



# Informe de prácticas

Tratamiento Inteligente de Datos

28 de enero de 2019

Felipe Peiró Garrido - [felipepg@correo.ugr.es](mailto:felipepg@correo.ugr.es)

José Andrés Bonilla - [jabonillab@correo.ugr.es](mailto:jabonillab@correo.ugr.es)

Juan Carlos Serrano Pérez - [jcsp0003@correo.ugr.es](mailto:jcsp0003@correo.ugr.es)

Pedro Manuel Gómez-Portillo López - [gomezportillo@correo.ugr.es](mailto:gomezportillo@correo.ugr.es)

# Índice

<b>1. Introducción</b>	<b>3</b>
<b>2. Dataset</b>	<b>4</b>
<b>3. Herramientas utilizadas</b>	<b>6</b>
3.1. KNIME	6
3.2. R	6
3.3. Python	6
<b>4. Preprocesamiento de datos</b>	<b>7</b>
4.1. Datos estadísticos	7
4.2. Diagrama de cajas	8
4.3. Correlación lineal	11
<b>5. Agrupamiento</b>	<b>12</b>
5.1. K-means	12
5.2. Clustering jerárquico	15
<b>6. Clasificación</b>	<b>17</b>
6.1. Árbol de decisión	17
6.2. Naive-Bayes	17
6.3. C4.5 de Quinlann	18
6.4. Random forest	18
6.5. Random forest con clústeres	19
<b>7. Regresión</b>	<b>20</b>
7.1. Correlación	20
7.2 Scatter Matrix	20
<b>8. Minería de texto</b>	<b>23</b>
8.1. Frecuencia de las palabras	23
8.2. Nube de palabras	25
8.3. Asociación de palabras	26
8.3.1. Like	26
8.3.2. Tast	27
8.3.3. Flavor	28
<b>Bibliografía</b>	<b>29</b>

# 1. Introducción

Amazon es una de las 500 mayores empresas de EE.UU. La compañía, con sede en Seattle, Washington, es el líder global en el comercio electrónico. Desde que Jeff Bezos lanzó Amazon.com en 1995, se ha hecho un progreso significativo en la oferta, en los sitios web y en la red internacional de distribución y servicio al cliente.



En la actualidad, Amazon ofrece gran variedad de productos, desde libros o productos electrónicos, hasta raquetas de tenis o diamantes. Tienen una presencia directa en Estados Unidos, Reino Unido, Alemania, Francia, Italia, España, Japón, Canadá y China, pero además pueden servir a los clientes en la mayoría de los países del mundo.

En este trabajo analizaremos los datos de 10.000 pedidos realizados en Amazon con técnicas de tratamiento inteligente de datos. Estos datos están compuestos por información de un total de 568.454 de reseñas que diferentes usuarios realizaron sobre alimentos hasta 2012.

## 2. Dataset

El conjunto con el que vamos a trabajar consta de más de 500.000 entradas, que serán reducido a 10.000 para facilitar su procesamiento. El conjunto de datos ha sido obtenido de la web *Kaggle*<sup>1</sup>. Esta web es una plataforma online para realizar competiciones de Data Mining y proporciona un repositorio para que las compañías publiquen sus datos y desde ahí comienza un concurso abierto para que los expertos en Data Mining de todo el mundo los descarguen, trabajen con ellos y propongan soluciones a los problemas de la compañía en cuestión.

El conjunto en cuestión puede ser descargado en el siguiente enlace de Kaggle.

<https://www.kaggle.com/snap/amazon-fine-food-reviews>

Como introducción a los datos tenemos que:

- Los comentarios fueron realizados entre octubre de 1999 y octubre de 2012.
- Hay un total de 568.454 comentarios.
- Los comentarios han sido realizados por un total de 256.059 usuarios distintos.
- Los comentarios son de 74.258 alimentos distintos.
- 260 usuarios realizaron más de 50 comentarios.

