Universidad Autónoma de Nuevo León

Facultad de Ciencias Físico Matemáticas

Maestría en Ciencia de Datos

Tarea 1 - Análisis de la descripción de productos audiovisuales en la plataforma de Netflix

Autor:
Lic. Leobardo García Reyes

Supervisado por:
Dr. Mayra Cristina Berrones Reyes

19 de mayo del 2022

Análisis de la descripción de productos audiovisuales en la plataforma de Netflix

Introducción

El preprocesamiento de texto es uno de los componentes más importantes para la clasificación de texto. En este documento, se tiene como objetivo analizar el impacto del preprocesamiento en la descripción de películas/series de la plataforma de Netflix.

También, analizar las palabras más recurrentes dependiendo a que géneros este clasificada la película/-serie, todo esto a partir de gráficas de frecuencias y nubes de palabras. Todo esto utilizando métodos de limpieza de texto, eliminando palabras que no aporten información importante y transformando las palabras a su raíz utilizando 2 tipos de stemming.

Datos

El conjunto de datos que se seleccionó para aplicar preprocesamiento de texto y analizar, se obtuvo a través del sitio web "Kaggle"[2]. El conjunto de datos proviene de la plataforma Netflix, el cual provee de contenido audiovisual en donde se puede encontrar más de 8,000 películas/series disponibles, a mediados del 2021.

El conjunto de datos se proporciona de manera tabulada de la siguiente manera (Fig. 1):

- **show_id:** ID de las película/serie.
- **type:** Tipo de producto audiovisual (película o serie).
- **tittle:** Nombre de la película/serie.
- **director:** Nombre del director.
- **cast:** Nombre de los actores involucrados en la película/serie.
- **country:** País donde se produjo de la película/serie.
- **date_add:** Fecha en la que se agrego la película/serie a la plataforma de Netflix.
- **release_year:** Fecha de lanzamiento de la película/serie.
- **rating:** Calificación TV de la película/serie.
- **duration:** Duración de la película en minutos o total de temporadas de la serie.
- **listed_in:** Géneros a los que pertenece la película/serie.
- **description:** Descripción de la película/serie.

show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
s1	Movie	Dick Johr	Kirsten Johnson		United Stat	September 25, 2	2020	PG-13	90 min	Document	As her father near
s2	TV Show	Blood & Water		Ama Qa	South Afric	September 24, 2	2021	TV-MA	2 Seasons	Internatior	After crossing patl
s3	TV Show	Ganglanc	Julien Lecle	Sami Bouajila, Tracy		September 24, 2	2021	TV-MA	1 Season	Crime TV S	To protect his fam
s4	TV Show	Jailbirds New Orleans				September 24, 2	2021	TV-MA	1 Season	Docuseries	Feuds, flirtations a
s5	TV Show	Kota Factory		Mayur M	India	September 24, 2	2021	TV-MA	2 Seasons	Internatior	In a city of coachir
s6	TV Show	Midnigh	Mike Flanag	Kate Siegel, Zach Gi		September 24, 2	2021	TV-MA	1 Season	TV Dramas	The arrival of a ch
s7	Movie	My Little	Robert Culle	Vanessa Hudgens, I		September 24, 2	2021	PG	91 min	Children &	Equestria's divide
s8	Movie	Sankofa	Haile Gerim	Kofi Gha	United Stat	September 24, 2	1993	TV-MA	125 min	Dramas, In	On a photo shoot
s9	TV Show	The Grea	Andy Devon	Mel Gie	United King	September 24, 2	2021	TV-14	9 Seasons	British TV S	A talented batch c

Figura 1: Conjunto de datos en disposición.

Metodología

Pre-procesamiento

En esta sección se describirá el preprocesamiento implementado en la descripción de las películas/series de la plataforma de Netflix. La finalidad del preprocesamiento, es saber que palabras son las más recurrentes dentro de la descripción de las películas/series, y también, obtener información sobre cuales son las palabras más importantes dependiendo a los géneros que está clasificado la película/serie.

Primero, en cada descripción de películas/series, se removió el texto que no sea relevante. Para esto, se siguieron los siguientes pasos (Fig. 2):

1. **Texto en minúscula:** La descripción de la película/serie se convirtió en minúscula para que palabras como "The" y "the", sean consideradas iguales.
2. **Eliminación de números:** Se eliminó de la descripción los números.
3. **Eliminación de texto en paréntesis:** Se eliminó el texto que este entre paréntesis.
4. **Eliminación de signos:** Se eliminó signos de puntuación, paréntesis, corchetes, etc.
4. **Eliminación de espacios en blanco:** Se eliminó los espacios en blanco dejando solo un espacio entre palabra.

