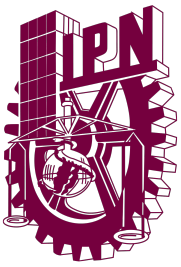

Instituto Politécnico Nacional

Escuela Superior de Cómputo

Práctica 3: Extracción automática de texto

R E P O R T E

Materia:

Tecnologías de Lenguaje Natural

presenta:

Luis Fernando Rodríguez-Domínguez

Profesor:

Ituriel Enrique Flores Estrada

Índice general

Capítulo 1

Introducción

El presente reporte documenta la realización de la Práctica 3 correspondiente a la asignatura de Tecnologías de Lenguaje Natural. El objetivo principal de esta práctica fue implementar, ejecutar y comparar diversas técnicas de resumen automático de tipo extractivo. El resumen extractivo se enfoca en seleccionar y concatenar las oraciones o frases más significativas directamente del texto original para formar un resumen coherente.

Para llevar a cabo este estudio comparativo, se utilizó como corpus de documentos las tres primeras cartas contenidas en la novela clásica "Frankenstein"de Mary Shelley, obtenida del Proyecto Gutenberg [1]. Cada carta fue tratada como un documento individual sobre el cual se aplicaron los algoritmos de resumen.

Antes de aplicar las técnicas de resumen, se realizó un proceso de normalización de texto sobre cada carta. Este paso es fundamental para reducir el ruido y mejorar la calidad de la entrada a los algoritmos, e incluyó tareas como la tokenización, la conversión a minúsculas y la eliminación de palabras vacías (stopwords) y caracteres no alfanuméricos utilizando NLTK.

Se implementaron y evaluaron seis algoritmos distintos, abarcando desde enfoques clásicos basados en estadísticas de términos hasta métodos más modernos que emplean modelos de lenguaje avanzados:

- TF-IDF (utilizando NLTK y Scikit-learn) [2].
- Frecuencia de palabras normalizada [3–5].
- RAKE (Rapid Automatic Keyword Extraction, con NLTK) [6].
- TextRank (implementado a través de la biblioteca Sumy [7, 8]).
- LSA (Latent Semantic Analysis, también con Sumy [7, 9]).
- BERT (Bidirectional Encoder Representations from Transformers, usando la biblioteca `bert-extractive-summarizer`) [10, 11].

Para cada algoritmo y cada carta, se generó un resumen compuesto por las cuatro frases consideradas más representativas. Además de la generación de los resúmenes, un aspecto clave de la práctica fue la medición y comparación de los tiempos de ejecución requeridos por cada algoritmo en el equipo de cómputo utilizado.

Capítulo 2

Desarrollo

2.1. Generación del Cuerpo de Documentos

El primer paso consistió en obtener el texto base para el análisis. Se descargó la novela "Frankenstein"de Mary Shelley desde el Proyecto Gutenberg (Ebook #84). De este texto completo, se requería extraer las tres primeras cartas para conformar un corpus de tres documentos distintos.

Para realizar esta extracción, se implementó la función `extraer_cartas_individuales` en Python. Esta función lee el archivo de texto (denominado `frankenstein.txt`), busca los marcadores "Letter 1", "Letter 2", y "Letter 3"que indican el inicio de cada carta, y separa el contenido correspondiente a cada una. Se asumió que el texto estaba codificado en UTF-8. La Figura 2.1 muestra el log de ejecución de esta función, confirmando la extracción exitosa.

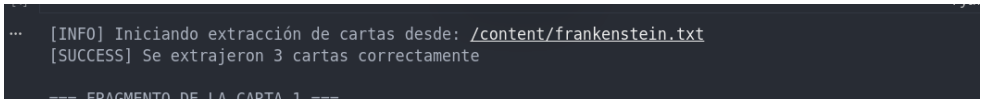

Figura 2.1: Log de la ejecución de la extracción de cartas.

Para verificar que la extracción fue correcta, se imprimieron las cartas completas. Las Figuras 2.2, 2.3 y 2.4 muestran el contenido de cada una de las cartas extraídas. (Nota: Por brevedad, en el notebook se mostraron fragmentos, pero aquí se asume que se capturó/mostrará el contenido relevante).

```
=== FRAGMENTO DE LA CARTA 1 ===
Letter 1

_To Mrs. Saville, England._


St. Petersburgh, Dec. 11th, 17—.


You will rejoice to hear that no disaster has accompanied the
commencement of an enterprise which you have regarded with such evil
forebodings. I arrived here yesterday, and my first task is to assure
my dear sister of my welfare and increasing confidence in the success
of my undertaking.

I am already far north of London, and as I walk in the streets of
Petersburgh, I feel a cold northern breeze play upon my cheeks, which
braces my nerves and fills me with delight. Do you understand this
feeling? This breeze, which has travelled from the regions towards
which I am advancing, gives me a foretaste of those icy climes.
Inspirited by this wind of promise, my daydreams become more fervent
and vivid. I try in vain to be persuaded that the pole is the seat of
frost and desolation; it ever presents itself to my imagination as the
region of beauty and delight. There, Margaret, the sun is for ever
visible, its broad disk just skirting the horizon and diffusing a
perpetual splendour. There—for with your leave, my sister, I will put
some trust in preceding navigators—there snow and frost are banished;
```

