Universidad del Valle de Guatemala
Data Science

25 de agosto 2020

**Laboratorio 5:**
**Análisis de Sentimientos**

**Análisis exploratorio:**

**- Se elaboró un análisis exploratorio en el que se explican los cruces de variables, hay gráficos explicativos y análisis que permiten comprender el conjunto de datos.**


**Figura 1:** Cantidad de columnas y filas


**Figura 2:** Columnas que se encuentran en el conjunto de datos

El conjunto de datos contiene información sobre reseñas de usuarios hacia diferentes productos. En la figura 1 podemos ver que el conjunto de datos cuenta con 25 filas y 71044 filas. El conjunto de datos cuenta con columnas que no necesariamente son necesarias o útiles para conocer el sentimiento expresado en las reseñas, pero probablemente sean interesantes para el análisis exploratorio.

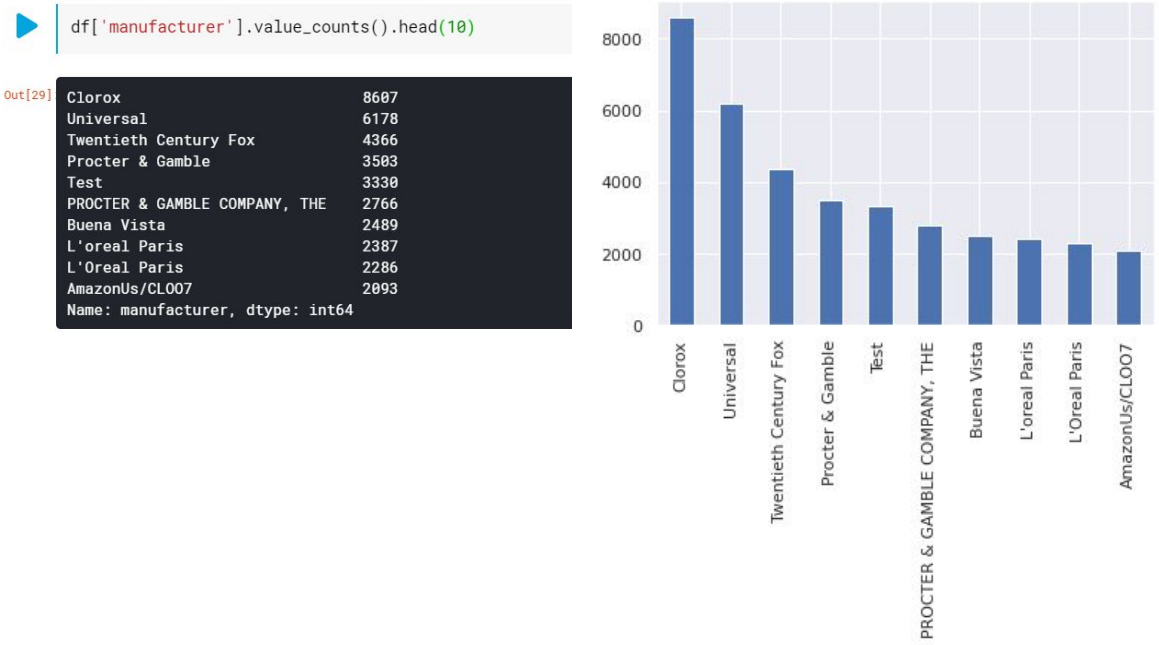


**Figura 3 y 4:** Cantidad de reseñas según el fabricante

En el conjunto de datos tenemos la columna fabricantes (*manufacturer)*, vamos a investigar la cantidad de fabricantes que existen y cuantas reseñas hay por cada uno de los fabricantes. En la figura 3 y 4, se puede ver la que Clorox, Universal y TCF son los mayores fabricantes de productos. Es importante aclarar que dentro del conjunto existen 463 diferentes fabricantes.

**(20 puntos) Limpieza y preprocesamiento de los datos:**
**- Se documentan las tareas de limpieza, incluyendo los paquetes/módulos que se usaron.**

**Eliminar columnas innecesarias.**

A menudo, encontrará que no todas las categorías de datos de un conjunto de datos le resultan útiles. Por ello, se eliminan las columnas que no sirvan o den información relevante para el análisis de sentimientos de las reseñas. Eliminaremos las columnas:

- id: no podemos obtener ninguna información de este, a parte se encuentra en un formato extremadamente horroroso.

