Universidad del Valle de Guatemala
Data Science
Catedrática: Lynette García
Pablo Viana 16091
José Martínez 15163
Sergio Marchena 16387

Laboratorio 6: Análisis de Sentimientos

Preprocesamiento

En esta etapa se realizó una limpieza de datos para las columnas “reviews.text” y “reviews.title”. En la limpieza se trabajaron los siguientes aspectos:

- Se convirtió todo el texto a minúsculas.
- Se removieron caracteres especiales como “@”, “#” o apóstrofos.
- Se removieron *stopwords* para dejar palabras que sean significativas para el análisis.
- Se removieron signos de puntuación.
- Se removieron números.
- Se removieron *urls*.

Para la limpieza se utilizó la librería *tm*. Esta librería provee una lista de *stopwords* en inglés que ayudaron para eliminar las mismas en el dataset. También se utilizó esta librería para remover número y signos de puntuación.

Análisis Exploratorio

El conjunto de datos contiene 71 045 reviews de 1000 productos diferentes, se incluye el título y el texto de los reviews, el nombre del productor, los metadatos del cliente, y más.

Para comenzar el análisis exploratorio buscaremos las palabras que más se repite en el texto del review por cada fila. El problema es que el texto no es lo suficientemente rico en palabras para poder hacer 1-gramas y encontrar las palabras que más se repiten. El plan de acción es como sigue: primero, se convertirá a factor el nombre del producto para saber su frecuencia. Si se determina que existen varias instancias de cada producto, se unirán sus textos de review. Esto para crear un solo párrafo que contenga todos los reviews del producto. Luego, sobre estos párrafos unificados se buscarán n-gramas, que corresponden a las palabras o frases que más se repiten en todos los reviews.

Una vez efectuada esta operación, es posible determinar las palabras y frases que más se repiten a lo largo de todos los reviews.

Tablas y Gráficos.

Top 10 de productos más vendidos.

	Var1	Freq
126	Clorox Disinfecting Wipes Value Pack Scented 150 Ct Total	8606
549	The Secret Life Of Pets (4k/uhd + Blu-Ray + Digital)	5510
219	Independence Day Resurgence (4k/uhd + Blu-Ray + Digital)	3609
554	Tide Original Liquid Laundry Detergent - 100 Oz	3498
178	Godzilla 3d Includes Digital Copy Ultraviolet 3d/2d Blu-Ray/...	3330
383	Olay Total Effects Daily Moisturizer, 7-In-1 Anti-Aging, 0.5oz	2766
544	The Jungle Book (blu-Ray/dvd + Digital)	2489
264	L'or233al Paris Elvive Extraordinary Clay Rebalancing Conditi...	2288
272	L'oreal Paris Revitalift Triple Power Deep-Acting Moisturizer	2286
125	Clorox Disinfecting Bathroom Cleaner	2093

Top 10 usuarios con más reviews registradas

	Var1	Freq
2271	An anonymous customer	421
2864	Anonymous	113
36349	Mike	105
1		95
10778	Chris	88
7917	ByAmazon Customer	58
13813	Dave	57
25920	John	49
45393	Rick	41
23277	James	32

Top 10 de marcas con reviews

	Var1	Freq
81	Clorox	10700
363	Universal Home Video	6178
346	Tide	5384
116	FOX	4498
99	Disney	3692
245	Olay	3420
375	Warner Home Video	3330
181	L'Oreal Paris	2683
180	L'oreal Paris	2288
318	Sony Pictures	2161

Top 10 de productores con más reviews

	Var1	Freq
81	Clorox	8607
426	Universal	6178
420	Twentieth Century Fox	4366
317	Procter & Gamble	3503
398	Test	3330
319	PROCTER & GAMBLE COMPANY, THE	2766
56	Buena Vista	2489
210	L'oreal Paris	2387
211	L'Oreal Paris	2286
13	AmazonUs/CLOO7	2093

