Práctica 3. Extracción automática de texto.

Desarrollado por Hernández Jiménez Erick Yael
Escuela Superior de Cómputo
Ingeniería en Inteligencia Artificial

Para la materia de Tecnologías de Lenguaje Natural
Impartida por Ituriel Enrique Flores Estrada
09 de noviembre de 2024

Extracción automática de texto.

Ilustraciones

No se encuentran elementos de tabla de ilustraciones.

Fórmulas

Resultados

Extracción automática de texto.

Resumen

En la presente se aplican técnicas de Procesamiento de Lenguaje Natural (PLN) para generar resúmenes automáticos de textos mediante algoritmos extractivos. A partir de tres cartas iniciales de la novela *"Frankenstein"*, se procesan los textos para identificar las frases clave y crear un resumen conciso.

Se utilizan herramientas como *NLTK* para la tokenización y eliminación de palabras vacías, y se implementan varios modelos de resumen, incluidos *TextRank*, *LSA*, *RAKE* y *BERT*. Estos algoritmos permiten extraer la información más relevante de los documentos, demostrando cómo los modelos de lenguaje facilitan la síntesis y análisis semántico de textos largos.

Palabras clave: Resumen automático, procesamiento de lenguaje natural, *TextRank*, *LSA*, *RAKE*, *BERT*, frases clave, tokenización, normalización, análisis semántico, algoritmos extractivos.

Extracción automática de texto.

Introducción

El presente reporte aborda los fundamentos del **procesamiento de lenguaje natural** (PLN) y su aplicación en la generación automática de resúmenes de texto. Entre las técnicas de resumen más utilizadas se encuentran los métodos extractivos, que seleccionan directamente las frases más representativas del texto original en función de su relevancia.

Para ello, se emplean algoritmos como *TextRank*, un modelo basado en gráficos que evalúa la importancia de frases o palabras en función de sus relaciones dentro del texto, y *Latent Semantic Analysis* (LSA), que utiliza descomposición matemática para identificar patrones temáticos. Además, el modelo *RAKE* permite identificar rápidamente frases clave, mientras que los modelos basados en *BERT* aportan capacidades de comprensión contextual avanzada mediante redes neuronales profundas. En conjunto, estas técnicas permiten sintetizar grandes volúmenes de información de manera eficiente y precisa, mejorando la accesibilidad y utilidad de los datos textuales en diversos campos.

Técnicas de resumen

Las técnicas de resumen de texto se dividen principalmente en dos enfoques: extractivo y abstractivo. En el **resumen extractivo**, el proceso consiste en seleccionar directamente las oraciones o frases más relevantes del texto original, basándose en métricas como frecuencia de términos, relevancia, y relaciones estructurales entre las palabras [1]. Entre los algoritmos extractivos más populares se encuentran *TextRank*, que emplea un modelo basado en gráficos para identificar las frases clave mediante relaciones de coocurrencia, y *Latent Semantic Analysis* (LSA), que utiliza descomposición de valores singulares para revelar patrones temáticos latentes en el texto. Otra técnica extractiva es *RAKE* (*Rapid Automatic Keyword Extraction*), que se enfoca en extraer palabras clave mediante un análisis rápido de la coocurrencia de términos.

En contraste, el **resumen abstractivo** implica la generación de nuevas frases que capturan la esencia del texto original, un enfoque que requiere una comprensión profunda del contexto y suele ser implementado mediante redes neuronales avanzadas, como *BERT* y otros modelos transformadores. Estos modelos permiten

una representación semántica detallada, capturando el significado subyacente del texto y creando resúmenes más coherentes y naturales. Cada técnica tiene sus aplicaciones según el tipo de texto y el nivel de precisión deseado en el resumen, contribuyendo a la accesibilidad y comprensión de grandes volúmenes de información.