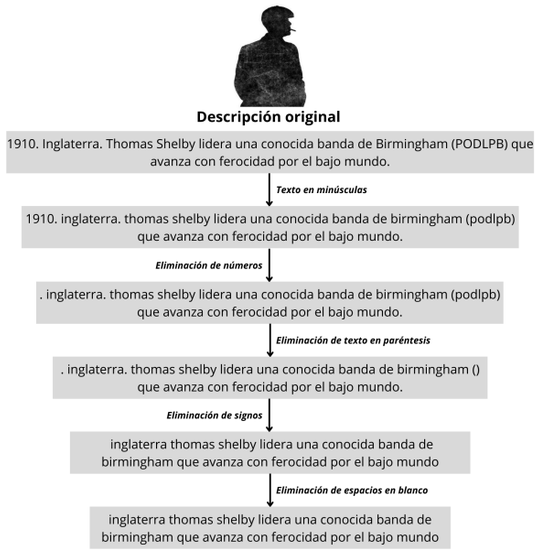


Figura 2: Ejemplo de remoción de texto no relevante en la descripción de la serie Peaky Blinders.

Una vez realizado una primera limpieza, se dividió la descripción por palabra, que se conoce como tokenización. Este proceso sirve para eliminar palabras que no sean relevantes o no aporten a la descripción como son las stopwords. Las stopwords (Fig. 3), son palabras que por su frecuencia y/o semántica no poseen valor discriminatorio alguno. Se trata de artículos, pronombres, preposiciones, etc.

Figura 3: Nube de palabras sobre stopwords.

Algo importante a destacar, es que la paquetería "nltk", tiene por defecto ciertas stopwords, pero es posible añadir más en caso de que así se requiera. Más adelante, esta opción se utilizó para eliminar palabras que pasaron por alto.

Para lo siguiente, es importante destacar dos conceptos: sintaxis y semántica. La sintaxis se refiere a la estructura de la oración, incluida la gramática y las partes del discurso. La semántica, por otro lado, se refiere al significado de la oración.

Dentro de la semántica, se encuentran dos conceptos relacionados con el "part-of-speech", que más adelante se verá, y son: sinonimia y polisemia. La sinonimia se refiere a dos palabras diferentes que tienen

el mismo significado. La polisemia se refiere a una sola palabra que tiene múltiples significados.
.

En la descripción de las películas/series, se puede dar el inconveniente de que aparezcan las palabras "dog" y "dogs". En este caso, ocurriría lo mismo con las palabras que empiezan con mayúscula y otras en minúscula, se tomarían como diferentes. Entonces, se busca que las palabras se reduzcan al mismo término, para esto se utilizó las paqueterías Stemming y Lemmatization.

Stemming implica la eliminación del sufijo de una palabra para reducir el tamaño del vocabulario y aquí se desglosan dos desventajas importantes. El overstemming, donde términos con diferente significado son transformados a un mismo término. El understemming, donde términos con similar significado no son reducidos a un mismo término. Mientras tanto, la Lemmatization es similar al Stemming, excepto que incorpora información sobre la parte del discurso del término utilizando diccionarios y un análisis morfológico mas complejo.

El siguiente diagrama muestra como se implementó lo anteriormente mencionado (Fig. 4) sobre la eliminación de stopwords y la transformación de las palabras a su raíz:

Figura 4: Diagrama para la eliminación de stopwords y transformación de palabras a su raíz.

Como ya se había mencionado, otro enfoque que se le dio al análisis, es sobre las palabras más frecuentes en la descripción de películas/series dependiendo a los géneros al que pertenecen. Como cada película/serie puede tener más de un género, se usó del método one hot encoding múltiple (Fig. 5). Lo que hizo este método, es convertir cada género en columnas y dependiendo si la película/serie pertenece a ese género le colocara 1 y 0 en caso de que no pertenezca a ese género.

Original	One Hot Encoding Múltiple					
Género	Documentaries	International TV Shows	TV Dramas	TV Mysteries	Crime TV Shows	TV Action & Adventure
Documentaries	1	0	0	0	0	0
International TV Shows, TV Dramas, TV Mysteries	0	1	1	1	0	0
Crime TV Shows, International TV Shows, TV Action & Adventure	0	1	0	0	1	1

Figura 5: Ejemplo del resultado del método one hot encoding múltiple.