Top 10 Ciudades donde hacen más reviews

	Var1	Freq
1		65634
365	Chicago	68
1381	New York	62
901	Houston	57
1134	Los Angeles	56
466	Dallas	42
77	Atlanta	40
1244	Miami	39
1558	Philadelphia	38
1779	San Diego	37

Modelos y Clasificación de Términos

Para la clasificación de términos y análisis de sentimientos, se utilizó la librería de R llamada “sentimentr”, la cual ya trae funciones especiales para que dado un texto (string) se puede extraer varias cosas:

- El análisis general del texto, es decir si es positivo o negativo.
- El número de palabras y oraciones en el texto dado.
- La lista de palabras negativas, neutras y positivas según el texto dado.

Pruebas con los datos proporcionados en la columna de “reviews.text”:

Para esta prueba se usó el texto del review con “reviews.id” = 148314686, el cual tiene el siguiente “reviews.text”: [1] "My husband and I bought this for some extra fun. We werevboth extremely disappointed. Especially for the price! Do not waste your money on this product. We felt nothing but a sticky mess from it."

El análisis general del texto.

```
> sentiment_by(text, var = NULL)
   element_id word_count        sd ave_sentiment
1:          1         35 0.5484875    -0.2552077
```

La variable de “ave_sentiment” es la que nos indica si es negativa o positivo el texto. En este caso como es -0.25, es un número negativo, entonces nos dice que el texto esta clasificado como negativo.

El análisis detallado del texto.

```
> sentiment(text)
   element_id sentence_id word_count   sentiment
1:          1           1         10  0.36366193
2:          1           2          4 -0.90000000
3:          1           3          4  0.00000000
4:          1           4          8  0.05303301
5:          1           5          9 -0.75000000
```

Se puede ver que el texto proporcionado consiste en 5 oraciones. Cada oración tiene 10,4,4,8 y palabras respectivamente y se puede ver que cada oración tiene una clasificación de sentimientos, donde la oración 1 es positiva, las oraciones 2 y 5 son negativas y las oraciones 3 y 4 son neutras, por estar cercanas a 0.

Palabras negativas, positivas y neutras en el texto proporcionado.

```
> extract_sentiment_terms(text)
   element_id sentence_id     negative  positive
1:          1           1               extra,fun
2:          1           2 disappointed
3:          1           3
4:          1           4        waste      money
5:          1           5  sticky,mess
```

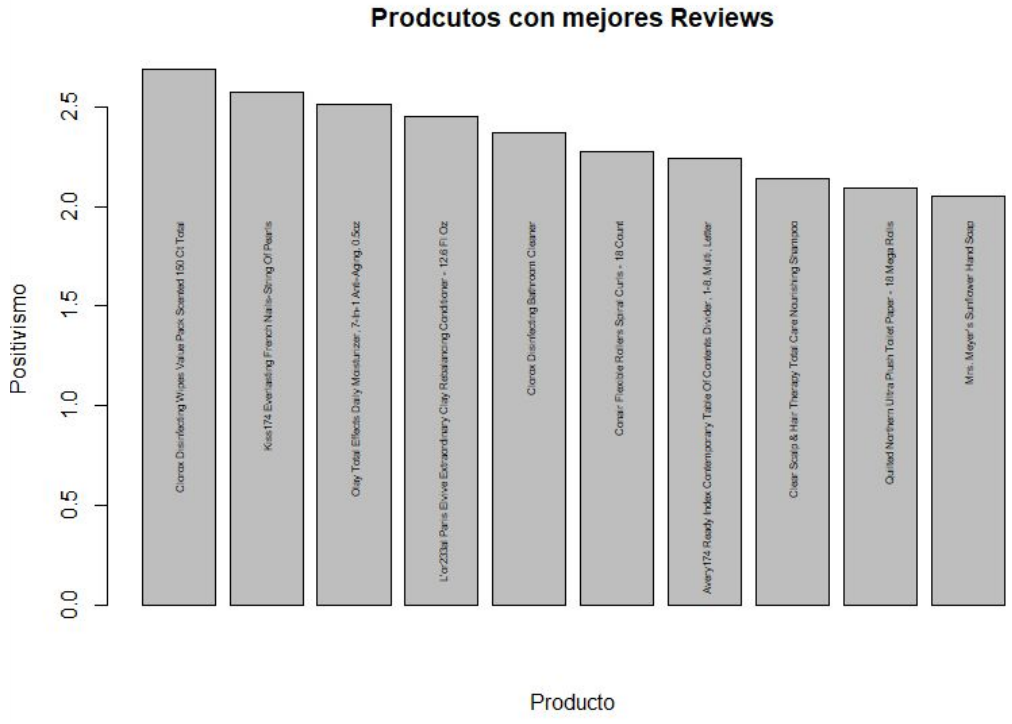
Se puede ver que esta ordenado por oraciones y que en la oración 1 y 4, la función encontró que hay 3 palabras positivas las cuáles son “extra”, “ fun” y “money”, que en las oraciones 2, 4 y 5 la función encontró 4 palabras negativas, las cuales son “disappointed”, “waste”, “sticky” y “mess”, por último la oración 3 no tenía palabras negativas ni positivas. Además se puede ver que en la oración 4 hay palabras positivas y negativas.