Term Frequency – Inverse Document Frequency

El **resumen por TF-IDF** (*Term Frequency-Inverse Document Frequency*) es una técnica extractiva que identifica las oraciones más relevantes de un texto en función de la importancia de sus palabras. Este método combina dos métricas clave [2]:

1. **Frecuencia de Término (TF)**: Mide la frecuencia con la que una palabra aparece en un documento, indicando su relevancia en el contexto de ese texto específico.

2. **Frecuencia Inversa de Documento (IDF)**: Valora la rareza de una palabra en un conjunto de documentos, asignando mayor peso a términos que aparecen con menor frecuencia en otros documentos del corpus. Esto reduce la importancia de palabras comunes y ayuda a identificar aquellas que son más representativas.

$$\mathrm{idf}_t = log\frac{N}{\mathrm{df}_t}$$

Fórmula 1. IDF.

Donde:

- **N**: es número de documento en el *corpus*.
- **df**: es el número de oraciones en el que aparece el término ***t***

El proceso de resumen por TF-IDF consiste en calcular el valor TF-IDF de cada palabra en el texto y luego sumar estos valores para cada oración. Las oraciones con los valores TF-IDF más altos se consideran más relevantes y se seleccionan para el resumen final. Este método permite capturar la esencia del documento original y es particularmente útil cuando se trabaja con textos largos o con varios documentos, ya que resalta el contenido más informativo y evita los términos triviales o repetitivos.

$$w_{t,d} = \mathrm{tf}_{t,d} \cdot log \frac{N}{\mathrm{df}_t}$$

Fórmula 2. Peso de término t en el documento d.

Donde:

- **tf**: es la frecuencia del término

Por otro lado, el suavizado del IDF se utiliza para evitar que los términos que aparecen en casi todos los documentos del corpus tengan un valor de IDF igual a cero, lo cual podría hacer que pierdan relevancia en el cálculo del TF-IDF. Esto es especialmente importante para los términos extremadamente comunes, como artículos, preposiciones o palabras funcionales, que aparecen en la mayoría de los documentos, pero aún pueden ser relevantes en ciertos contextos. Al introducir un valor constante en el denominador de la fórmula del IDF, se asegura que incluso estos términos frecuentes tengan un valor mínimo de IDF, lo que les permite contribuir de manera más adecuada a la ponderación de la importancia de los términos en los documentos. Este suavizado mejora la estabilidad y efectividad del modelo de TF-IDF, especialmente en tareas como la clasificación de texto y la recuperación de información.

$$w_{t,d} = \mathrm{tf}_{t,d} \cdot log \frac{N}{1 + \mathrm{df}_t}$$

Fórmula 3. Peso de término t en el documento d con suavizado.

Frecuencias normalizadas

El **método de frecuencias normalizadas** [3] [4] es una técnica de resumen extractivo que identifica las oraciones más relevantes de un texto mediante la frecuencia de aparición de sus palabras, ajustada por el contexto de ese mismo documento. A diferencia de métodos que consideran múltiples documentos, este enfoque solo se enfoca en el contenido del texto en cuestión.

La técnica comienza calculando la frecuencia de cada palabra dentro del texto y luego normaliza estos valores dividiendo cada frecuencia por la longitud del documento, de modo que las palabras más repetidas se destaquen proporcionalmente al tamaño del

texto. Las oraciones que contienen una mayor proporción de estas palabras de alta frecuencia se consideran clave y se seleccionan para el resumen.

Este enfoque es eficaz para identificar términos significativos en textos largos y reducir la redundancia al dar prioridad a frases con alto valor informativo en el contexto del mismo documento, lo que lo convierte en una opción útil para resúmenes simples y eficientes.

Rapid Automatic Keyword Extraction

RAKE (*Rapid Automatic Keyword Extraction*) es un algoritmo no supervisado para extraer palabras clave de un texto. Funciona al identificar frases que no contienen palabras comunes ("stop words"), asignar puntuaciones a las palabras basadas en su frecuencia y coocurrencia, y luego calcular una puntuación total para las frases, permitiendo así extraer las palabras clave más importantes.