- categories: eliminamos esta columna al no tener la información que necesitamos para analizar el texto

- dateAdded, dateUpdated: ya que estas son solo fechas sin mucha información que nos pueda ser útil es que las eliminamos.

- ean: debido a que esta da un número que indica la variación del producto.

- Keys: es una lista de identificadores internos de DataFiniti para este producto.

- manufacturerNumber: esto es porque tenemos el manufactured, que es lo mismo.

- name: esta variable redunda con la de manufacturer, además de ser más larga, cuando la otra es más clara.

- reviews.didPurchase: una variable que solo indica si se compró el producto o no y sentimos que no tiene gran relevancia por el concepto en el que estamos.

- reviews.id: este es solo el id de los reviews, los cuales no nos dan ninguna información útil.
- reviews.numHelpful y reviews.rating: tenemos un par de variables numéricas, donde la primera está muy incompleta con muchos vacíos y la otra es algo difícil de ver su utilidad en el análisis de texto, por ser muy simplista.

- reviews.sourceURLs: luego está el url de donde se obtuvo la reseña y que eliminaremos por no ser realmente útil, ya que es un url.

- reviews.userCity y reviews.userProvince: solo son nombres de la ciudad y de la provincia de quién hizo el review, lo cual no creemos sea de mucha utilidad para el análisis.

- upc: una variable numérica que no vemos una gran utilidad al agregarla a nuestro análisis, así que por eso mismo es que la descartamos.

**Limpieza del Dataset:**

Para el procesamiento de la data hicimos uso de las siguientes librerías listadas a continuación:

- ➔ library(stringi)
- ➔ library(dplyr)
- ➔ library(qdapRegex)
- ➔ install.packages("qdapRegex")
- ➔ install.packages("stopwords")
- ➔ library("stopwords")
- ➔ library('tm')
- ➔ library(xlsx)

Con estas librerías se llevó a cabo la limpieza de la data haciendo los siguientes procesos:

- Convertir el texto a mayúsculas o a minúsculas
- Quitar los caracteres especiales que aparecen como “#”,”@” o los apóstrofes.

- Quitar las url
- Revisar si hay emoticones y quitarlos (a menos que le den información)
- Quitar los signos de puntuación
- Quitar los artículos, preposiciones y conjunciones (stopwords)
- Quitar números que interferirá en las predicciones.
- eliminar las filas duplicadas.

**Clasificación de las palabras:**
**- Clasificación de las palabras en positivas, negativas y neutrales. Explicación de las fuentes de datos o diccionarios utilizados. Gráficas y nubes de palabras**