Se seguirá esta lógica y esta librería para poder determinar los productos y su análisis de sentimientos.

Preguntas para analizar.

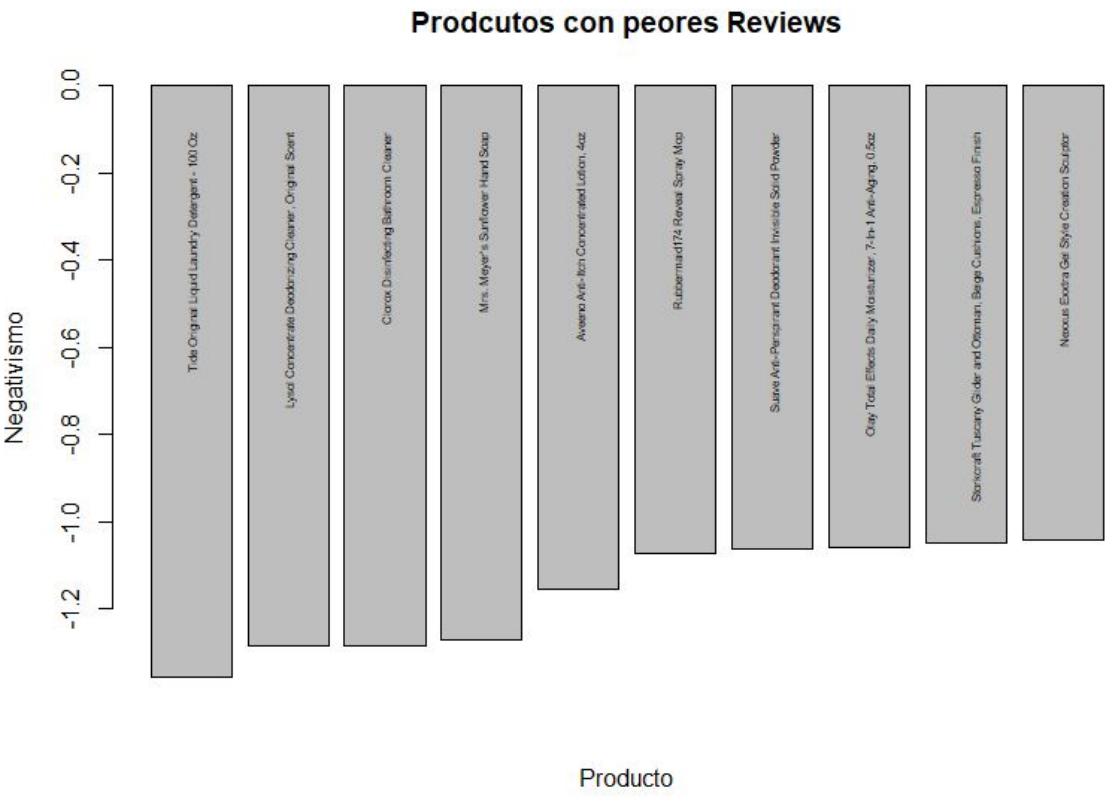
- Cuáles son los 10 productos de mejor calidad dado su review.