Para acabar esta introducción al conjunto de datos, presentamos las distintas variables de las que está compuesto, así como una breve descripción:

- **IdRow**: identificador de la columna.
- **ProductIdUnique**: identificador único del producto comprado por el usuario.
- **UserIdUnqiue**: identificador único del usuario que ha realizado el comentario.
- **ProfileNameProfile**: nombre del usuario que ha realizado el comentario.
- **HelpfulnessNumeratorNumber**: Número total de usuarios que consideraron útil la revisión.
- **HelpfulnessDenominatorNumber**: Número total de usuarios que indicaron si la revisión les resultó útil o no.

---

<sup>1</sup> <https://www.kaggle.com>

- **ScoreRating:** Puntuación entre 1 y 5 que los usuarios dejaron en la web.
- **TimeTimestamp:** Marca de tiempo UNIX del comentario. Esta marca indica los segundos que han pasado desde el 1 de enero de 1970 hasta la fecha del comentario.
- **SummaryBrief:** breve resumen del comentario del usuario.
- **Text:** texto completo del comentario del usuario.

Un ejemplo de una entrada del conjunto de datos con el que hemos trabajado sería el siguiente:

- **IdRow:** 16
- **ProductIdUnique:** B001GVISJM
- **UserIdUnique:** A1CZX3CP8IKQIJ
- **ProfileNameProfile:** Brian A. Lee
- **HelpfulnessNumeratorNumber:** 4
- **HelpfulnessDenominatorNumber:** 5
- **ScoreRating:** 5
- **TimeTimestamp:** 1262044800
- **SummaryBrief:** Lots of twizzlers, just what you expect.
- **Text:** My daughter loves twizzlers and this shipment of six pounds really hit the spot.  
It's exactly what you would expect...six packages of strawberry twizzlers.

## 3. Herramientas utilizadas

Para la realización de la práctica hemos utilizado tres herramientas principales. Aunque el guión de la práctica daba a elegir, principalmente, entre KNIME o R, hemos optado por usar las dos, una en cada parte, para aprender a utilizarlas.

### 3.1. KNIME<sup>2</sup>

KNIME (*Konstanz Information Miner*) es una plataforma *open software* de minería de datos que permite el desarrollo de modelos en un entorno visual que fue desarrollada en Java en el departamento de bioinformática y minería de datos de la Universidad de Constanza, Alemania.

### 3.2. R<sup>3</sup>

R es un lenguaje *open source* de computación estadística. Se puede utilizar el programa R Studio como IDE para trabajar con R, o que permite trabajar con este lenguaje de manera más cómoda.

Hemos utilizado R para la parte de procesamiento de lenguaje natural.

### 3.3. Python<sup>4</sup>

Python es un lenguaje de scripting que permite un prototipado rápido, lo que es perfecto para adaptar el archivo CSV con los datos que elegimos a las necesidades que nos fueron surgiendo.