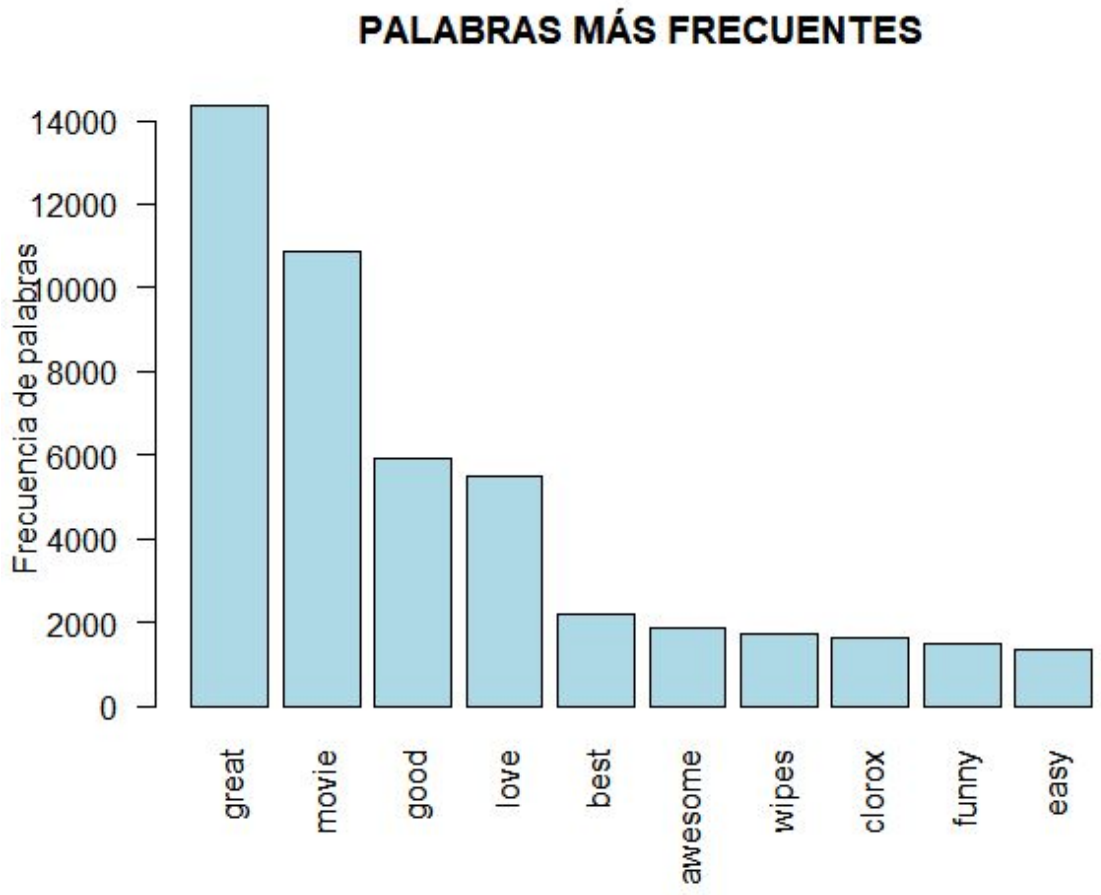


**Figura 5:** Histograma de las palabras más usadas


**Figura 6:** nubes de palabras más usadas en los reviews

Para poder determinar si una palabra era negativa o positiva, se utilizó la librería de “sentimentr” en R. Esta librería permite el análisis de sentimientos, ya sean positivos o negativos, además de permitir clasificar las palabras de un texto. Esto nos dio el siguiente output de datos que permiten ver cuál fue la clasificación de las palabras según el algoritmo.