sentiment	name
2.687936	Clorox Disinfecting Wipes Value Pack Scented 150 Ct Total
2.575728	Kiss174 Everlasting French Nails-String Of Pearls
2.510229	Olay Total Effects Daily Moisturizer, 7-In-1 Anti-Aging, 0.5oz
2.451814	L'or233al Paris Elvive Extraordinary Clay Rebalancing Conditi...
2.367570	Clorox Disinfecting Bathroom Cleaner
2.278025	Conair Flexible Rollers Spiral Curls - 18 Count
2.240119	Avery174 Ready Index Contemporary Table Of Contents Divi...
2.138998	Clear Scalp & Hair Therapy Total Care Nourishing Shampoo
2.093082	Quilted Northern Ultra Plush Toilet Paper - 18 Mega Rolls
2.053685	Mrs. Meyer's Sunflower Hand Soap



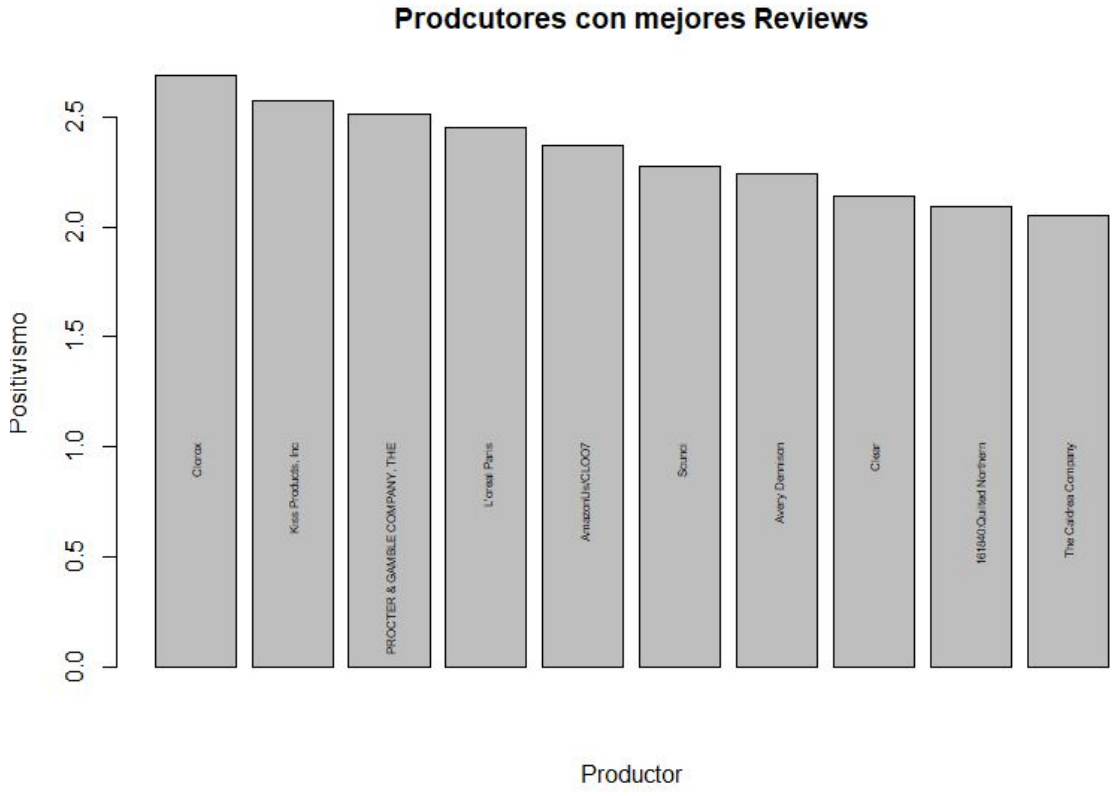
- Cuáles son los 10 productos de menor calidad dado su review