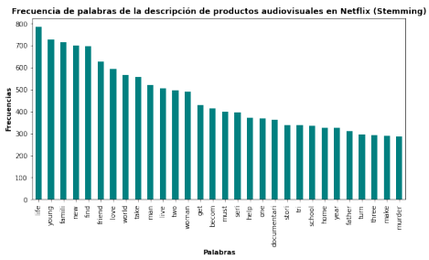
Resultados

En este apartado, se presentan los resultados obtenidos tras realizar la limpieza de texto correspondiente. Para identificar cuales son las diferencias entre los dos métodos de stemming, se gráfico la frecuencia de las primeras 30 palabras. La primera diferencia notable, es que la palabra "life" tiene mayor aparición en el método de Lemmatization (Fig. 6b), esto se debe a que por defecto el método intentará encontrar el sustantivo más cercano. Esto se puede modificar añadiendo el parámetro "pos" dándole un valor de "a" para adjetivos, "v" para verbos y "n" para sustantivos como se muestra en la siguiente linea de código:

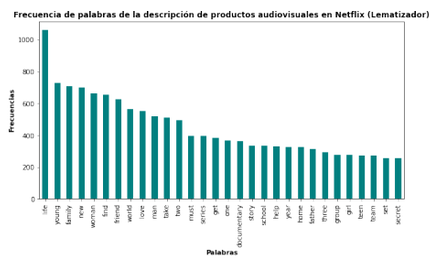
```
metodo_lematizador = WordNetLemmatizer()

palabra_raiz2 = []
for cadena in sin_sw:
    palabra_raiz2.append(
    metodo_lematizador.lemmatize(cadena, pos = 'v'))
```

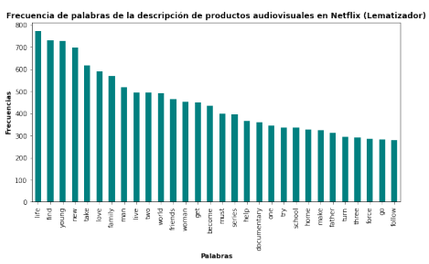
En donde, la variable "metodo_lematizador" esta guardando el método stemming que se va utilizar, la variable "palabra_raiz2" es una lista vacía donde se irán guardando cada palabra ya transformada a su raíz y por último, en la función "lemmatize()" se añade el texto que va transformar y con el parámetro "pos" se le indica lo que se quiere considerar como ya se menciono.



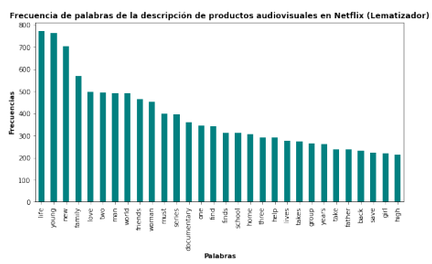
(a) Frecuencia de palabras de la descripción de productos audiovisuales de la plataforma de Netflix con el método de Stemming.



(b) Frecuencia de palabras de la descripción de productos audiovisuales de la plataforma de Netflix con el método de Lemmatization considerando pos = "n".



(c) Frecuencia de palabras de la descripción de productos audiovisuales de la plataforma de Netflix con el método de Lemmatization considerando pos = "v".

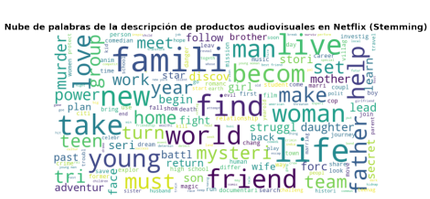


(d) Frecuencia de palabras de la descripción de productos audiovisuales de la plataforma de Netflix con el método de Lemmatization considerando pos = "a".

Figura 6: Gráficos de frecuencias de palabras con los métodos de Stemming y Lemmatization.

Comparando la Fig. 6a con la Fig. 6b, Fig. 6c y Fig. 6d, se puede ver como el método Stemming cambia las palabras como "family" a "famili", "series" a "seri" o "story" a "stori", esto es una de las desventajas de usar Stemming, que la palabra raíz no esta bien escrita y puede traer problemas sino se llegan a considerar. Por otro lado, la Fig. 6d y la Fig. 6c separa como diferente la palabra "find" y "finds", donde se sume que las separa por la contextualización de la descripción.

Entre los diferentes desgloses del método Lemmatization, el que más se parece al método Stemming, es la Fig. 6c. Otra forma de ver esta comparativa es con la Fig. 7, donde se puede apreciar que las palabras más grandes tienen mayor frecuencia.