---

<sup>2</sup> <https://www.knime.com/>

<sup>3</sup> <https://www.r-project.org/>

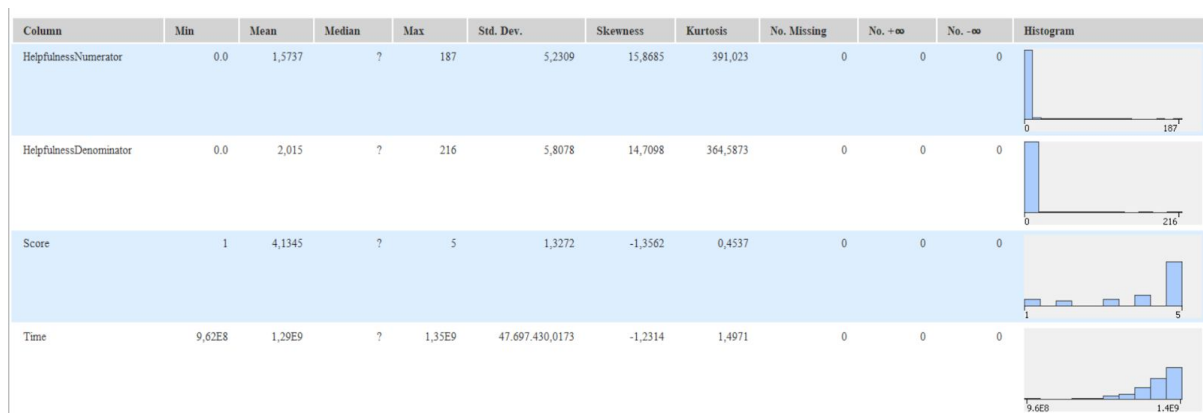
<sup>4</sup> <https://www.python.org/>

## 4. Preprocesamiento de datos

El preprocesamiento de datos es un método de análisis de datos para tratamiento estadístico al que se someten las muestras recogidas durante un proceso de investigación en cualquier campo científico.

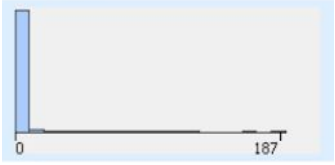

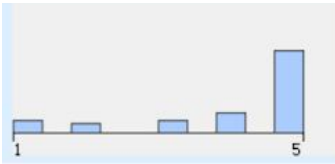
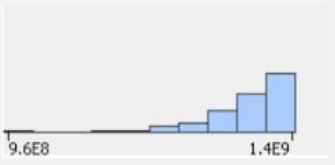
### 4.1. Datos estadísticos

Haciendo uso del nodo *Statistics* de Knime hemos extraído valores como el mínimo, la media, el máximo, desviación típica, asimetría estadística, curtosis, datos perdidos, datos con valor infinito así como un histograma de los atributos de tipo numérico.



Como la resolución no es muy buena, a continuación se presentan los mismos dato en forma de tabla.

Colums	Min	Mean	Median	Max	Std. Dev.
HelpfulnessNumerator	0.0	1,5737	?	187	5,2309
HelpfulnessDenominator	0.0	2,015	?	216	5,8078
Score	1	4,1345	?	5	1,3272
Time	9,6E8	1,29E9	?	1.35E9	47.697.430,0173

Skewness	Kurtosis	No. Missing	No. $+\infty$	No. $-\infty$	Histogram
15,8685	391,023	0	0	0	
15,7098	364,587	0	0	0	
-1,3562	0,4537	0	0	0	
-1,2314	1,4971	0	0	0	

A raíz de lo anterior podemos observar cómo ambos atributos *Helpfulness* generalmente toman un valor 0 debido a que normalmente no se puntúan los comentarios de los productos, así como que el *HelpfulnessDenominator* toma siempre valores mayores o iguales que el *HelpfulnessNumerator*.