name	sentiment
Tide Original Liquid Laundry Detergent - 100 Oz	-1.355293
Lysol Concentrate Deodorizing Cleaner, Original Scent	-1.285439
Clorox Disinfecting Bathroom Cleaner	-1.284131
Mrs. Meyer's Sunflower Hand Soap	-1.272615
Aveeno Anti-Itch Concentrated Lotion, 4oz	-1.156590
Rubbermaid174 Reveal Spray Mop	-1.071866
Suave Anti-Perspirant Deodorant Invisible Solid Powder	-1.062563
Olay Total Effects Daily Moisturizer, 7-In-1 Anti-Aging, 0.5oz	-1.058750
Storkcraft Tuscany Glider and Ottoman, Beige Cushions, Esp...	-1.047947
Nexxus Exxtra Gel Style Creation Sculptor	-1.042983



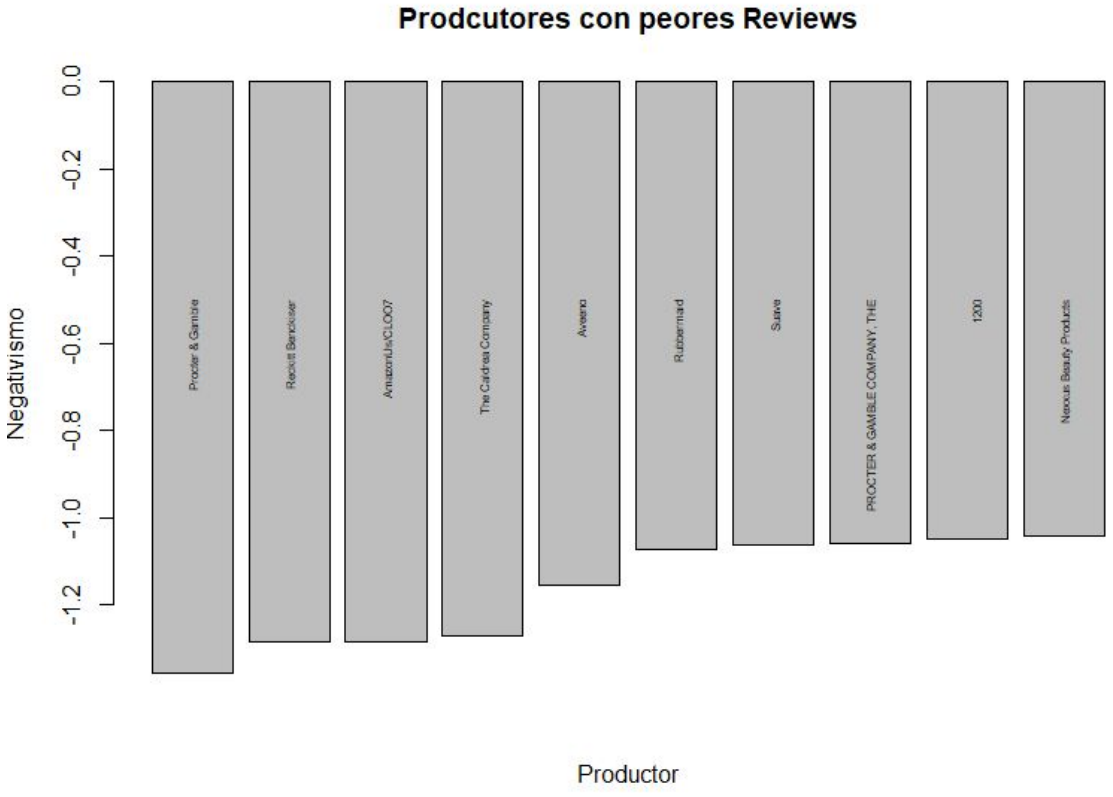
- Cuáles son los productores que tienen productos de mejor calidad.

