Preprocesamiento De Texto

AMÉRICA VICTORIA RAMÍREZ CÁMARA
Facultad de Ciencias Físico Matemáticas
Universidad Autónoma de Nuevo León
Nuevo León, México
america.ramirezcm@uanl.edu.mx

I. Introducción

El preprocesamiento de datos es un paso en el proceso de minería de datos y análisis de datos que toma datos sin procesar y los transforma en un formato que puede ser entendido y analizado por computadoras y aprendizaje automático.

Los datos en bruto del mundo real en forma de texto, imágenes, video, etc., son desordenados. No solo puede contener errores e inconsistencias, sino que a menudo es incompleta y no tiene un diseño regular y uniforme.

A las máquinas les gusta procesar información agradable y ordenada - leen los datos como 1s y 0s. Así que calcular datos estructurados, como números enteros y porcentajes es fácil. Sin embargo, los datos no estructurados, en forma de texto e imágenes, primero deben limpiarse y formatearse antes del análisis.

Al usar conjuntos de datos para entrenar modelos de aprendizaje automático, a menudo escuchará la frase ”basura dentro, basura fuera” Esto significa que si usa datos malos o ”sucios” para entrenar a su modelo, terminará con un modelo malo e inadecuadamente entrenado que en realidad no será relevante para su análisis.

Los datos preprocesados es aún más importante que los algoritmos más potentes, hasta el punto de que los modelos de aprendizaje automático entrenados con datos malos en realidad podría ser perjudicial para el análisis que está tratando de hacer - dándole resultados de ”basura”.

Dependiendo de sus técnicas y fuentes de recopilación de datos, puede terminar con datos que están fuera de rango o que incluyen una característica incorrecta. Su conjunto podría tener valores o campos faltantes. O datos de texto, por ejemplo, a menudo tendrán palabras mal escritas y símbolos irrelevantes, URL, etc.

En esta tarea, se realiza un preprocesamiento de información utilizando la limpieza de información, la extracción de palabras clave, stopwords, lematización, y stemming.

II. Planteamiento del problema

El dataset analizado es: ”rotten_tomatoes_critic_reviews.csv”, en el cual, contiene todas las críticas para la película disponible en Rotten Tomatoes. El dataset mencionado fué descargado de la página Kaggle, sin embargo, la fuente oficial es de la página ”Rotten Tomatoes”, a continuación se expone la definición de estos conceptos:

- Extracción de palabras clave: La extracción de palabras clave se encarga de la identificación automática de los términos que mejor describen el tema de un documento. Las frases clave, los términos clave, los segmentos clave o solo las palabras clave son la terminología que se utiliza para definir los términos que representan la información más relevante contenida en el documento. Aunque la terminología es diferente, la función es la misma: caracterización del tema discutido en un documento. La tarea de extracción de palabras clave es un problema importante en la minería de texto, extracción de información, recuperación de información y procesamiento de lenguaje natural (PNL).
- Stopwords: Las palabras de parada son cualquier palabra en una lista de parada (o stoplist o diccionario negativo) que se filtra (es decir, se detiene) antes o después del procesamiento de los datos del lenguaje natural (texto). No existe una lista universal única de palabras clave utilizadas por todas las herramientas de procesamiento del lenguaje natural, ni ninguna regla acordada para identificar las palabras clave, y de hecho no todas las herramientas incluso utilizan dicha lista. Por lo tanto, cualquier grupo de palabras puede ser elegido como las palabras clave para un propósito determinado. La ”tendencia general en los sistemas de [recuperación de información] a lo largo del tiempo ha sido desde el uso estándar de listas de parada bastante grandes (200-300 términos) a listas de parada muy pequeñas (7-12 términos) a ninguna lista de parada en absoluto.
- Lematización: Es una técnica en la recuperación de datos en los sistemas de información (RDSI), esta técnica sirve para reducir variantes morfológicas de la formas de una palabra a raíces comunes o lexemas; con el fin de mejorar la habilidad de los motores de búsqueda para mejorar las consultas en documentos. Básicamente, este consiste en remover el plural, el tiempo, o los atributos finales de la palabra.
- Stemming: Es un método para reducir una palabra a su raíz o (en inglés) a un stem. Hay algunos algoritmos de stemming que ayudan en sistemas de recuperación de información. Stemming aumenta el recall que es una medida sobre el número de documentos que se pueden encontrar con una consulta. Por ejemplo una consulta sobre ”bibliotecas” también encuentra documentos en los que solo aparezca ”bibliotecario” porque el stem de las dos palabras es el mismo (”bibliotec”).