| element_id | sentence_id | negative | neutral | positive | sentence |
|---|---|---|---|---|---|
| 1 | 1 | c("hype", "crazy") | c("album", "hip", "hop", "side", "current", "pop", "sound", "lis... | c("love", "good", "star") | love album good hip hop side current pop sound hype liste... |
| 2 | 1 | character(0) | c("flavor", "review", "collected", "part") | c("good", "promotion") | good flavor review collected part promotion |
| 3 | 1 | character(0) | flavor | good | good flavor |
| 4 | 1 | c("disappointed", "disappointed", "messy", "difficult", "lacke... | c("read", "read", "reviews", "reviews", "looking", "buying", "o... | c("enhanced", "captivating", "captivating", "sensation", "sens... | read reviews looking buying one couples lubricants ultimate... |
| 5 | 1 | c("irritation", "burning") | c("husband", "bought", "gel", "gel", "gel", "us", "caused", "fel... | c("like", "recommend") | husband bought gel us gel caused irritation felt like burning... |
| 6 | 1 | disappointed | c("boyfriend", "bought", "bought", "spice", "things", "bedroo... | love | boyfriend bought spice things bedroom highly disappointe... |
| 7 | 1 | disappointed | c("bought", "earlier", "today", "check", "based", "product", "d... | c("excited", "like", "fans") | bought earlier today excited check based product descriptio... |
| 8 | 1 | c("tingling", "sticky", "disappointing") | c("bought", "product", "husband", "try", "warming", "left", "u... | c("impressed", "loved") | bought product husband try impressed tingling warming lef... |
| 9 | 1 | c("disappointed", "waste", "sticky", "mess") | c("husband", "bought", "werevboth", "extremely", "especially... | c("extra", "fun", "money") | husband bought extra fun werevboth extremely disappointe... |
| 10 | 1 | character(0) | c("got", "husband", "nothing", "just", "just", "lube", "lube", "g... | c("surprise", "special", "save", "money", "wish", "refund") | got surprise husband nothing special just lube save money ... |
| 11 | 1 | character(0) | c("tried", "husband", "felt", "different", "effect", "change", "fir... | character(0) | tried husband felt different effect change first te using ky ge... |
| 12 | 1 | waste | c("purchased", "thinking", "sort", "warming", "effect", "origin... | money | purchased thinking sort warming effect original dont waste ... |
| 13 | 1 | disappointed | c("bought", "time", "bit", "absutely", "nothing", "especially", ... | enhance | bought enhance time bit absutely nothing disappointed es... |
| 14 | 1 | disappointed | c("bought", "really", "n", "nothing", "product") | c("liking", "variety", "like") | bought really liking n variety nothing like disappointed prod... |
| 15 | 1 | sticky | c("first", "time", "type", "product", "maybe", "expected", "mu... | c("purchase", "feeling") | first time purchase type product maybe expected much thin... |
| 16 | 1 | character(0) | c("bought", "tried", "wife", "feel", "anything", "product", "pro... | c("enjoy", "enhancement") | bought tried wife enjoy feel anything enhancement product... |
| 17 | 1 | character(0) | c("bought", "product", "spice", "things", "fianc", "didnt", "any... | passion | bought product spice things fianc didnt anything passion wi... |
| 18 | 1 | disappointing | c("bought", "try", "spice", "things", "felt", "nothing") | honest | bought try spice things felt nothing disappointing honest |
| 19 | 1 | character(0) | c("bought", "reviews", "jelly", "feel", "didnt", "us", "warmfeeli... | c("better", "like", "work", "well", "kind") | bought better reviews jelly feel like didnt work well us kind ... |
| 20 | 1 | wrong | c("first", "first", "time", "time", "time", "using", "ky", "didnt", "... | like | first time using ky didnt anything opinion felt like every tim... |
| 21 | 1 | c("disappear", "drawback", "sticky", "sticky", "irritation") | c("noticed", "product", "product", "store", "shop", "frequenti... | c("clearance", "love", "like", "like", "better", "pleasurable", "a... | noticed product clearance store shop frequently decided gi... |
| 22 | 1 | character(0) | c("used", "product", "couple", "times", "since", "recived", "try... | c("free", "good", "good", "warm", "pretty", "sensation", "reco... | used product couple times since recived free try good exper... |
| 23 | 1 | c("sticky", "bad") | c("used", "times", "stays", "bit", "products") | better | used times stays bit sticky bad better products |
| 24 | 1 | character(0) | c("review", "collected", "part") | c("loved", "promotion") | loved review collected part promotion |
| 25 | 1 | character(0) | c("product", "will", "give") | c("good", "great", "feeling") | good product will give great feeling |
| 26 | 1 | c("tired", "boring", "shake", "disappoint", "sticky", "tacky") | c("really", "using", "product", "product", "product", "husband... | c("enjoyed", "marriage", "awesome", "sensation", "excellent",... | really enjoyed using product husband years marriage things... |
| 27 | 1 | sticky | c("product", "product", "couples", "lubricant", "lot", "lubrican... | c("awesome", "warm", "feeling", "like", "easy", "clean", "recei... | awesome product couples warm feeling lubricant sticky like ... |
| 28 | 1 | c("slimy", "shake") | c("product", "warming", "experience", "partners", "least", "lif... | c("exceptional", "smooth", "sensation", "share", "great") | exceptional product smooth slimy warming sensation experi... |
| 29 | 1 | sticky | c("used", "first", "time", "time", "boyfriend", "just", "much", "... | c("loved", "great") | used first time boyfriend loved just much didnt get sticky fel... |
| 30 | 1 | character(0) | c("product", "exactly", "says") | c("great", "works", "well") | great product exactly says works well |
| 31 | 1 | c("lemon", "enough", "smells") | c("scent", "frosting", "wont", "able", "get", "lotion", "lotion", "... | c("yes", "best", "like", "like", "natural", "major", "completely", ... | yes scent best like lemon frosting wont able get enough loti... |
| 32 | 1 | character(0) | c("handcreamlemoncream", "fragrance", "last", "day", "hand... | c("refreshing", "wish", "wish", "soothing", "compliments", "gr... | handcreamlemoncream refreshing fragrance wish last day s... |
| 33 | 1 | c("smells", "lemon") | c("cookie", "absolutely", "kids") | c("like", "love", "love") | smells like lemon cookie absolutely love kids love |
| 34 | 1 | character(0) | c("will", "lotion", "every", "time", "put") | c("love", "smile") | will love lotion smile every time put |

**Figura 7:** Clasificado de las palabras neutras, positivas y negativas

Claramente podemos ver que gracias a uno de los recursos que proveo Lynnet para hacer este análisis, nos permite obtener una clasificación muy buena en base a esta librería que se encargó de obtener la clasificación de todas las palabras.

