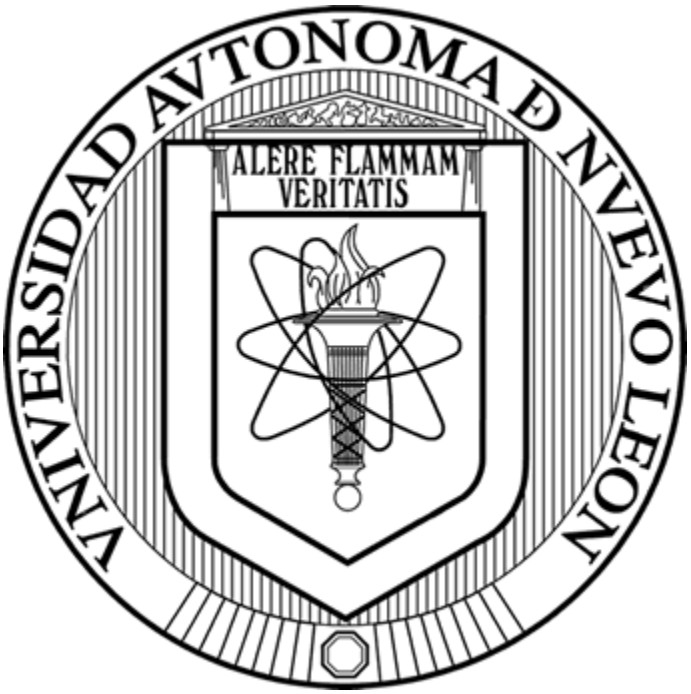
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO-MATEMÁTICAS



REPORTE I

PREPROCESAMIENTO DE TEXTO Y EXTRACCIÓN DE PALABRAS CLAVE

Autor 1: Juan de Jesús Aguilar Solano Matrícula: 1576327

Martes 17 de mayo de 2022

INTRODUCCIÓN

La clasificación de texto se ha vuelto un campo de estudio en constante crecimiento, pues sus aplicaciones son muy extensas, algunas de ellas pueden ser la clasificación de artículos, libros u otros objetos a partir de su descripción o contenido de éstos.

En este proyecto se analizará un libro de texto a través del preprocesamiento de texto para la extracción de información y sus posibles aplicaciones.

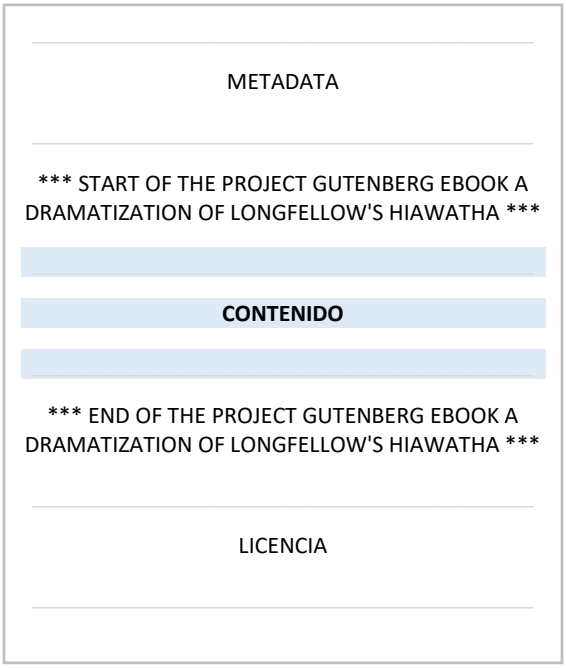
EXPERIMENTACIÓN

Para este proyecto se utilizó el libro "A dramatization of Longfellow's Hiawatha. A spectacular drama in six acts por Henry Wadsworth Longfellow", el cual fue obtenido a partir del proyecto Gutenberg [2]. De este libro se extraerán palabras clave que nos brinden información relevante acerca del contenido de éste y así obtener atributos importantes para su posterior clasificación, el criterio utilizado para asignar el nivel de importancia está determinado por la frecuencia de aparición de las palabras. El proceso consiste en una serie de pasos que se describirán a continuación.

Lectura de datos

Para la lectura de datos se utilizó la función open() de Python y se asignó el contenido a una variable, de modo que todo el texto contenido en el archivo ahora es una cadena de texto.

El contenido de la variable contiene texto que no es útil para su análisis, dicho texto proviene de la información generada por el proyecto Gutenberg y dichos datos contienen información acerca de uso de licencia, editorial, etc. De modo que se recorto el texto para quedarnos solo con la información que contiene la información para el análisis, para hacer esto debemos ver que los libros contenidos en el proyecto Gutenberg siguen la siguiente estructura



De modo que podemos partir el texto del libro en tres partes con tan solo buscar los caracteres "***" y quedarnos solo con la parte central que es correspondiente al cuerpo principal del libro.

Preprocesamiento

En esta fase del procesamiento de texto se procedió a realizar procesos básicos, los cuales son los siguientes:

- Convertir todos los caracteres a minúsculas.
- Remover signos de puntuación y caracteres especiales, los cuales son los siguientes:

[¿, ?, ¡, ‘, “, #, :, “, ”, _, ., ,), (, \, /, *, y –]

- Normalizar los espacios, es decir, el espaciado doble, triple, cuádruple, etc. Es sustituido por un espaciado simple.
- Separar la cadena de texto en palabras y esta es guardada en una lista, tomando como referencia de separación el espaciado simple.
- Removemos las “stopwords” de la lista de palabras, las “stopwords” se obtuvieron de la biblioteca nltk [3].
- Lemmatización de las palabras. Este proceso se lleva a cabo mediante el lematizador WordNetLemmatizer, también contenido en la biblioteca nltk.