NLTK ofrece una implementación de RAKE a través de "*rake-nltk"*, que permite tokenizar textos y asignar puntuaciones automáticamente a las palabras clave mediante técnicas similares, simplificando el proceso de análisis. [5]

TextRank

TextRank es un algoritmo inspirado en el sistema de *PageRank* de *Google*, diseñado para seleccionar frases o palabras clave importantes a partir de un texto mediante una representación en forma de grafo. Se basa en asignar una puntuación iterativa a palabras/frases según su conectividad y relevancia. [6]

La implementación de *TextRank* en "Sumy" (biblioteca para resumen de texto) genera resúmenes extractivos. Primero, construye un grafo con frases como nodos y relaciones basadas en similitud de contenido. Luego, aplica el algoritmo iterativo para determinar qué frases tienen mayor peso, seleccionando las más relevantes para el resumen. [7]

BERT

BERT (*Bidirectional Encoder Representations from Transformers*) es un modelo basado en transformers que comprende el contexto bidireccional de las palabras, lo

que lo hace poderoso para análisis contextual y resumen de textos. Funciona al generar representaciones ricas de frases para tareas de procesamiento de lenguaje natural [8].

Por otro lado, *BART* (*Bidirectional and Auto-Regressive Transformer*), en su implementación, utiliza *BERT* para codificar texto y genera un resumen al estilo de un decodificador autorregresivo, logrando resúmenes extractivos y abstracticos. *BART* combina el aprendizaje bidireccional con técnicas de generación, mejorando la precisión en resúmenes [9].

El uso de *BART* para resúmenes presenta ventajas clave sobre otros modelos basados en *BERT*. *BART* es una arquitectura secuencia-a-secuencia que combina el poder de la codificación bidireccional (como *BERT*) con capacidades de generación autorregresiva. Esto lo hace sobresalir tanto en resúmenes extractivos como en los abstracticos, proporcionando resúmenes más cohesionados y precisos. Además, al estar diseñado para tareas de "*denoising*" (corrección de ruido), *BART* es más robusto para transformar texto y manejar contextos complejos, superando a modelos exclusivamente extractivos o solo basados en codificación contextual.

Hugging Face implementa este modelo en su biblioteca '*transformers*' con la clase '*pipeline*'. Debido a las características y restricciones de la clase, el texto propuesto como entrada en el desarrollo de la presente supera los límites de procesamiento, desbordando la memoria, por lo que es una consideración con la que se lidiará en el desarrollo. [10]

LSA

LSA (Latent Semantic Analysis) es una técnica de reducción de dimensionalidad basada en la descomposición en valores singulares (SVD). En el contexto de resúmenes, LSA identifica las relaciones semánticas entre palabras a través de matrices de términos y documentos, y luego reduce las dimensiones para resaltar los conceptos más significativos. Aunque eficaz para extraer temas clave, LSA puede resultar en resúmenes más mecánicos o genéricos debido a su enfoque en la extracción de conceptos en lugar de la generación de texto nuevo o fluido [11]. '*Sumy*' implementa este algoritmo. [7]

Extracción automática de texto.

Desarrollo

Generación de cuerpos de documentos

Como parte de las indicaciones, se guardaron las 3 primeras cartas del libro de "*Frankenstein*" en archivos de texto con los nombres: "Carta_1.txt", "Carta_2.txt" y "Carta_3.txt". Estos se guardaron en las variables homónimas y se tokenizarán en enunciados para facilitar la normalización siguiente y, así, evitar problemas en la definición de las oraciones.

Esta tokenización se hizo con el método '*sent_tokenize*' de '*NLTK*'. Finalmente, se juntan los tokens en una misma cadena de texto '*corpus*'. Esto facilita la manipulación y separación de los tokens tanto por palabras como por enunciados.