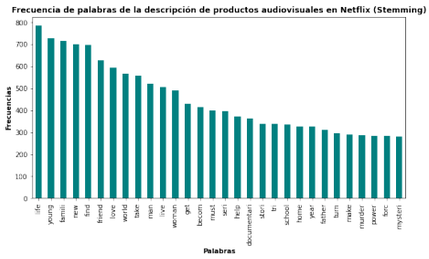


(a) Nube de palabras de la descripción de productos audiovisuales de la plataforma de Netflix con el método de Stemming.

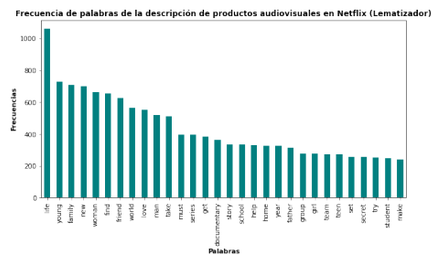
(b) Nube de palabras de la descripción de productos audiovisuales de la plataforma de Netflix con el método de Lemmatization

Figura 7: Nube de palabras con los métodos de Stemming y Lemmatization.

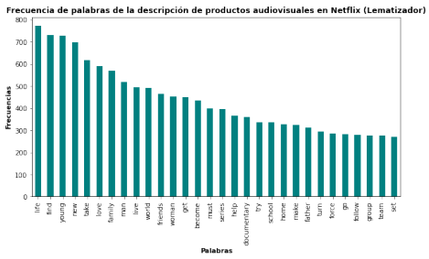
También, algo que se encontró en las gráficas de frecuencia (Fig. 6), es la aparición de números, pero en texto. Por ende, se decidió eliminar estos números hasta el 10 en la parte de los stopwords para que de esta manera, abra paso a nuevas palabras que si puedan dar información adicional, que dando así como se muestra en la Fig. 8.



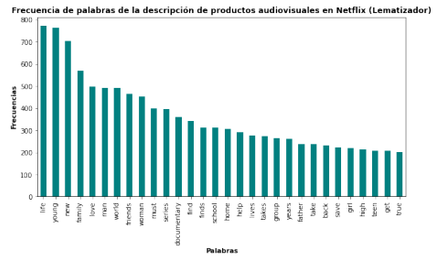
(a) Frecuencia de palabras de la descripción de productos audiovisuales de la plataforma de Netflix con el método de Stemming.



(b) Frecuencia de palabras de la descripción de productos audiovisuales de la plataforma de Netflix con el método de Lemmatization considerando pos = "n".



(c) Frecuencia de palabras de la descripción de productos audiovisuales de la plataforma de Netflix con el método de Lemmatization considerando pos = "v".



(d) Frecuencia de palabras de la descripción de productos audiovisuales de la plataforma de Netflix con el método de Lemmatization considerando pos = "a".

Figura 8: Gráficos de frecuencias de palabras con los métodos de Stemming y Lemmatization sin números en texto.

Otro resultado a mostrar, es sobre las palabras más frecuentes en la descripción de películas/series dependiendo a los géneros al que pertenecen. Como se mencionó anteriormente se implemento un código para separar los géneros al que puede pertenecer las películas/series con el método one hot encoding múltiple. En la Fig. 9 se muestran los 10 géneros más populares en la plataforma de Netflix en donde destaca los géneros International Movies, Drama y Comedies.

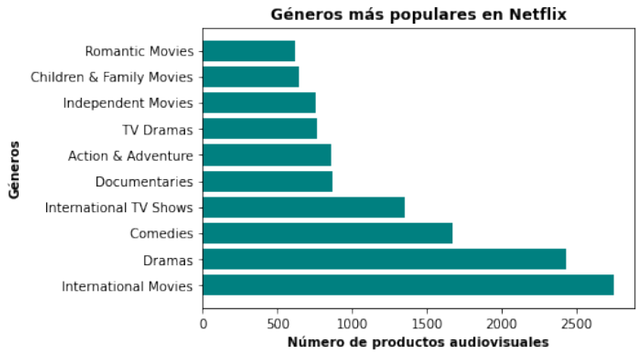
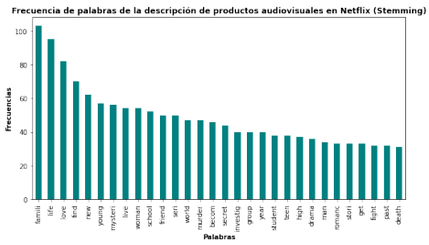


Figura 9: Los 10 géneros más populares en la plataforma de Netflix.