Función de preprocesamiento

Todos los pasos descritos durante este apartado se introdujeron en una sola función llamada preprocesamiento_words()

```
def preprocesamiento_words(texto):
    # Palabras a minusculas
    texto = texto.lower()

    # Removing Punctuations u otros caracteres
    texto = re.sub(r'[¿|?|!|\'|\"|#|:|“|”|_*]', r'', texto)
    texto = re.sub(r'[.|,|)|(|\|/|\**|\-*]', r'', texto)

    # Normaliza los espacios
    texto = re.sub(r' +', r' ', texto)

    # Separa oracion en palabras, remueve stopwords y lematiza
    words = [lemmatizer.lemmatize(word) for word in texto.split() if word
             not in stop]
    return words
```

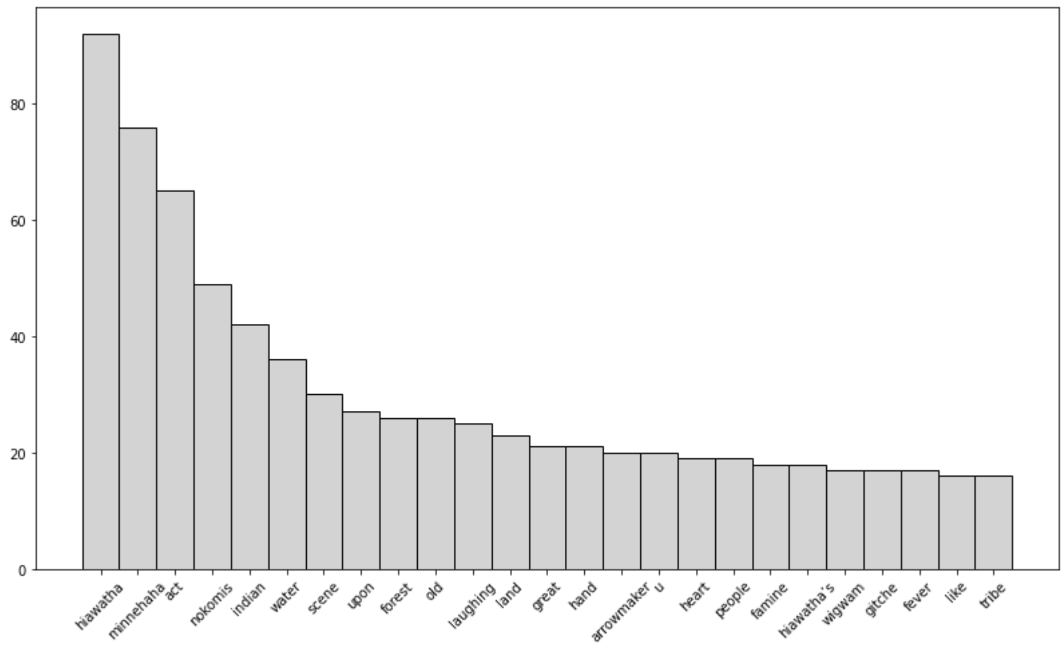
El resultado de la función es una lista de palabras ya procesadas contenidas en el libro.

Conteo de palabras

En esta parte del contenido se obtiene una tabla de frecuencias para todas las palabras incluidas en el listado descrito anteriormente, para llevar a cabo este conteo utilizamos la función Counter() de la biblioteca collections, el resultado es un diccionario con la palabra y su frecuencia, misma que es procesada mediante un dataframe de la biblioteca pandas para ser ordenado por la frecuencia de la palabra.

RESULTADOS

El resultado del preprocesamiento es una lista de palabras con 1636 palabras y su respectiva frecuencia de aparición. A continuación, se muestran las primeras 40 palabras principales que representan el 21.45% de todas las palabras obtenidas.



Nuble de palabras

Para mostrar la relevancia de resultados se muestra una nuble de palabras que contiene las 200 palabras principales.

Descripción general del libro a partir de los datos

Como puede verse, las palabras de mayor peso son hiawatha y minnehaha, también aparece mucho la palabra indian, lo cual puede hacer referencia a que el libro trata sobre nativos americanos y hiawatha y minehaha parecieran los nombres de los protagonistas o tribus relevantes para la historia. Además, aparecen otras palabras como arrowmaker, wáter, forest o act, lo cual podrían referenciar el estilo de vida que llevaban estas personas.

Esto tan solo es una suposición dicha a partir de las palabras más importantes del libro, pero podríamos asociar palabras clave a éste para agruparlo en libros que compartan características similares y así ser clasificados.

CONCLUSIÓN

El preprocesamiento de datos es una parte esencial en cualquier análisis de datos, el preprocesamiento de texto es vital para la extracción de conocimientos, en este proyecto se llevó a cabo una serie de procesos que dieron como resultado palabras clave contenidas en un libro analizado, dichas palabras pueden darnos pistas acerca de que trata el libro, además pueden utilizarse en algoritmos de clasificación o agrupación para diferentes propósitos.

BIBLIOGRAFÍA

[1]. *Project Gutenberg*. (s. f.). Project Gutenberg. Recuperado 17 de mayo de 2022, de https://www.gutenberg.org/

[2]. Wadsworth, H. *"A dramatization of Longfellow's Hiawatha. A spectacular drama in six acts"*. Project Gutenberg, January 12, 2022. https://www.gutenberg.org/ebooks/67148

[3]. Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc.

REPOSITORIO

Los archivos y códigos se encuentran en el repositorio cuyo enlace es:

https://github.com/juanagsolano/Procesamiento-y-clasificaci-n-de-datos