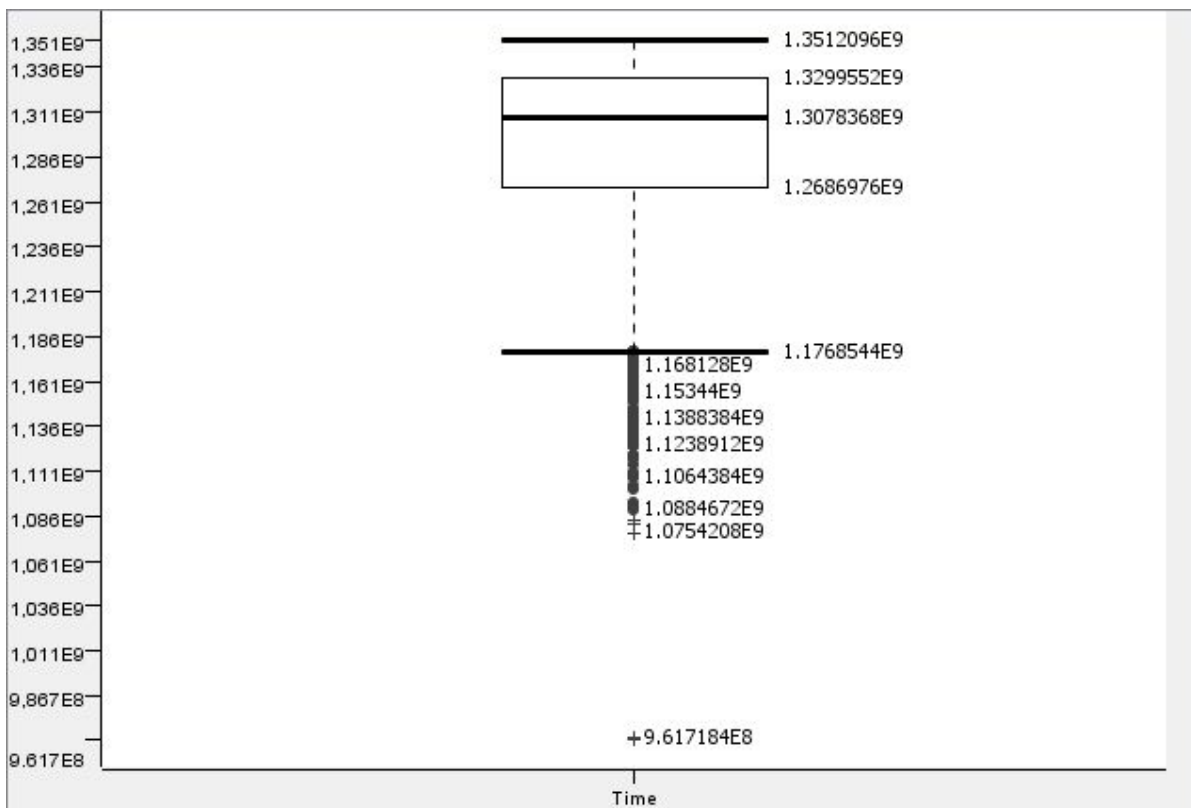
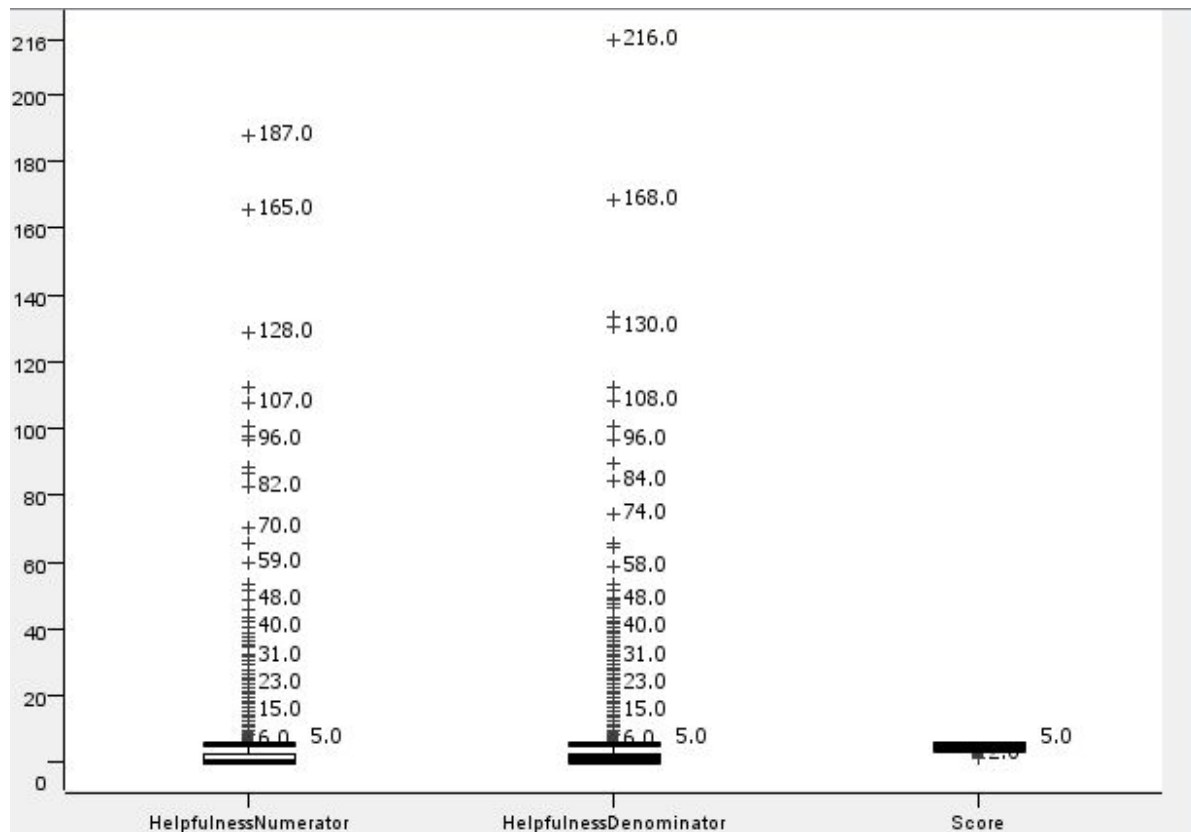
Por otro lado vemos que el valor *Score* toma valores entre 1 y 5, predominando este último. Finalmente vemos en el atributo *Time* que predominan los comentarios más antiguos.

