Análisis de texto

Leslye Hernández
1819111
Universidad Autónoma de Nuevo León
Facultad de Ciencias Físicas Matemáticas
Maestría en Ciencias de datos

Resumen

Este trabajo presenta un análisis de texto de reseñas extraídas de la plataforma de ventas de e-commerce; Amazon, para la categoría de "All Beauty".
Por medio de técnicas de procesamiento de lenguaje natural, estadistíca descriptiva, se pretende presentar las principales características de las reseñas, la frecuencia de palabras, signos de puntuación, lematización y características que distinguen a las reseñas de un usuario o producto contra otra, esto con el uso de paquetes y librerías del lenguaje Python.

Introducción

El catedrático del Instituto Cervantes, Agustín Vera Luján describe al texto lingüístico como un acto de la comunicación por el que dos o mas personas se transmiten información, o a la unidad a través de la que se lleva a cabo tal acto de comunicación, ya sea, por oral o escrita.

Enfocándonos al tipo de texto escrito, este es contituído por un objetivo, lo que distingue su contenido, estructura y audiencia a la cual va dirigida, es así en como se pueden agrupar en diferentes tipos de textos:

- **Texto expositivo:** Contiene una estructura textual que está ligada al análisis y la síntesis, su principal objetivo es explicar o exponer algo en términos de su tema principal, considera las relaciones causa-efecto o de comparación, entre otros aspectos.
  Algunos ejemplos de este tipo son los Folletos, libros de texto, enciclopedias, manuales, exposiciones orales o museísticas, disertación, conferencia, ponencia, informes, apuntes, exámenes.
- **Texto narrativo:** Texto en el que predomina la narración, se distingue porque desarrolla una historia en un tiempo y un espacio determinado (real o imaginario) y además las acciones con personajes tienen un desarrollo dentro de la historia, estos personajes pueden ser personas, animales o cosas humanizadas.
  Algunos géneros del texto narrativos son: cuento, leyenda, novela, noticia, crónica, chiste, historieta, textos de historia, cartas familiares, biografía, videoclip, videojuego, testimonio, película de ficción o documental.
- **Texto argumentativo:** Se distingue por una organización textual centrada en el juicio y en la toma de posición respecto de algún asunto polémico. Este tipo de texto contiene la defensa de una un tema sustentada con argumentos, considera una crítica sobre un hecho, causa o circunstancia.
  Entre los géneros de este tipo están: asambleas, debates, texto político, texto publicitario, artículo de opinión, entrevista, cartón político, editorial, columna, reseña crítica, etc.

En el documento se enfocará en el análisis de este último, ya que las reseñas son del tipo argumentativo, por lo que por medio de la estadítica descritiva se irán exponiendo la crítica principal de la resña así como caracterísiticas puntuales de la misma.

Estructura, análisis estadístico

La reseña fue extraída de `https://amazon-reviews-2023.github.io` , en csv. El extracto contiene untotal de 633,693 reseñas con 3 columnas las cuales se muestran en el *Cuadro 1*.

Para comprender mejor los datos del conjunto de reseñas, se muestra la figura 1, la cual muestra las palabras que existen dentro de las reñenas, además la *figura 2* exhibe a los productos que recibieron reseñas más extensas (con mayor número de palabras), el porcentaje hace referencia a la frecuencia relativa considerando todas las reseñas extraídas de la categoría, que son 633,693 reseñas, mientras el número que aparece entre paréntesis es la cantidad de palabras que tiene el producto en sus reseñas.

Palabras

En la *figura 3*, se muestran las palabras más repetidas dentro de las reseñas, se puede resaltar que las que mayormente tienen relevancias son palabras como "great", "hair", "product", lo que se puede asumir que la mayoría de las reseñas son positivas de los productos. Asimismo se encuentran verbos en pasado, pues las reseñas se escriben con la experiencia que el usuario experimentó.

Signos de puntuación

La *figura 4,* muestra el top 5 de los signos de puntuación que se encuentran dentro de las reseñas. Entre estos se destacan los signos de coma, punto y paréntesis, además a diferencia de otros tipo de textos, en la reseña se puede hacer uso de más signos de puntuación, pues estos pueden estar asociados a escribir alguna expresión de "emoji" .

Emoji

Un emoji es un pictograma que hace referencia a una emoción, hoy en día es utilizado dentro de las redes sociales o textos informales para dar más sentimiento al texto que se redacta. Dentro de las reseñas, se encuentran algunas que contienen diferentes emojis, los más repetidos estan asociados con emociones positivas.