manufacturer	sentiment
Clorox	2.687936
Kiss Products, Inc	2.575728
PROCTER & GAMBLE COMPANY, THE	2.510229
L'oreal Paris	2.451814
AmazonUs/CLOO7	2.367570
Scunci	2.278025
Avery Dennison	2.240119
Clear	2.138998
161840:Quilted Northern	2.093082
The Caldrea Company	2.053685



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

manufacturer	sentiment
Procter & Gamble	-1.355293
Reckitt Benckiser	-1.285439
AmazonUs/CLOO7	-1.284131
The Caldrea Company	-1.272615
Aveeno	-1.156590
Rubbermaid	-1.071866
Suave	-1.062563
PROCTER & GAMBLE COMPANY, THE	-1.058750
1200	-1.047947
Nexxus Beauty Products	-1.042983

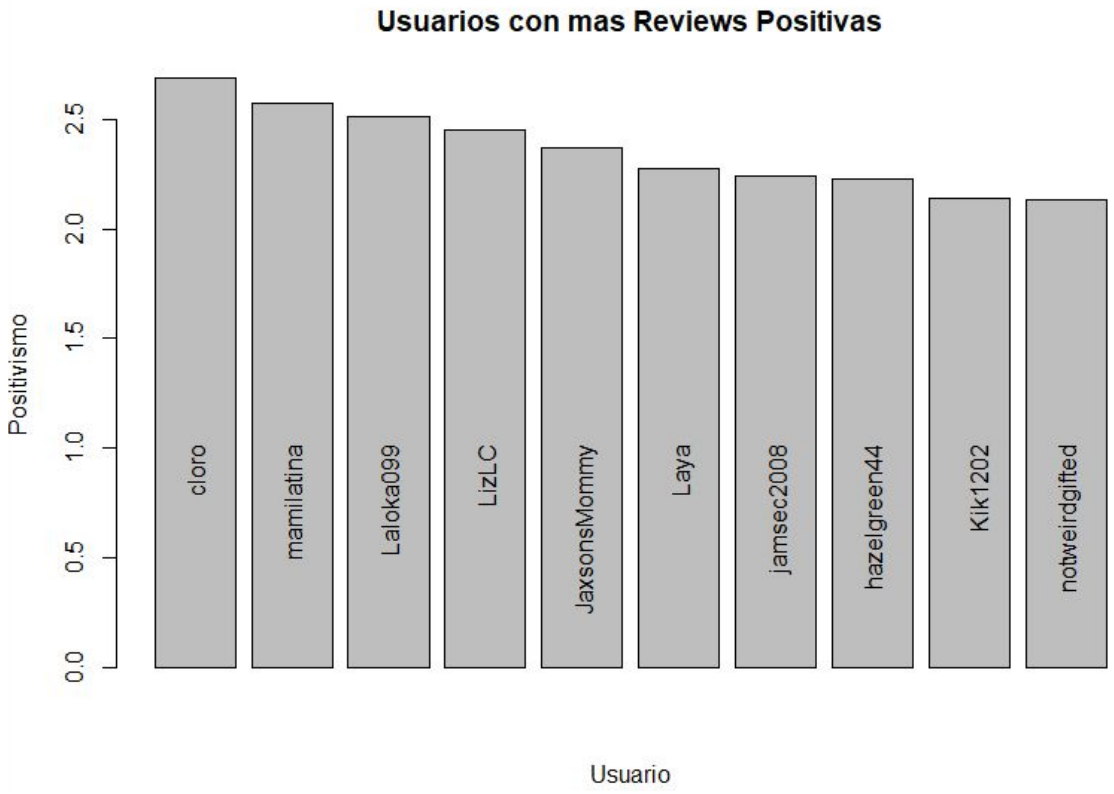


- Cuáles son los usuarios que dan la mayor cantidad de reviews a distintos productos.