Normalización de textos

El enfoque que se tendrá en esta práctica será en la extracción más precisa del contenido del resumen correspondiente para que, al leer estos resúmenes, se pueda rescatar la información más destacada que se encuentra en estas 3 cartas. El flujo que se usará, y su justificación se explicará a continuación:

1. Conversión a minúsculas: para evitar redundancias en el contenido significativo del cuerpo
2. Eliminación de espacios, números, signos de puntuación y el carácter '—': Estos caracteres se encuentra en los 3 archivos originales, siendo de carácter visual para separar diálogos y contextos en las frases. No aporta contenido al proceso de ninguno de los algoritmos por lo que su eliminación reducirá el análisis. Cabe mencionar que el carácter '—' es distinto de '-', siendo el último relevante para la generación del resumen debido a que, en el inglés, cambia el significado de las palabras adyacentes, por lo que se mantiene en el cuerpo.
3. Filtrado de palabras más cortas de 3 letras: En el inglés, y particularmente en estas cartas, las abreviaciones se usan para honoríficos, identificadores de nombres y ordinales que no nos interesarán en este cuaderno.
4. Tokenización tanto por palabras como por enunciados: Esto para mantener la relación entre los enunciados y sus tokens correspondientes.

Con este flujo de normalización simple evitamos redundar en palabras y en aplicación de métodos para enfocarnos en el contenido de las palabras y su contexto en el caso de resumidores más complejos como el que se realiza con *BERT*, *LSA* o *TextRank*.

Por otro lado, no se aplicaron normalizaciones que modificaran la forma de la palabra ya que esto podría generar redundancias y exagerando los pesos en métodos de resumen simples como en TF-IDF, Frecuencias normalizadas o *RAKE*, o perderían el significado completo que tienen perdiendo peso en métodos como con los Transformers, *LSA* o *TextRank*.

Una muestra de los resultados de este proceso es el siguiente texto:

to mrs saville england
['mrs', 'saville', 'england']

st petersburgh dec th
['petersburgh', 'dec']

you will rejoice to hear that no disaster has accompanied the commencement of an enterprise which you have regarded with such evil forebodings
['you', 'will', 'rejoice', 'hear', 'that', 'disaster', 'has', 'accompanied', 'the', 'commencement', '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
['arrived', 'here', 'yesterday', 'and', 'first', 'task', 'assure', 'dear', 'sister', 'welfare', 'and', 'increasing', 'confidence', 'the', 'success', '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
['already', 'far', 'north', 'london', 'and', 'walk', 'the', 'streets', 'petersburgh', 'feel', 'cold', 'northern', 'breeze', 'play', 'upon', 'cheeks', 'which', 'braces', 'nerves', 'and', 'fills', 'with', 'delight']

do you understand this feeling
['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
['this', 'breeze', 'which', 'has', 'travelled', 'from', 'the', 'regions', 'towards', 'which', 'advancing', 'gives', 'foretaste', 'those', 'icy', 'climes']

inspirited by this wind of promise my daydreams become more fervent and vivid

Extracción automática de texto.

['inspirited', 'this', 'wind', 'promise', 'daydreams', 'become', 'more', 'fervent', 'and', 'vivid']

Resultado 1. Muestra de corpus normalizado.

Resumen automático extractivo de texto

Tal como se indica en las instrucciones, se obtendrán las 5 frases más significativas de las 3 cartas en conjunto con cada método de resumen.

TF-IDF

Para almacenar las frecuencias de cada enunciado con respecto a todo el documento, se generó una clase llamada TF-IDF que almacena tanto la relevancia del enunciado como como las relevancias de todo el corpus pero que permitía calcular todas las frecuencias en los métodos '*calcularTF*', '*calcularIDF*' y '*calcular_TF_IDF*', siendo este último el método que junta los anteriores. Posteriormente se aplica el TF-IDF por cada enunciado. Dado que a todos se les daba el mismo corpus, el resultado se apegaba a un peso de todo el corpus, asegurando que los resultados fueran correctos.