Estos conceptos se aplican en la solución propuesta para realizar un análisis de las críticas sobre tres películas estrenadas en diferentes años(2002, 2012 y 2017) con diferentes actores y actrices, contando la misma historia sobre un superheroe del MCU (Marvel Cinematic Universe): Spider-Man.

Spider-Man (conocida como El Hombre Araña en Hispanoamérica) es una película de superhéroes estadounidense de 2002 basada en el personaje del mismo nombre de Marvel Comics. Dirigida por Sam Raimi a partir de un guion de David Koepp, es la primera entrega de la trilogía de Spider-Man y está protagonizada por Tobey Maguire como el personaje principal, junto a Willem Dafoe, Kirsten Dunst, James Franco, Cliff Robertson y Rosemary Harris. La película se centra en el genio adolescente marginado Peter Parker, quien desarrolla habilidades sobrehumanas parecidas a las de una araña después de ser mordido por una araña genéticamente alterada.

The Amazing Spider-Man es una película de superhéroes estadounidense de 2012 basada en el personaje de Marvel Comics Spider-Man y compartiendo el título de la serie de cómics del mismo nombre. Es la cuarta película teatral de Spider-Man producida por Columbia Pictures y Marvel Entertainment, un reinicio de la serie después de la trilogía Spider-Man de 2002-2007 de Sam Raimi, y la primera de las dos películas de The Amazing Spider-Man. La película fue dirigida por Marc Webb y escrita por James Vanderbilt, Alvin Sargent y Steve Kloves de una historia de Vanderbilt, y protagonizada por Andrew Garfield como Peter Parker / Spider-Man junto a Emma Stone, Rhys Ifans, Denis Leary, Campbell Scott, Irrfan Khan, Martin Sheen y Sally Field. En la película, después de que Parker es mordido por una araña genéticamente alterada, adquiere nuevos poderes y se aventura a salvar la ciudad de las maquinaciones de un misterioso enemigo reptil.

Spider-Man: Homecoming es una película de superhéroes estadounidense basada en el personaje de Marvel Comics Spider-Man, coproducida por Columbia Pictures y Marvel Studios, y distribuida por Sony Pictures Releasing. Es el segundo reinicio de la película de Spider-Man y la película 16 en el Universo Cinematográfico de Marvel (MCU). La película fue dirigida por Jon Watts, a partir de un guion de los equipos de escritura de Jonathan Goldstein y John Francis Daley, Watts y Christopher Ford, y Chris McKenna y Erik Sommers. Tom Holland interpreta a Peter Parker / Spider-Man, junto a Michael Keaton, Jon Favreau, Gwyneth Paltrow, Zendaya, Donald Glover, Jacob Batalon, Laura Harrier, Tony Revolori, Bokeem Woodbine, Tyne Daly, Marisa Tomei y Robert Downey Jr. En Spider-Man: Homecoming, Peter Parker intenta equilibrar la vida de la escuela secundaria con ser Spider-Man mientras se enfrenta al Buitre (Keaton).

III. Solución Propuesta

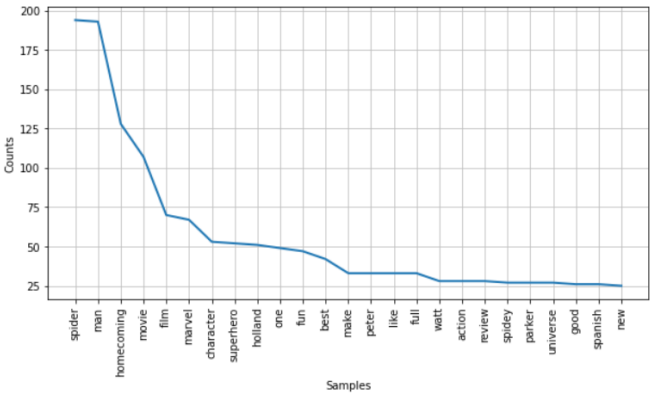
La solución propuesta para realizar este preprocesamiento de datos de texto es implementar; limpieza de información,

extracción de palabras clave, stopwords, lematización, y stemming. Solo la limpieza de información se realizó de manera general, el resto del proceso se realizó tres veces, debido a que el análisis es para cada película.