Gramática

Usando el paquete `spacy` , se pueden extraer distintas características, por ejemplo, considerando la siguiente reseña:

"It was very easy to use and worked very quickly This product worked amazingly well. It was very easy to use and worked very quickly."

Se pueden extraen las siguientes características:

- Verbos encontrados: 'use', 'worked', 'worked', 'use', 'worked'.
- Raíces de los verbos: 'use', 'work', 'work', 'use', 'work'.
- Desinencias aproximadas: 'se', 'ed', 'ed', 'se', 'ed'.
- Adverbios encontrados: 'very', 'very', 'quickly', 'amazingly', 'well', 'very', 'very', 'quickly'.
- Raíces de los adverbios: 'very', 'very', 'quickly', 'amazingly', 'well', 'very', 'very', 'quickly'.
- Desinencias aproximadas: 'ry', 'ry', 'ly', 'ly', 'll', 'ry', 'ry', 'ly'.

Esta librería es de gran utilidad cuando se requiere analizar el texto y traducir a diferentes idiomas.

Tema principal

Con el paquete `sklearn` se pueden obtener los Unigramas, bigramas y trigramas, estos son importantes para obtener información del texto, como el tema principal, algunos sentimientos del personaje, etc. Se muestra un ejemplo de un bigrama, considerando 3 reseñas elegidas aleatoriamente en el *cuadro 2*.

Se observa que las frases que más aparecen en las reseñas son las frases A y E, pues las podemos encontrar al menos dos veces en el extracto del texto.

Sentimiento.

Con el paquete `textblob`, se extrajó la polaridad, la cual está asociada a un sentimiento positivo (Valores cercanos a 1) y negativos (valores cercanos a -1) y subjetividad del texto, para este caso, se utilizó la frase propuesta en el apartado de grámatica y se obtuvieron los siguientes resultados:

- Polaridad: 0.51
- Subjetividad: 0.84

Puntos a destacar.

Por medio del análisis que se obtuvo en la sección anterior se pueden destacar algunas caracteristicas.

Propósito y audiencia

En primer lugar el próposito o el fin de una reseña es informar y criticar a un producto o una situación. Esto lo deja en claro la frecuencia de las palabras y el contenido de la misma, que se distingue por el uso de adjetivos que califican o describen al sustantivo.

Gramática

La grámatica juega un papel muy importante en la construcción de las reseñas, pues se puede destacar que reseña tiene más variedad en los signos de puntuación, apoyándonos en la *figura 3* se puede observar en que uno de los signos más utilizados es el apóstrofe, pues este signo es comúnmente utilizado en el lenguaje informal.

La coherencia y la gramática forma un rol de alta relevancia que muy pocas veces se considera al escribir una reseña en un sitio web, por lo que típicamente en estas se encuentran errores de ortografía.

Conclusiones finales

Conforme avanza la tecnología y el e-commerce ligado a ésta, es más común encontrar reviews de todo tipo en internet.

Las personas se han ido acostumbrando en aceptar la grámatica de este tipo de textos, encontrar palabras contraídas, slangs o los hasta emojis que se vuelven una forma común al leer un texto en la red, pues normalmente el enfoque de las reseñas está dirigido al valor de experiencia que aporta.

Si bien las caracteristicas que distinguen a un argumentativo son evidentes, cabe señalar en la facilidad que se tiene para detectar estas características por medio de tecnologías y lenguajes de programación, como los usados en este trabajo.

A través de este, se puede inducir el tema principal de cualquiera de los dos textos, por medio de la frecuencia de las palabras o también en que tiempo están escritos, esto por medio de los verbos y las librerías que extraen estos mismos. Asímismo se realizan análisis de sentimientos para detectar con que sentimiento está asociado el texto.

Este tipo de análisis de texto es crucial para la detección de patrones, detectar muletillas del redactor o también en un enfoque más avanzado, para traducir de un idioma a otro. Lo importante es darle el enfoque correcto y la interpretación adecuada.

Referencias

- Vera L. A. (2024). "¿Qué es un texto?" Disponible en: `https://www.cerasa.es/media/areces/files/book-attachment-3498.pdf`.
- Gracida, M. Y. (2011). "Tipos de textos". En Textos Modelo. Portal Académico del CCH, UNAM. Disponible en: `https://portalacademico.cch.unam.mx/alumno/tlriid2/unidad1/textosmodelo/tiposDeTextos`.
- Hernández B. L. (2025). "Análisis de texto - Reseñas" Disponible en: `https://github.com/LeslyeHdz13/PCD/blob/main/Tareas/An%C3%A1lisis%20de%20texto/AnalisisDeTexto_Rese%C3%B1as.ipynb`

Anexos

Cuadro 1: Variables.