Finalmente, para el resumen, creamos una lista con las relevancias por enunciados ordenados descendentemente e imprimimos los tokens de los primeros 5 enunciados que tienen la mayor relevancia. Los resultados son los siguientes:

commenced inuring body hardship	4.07753744390572
july	4.07753744390572
archangel march	3.619392077968642
yet second step taken towards enterprise	3.4234631896679355
heaven bless beloved sister	3.388327352587809

Resultado 2. Frase (izquierda) y resultado TF-IDF (derecha).

Este resumen, por la simpleza de su método, 3 frases de los resultados no contienen información relevante más allá del patrón que seguían las cartas: la fecha de la carta y la despedida que se usaba en cada una de ellas. Por ello, de este resultado, los más relevantes son:

- "commenced inuring body hardship"
- "yet second step taken towards enterprise"

Extracción automática de texto.

Frecuencia de palabras normalizadas

Similar al método anterior, se creó una clase para obtener las relevancias de cada enunciado con respecto al cuerpo: '*Frecuencias_normalizadas*'. Así, el método que calcula los pesos es '*calcular_frecuencias*'.

El procedimiento de aplicación fue similar al caso anterior, por lo que se procederá a mostrar los resultados:

```
---------- Resumen: ----------
Enunciado 1:
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

Enunciado 2:
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

Enunciado 3:
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

Enunciado 4:
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

Enunciado 5:
therefor 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
```

Resultado 3. Frecuencias Normalizadas.

Comparando con los resultados del método *TF-IDF*, este resumen logra extraer con mejor eficacia los enunciados, aunque largos, más relevantes, siendo que no hay enunciados irrelevantes y son las que más detallan el discurso de las cartas.

RAKE – NLTK

La implementación de este método fue hecha con la guía del contenido de Manmohan Singh [12].

1. Se importa el algoritmo Rake
2. Se guarda en una variable
3. Se aplica el algoritmo sobre el texto original
4. Se obtiene el resumen

Para ello se importa el constructor '*Rake*' de '*nltk*'. Dado que el método ya extrae las palabras clave de los enunciados, lo alimentamos con la unión de las cadenas con los enunciados que conforman el corpus y, posteriormente, obtenemos los enunciados con los mejores 5 puntajes. Los resultados son:

```
['voluntarily endured cold famine thirst', 'succeed many many months perhaps
years', 'old man decidedly refused thinking', 'cold northern breeze play upon',
'beauty every region hitherto discovered']
```

Resultado 4. Resultados parciales de RAKE.

Y, por lo tanto, las frases completas son:

```
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

if i succeed many many months perhaps years will pass before you and i may meet

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

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

therefor 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
```

Resultado 5. RAKE.

Extracción automática de texto.

En este caso, podemos ver que, en comparación con los resultados del método anterior, tienen en común las siguientes frases:

> *“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”*

> *“therefor 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”*

Esto indica que estas dos frases contienen el mayor peso del mensaje de las tres cartas en general o que pertenecen al resumen general.

TextRank

Como se mencionó en el marco teórico, se hará uso del constructor y método que la biblioteca ‘*sumy*’ incluye. Primero se inicializa el constructor con el método, posteriormente se le alimenta con las stopwords a considerar por el idioma inglés. Para que el texto pueda ser procesado adecuadamente, se le tiene que dar el formato ajustado al método con el constructor ‘*PlaintextParser*’, alimentado con la unión de los enunciados normalizados. Finalmente, se le alimenta al algoritmo con este nuevo contenido y se le indica que se buscan las 5 frases más relevantes.

Los resultados son los siguientes:

i shall satiate my ardent curiosity with the sight of a part of the world never before visited and may tread a land never before imprinted by the foot of man.
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.
you may deem me romantic my dear sister but i bitterly feel the want of a friend.