## 4.2. Diagrama de cajas

Mediante un diagrama de cajas podemos representar gráficamente los datos y conocer sus valores mínimos, máximos, la mediana y los cuartiles posibilitando de esta manera conocer los valores atípicos en el conjunto de datos.

El diagrama de cajas para este conjunto de datos es el siguiente.





Los valores calculados son los siguientes:

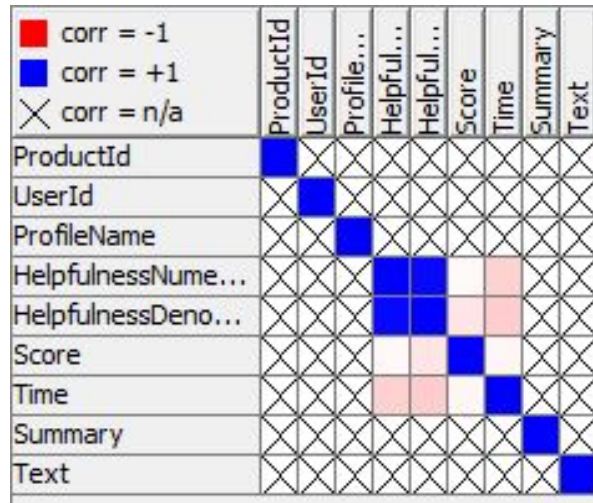
Valor	<i>HelpfulnessNum</i>	<i>HelpfulnessDenom</i>	<i>Score</i>	<i>Time</i>
Mínimo	0	0	1	961718400 (23/06/2000)
Más pequeño	0	0	3	1176854400 (18/04/2007)
Cuartil inferior	0	0	4	1268697600 (16/03/2010)
Mediana	0	1	5	130783680 (12/06/2011)
Cuartil superior	2	2	5	1329955200 (23/02/2012)
Más grande	5	5	5	1351209600 (26/10/2012)
Máximo	187	216	5	1351209600 (26/10/2012)

De esta manera podemos llegar a las siguientes conclusiones.

- *HelpfulnessNumerator* normalmente toma valores de 0 a 5 por lo que los valores 165 y 187 podrían considerarse anómalos.
- *HelpfulnessDenominator* normalmente toma valores de 0 a 5 por lo que los valores 168 y 216 podrían considerarse anómalos.
- *Score* normalmente toma valores de 3 a 5 por lo que parece no tener valores anómalos.
- *Time* normalmente toma valores de 1176854400 (18/04/2007) a 1351209600 (26/10/2012) por lo que el valor 961718400 (23/06/2000) podría considerarse anómalo.