Empleando lo que se modificó anteriormente para la limpieza de texto, se eligió el género "TV Dramas" para detectar las palabras más recurrentes dentro de este género tomando en cuenta los 2 métodos de stemming. En la Fig. 10, se puede observar como la aparición de las palabras como "love", "romance", "struggle" pueden generar más drama a la descripción de la película/serie.



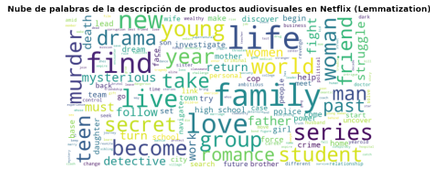
(a) Frecuencia de palabras del género "TV Dramas" con el método de Stemming.

(b) Frecuencia de palabras del género "TV Dramas" con el método de Lemmatization considerando pos = "v".

Figura 10: Frecuencia de palabras del género "TV Dramas" con el método de Stemming y Lemmatization.

Lo anterior mencionado también se puede observar en la Fig. 10 con la aparición de más palabras que pueden distinguir a este género. De igual forma, se pueden encontrar diferencia entre los 2 métodos de stemming, pero sin alejarse demasiado.

(a) Nube de palabras del género "TV Dramas" con el método de Stemming.



(b) Nube de palabras del género "TV Dramas" con el método de Lemmatization considerando pos = "v".

Figura 11: Nube de palabras del género "TV Dramas" con el método de Stemming y Lemmatization.

Conclusiones

La extracción de información dentro de documentos, sitios de internet, etc., requiere de diferentes técnicas de preprocesamiento que ayuden a la limpieza de texto, la eliminación de palabras vacías y stemming, para que de esta forma puedan revelar información importante que ayude a la investigación o toma de decisiones.

Al realizar este trabajo se puede dar cuenta que es necesario volver a correr las veces necesarias un código para detectar a tiempo nuevas inconsistencias cuando se este realizando el análisis. Así pues, se tiene menos errores y mejor confiabilidad de la información obtenida.

Otra cosa importante a destacar, es el método a implementar para el análisis, ya que aquí se utilizaron 2 métodos, existen aún más que pueden dar una mejor visión de lo que se necesita. En mi opinión, el método Lemmatization, tuvo mejores resultados, ya que una de las ventajas que tiene de usarla es la parte del discurso de la palabra, porque produce palabras reales del diccionario.

Referencias

[1] https://github.com/Zarcklet/ProcesamientoClasificacionDatos/blob/main/README.md

[2] Bansal, S. (s/f). Netflix Movies and TV Shows [Data set]. En Netflix Movies and TV Shows. Recuperado el 13 de mayo de 2022, de https://www.kaggle.com/datasets/shivamb/netflix-shows.

[3] Anandarajan, M., Hill, C., Nolan, T. (2019). Text Preprocessing. In: Practical Text Analytics. Advances in Analytics and Data Science, vol 2. Springer, Cham. https://doi.org/10.1007/978-3-319-95663-3_4.

[4] Elia, F. (24 de junio del 2020). Stemming vs Lemmatization. Baeldung. https://www.baeldung.com/cs/stemming-vs-lemmatization.

[5] Fernández, L. A. U. (4 de mayo del 2019). Reducir el número de palabras de un texto: lematización y radicalización (stemming) con Python. Qu4nt. https://medium.com/qu4nt/reducir-el-n%C3%BAmero-de-palabras-de-un-texto-lematizaci%C3%B3n-y-radicalizaci%C3%B3n-stemming-con-python-965bfd0c69fa.

[6] Python. Stemming words with NLTK (18 de mayo del 2022). GeeksforGeeks. https://www.geeksforgeeks.org/python-stemming-words-with-nltk/.

[7] Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. Information Processing & Management, 50(1), 104–112. https://doi.org/10.1016/j.ipm.2013.08.006.

[8] Vijayarani et al, S. (2015). Preprocessing Techniques for Text Mining - An Overview. International Journal of Computer Science & Communication Networks, 5(1), 7–16. https://www.researchgate.net/profile/Vijayarani-Mohan/publication/339529230_Preprocessing_Techniques_for_Text_Mining_-_An_Overview/links/5e57a0f7299bf1bdb83e7505/Preprocessing-Techniques-for-Text-Mining-An-Overview.pdf