Figura 2.2: Contenido de la Carta 1.

```
=== FRAGMENTO DE LA CARTA 2 ===
Letter 2

_To Mrs. Saville, England._

Archangel, 28th March, 17—.


How slowly the time passes here, encompassed as I am by frost and snow!
Yet a second step is taken towards my enterprise. I have hired a
vessel and am occupied in collecting my sailors; those whom I have
already engaged appear to be men on whom I can depend and are certainly
possessed of dauntless courage.

But I have one want which I have never yet been able to satisfy, and the
absence of the object of which I now feel as a most severe evil, I have no
friend, Margaret: when I am glowing with the enthusiasm of success, there
will be none to participate my joy; if I am assailed by disappointment, no
one will endeavour to sustain me in dejection. I shall commit my thoughts
to paper, it is true; but that is a poor medium for the communication of
feeling. I desire the company of a man who could sympathise with me, whose
eyes would reply to mine. You may deem me romantic, my dear sister, but I
bitterly feel the want of a friend. I have no one near me, gentle yet
courageous, possessed of a cultivated as well as of a capacious mind, whose
tastes are like my own, to approve or amend my plans. How would such a
friend repair the faults of your poor brother! I am too ardent in execution
```

Figura 2.3: Contenido de la Carta 2.

```
=== FRAGMENTO DE LA CARTA 3 ===
Letter 3

_To Mrs. Saville, England._

July 7th, 17–.


My dear Sister,

I write a few lines in haste to say that I am safe—and well advanced
on my voyage. This letter will reach England by a merchantman now on
its homeward voyage from Archangel; more fortunate than I, who may not
see my native land, perhaps, for many years. I am, however, in good
spirits: my men are bold and apparently firm of purpose, nor do the
floating sheets of ice that continually pass us, indicating the dangers
of the region towards which we are advancing, appear to dismay them. We
have already reached a very high latitude; but it is the height of
summer, and although not so warm as in England, the southern gales,
which blow us speedily towards those shores which I so ardently desire
to attain, breathe a degree of renovating warmth which I had not
expected.

No incidents have hitherto befallen us that would make a figure in a
letter. One or two stiff gales and the springing of a leak are
accidents which experienced navigators scarcely remember to record, and
I shall be well content if nothing worse happen to us during our voyage.
```

Figura 2.4: Contenido de la Carta 3.

2.2. Normalización de Textos

Antes de aplicar los algoritmos de resumen, es crucial normalizar el texto de cada carta. La normalización reduce la variabilidad léxica y elimina elementos no informativos (ruido), permitiendo que los algoritmos se enfoquen en el contenido semántico relevante y mejoren la calidad de los resúmenes generados.

Se implementó la función `normalizar_texto`, la cual aplica las siguientes técnicas justificadas a continuación:

- **Tokenización en oraciones:** El resumen extractivo se basa en seleccionar oraciones completas. Se utilizó `nltk.sent_tokenize` para dividir cada carta en sus oraciones constituyentes.
- **Tokenización en palabras:** Cada oración se segmentó en palabras individuales usando `nltk.word_tokenize`.
- **Conversión a minúsculas:** Para tratar palabras como "Thez "the"de manera idéntica, todo el texto se convirtió a minúsculas.
- **Eliminación de stopwords:** Palabras comunes como .ª", "the", ïs", ïn", que aportan poco significado discriminativo, fueron eliminadas utilizando la lista de stopwords en inglés de NLTK.
- **Filtrado de no alfanuméricos:** Se eliminaron signos de puntuación y otros caracteres no alfanuméricos, ya que generalmente no son relevantes para el análisis de frecuencia o semántico a nivel de palabra. Se conservaron únicamente los tokens que contenían caracteres alfanuméricos (`isalnum()`).

La función devuelve tanto la lista de oraciones originales (necesarias para construir el resumen final) como la lista de oraciones normalizadas (utilizadas como entrada para la mayoría de los algoritmos). La ejecución de esta función se muestra en la Figura 2.5.

```
[INFO] Iniciando normalización de texto (6848 caracteres)
[SUCCESS] Normalización completada en 0.0723 segundos. 49 oraciones procesadas.
[INFO] Iniciando normalización de texto (7358 caracteres)
[SUCCESS] Normalización completada en 0.0143 segundos. 49 oraciones procesadas.
[INFO] Iniciando normalización de texto (1707 caracteres)
[SUCCESS] Normalización completada en 0.0046 segundos. 19 oraciones procesadas.
```

Figura 2.5: Log de la normalización de la Carta 1.

La Figura 2.6 ilustra el efecto de la normalización sobre las primeras oraciones de la Carta 1.

```
Oración original: You will rejoice to hear that no disaster has accompanied the
commencement of an enterprise which you have regarded with such evil
forebodings.
Oración normalizada: rejoice hear disaster accompanied commencement enterprise regarded evil forebodings

Oración original: I arrived here yesterday, and my first task is to assure
my dear sister of my welfare and increasing confidence in the success
of my undertaking.
Oración normalizada: arrived yesterday first task assure dear sister welfare increasing confidence success undertaking
```

Figura 2.6: Ejemplo de oraciones originales vs. normalizadas (Carta 1).

2.3. Resumen Automático Extractivo

El núcleo de la práctica consistió en generar un resumen para cada una de las tres cartas utilizando seis algoritmos diferentes. Cada resumen se construyó seleccionando las cuatro oraciones consideradas más representativas por el algoritmo correspondiente. Se midió el tiempo de ejecución para cada proceso de resumen.

