Autor: Jeisson Poveda.

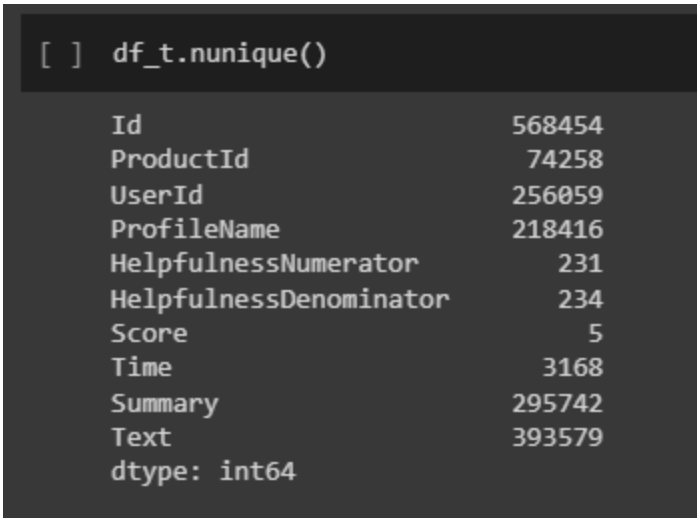
Análisis de sentimiento Amazon reviews.

- Preprocesamiento y EDA.

Se observa que el conjunto de datos tiene las siguientes características:

El conjunto de datos de Amazon Fine Food Reviews consta de reseñas de alimentos finos de Amazon.

- Número de reseñas: 568.454
- Número de usuarios: 256.059
- Número de productos: 74.258
- Intervalo de tiempo: octubre de 1999 — octubre de 2012
- Número de atributos/columnas en datos: 10



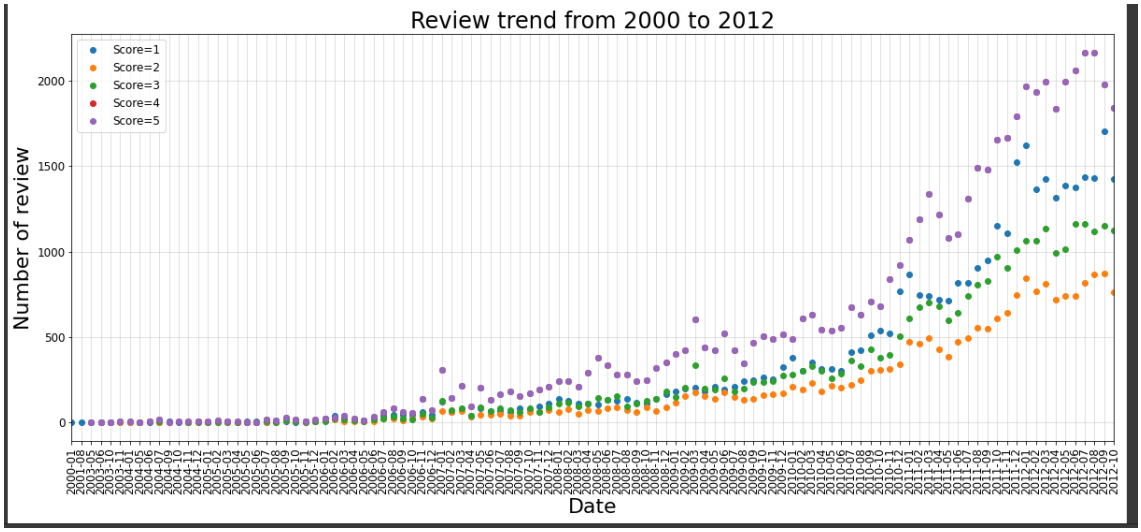
```
[ ]  df_t.nunique()

Id                      568454
ProductId                74258
UserId                  256059
ProfileName             218416
HelpfulnessNumerator       231
HelpfulnessDenominator     234
Score                        5
Time                      3168
Summary                 295742
Text                    393579
dtype: int64
```