Nombre de la variable	Tipo de dato	Descripción
productid	string	Asocia cada producto con un id único
prodDescription	string	Describe al producto que se hace referencia
titleReview	string	Reseña escrita por el usuario

Figura 1: Nube de palabras - Elaboración propia.

Cuadro 2: Bigrama

No.	A	B	C	D	E
1	2	0	1	1	2
2	0	1	0	0	1
3	0	0	0	0	0

A: not great
B:thank you
C: the consistency
D: thick wound
E: this product

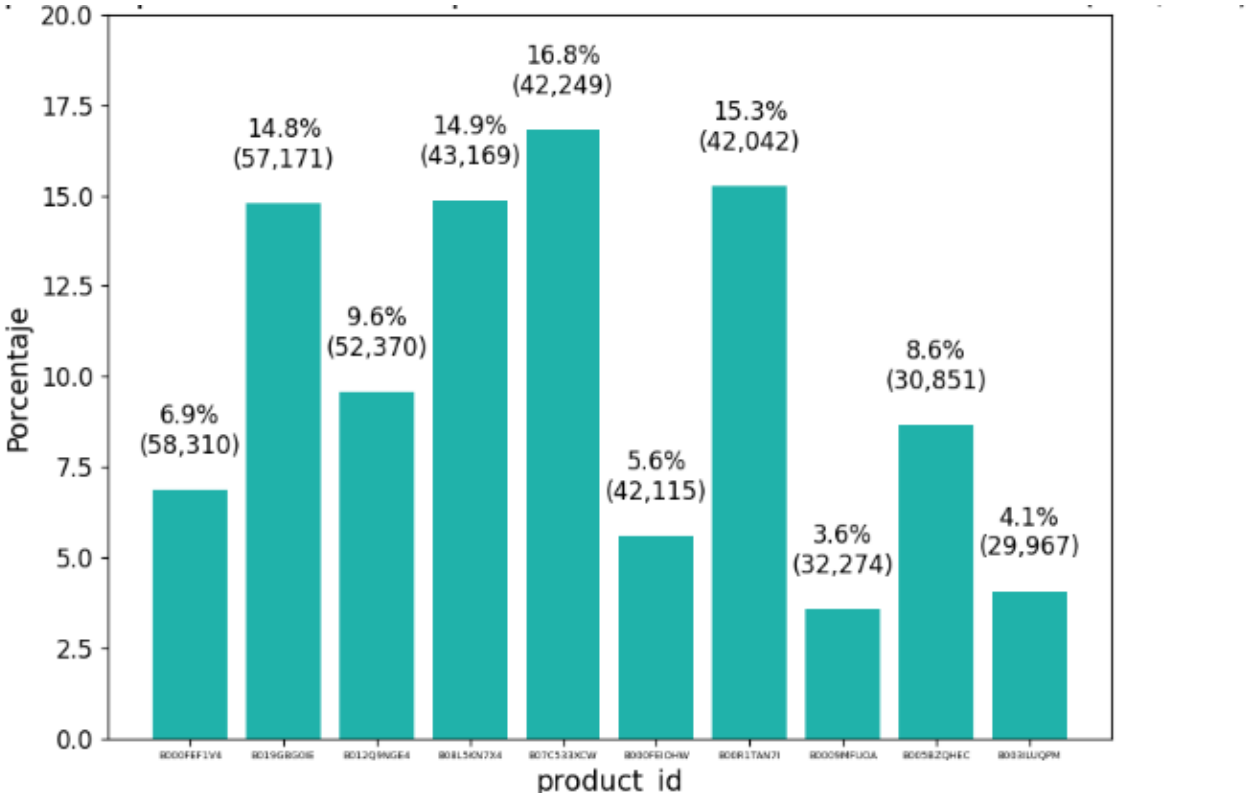

Figura 2: Top 10 de los productoo con más palabras dentro de sus reseñas - Elaboración propia.