2.3.1. TF-IDF (Term Frequency-Inverse Document Frequency)

Este método asigna un peso a cada palabra en función de su frecuencia en una oración (TF) y su inversa de frecuencia en el conjunto de todas las oraciones (IDF). Las oraciones se puntúan sumando los pesos TF-IDF de sus palabras normalizadas. Se seleccionan las n oraciones con mayor puntuación. Se implementó usando `TfidfVectorizer` de Scikit-learn en la función `resumen_tfidf`. Los logs y resultados se muestran en las Figuras 2.7 a 2.9.

```
===== RESUMEN TF-IDF: CARTA 1 =====
[INFO] Iniciando resumen TF-IDF (4 oraciones)
[SUCCESS] Resumen TF-IDF completado en 0.0066 segundos
- There—for with your leave, my sister, I will put
some trust in preceding navigators—there snow and frost are banished;
and, sailing over a calm sea, we may be wafted to a land surpassing in
wonders and in beauty every region hitherto discovered on the habitable
globe.
- But supposing all these conjectures to be false, you
cannot contest the inestimable benefit which I shall confer on all
mankind, to the last generation, by discovering a passage near the pole
to those countries, to reach which at present so many months are
requisite; or by ascertaining the secret of the magnet, which, if at
all possible, can only be effected by an undertaking such as mine.
- These reflections have dispelled the agitation with which I began my
letter, and I feel my heart glow with an enthusiasm which elevates me
to heaven, for nothing contributes so much to tranquillise the mind as
a steady purpose—a point on which the soul may fix its intellectual
eye.
- I
accompanied the whale-fishers on several expeditions to the North Sea;
I voluntarily endured cold, famine, thirst, and want of sleep; I often
worked harder than the common sailors during the day and devoted my
nights to the study of mathematics, the theory of medicine, and those
branches of physical science from which a naval adventurer might derive
the greatest practical advantage.
```

Figura 2.7: Resumen TF-IDF: Carta 1.

```
===== RESUMEN TF-IDF: CARTA 2 =====
[INFO] Iniciando resumen TF-IDF (4 oraciones)
[SUCCESS] Resumen TF-IDF completado en 0.0045 segundos
- But I have one want which I have never yet been able to satisfy, and th
absence of the object of which I now feel as a most severe evil, I have n
friend, Margaret: when I am glowing with the enthusiasm of success, there
will be none to participate my joy; if I am assailed by disappointment, n
one will endeavour to sustain me in dejection.
- A youth passed in solitude, my best years
spent under your gentle and feminine fosterage, has so refined the
groundwork of my character that I cannot overcome an intense distaste to
the usual brutality exercised on board ship: I have never believed it to
necessary, and when I heard of a mariner equally noted for his kindliness
of heart and the respect and obedience paid to him by his crew, I felt
myself peculiarly fortunate in being able to secure his services.
- He had already bought a farm with his
money, on which he had designed to pass the remainder of his life; but he
bestowed the whole on his rival, together with the remains of his
prize-money to purchase stock, and then himself solicited the young
woman’s father to consent to her marriage with her lover.
- I am going to unexplored regions, to “the
land of mist and snow,” but I shall kill no albatross; therefore do not
be alarmed for my safety or if I should come back to you as worn and
woeful as the “Ancient Mariner.” You will smile at my allusion, but I
will disclose a secret.
```

Figura 2.8: Resumen TF-IDF: Carta 2.

```
===== RESUMEN TF-IDF: CARTA 3 =====
[INFO] Iniciando resumen TF-IDF (4 oraciones)
[SUCCESS] Resumen TF-IDF completado en 0.0029 segundos
- This letter will reach England by a merchantman now on
its homeward voyage from Archangel; more fortunate than I, who may not
see my native land, perhaps, for many years.
- I am, however, in good
spirits: my men are bold and apparently firm of purpose, nor do the
floating sheets of ice that continually pass us, indicating the dangers
of the region towards which we are advancing, appear to dismay them.
- We
have already reached a very high latitude; but it is the height of
summer, and although not so warm as in England, the southern gales,
which blow us speedily towards those shores which I so ardently desire
to attain, breathe a degree of renovating warmth which I had not
expected.
- One or two stiff gales and the springing of a leak are
accidents which experienced navigators scarcely remember to record, and
I shall be well content if nothing worse happen to us during our voyage.
```

Figura 2.9: Resumen TF-IDF: Carta 3.

2.3.2. Frecuencia de Palabras Normalizada

Un método más simple donde se calcula la frecuencia de cada palabra normalizada en todo el documento (carta). La puntuación de una oración es la suma de las frecuencias de las palabras que contiene. Se seleccionan las n oraciones con mayor puntuación. Implementado en la función `resumen_frecuencia`. Los logs y resultados se muestran en las Figuras 2.10 a 2.12.

```
===== RESUMEN POR FRECUENCIA: CARTA 1 =====
[INFO] Iniciando resumen por Frecuencia de Palabras (4 oraciones)
[SUCCESS] Resumen por Frecuencia completado en 0.0014 segundos
- There—for with your leave, my sister, I will put
some trust in preceding navigators—there snow and frost are banished;
and, sailing over a calm sea, we may be wafted to a land surpassing in
wonders and in beauty every region hitherto discovered on the habitable
globe.
- I
may there discover the wondrous power which attracts the needle and may
regulate a thousand celestial observations that require only this
voyage to render their seeming eccentricities consistent for ever.
- But supposing all these conjectures to be false, you
cannot contest the inestimable benefit which I shall confer on all
mankind, to the last generation, by discovering a passage near the pole
to those countries, to reach which at present so many months are
requisite; or by ascertaining the secret of the magnet, which, if at
all possible, can only be effected by an undertaking such as mine.
- I
accompanied the whale-fishers on several expeditions to the North Sea;
I voluntarily endured cold, famine, thirst, and want of sleep; I often
worked harder than the common sailors during the day and devoted my
nights to the study of mathematics, the theory of medicine, and those
branches of physical science from which a naval adventurer might derive
the greatest practical advantage.
```