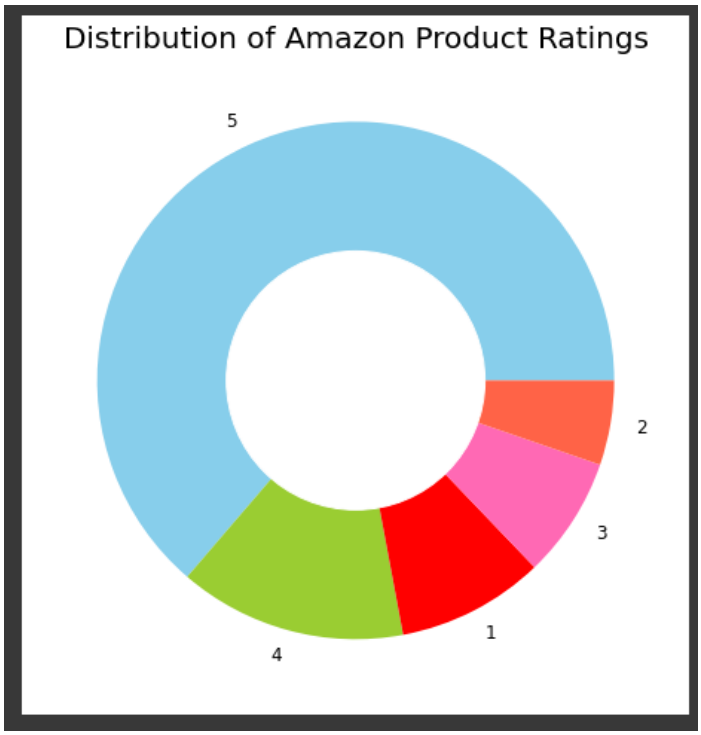
Atributos.

1. Id:Identificación
2. ProductId — identificador único para el producto
3. UserId — identificador único para el usuario
4. ProfileName: Nombre de perfil
5. Helpfulness Numerator: número de usuarios que encontraron útil la reseña
6. HelpfullnessDenominator: número de usuarios que indicaron si la reseña les resultó útil o no.
7. Score — calificación entre 1 y 5
8. Time: marca de tiempo para la revisión
9. Summary — breve resumen de la reseña
10. Text — texto de la reseña

Se observa que la tendencia de score 5 de los reviews de los usuarios con el tiempo, lo cual es bueno, pues indica que esta recibiendo calificaciones positivas.



En el diagrama se observa que la mayoría de reviese son positivas, con calificaciones de 5.



Score	col_0 frecuencia
1	36306
2	20802
3	29769
4	56094
5	250962

Se ajusta expresiones de abreviaciones de palabras, como se indica a continuacion.

```
# specific
phrase = re.sub(r"won't", "will not", phrase)
phrase = re.sub(r"can\'t", "can not", phrase)

# general
phrase = re.sub(r"n\'t", " not", phrase)
phrase = re.sub(r"\'re", " are", phrase)
phrase = re.sub(r"\'s", " is", phrase)
phrase = re.sub(r"\'d", " would", phrase)
phrase = re.sub(r"\'ll", " will", phrase)
phrase = re.sub(r"\'t", " not", phrase)
phrase = re.sub(r"\'ve", " have", phrase)
phrase = re.sub(r"\'m", " am", phrase)
```

Stopwords. Se realizar el filtro de palabras sin significado como artículos, pronombres, preposiciones, etc.

```
stopwords= set(['br', 'the', 'i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've",\
            "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', \
            'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their',\
            'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', \
            'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', \
            'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', \
            'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after',\
            'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further',\
            'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more',\
            'most', 'other', 'some', 'such', 'only', 'own', 'same', 'so', 'than', 'too', 'very', \
            's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', \
            've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn',\
            "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn',\
            "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", \
            'won', "won't", 'wouldn', "wouldn't"])
```

Se usa para técnicas para procesar el lenguaje natural:

- Bag of Words
- Tfidf
- Word2vec
- Average Word2vec

Luego permiitira comparar con el modelo, cual premite un mejor desempeño.

2. Mejoras Futuras.

Usar más features del modelo para la predicción y aprovechar más características del conjunto de datos como la descripción del producto.

Recopilar más datos de la web de Amazon mediante técnicas de web scrapping para aumentar la cantidad de datos para entrenar y mejorar el proceso de cálculo

Se pueden mostrar algunos análisis más entre la descripción del producto y otras características del conjunto de datos.

Usar el parámetro de tiempo, para obsevar si tiene alguna implicación en la predicción del score del producto.