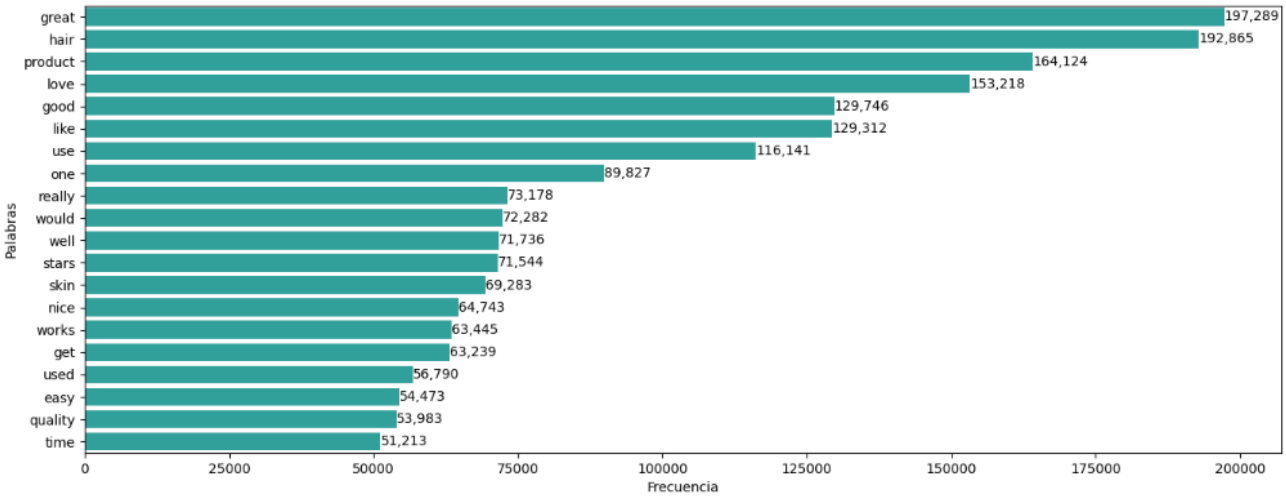

Figura 3: Top 20 de palabras más frecuentes en el texto - Elaboración propia.

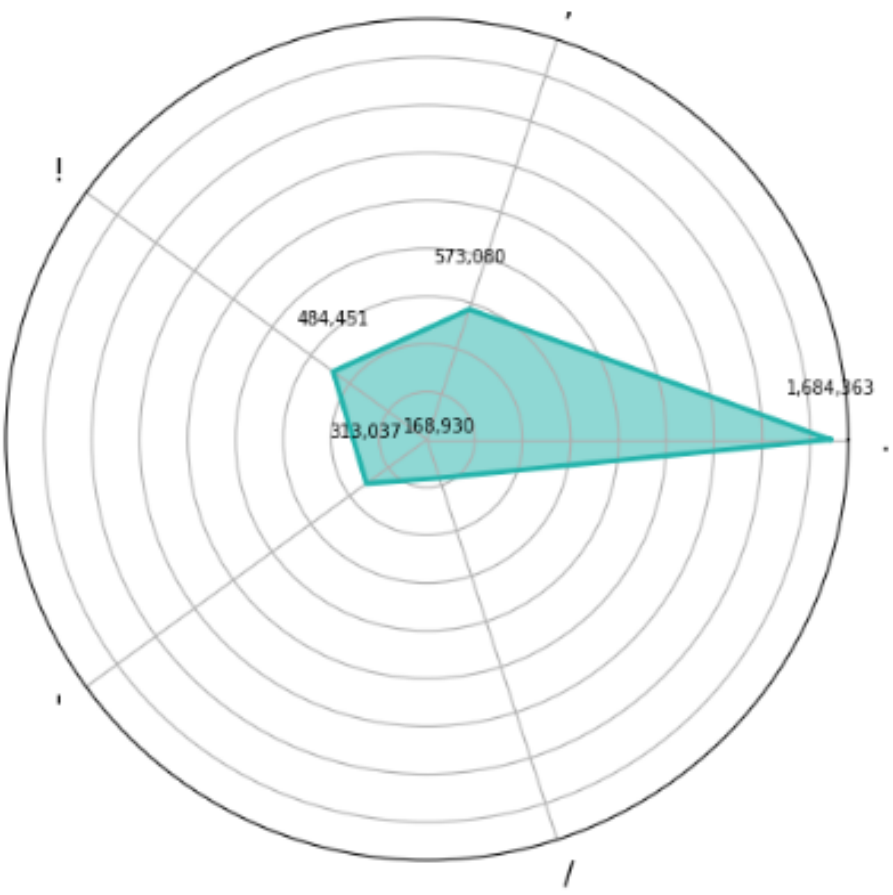

Figura 4: Top 5 de signos más frecuentes utilizados en la reseña - Elaboración propia.