and the dominican republic.
meanwhile, havana radio reported today that 25,000 people were evacuated from guantanamo province on cuba's southeastern coast as strong winds fanning out from gilbert began brushing the island.
``right now it's actually moving over jamaica,'' said bob sheets, director of the national hurricane center in miami.
prime minister edward seaga of jamaica alerted all government agencies, saying sunday night: ``hurricane gilbert appears to be a real threat and everyone should follow the instructions and hurricane precautions issued by the office of disaster preparedness in order to minimize the danger''.
forecasters say gilbert was expected to lash jamaica throughout the day and was on track to later strike the cayman islands, a small british dependency northwest of jamaica.

3.2 Configuración (M, F, TextRank, best)

hurricane gilbert, packing 110 mph winds and torrential rain, moved over this capital city today after skirting puerto rico, haiti and the dominican republic.
meanwhile, havana radio reported today that 25,000 people were evacuated from guantanamo province on cuba's southeastern coast as strong winds fanning out from gilbert began brushing the island.
``right now it's actually moving over jamaica,'' said bob sheets, director of the national hurricane center in miami.
prime minister edward seaga of jamaica alerted all government agencies, saying sunday night: ``hurricane gilbert appears to be a real threat and everyone should follow the instructions and hurricane precautions issued by the office of disaster preparedness in order to minimize the danger''.
forecasters say gilbert was expected to lash jamaica throughout the day and was on track to later strike the cayman islands, a small british dependency northwest of jamaica.
hurricane gilbert's growth from a harmless low pressure zone off africa to a ferocious killer in the gulf of mexico was fueled by a combination of heat, moisture and wind that baffles forecasters.
heavy rain and stiff winds downed power lines and caused flooding in the dominican republic on sunday night as the hurricane's center passed just south of the barahona peninsula, then less than 100 miles from neighboring haiti.
the associated press' caribbean headquarters in san juan, puerto rico, was unable to get phone calls through to kingston, where high winds and heavy rain preceding the storm drenched the capital overnight, toppling trees, causing local flooding and littering streets with branches.
all jamaica-bound flights were canceled at miami international airport, while flights from grand cayman, the main island of the three-island chain, arrived packed with frightened travelers.
flash flood warnings were issued for the parishes of portland on the northeast and st. mary on the north.

4 Resultados de los experimentos

4.1 Experimento (M, F, TextRank, 5baseline+best)

---

1 ROUGE-1 Average_R: 0.48798 (95%-conf.int. 0.46823 - 0.50796)
1 ROUGE-1 Average_P: 0.48847 (95%-conf.int. 0.46799 - 0.50905)
1 ROUGE-1 Average_F: 0.48813 (95%-conf.int. 0.46761 - 0.50846)

---

1 ROUGE-2 Average_R: 0.24368 (95%-conf.int. 0.21271 - 0.27486)
1 ROUGE-2 Average_P: 0.24379 (95%-conf.int. 0.21267 - 0.27481)
1 ROUGE-2 Average_F: 0.24369 (95%-conf.int. 0.21255 - 0.27478)

---

1 ROUGE-SU4 Average_R: 0.27780 (95%-conf.int. 0.25000 - 0.30644)
1 ROUGE-SU4 Average_P: 0.27796 (95%-conf.int. 0.24963 - 0.30636)
1 ROUGE-SU4 Average_F: 0.27783 (95%-conf.int. 0.24983 - 0.30622)

4.2 Experimento (M, F, TextRank, best)

---

1 ROUGE-1 Average_R: 0.43856 (95%-conf.int. 0.42189 - 0.45536)
1 ROUGE-1 Average_P: 0.45154 (95%-conf.int. 0.43483 - 0.46876)
1 ROUGE-1 Average_F: 0.44486 (95%-conf.int. 0.42793 - 0.46178)

---

1 ROUGE-2 Average_R: 0.16510 (95%-conf.int. 0.14293 - 0.18883)
1 ROUGE-2 Average_P: 0.16980 (95%-conf.int. 0.14707 - 0.19390)
1 ROUGE-2 Average_F: 0.16738 (95%-conf.int. 0.14489 - 0.19149)

---

1 ROUGE-SU4 Average_R: 0.20610 (95%-conf.int. 0.18639 - 0.22648)
1 ROUGE-SU4 Average_P: 0.21213 (95%-conf.int. 0.19222 - 0.23261)
1 ROUGE-SU4 Average_F: 0.20903 (95%-conf.int. 0.18905 - 0.22973)