Extracción automática de texto.

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. 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.

Resultado 6. TextRank.

En este caso, los resultados que da *TextRank* no coinciden con los de *RAKE*, pero sí concuerda con las Frecuencias Normalizadas en las siguientes sentencias:

> *"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"*

> *"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."*

Con esto, podemos decir que estas dos frases pueden estar incluidas en el resultado del resumen de las cartas.

BERT – Transformers

Para este método, no se encontró un modelo estándar (que incluyera la biblioteca de '*Hugging Face*') como base directa a BERT, por lo que se usó el modelo BART que se encuentra disponible en la misma biblioteca y que está basado en BERT [9].

Primero se cargó el modelo en el programa con el constructor '*pipeline*' para ser usado. Cabe mencionar que, como el modelo de resumen tiene límite de tokens de entrada, dividiremos el corpus en 4 partes, cada una de aproximadamente 3895 caracteres o hasta el punto más cercano. Se pueden ver los detalles del corte en el cuaderno de Jupyter en la sección correspondiente.

Extracción automática de texto.

Posteriormente, se alimenta al modelo con cada porción y se guarda el resultado en la lista de resultados '*textos_bert*'. Los resultados parciales son:

Resultado 1:
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. 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. therefor with your leave my sister i will put some trust in preceding navigatorsthere snow and frost are banished and sailing over a calm sea we may be wafted to a land surpassing in wonders.

Resultado 2:
six years have passed since i resolved on my present undertaking. i can even now remember the hour from which i dedicated myself to this great enterprise. i accompanied the whale-fishers on several expeditions to the north sea. i voluntarily endured cold famine thirst and want of sleep. and now dear margaret do i not deserve to accomplish some great purpose. my courage and my resolution is firm but my hopes fluctuate and my spirits are often depressed.

Resultado 3:
i am too ardent in execution and too impatient of difficulties. but it is a still greater evil to me that i am self-educated for the first 14 years of my life. i am in reality more illiterate than many schoolboys of fifteen. i greatly need a friend who would have sense enough not to despise me as romantic. well these are useless complaints i shall certainly find no friend on the wide ocean nor even here in archangel among merchants and seamen.

Resultado 4:
i cannot describe to you my sensations on the near prospect of my undertaking. it is impossible to communicate to you a conception of the trembling sensation half pleasurable and half fearful with which i am preparing to depart. 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. 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 for many years.

Resultado 7. Parciales BART.

Posteriormente se le alimenta al modelo con la unión de estos resultados parciales y obtenemos el resumen final generado.

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.

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.

with your leave my sister i will put some trust in preceding navigators there snow and frost are banished.

Resultado 8. Resultados BART

Con estos resultados, primero debemos notar que el dividir el texto en fragmentos ocasiona perdida de información contextual pero que, a causa de la longitud principalmente de la primera carta y lo límites en los recursos de capacidad de procesamiento bajo el que fue desarrollada la presente práctica, fue algo necesario para obtener un resultado y, ya que se procesa dos veces obtenemos un resultado no tan bueno como se esperaría pero que sigue siendo congruente.

Una vez mencionado lo anterior, de los resultados finales y parciales del modelo podemos ver que concuerda la siguiente frase con el método RAKE:

> *"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."*

LSA – sumy

Tal como se mencionó antes, se ocupará '*sumy*' para aplicar el método *LSA*. Primero se inicializa el método con el constructor '*LsaSummarizer*'. Posteriormente se le alimentará con el documento adaptado anteriormente en *TextRank* y con la instrucción de que encuentre las 5 frases principales. Los resultados son los siguientes:

Enunciado 1:
these visions faded when i perused for the first time those poets whose effusions entranced my soul and lifted it to heaven.

Enunciado 2:

Extracción automática de texto.

i desire the company of a man who could sympathise with me whose eyes would reply to mine.