Figura 2.10: Resumen Frecuencia: Carta 1.

```
===== RESUMEN POR FRECUENCIA: CARTA 2 =====
[INFO] Iniciando resumen por Frecuencia de Palabras (4 oraciones)
[SUCCESS] Resumen por Frecuencia completado en 0.0004 segundos
- But I have one want which I have never yet been able to satisfy, and the
absence of the object of which I now feel as a most severe evil, I have no
friend, Margaret: when I am glowing with the enthusiasm of success, there
will be none to participate my joy; if I am assailed by disappointment, no
one will endeavour to sustain me in dejection.
- It is true that I have thought more and that my
daydreams are more extended and magnificent, but they want (as the painters
call it) _keeping;_ and I greatly need a friend who would have sense
enough not to despise me as romantic, and affection enough for me to
endeavour to regulate my mind.
- A youth passed in solitude, my best years
spent under your gentle and feminine fosterage, has so refined the
groundwork of my character that I cannot overcome an intense distaste to
the usual brutality exercised on board ship: I have never believed it to be
necessary, and when I heard of a mariner equally noted for his kindliness
of heart and the respect and obedience paid to him by his crew, I felt
myself peculiarly fortunate in being able to secure his services.
- He saw
his mistress once before the destined ceremony; but she was bathed in
tears, and throwing herself at his feet, entreated him to spare her,
confessing at the same time that she loved another, but that he was poor,
and that her father would never consent to the union.
```

Figura 2.11: Resumen Frecuencia: Carta 2.

```
===== RESUMEN POR FRECUENCIA: CARTA 3 =====
[INFO] Iniciando resumen por Frecuencia de Palabras (4 oraciones)
[SUCCESS] Resumen por Frecuencia completado en 0.0001 segundos
- This letter will reach England by a merchantman now on
its homeward voyage from Archangel; more fortunate than I, who may not
see my native land, perhaps, for many years.
- I am, however, in good
spirits: my men are bold and apparently firm of purpose, nor do the
floating sheets of ice that continually pass us, indicating the dangers
of the region towards which we are advancing, appear to dismay them.
- We
have already reached a very high latitude; but it is the height of
summer, and although not so warm as in England, the southern gales,
which blow us speedily towards those shores which I so ardently desire
to attain, breathe a degree of renovating warmth which I had not
expected.
- One or two stiff gales and the springing of a leak are
accidents which experienced navigators scarcely remember to record, and
I shall be well content if nothing worse happen to us during our voyage.
```

Figura 2.12: Resumen Frecuencia: Carta 3.

2.3.3. RAKE (Rapid Automatic Keyword Extraction)

RAKE identifica frases clave analizando patrones de co-ocurrencia de palabras, delimitadas por stopwords. Se utilizó la implementación `rake-nltk`. Se extrajeron las frases clave y sus puntuaciones. La puntuación de una oración se calculó sumando las puntuaciones de las frases clave que contenía. Implementado en `resumen_rake`. Logs y resultados en Figuras 2.13 a 2.15.

```
===== RESUMEN RAKE: CARTA 1 =====
[INFO] Iniciando resumen RAKE (4 oraciones)
[SUCCESS] Resumen RAKE completado en 0.0356 segundos
- I am already far north of London, and as I walk in the streets of
Petersburgh, I feel a cold northern breeze play upon my cheeks, which
braces my nerves and fills me with delight.
- There—for with your leave, my sister, I will put
some trust in preceding navigators—there snow and frost are banished;
and, sailing over a calm sea, we may be wafted to a land surpassing in
wonders and in beauty every region hitherto discovered on the habitable
globe.
- But supposing all these conjectures to be false, you
cannot contest the inestimable benefit which I shall confer on all
mankind, to the last generation, by discovering a passage near the pole
to those countries, to reach which at present so many months are
requisite; or by ascertaining the secret of the magnet, which, if at
all possible, can only be effected by an undertaking such as mine.
- I
accompanied the whale-fishers on several expeditions to the North Sea;
I voluntarily endured cold, famine, thirst, and want of sleep; I often
worked harder than the common sailors during the day and devoted my
nights to the study of mathematics, the theory of medicine, and those
branches of physical science from which a naval adventurer might derive
the greatest practical advantage.
```

Figura 2.13: Resumen RAKE: Carta 1.

```
===== RESUMEN RAKE: CARTA 2 =====
[INFO] Iniciando resumen RAKE (4 oraciones)
[SUCCESS] Resumen RAKE completado en 0.0317 segundos
- But I have one want which I have never yet been able to satisfy, and the
absence of the object of which I now feel as a most severe evil, I have no
friend, Margaret: when I am glowing with the enthusiasm of success, there
will be none to participate my joy; if I am assailed by disappointment, no
one will endeavour to sustain me in dejection.
- A youth passed in solitude, my best years
spent under your gentle and feminine fosterage, has so refined the
groundwork of my character that I cannot overcome an intense distaste to
the usual brutality exercised on board ship: I have never believed it to be
necessary, and when I heard of a mariner equally noted for his kindliness
of heart and the respect and obedience paid to him by his crew, I felt
myself peculiarly fortunate in being able to secure his services.
- He saw
his mistress once before the destined ceremony; but she was bathed in
tears, and throwing herself at his feet, entreated him to spare her,
confessing at the same time that she loved another, but that he was poor,
and that her father would never consent to the union.
- But the old
man decidedly refused, thinking himself bound in honour to my friend, who,
when he found the father inexorable, quitted his country, nor returned
until he heard that his former mistress was married according to her
inclinations.
```

Figura 2.14: Resumen RAKE: Carta 2.

```
===== RESUMEN RAKE: CARTA 3 =====
[INFO] Iniciando resumen RAKE (4 oraciones)
[SUCCESS] Resumen RAKE completado en 0.0035 segundos
- This letter will reach England by a merchantman now on
its homeward voyage from Archangel; more fortunate than I, who may not
see my native land, perhaps, for many years.
- I am, however, in good
spirits: my men are bold and apparently firm of purpose, nor do the
floating sheets of ice that continually pass us, indicating the dangers
of the region towards which we are advancing, appear to dismay them.
- We
have already reached a very high latitude; but it is the height of
summer, and although not so warm as in England, the southern gales,
which blow us speedily towards those shores which I so ardently desire
to attain, breathe a degree of renovating warmth which I had not
expected.
- One or two stiff gales and the springing of a leak are
accidents which experienced navigators scarcely remember to record, and
I shall be well content if nothing worse happen to us during our voyage.
```

Figura 2.15: Resumen RAKE: Carta 3.