**Algoritmo de clasificación:**

**- Se describe el algoritmo que se usó para clasificar el review en positivo, negativo o neutro.**

Para determinar qué tan positivo o negativo era una review, se utilizó también la librería de “sentimentr” para su clasificación. Esta librería permite poder hacer análisis de todo tipo de textos. ya sea por oraciones o aplicando tokenización a frases, párrafos, etc. Así que en esta ocasión lo hicimos como si fuera de párrafos. Así obtuvimos un índice que nos indicaba que tan positivo o negativo podía llegar a ser una reseña. Esto dado porque si sentiment és negativos, es porque fue una reseña negativas y positivo en el caso opuesto. a continuación mostraré un ejemplo de como se ver dicha clasificación:

| | element_id | sentence_id | word_count | sentiment |
|---|---|---|---|---|
| 1 | 1 | 1 | 2 | 0.42426407 |
| 2 | 2 | 1 | 1 | 0.75000000 |
| 3 | 3 | 1 | 1 | 0.75000000 |
| 4 | 4 | 1 | 1 | -1.00000000 |
| 5 | 5 | 1 | 1 | -0.75000000 |
| 6 | 6 | 1 | 1 | 0.75000000 |
| 7 | 7 | 1 | 1 | -1.00000000 |
| 8 | 8 | 1 | 1 | 0.75000000 |
| 9 | 9 | 1 | 1 | -1.00000000 |
| 10 | 10 | 1 | 2 | 0.00000000 |
| 11 | 11 | 1 | 2 | 0.53033009 |
| 12 | 12 | 1 | 2 | -0.10606602 |
| 13 | 13 | 1 | 2 | -0.10606602 |
| 14 | 14 | 1 | 2 | 0.53033009 |
| 15 | 15 | 1 | 1 | -1.00000000 |
| 16 | 16 | 1 | 1 | 0.00000000 |
| 17 | 17 | 1 | 1 | -1.00000000 |
| 18 | 18 | 1 | 1 | -1.00000000 |
| 19 | 19 | 1 | 2 | 0.88388348 |
| 20 | 20 | 1 | 3 | 0.00000000 |
| 21 | 21 | 1 | 1 | 0.75000000 |

**Figura 8:** Muestra de los resultados del clasificador de reviews

A continuación estarán los incisos pedidos para la clasificación de las reviews, según lo obtenido por el algoritmo:

- Cuáles son los 10 productos de mejor calidad dado su review.

```
> setDT(sentimientos)[order(-sentiment), .SD[1:10]]
    element_id sentence_id word_count sentiment
 1:       8232           1          4  2.275000
 2:      15919           1          5  2.236068
 3:      16395           1          6  2.122891
 4:      31033           1          5  1.896186
 5:      55309           1          3  1.870615
 6:      62659           1          3  1.818653
 7:      32597           1          3  1.732051
 8:       2138           1          3  1.662769
 9:       4061           1          3  1.662769
10:      49927           1          3  1.662769
> TopBest <- setDT(sentimientos)[order(-sentiment), .SD[1:10]]
> TopBest$element_id[0]
integer(0)
> TopBest$element_id[1]
[1] 8232
> DataClean$reviews.text[TopBest$element_id[1]]
[1] "clorox wipes convenient cleaning want get something really clean really quick review collected part promotion"
> DataClean$reviews.text[TopBest$element_id[1]-1]
[1] "easy use can trust items cleaned clean like fresh clean smell also review collected part promotion"
> TopBest <- setDT(sentimientos)[order(-sentiment), .SD[1:10]]
> DataClean$reviews.text[TopBest$element_id[1]-1]
[1] "easy use can trust items cleaned clean like fresh clean smell also review collected part promotion"
> DataClean$reviews.text[TopBest$element_id[2]-1]
[1] "grateful burts bees many products obsessed brand lip shimmer nutmeg one little color swatch website midleading shade actually beautiful coral glad found shade please never dis
continue perfect every way perfect tone size texture ingredients value thanks burts allowing switch parabenfree cosmetics easy end eating lip products inadvertently anyway natural"
> DataClean$reviews.text[TopBest$element_id[3]-1]
[1] "love rice"
> DataClean$reviews.text[TopBest$element_id[4]-1]
[1] "found remake jungle book entertaining stayed fairly true original awakened inner child meexcellent"
> DataClean$reviews.text[TopBest$element_id[5]-1]
[1] "purchased smead file folders based idea saw pinterest place childrens artwork file folder year life save file storage bin just got done creating happy product folders sturdy c
olors great easy creating labels placing inside plastic tabs confident will last happy now place store daughters projects"
> DataClean$reviews.text[TopBest$element_id[6]-1]
[1] "watch movie expecting humor seth rogens projects surprise yea sure vulgar adult animated move oh yea please let children watch movie well written funny gets better watch coupl
e times catch lil anecdotes also digital copy comes k uhd bluray k uhd copy sonys k ultra app great treat k tv without uhd bluray player"
> DataClean$reviews.text[TopBest$element_id[7]-1]
[1] "love nail polish drys super fast colors beautiful"
> DataClean$reviews.text[TopBest$element_id[8]-1]
[1] "moments give laugh glad didnt see theaters"
> DataClean$reviews.text[TopBest$element_id[9]-1]
[1] "product helped become organized file cabinet now less bulky"
> DataClean$reviews.text[TopBest$element_id[10]-1]
[1] "ive waiting finally arrived using days now surprised actually getting questions using face will honest little skeptical first arrived loyal fan tried true cream sure change pr
oduct actual ingredients certainly noticing something happening people commenting work lets see happens solid week review collected part promotion"
```

**Figura 9:** top 10 de de los mejores comentarios y su review

- Cuáles son los 10 productos de menor calidad dado su review.

```
    element_id sentence_id word_count sentiment
 1:       1803           1          3 -2.078461
 2:       3074           1          4 -1.500000
 3:      15003           1          2 -1.414214
 4:      44300           1          2 -1.414214
 5:      52830           1          2 -1.414214
 6:      58529           1          2 -1.414214
 7:      20644           1          3 -1.299038
 8:      21490           1          3 -1.299038
 9:      29813           1          3 -1.299038
10:      45150           1          3 -1.299038
> DataClean$reviews.text[Topworst$element_id[1]-1]
[1] "boring movie couldnt watch minutes"
> DataClean$reviews.text[Topworst$element_id[2]]
[1] "movie just needed never stop added much original"
> DataClean$reviews.text[Topworst$element_id[3]]
[1] "thing takes smell kitchen drains really dislike smell just put small amount cover drain awhile soak stopper"
> DataClean$reviews.text[Topworst$element_id[4]]
[1] "like strawberries salads helps keep fresher less soggy crisp buy"
> DataClean$reviews.text[Topworst$element_id[5]-1]
[1] "good concept poor design unlike negative reviews problems sprayer bottle worked fine problem like oregon mop head fell third time used kept reattaching keep fallin
g use frustrating closer inspection can see poor design mop head attached handle uses exerting heavy pressure mopping handle clip pushed open will close enough secure m
op head maybe dainty dont push attachment will last longer im lbs just trying clean kitchen floor replaced similar mop clorox solid clippin mop head attaches dont expec
t issue always think rubbermaid high quality big disappointment uses"
> DataClean$reviews.text[Topworst$element_id[6]-1]
[1] "usual second movie didnt match original based movie wasnt bad definitely good original"
> DataClean$reviews.text[Topworst$element_id[7]-1]
[1] "bring back old formula new one smells terrible switching another brand"
> DataClean$reviews.text[Topworst$element_id[8]-1]
[1] "used clear scalp shampoo conditioner couple months noticed developing sores scalp couple months ago went away recently came back stopped using scalp started cleari
ng shampoo made hair feel great left scalp bad condition"
> DataClean$reviews.text[Topworst$element_id[9]-1]
[1] "worst flakeing mascara ever ten minutes apply flakes will buy"
> DataClean$reviews.text[Topworst$element_id[10]-1]
[1] "switched normal brand try tide pods really hoping product work family using pods noticed blue streaks clothes brand new first time washing must say disapointed pro
duct will buy squeeze product water washer will happen however reason bought simplify life"
>
```