	Var1	Freq
2271	An anonymous customer	421
2864	Anonymous	113
36349	Mike	105
1		95
10778	Chris	88
7917	ByAmazon Customer	58
13813	Dave	57
25920	John	49
45393	Rick	41
23277	James	32

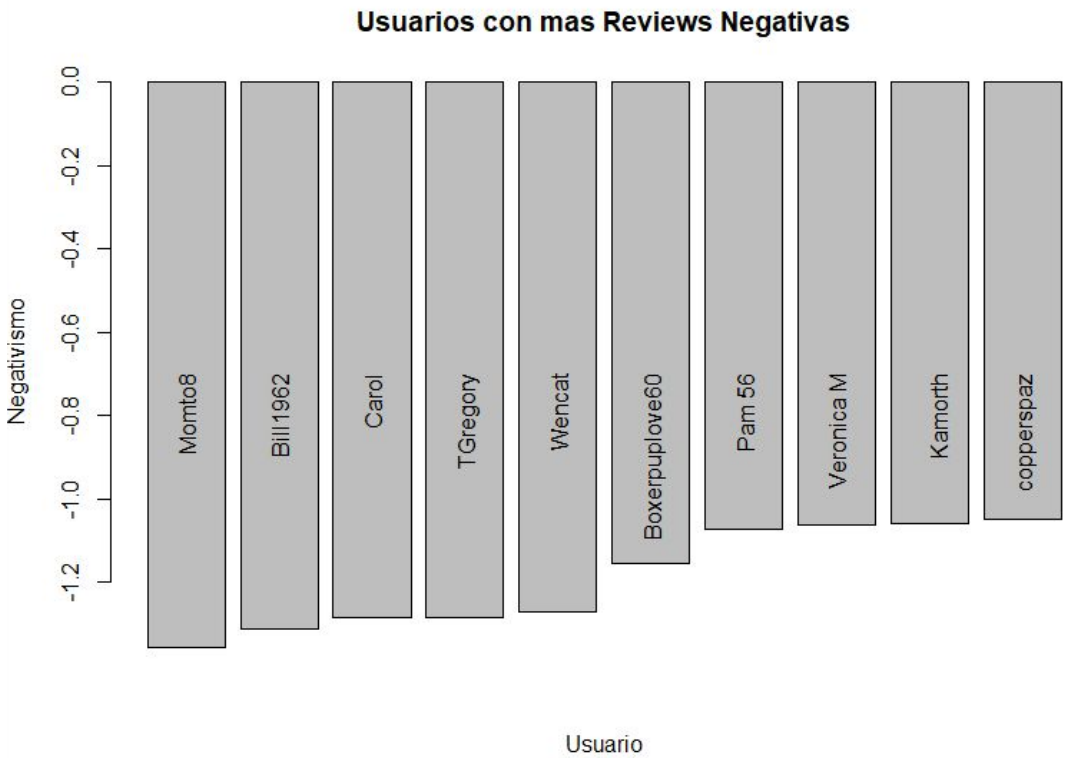
- Cuáles son los usuarios que más reviews negativos y positivos dan en promedio.
    - Positivos

reviews.username	sentiment
cloro	2.687936
mamilatina	2.575728
Laloka099	2.510229
LizLC	2.451814
JaxsonsMommy	2.367570
Laya	2.278025
jamsec2008	2.240119
hazelgreen44	2.231184
Kik1202	2.138998
notweirdgifted	2.136667



- Negativos

reviews.username	sentiment
Momto8	-1.355293
Bill1962	-1.310805
Carol	-1.285439
TGregory	-1.284131
Wencat	-1.272615
Boxerpuplove60	-1.156590
Pam 56	-1.071866
Veronica M	-1.062563
Kamorth	-1.058750
copperspaz	-1.047947



- Imagine que usted es analista de negocios y que está realizando este análisis para el productor que tiene más productos con malos reviews ¿Qué le propondría a esta empresa para mejorar sus productos?