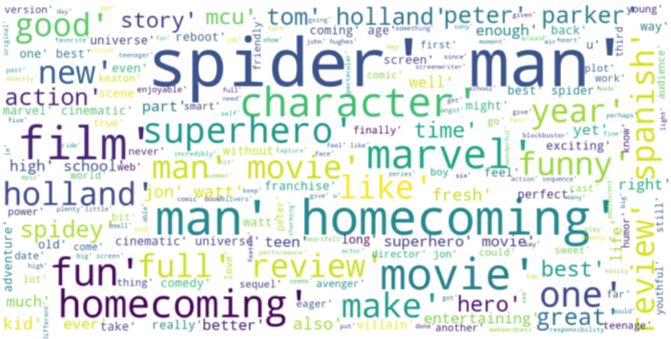
- Limpieza de información: Para este proceso, lo que se realizó fue filtrar el archivo csv, mencionado previamente, mediante la columna ”rotten_tomatoes_link” (enlace desde el que se han adquirido los datos de las películas) para elegir las tres películas a analizar: m/spider_man_homecoming, m/the_amazing_spider_man y m/spiderman, creando un nuevo archivo llamado dataset.csv. Una vez guardada esta información en el nuevo archivo, se eliminaron registros duplicados y valores nulos de la columna ”review_content” (comentarios del crítico) en el mismo archivo.
- Extracción de palabras clave: Antes de iniciar este proceso, se unió en un solo registró la columna ”review_content” filtrado por cada película y así realizar este proceso y el resto para un solo registro. En esta parte, se sustrae solamente palabras con letras de la a-z y A-Z. Después de esta sustracción, se cambian todas las palabras a minúsculas, para no contar con variedad de la misma palabra tanto en mayúsculas como en minúsculas.
- Stopwords: Se eliminan las stopwords en idioma inglés utilizado la librería NLTK.
- Stopwords: Se eliminan las stopwords en idioma inglés utilizado la librería NLTK.
- Stemming: Se utiliza la función stem utilizando la librería NLTK para cada palabra, sin embargo, los resultados en este proceso, en mi opinión, no hacen sentido para el análisis, por lo que se descartó para los resultados finales.
- Lematización: Se utiliza la función lematize utilizando la librería NLTK para cada palabra que no se encuentre en la lista de stopwords en idioma inglés y agregando las siguientes palabras: ’Amazing’, ’amazing’, ’AMAZING’. El motivo de agregar estas palabras fue porque en las críticas para la película ”The Amazing Spider-Man”, esta palabra es muy frecuente debido a que se menciona el título y no por ser una palabra positiva del crítico.

IV. Análisis de resultados

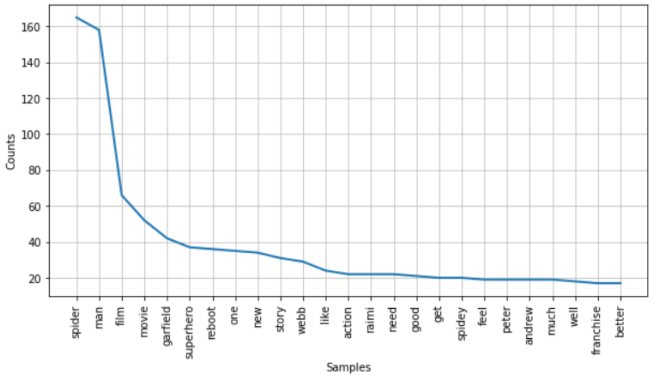
Para la película Spider-Man: Homecoming(2017) se puede observar en la gráfica de abajo, las palabras más frecuentes en las críticas como spider, man, homecoming, movie y film, sim embargo, se espera que muchas de ellas sean igual de frecuentes en las otras dos películas. Además, se observan otras palabras que nos pueden indicar que tal les pareció la película a los críticos, como: fun, best y good.



En la siguiente imagen, podemos apreciar con más visibilidad las palabras más frecuentes, variando en tamanio de estas, indicando entre mayor o menor frecuencia.