**Figura 10:** top 10 de de los peores comentarios y su review

- Cuáles son los usuarios que dan la mayor cantidad de reviews a distintos productos.

```
[9]: # 6.3
df_manufacturer_username = df[['manufacturer', 'reviews.username']]
df3 = df_manufacturer_username.drop_duplicates()
df3['reviews.username'].value_counts().head(20)
```

```
Out[9]: An anonymous customer    68
Anonymous                26
Mike                     25
Chris                    24
Lisa                     22
Sandy                    22
Rick                     22
John                     21
Laura                    18
James                    18
Melissa                  17
Robert                   17
Joey                     17
Cindy                    17
Debbie                   17
Susan                    17
Dave                     17
Mimi                     17
Mary                     17
Linda                    17
Name: reviews.username, dtype: int64
```

**Figura 11:** Fabricantes con más reviews a distintos productos

Los resultados indican que la mayor cantidad de personas realizan una reseña de tipo anónimo. A pesar de ello los usuarios Mike, Chris y Lisa han realizado más de 20 reseñas cada uno a distintos productos.

- Cuáles son los productores que tienen productos de mejor calidad.

Out[95]:

| | manufacturer | reviews.rating |
|---|---|---|
| 0 | AmazonUs/CLOO7 | 4.841854 |
| 1 | Buena Vista | 4.739654 |
| 2 | Clorox | 4.821424 |
| 3 | L'Oreal Paris | 4.456255 |
| 4 | L'oreal Paris | 4.456640 |
| 5 | PROCTER & GAMBLE COMPANY, THE | 4.077368 |
| 6 | Procter & Gamble | 3.934627 |
| 7 | Test | 4.349550 |
| 8 | Twentieth Century Fox | 4.095053 |
| 9 | Universal | 4.663807 |

+ Code + Markdown

**Figura 12:** Mejores fabricantes según la calidad

Sería inutil obtener el promedio de las reseñas para todos los fabricantes, debido a que existen muchos fabricantes y algunos de ellos tienen pocas reseñas. Por ello, obtendremos cuales son los fabricantes que tienen reseñas más positivas que se encuentren entre los 20 fabricantes con más reseñas, utilizando las columnas *manufacturer y reviews.rating*.

- Cuáles son los productores que tienen productos de peor calidad.

| | | |
|---|---|---|
| 19 | WarnerBrothers | 4.560510 |
| 6 | L'oreal Paris | 4.456640 |
| 5 | L'Oreal Paris | 4.456255 |
| 14 | Test | 4.349550 |
| 13 | Sony Pictures | 4.152607 |
| 16 | Twentieth Century Fox | 4.095053 |
| 9 | PROCTER & GAMBLE COMPANY, THE | 4.077368 |
| 10 | Procter & Gamble | 3.934627 |
| 11 | Rubbermaid | 2.801444 |
| 7 | Nexxus Beauty Products | 1.352691 |

**Figura 13:** Peores fabricantes según la calidad

Los resultados, en la figura 7, muestran los peores fabricantes según la calidad. Nexxus Beauty Products está siendo el peor fabricante.

**-Se elaboró una función que permite predecir las n posibles palabras que escribirá el usuario tras la frase ingresada.**

Para la creación de la función de predicción se inició con la creación de una matriz que asocia cada lista de términos con sus propios términos frecuentes, mediante la librería TDM. Una vez asociadas las listas de reviews con la matriz de frecuencias se aplicó el modelo KNN. Este hace una asociación entre los términos frecuentes de cada vector de palabras con su fila asociada en la matriz de palabras frecuentes.

**- Se elaboró una propuesta de estrategia para el productor que tiene más productos con reviews negativos.**

Para aumentar los reviews positivos para un producto se puede realizar un modelo de clasificación que podamos obtener todos los reviews negativos. Luego de esto podemos contar con un grupo de asesores de calidad que obtengan retroalimentación sobre en qué está fallando el producto. Es importante que menciones que lo mejor sería obtener solo reseñas de las personas que se sabe que compraron el producto.

Otra alternativa, siendo más minuciosos es saber clasificar los diferentes productos y categorías de este. De esta forma podemos ver si es alguna categoría en la que la calidad está bajando.