2.3.4. TextRank (Sumy)

Este algoritmo basado en grafos modela las oraciones como nodos y la similitud entre ellas como aristas ponderadas. Utiliza un algoritmo similar a PageRank para encontrar las oraciones más çentrales.º importantes. Se usó la implementación `TextRankSummarizer` de la biblioteca `Sumy` en la función `resumen_textrank_sumy`. Logs y resultados en Figuras 2.16 a 2.18.

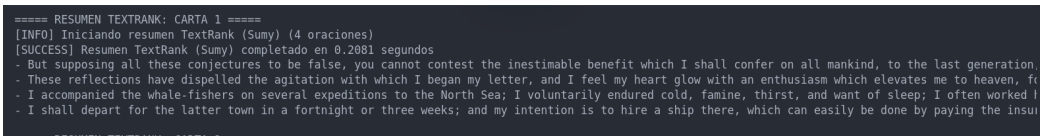

Figura 2.16: Resumen TextRank: Carta 1.

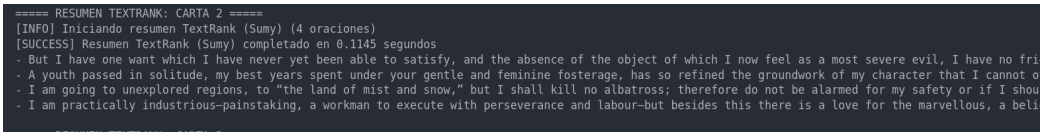

Figura 2.17: Resumen TextRank: Carta 2.

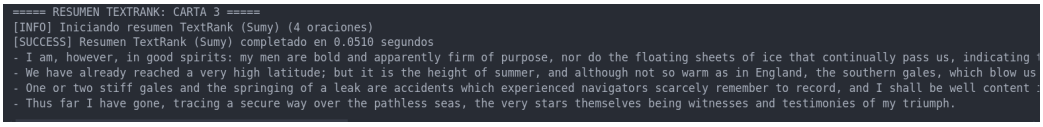

Figura 2.18: Resumen TextRank: Carta 3.

2.3.5. LSA (Latent Semantic Analysis, Sumy)

LSA utiliza la descomposición en valores singulares (SVD) de la matriz término-documento (o en este caso, término-oración) para identificar conceptos latentes. Las oraciones que mejor representan los conceptos más importantes son seleccionadas para el resumen. Se empleó `LsaSummarizer` de `Sumy` en la función `resumen_lsa_sumy`. Logs y resultados en Figuras 2.19 a 2.21.

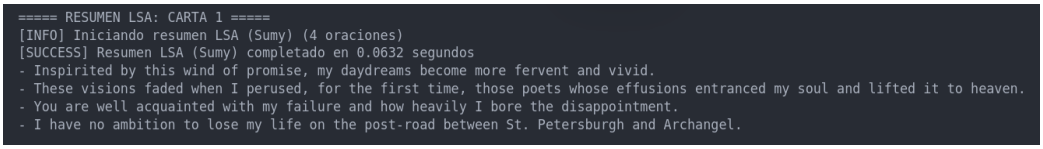

Figura 2.19: Resumen LSA: Carta 1.

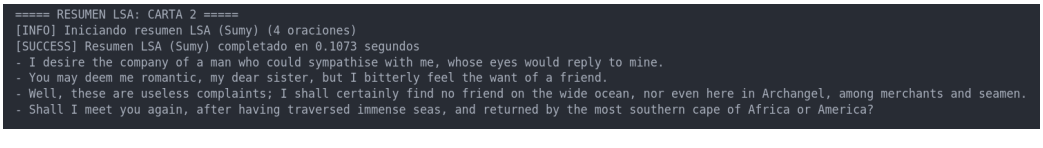

Figura 2.20: Resumen LSA: Carta 2.

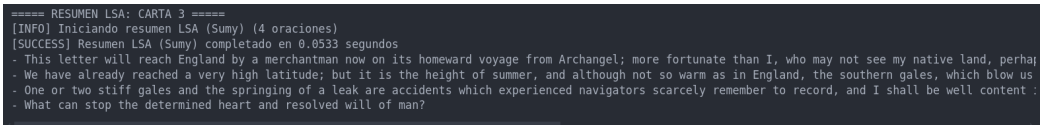

Figura 2.21: Resumen LSA: Carta 3.

2.3.6. BERT (bert-extractive-summarizer)

Este enfoque moderno utiliza un modelo BERT pre-entrenado para obtener embeddings contextuales de las oraciones. Luego, aplica un algoritmo de clustering (como K-Means) para identificar los centroides de los temas principales y selecciona las oraciones más cercanas a estos centroides. Se usó la biblioteca `bert-extractive-summarizer` en la función `resumen_bert`. Logs y resultados en Figuras 2.22 a 2.24.