Enunciado 3:
well these are useless complaints i shall certainly find no friend on the wide ocean nor even here in archangel among merchants and seamen.

Enunciado 4:
shall i meet you again after having traversed immense seas and returned by the most southern cape of africa or america.

Enunciado 5:
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.

Resultado 9. LSA.

De estos resultados, la frase que concuerda con el resto es:

> "*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.*"

Conclusiones

Con los resultados anteriores, y por la frecuencia de coincidencias, la siguiente frase forma parte del resumen:

> "*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.*"

Seguido de esta frase siguen:

> *"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."*

> *"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*

Extracción automática de texto.

> *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"*
>
> *"therefor 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 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"*

Con estas 5 frases podemos decir que son el resumen de las 3 cartas.

Así, en este caso particular, Frecuencias Normalizadas, *RAKE* y *TextRank* fueron los métodos que mejores resultados arrojaron. Esto debido a la normalización por la que pasó el texto antes de que fueran aplicados. En el caso de *BERT*, un mejor acercamiento a un resultado óptimo sería la tokenización por estemas ya que el esto permitiría al modelo interpretar las palabras que partan del mismo estema como un mismo token y así le de más peso a aquellas que contengan variantes del mismo estema, aunque para su interpretación requeriría de un posprocesamiento que, a partir de la serie de tokens que arroje, se encuentre la frase que le corresponde.

En conclusión, la efectividad de cada modelo dependerá de la normalización que reciba el texto, los patrones evidentes que tenga y el análisis detrás de la elección de este preprocesamiento. Así mismo, entre mayor sea la complejidad del modelo, mejores resultados podrán dar pero se requerirá de mejores métodos de preprocesamiento y de más información, de lo contrario, se obtendrían resultados no tan precisos como en la aplicación presente de *BART* o no tan fáciles de interpretar.

Extracción automática de texto.

Bibliografía

[1] M. E. Mendoza Becerra y E. Leon Guzman, «Una Revisión de la Generación Automática de Resúmenes Extractivos,» *Revista UIS Ingenierías,* pp. 7-27, 12 2013.

[2] D. Kauchak, «TF-IDF,» 13 09 2009. [En línea]. Available: http://www.cs.pomona.edu/~dkauchak/classes/f09/cs160-f09/lectures/lecture5-tfidf.pdf.

[3] M. Mayo, "Getting Started with Automated Text Summarization," KDnuggets, 26 10 2022. [Online]. Available: https://www.kdnuggets.com/2019/11/getting-started-automated-text-summarization.html. [Accessed 07 11 2024].

[4] D. Sblendorio, «How to do text summarization with deep learning and Python,» ActiveState, 23 12 2021. [En línea]. Available: https://www.activestate.com/blog/how-to-do-text-summarization-with-python/. [Último acceso: 07 11 2024].

[5] V. Sharma, «rake-nltk 1.0.6,» pypi, 15 09 2021. [En línea]. Available: https://pypi.org/project/rake-nltk/. [Último acceso: 07 11 2024].

[6] R. Mihalcea y P. Tarau, «TextRank: Bringing Order into Texts,» University of North Texas, Barcelona, 2004.

[7] M. Belica, «sumy 0.11.0,» pypi, 23 10 2022. [En línea]. Available: https://pypi.org/project/sumy/. [Último acceso: 07 11 2024].

[8] J. Devlin, M.-W. Chang, K. Lee y K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019.

[9] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov y L. Zettlemoyer, BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, 2019.

[10] Hugging Face, «Pipelines,» Transformers, [En línea]. Available: https://huggingface.co/docs/transformers/v4.46.2/en/main_classes/pipelines#transformers.SummarizationPipeline.

[11] N. E. Evangelopoulos, «Latent semantic analysis,» *WIRES-Cogn Sci,* pp. 683-692, 6 12 2013.