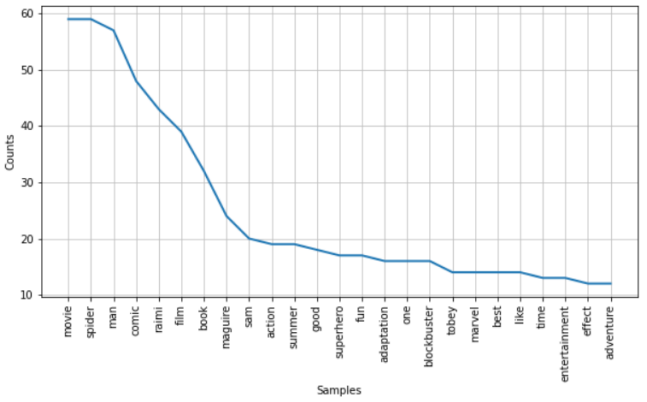


Para la película The Amazing Spider-Man(2012) se puede observar en la gráfica de abajo, las palabras más frecuentes en las críticas como spider, man, film, movie y garfield. Además, se observan otras palabras que nos pueden indicar que tal les pareció la película a los críticos, como: well, better y good. También se observa que el apellido del actor ”garfield” es muy frecuente, por lo que claramente los críticos comentaron acerca del actor principal.

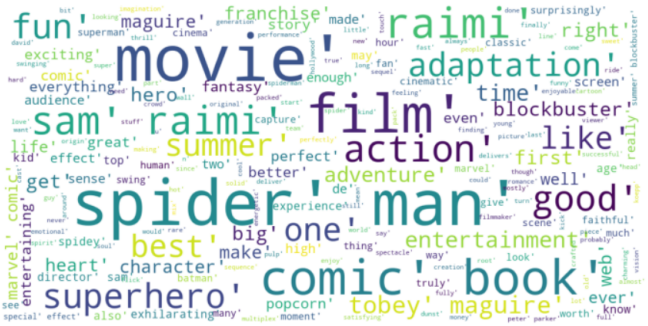


En la siguiente imagen, podemos apreciar con más visibilidad las palabras más frecuentes, variando en tamanio de estas, indicando entre mayor o menor frecuencia.

Para la película Spider-Man(2002) se puede observar en la gráfica de abajo, las palabras más frecuentes en las críticas como spider, man, comic, movie y raimi. Además, se observan otras palabras que nos pueden indicar que tal les pareció la película a los críticos, como: fun, best y good. También se observa que el apellido del director ”raimi” es muy frecuente, por lo que claramente los críticos comentaron acerca del director de la película.



En la siguiente imagen, podemos apreciar con más visibilidad las palabras más frecuentes, variando en tamanio de estas, indicando entre mayor o menor frecuencia.



V. Conclusiones

El preprocesamiento de texto ayudó a resumir las críticas de las tres películas con las técnicas descritas previamente, de esta manera, el análisis fue más sencillo, sin embargo, a pesar de que el steamming, no se consideró, los resultados obtenidos fueron buenos.

En general, de acuerdo a los resultados, las tres peliculas son buenas y divertidas. En las gráficas de la pelicula ”Spider-Man: Homecoming”(2017) se hace mención del actor principal ”Holland”, sin embargo, comparado con la película ”The Amazing Spider-Man”(2012), el actor principal ”Garfield” es mayormente mencionado en las críticas. A diferencia de estas observaciones, en la película de ”Spider-Man”(2002) la crítica fue más para el director de la película ”Raimi” que del actor principal ”Maguire”.

References

[1] Programador Clic. (2020) Un módulo de Python cada semana—multiprocesamiento. [Online]. Available: https://programmerclick.com/article/4426970514/

[2] Saket Thavanani. (2020) TF-IDF Calculation Using Map-Reduce Algorithm in PySpark. [Online]. Available: https://towardsdatascience.com/tf-idf-calculation-using-map-reduce-algorithm-in-pyspark-e89b5758e64c

[3] freeCodeCamp. (2018) How to process textual data using TF-IDF in Python. [Online]. Available: https://www.freecodecamp.org/news/how-to-process-textual-data-using-tf-idf-in-python-cd2bbc0a94a3/

[4] GeeksforGeeks. (2021) Read JSON file using Python. [Online]. https://www.geeksforgeeks.org/read-json-file-using-python/?ref=lbp

[5] Ernesto Rico Schmidt. (2018) Implementando MapReduce. [Online]. Available: https://rico-schmidt.name/pymotw-3/multiprocessing/mapreduce.html