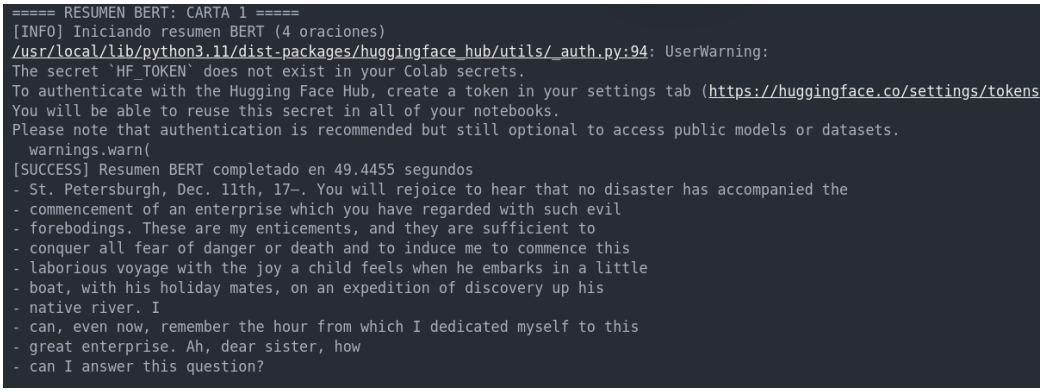

Figura 2.22: Resumen BERT: Carta 1.

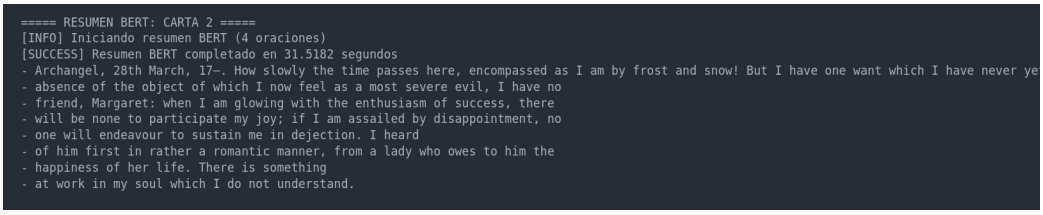

Figura 2.23: Resumen BERT: Carta 2.

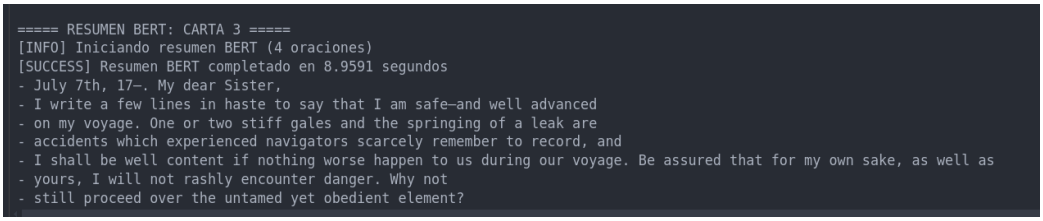

Figura 2.24: Resumen BERT: Carta 3.

2.4. Comparación de Tiempos de Ejecución

Para evaluar la eficiencia computacional de cada algoritmo, se registraron los tiempos de ejecución al generar el resumen de cada carta. Estos tiempos se visualizaron mediante gráficos de barras para facilitar la comparación. La función `graficar_tiempos` se encargó de generar estas visualizaciones. Los resultados se presentan en las Figuras 2.25, 2.26 y 2.27.

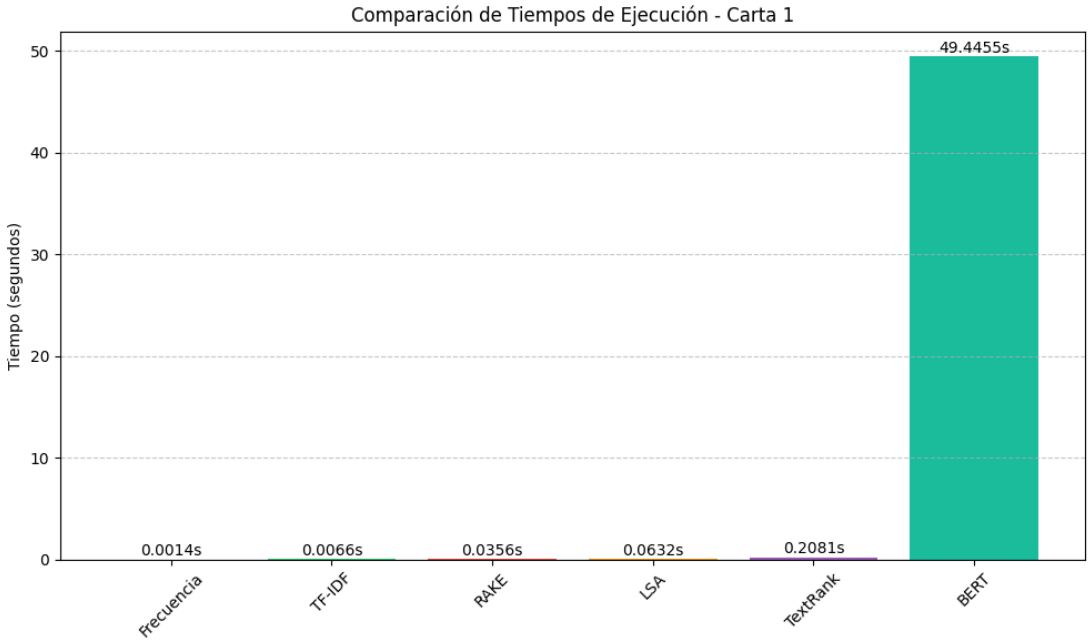

Figura 2.25: Comparación de Tiempos de Ejecución - Carta 1.