Como primer paso se seleccionó la empresa con el promedio más bajo de sentimientos encontrados en sus reviews. Las empresas con menos de 15 reviews no se tomaron en cuenta ya que se consideró qué sí solo venden pocos productos y son malos reviews, su calificación sería muy baja, es por eso que solo se miraron empresas con más de 15 reviews para poder tener reviews más variados. La empresa seleccionada fue “Aveeno”, la cual tiene un promedio de “aceptación” de 0.005412462 siendo este valor muy cercano a ser negativo, lo cual es muy malo.

Como segundo paso se aislaron todas las reviews de esta empresa y se procedió a analizar los reviews para cada producto. Usando la librería descrita se pueden ver las palabras negativas por cada review. Estas son:

neg	list [36]	List of length 36
[[1]]	character [8]	'itch' 'sticky' 'tacky' 'uncomfortable' 'friction' 'itchy' ...
[[2]]	character [8]	'mosquito' 'itch' 'itch' 'itch' 'itch' 'itch' ...
[[3]]	character [1]	'discontinued'
[[4]]	character [4]	'allergic' 'poison' 'poison' 'sorry'
[[5]]	character [1]	'miss'
[[6]]	character [3]	'stops' 'itching' 'irritation'
[[7]]	character [1]	'allergies'
[[8]]	character [4]	'itchy' 'itchy' 'itchy' 'stopped'
[[9]]	character [4]	'suffer' 'bad' 'itchy' 'drawback'
[[10]]	character [5]	'irritated' 'hurt' 'rash' 'suffering' 'itchy'
[[11]]	character [5]	'itching' 'stop' 'anti' 'itch' 'stopped'
[[12]]	character [5]	'hate' 'itch' 'uncomfortable' 'itching' 'stop'
[[13]]	character [4]	'cancer' 'itchy' 'itchy' 'greasy'
[[14]]	character [3]	'itchy' 'anti' 'itch'
[[15]]	character [5]	'mosquito' 'poison' 'itchy' 'unfortunately' 'hard'
[[16]]	character [1]	'itchy'
[[17]]	character [1]	'itchy'
[[18]]	character [5]	'itchy' 'itchy' 'anti' 'itch' 'itch'
[[19]]	character [2]	'bug' 'irritations'
[[20]]	character [6]	'itchy' 'rash' 'rash' 'anti' 'itch' 'poison'
[[21]]	character [1]	'itching'
[[22]]	character [1]	'itchy'
[[23]]	character [2]	'problems' 'harsh'
[[24]]	character [3]	'itchy' 'rash' 'bug'
[[25]]	character [7]	'bouts' 'poison' 'poison' 'smell' 'stops' 'aggravating' ...
[[26]]	character [1]	'symptoms'
[[27]]	character [20]	'rash' 'rash' 'rash' 'pain' 'pain' 'pain' ...
[[28]]	character [2]	'stopping' 'problems'
[[29]]	character [1]	'irritate'
[[30]]	character [3]	'rash' 'itch' 'itch'
[[31]]	character [3]	'expensive' 'disability' 'problem'
[[32]]	character [19]	'anti' 'anti' 'anti' 'itch' 'itch' 'suffers' ...
[[33]]	character [6]	'bug' 'repellent' 'stumbled' 'burning' 'scratch' 'scars'
[[34]]	character [3]	'terrible' 'allergic' 'miserable'
[[35]]	character [1]	'irritation'
[[36]]	character [1]	'rash'

Se puede ver que la palabra más repetida es “itchy”. Esto sugiere que los productos o el producto que se esta vendiendo está causando algún tipo de reacción en la piel de los usuarios, causandoles irritacion y picazon. La recomendación sería ver la fórmula y asegurarse de lo que se esta vendiendo esta bien hecho.