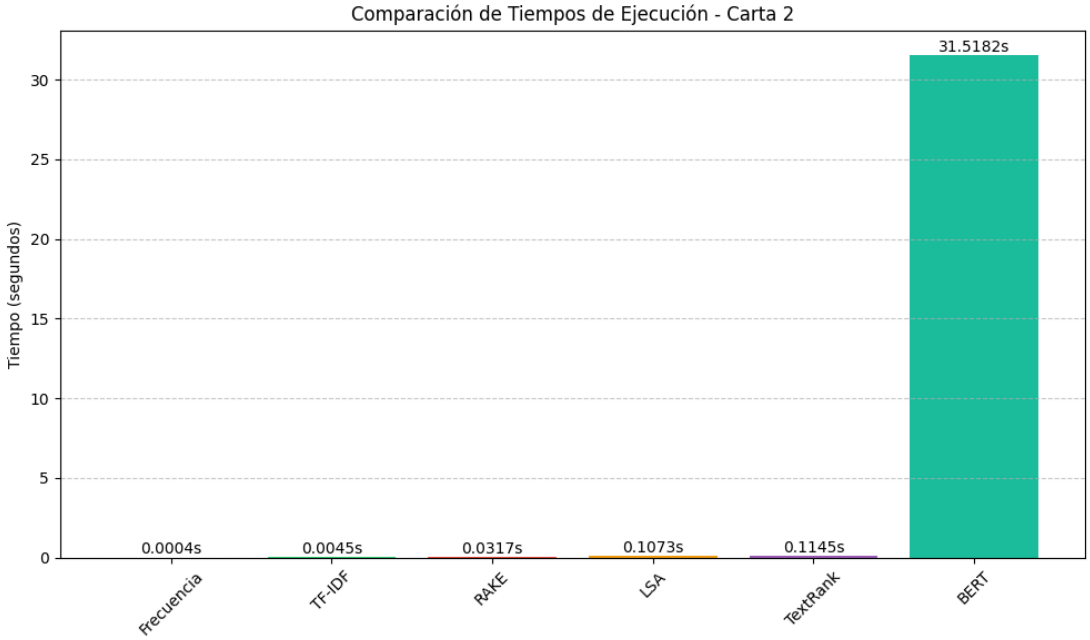

Figura 2.26: Comparación de Tiempos de Ejecución - Carta 2.

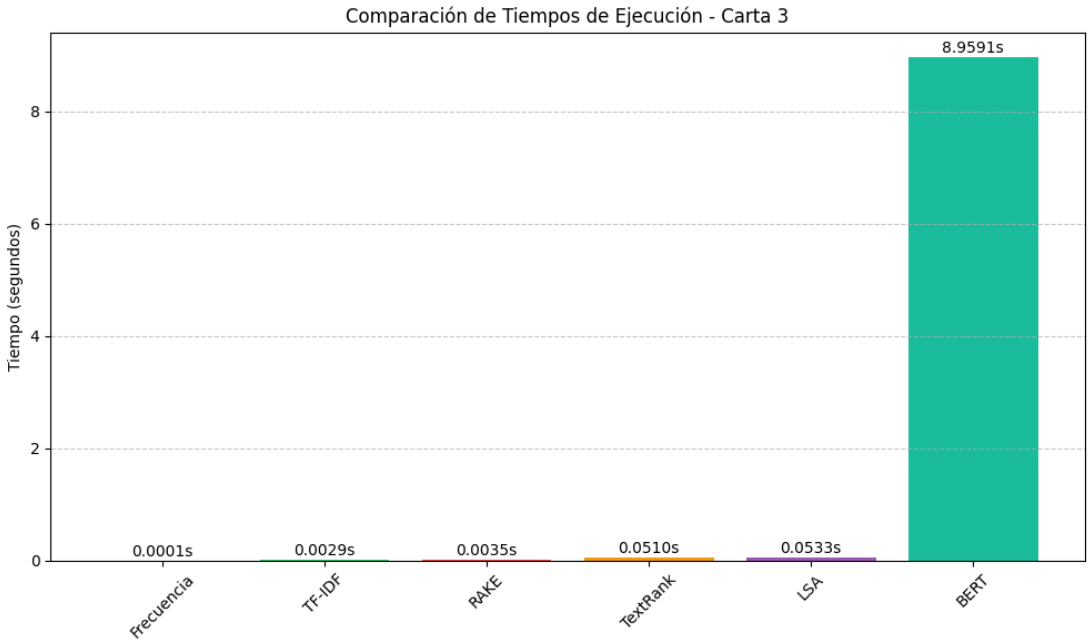

Figura 2.27: Comparación de Tiempos de Ejecución - Carta 3.

Nota: Todos los gráficos y evidencias presentados se obtuvieron a partir de la ejecución de scripts en Python, los cuales se encuentran documentados en el código fuente adjunto.

Capítulo 3

Conclusiones Generales

3.1. Análisis de Resultados

Como se detalló en el capítulo anterior, se aplicaron seis algoritmos distintos para generar un resumen de cuatro frases clave para cada una de las tres primeras cartas de "Frankenstein".

3.1.1. Análisis Cualitativo de los Resúmenes

Al examinar los resúmenes generados (mostrados en las Figuras 2.7 a 2.24), se observan diferencias interesantes que reflejan las metodologías subyacentes de cada algoritmo:

- **TF-IDF y Frecuencia Normalizada:** Estos métodos produjeron resúmenes casi idénticos (idénticos para la Carta 3), centrándose en oraciones con alta carga de palabras frecuentes después de la normalización. Capturaron bien los temas centrales explícitos (el viaje al polo, la necesidad de un amigo, los preparativos, los peligros). Son efectivos para extraer los puntos más repetidos léxicamente.
- **RAKE:** Similar a los métodos de frecuencia para las Cartas 1 y 3. En la Carta 2, dio más prominencia a la historia del capitán, sugiriendo que identifica bien frases temáticas específicas aunque no sean las más frecuentes globalmente.
- **TextRank:** Al basarse en la conectividad y similitud entre oraciones, a menudo generó resúmenes equilibrados, capturando diferentes facetas del texto. Por ejemplo, incluyó tanto la soledad como la ambición exploradora en la Carta 2, y el tono de determinación en la Carta 3.
- **LSA:** Este enfoque semántico a veces divergió significativamente. Para la Carta 1, eligió oraciones sobre sueños pasados y miedos, mientras que para la Carta 2 enfatizó la soledad. Parece capturar temas latentes, aunque no siempre coincidan con los puntos más evidentes del texto.
- **BERT:** Utilizando su comprensión contextual, BERT seleccionó oraciones que a menudo reflejaban puntos narrativos clave o cambios emocionales. Sus resúmenes

parecen semánticamente coherentes, aunque la selección específica es menos transparente que en los métodos estadísticos.

En general, los métodos estadísticos simples (TF-IDF, Frecuencia, RAKE) son predecibles y se centran en la prominencia léxica. TextRank y LSA introducen criterios de centralidad y semántica latente, respectivamente, ofreciendo perspectivas diferentes. BERT, con su base en modelos de lenguaje profundos, busca una representatividad basada en el significado contextual.

3.1.2. Análisis Cuantitativo de Tiempos de Ejecución

Los gráficos de comparación de tiempos (Figuras 2.25, 2.26 y 2.27) revelan diferencias drásticas en la eficiencia:

- **BERT:** Fue, con diferencia, el algoritmo más lento, requiriendo entre 9 y 49 segundos según la longitud de la carta. Su complejidad computacional es inherentemente alta.
- **Otros Algoritmos:** Los cinco métodos restantes fueron significativamente más rápidos.
  - **Frecuencia:** El más rápido consistentemente (milisegundos).
  - **TF-IDF y RAKE:** También extremadamente rápidos (milisegundos).
  - **TextRank y LSA (Sumy):** Ligeramente más lentos que los anteriores (decenas a cientos de milisegundos), debido a los cálculos de grafos y SVD.

Se evidencia un claro compromiso (*trade-off*) entre la complejidad (y potencial calidad semántica) del algoritmo y su velocidad de ejecución. BERT ofrece capacidades avanzadas a costa de un tiempo de procesamiento mucho mayor, mientras que los métodos más simples son casi instantáneos pero pueden basarse en características más superficiales del texto.

3.2. Conclusiones Finales

La práctica permitió implementar y comparar exitosamente seis técnicas de resumen automático extractivo sobre un corpus real. Se cumplieron los objetivos de extracción, normalización, generación de resúmenes y medición de tiempos.

Se constató que no existe un único algoritmo "mejor", sino que la elección depende del caso de uso específico. Para aplicaciones que requieren velocidad y buena identificación de temas explícitos, los métodos basados en frecuencia (Frecuencia, TF-IDF, RAKE) son adecuados. Si se busca capturar relaciones entre oraciones o temas latentes con buena eficiencia, TextRank y LSA son alternativas interesantes. BERT es la opción más potente en términos de comprensión del lenguaje, ideal para escenarios donde la calidad semántica es prioritaria y el tiempo de cómputo no es una restricción crítica.

Todos los métodos generaron resúmenes extrayendo oraciones textuales. Una limitación inherente de este enfoque es que la concatenación de estas oraciones no siempre resulta en

un texto completamente fluido o coherente, a diferencia de los métodos abstractivos (que generan texto nuevo).

Finalmente, la medición de tiempos de ejecución subraya la importancia de considerar la eficiencia computacional al seleccionar una técnica de TLN, especialmente para el procesamiento de grandes volúmenes de datos o en aplicaciones en tiempo real. Los avances en hardware (como GPUs) pueden mitigar parcialmente el costo de modelos complejos como BERT, pero la diferencia con los métodos estadísticos sigue siendo considerable.

Bibliografía

[1] Proyecto Gutenberg, "Frankenstein; or, the modern prometheus by mary wollstonecraft shelley (ebook #84)." https://www.gutenberg.org/ebooks/84, Accessed 2024.

[2] Spot Intelligence, "Keyword extraction using python," 2022.

[3] KDnuggets, "Getting started with automated text summarization," 2019. Ejemplo de resumen por frecuencia.

[4] ActiveState, "How to do text summarization with python." Ejemplo de resumen por frecuencia.

[5] Medium - 1 Hour Blog Series, "Automatic text summarization made simpler using python." Ejemplo de resumen por frecuencia.

[6] MarkovML Blog, "Text summarization with rake algorithm." Código y explicación de RAKE.

[7] Snyk Advisor, "sumy.summarizers.text_rank.textranksummarizer function." Referencia de la biblioteca Sumy (TextRank/LSA).

[8] DevGenius Blog, "Create precise text summaries with textrank algorithm in python," 2023. Ejemplo y librería TextRank.

[9] Turing.com, "5 powerful text summarization techniques in python." Implementaciones de LSA y otras técnicas.

[10] Analytics Vidhya - Medium, "Text summarization using bert, gpt2, xlnet," 2020. Tutorial de resumen con BERT.

[11] Exxact Corp Blog, "Extractive summarization with llm using bert," 2023. Tutorial de resumen con BERT.