### 4.3. Correlación lineal

La matriz de correlación lineal de los datos es la siguiente.



Esta matriz da a entender que los datos están muy débilmente correlacionados; los únicos datos que están correlacionados son *HelpfulnessNumerator* y *HelpfulnessDenominator*. Lo cual tiene sentido teniendo en cuenta que son los valores de una fracción que nos daría el valor de utilidad medio de ese comentario según la opinión de los usuarios.

Por otro lado, *Score* y *Time* están algo correlacionados con las dos anteriores variables.

Tras el procesamiento de los datos deberíamos considerar:

1. Eliminar los atributos *ProductId*, *UserId* y *ProfileName* ya que son valores de identificación de otros atributos de cuyos conjuntos de datos no disponemos para poder cruzar los resultados y obtener información acerca de éstos. Por ello sería conveniente eliminarlos debido a que no van a ser de utilidad y van a ocupar espacio innecesario.
2. Eliminar el atributo *Summary* debido a que es un resumen de *Text* y se considera que para el momento de aplicar la minería de texto será más útil realizarla con el comentario original para no perder información ideas que puedan ser importantes.
3. Se ha considerado la creación de un nuevo atributo llamado *Utility*, resultado de la división entre *HelpfulnessNumerator* y *HelpfulnessDenominator* pero se ha descartado debido a que hay comentarios que han podido pasar desapercibidos y al no haber recibido ningún valor *Helpfulness* sería NaN y darle un valor como puede ser 0 sería dar un valor falso.

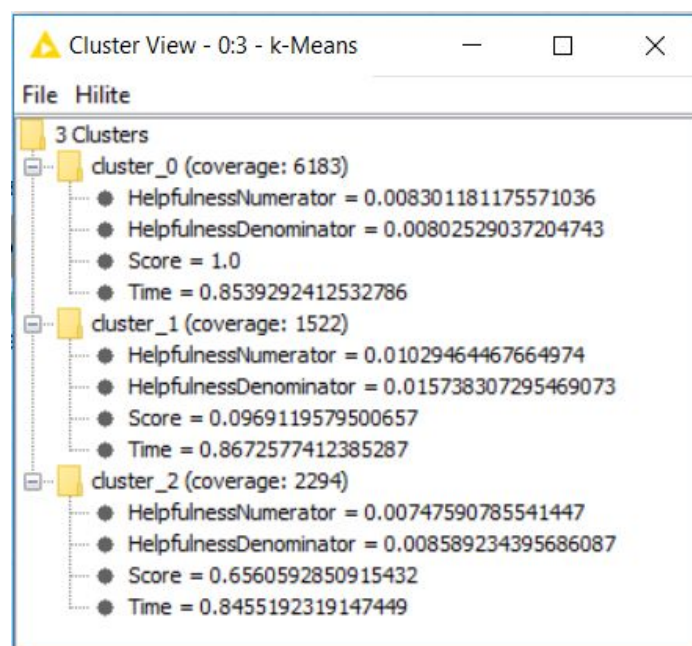
## 5. Agrupamiento

Un algoritmo de agrupamiento es un procedimiento de agrupación de una serie de vectores de acuerdo con un criterio. Esos criterios son por lo general distancia o similitud.

### 5.1. K-means

El primer método de agrupamiento que vamos a utilizar es el método de las k-medias usando el nodo *k-Means* de KNIME. Con el método k-medias dividiremos el conjunto de 9999 observaciones en 3 grupos en función del valor medio que obtenga cada una.

En la siguiente imagen vemos los valores medios de cada grupo.



A continuación vemos las estadísticas del proceso de agrupamiento en el que correctamente tenemos 9999 comentarios que han sido divididos en 3 grupos distintos. El agrupamiento realizado ha resultado con una entropía de 0.3657 y una calidad de 0.8425 la cual podemos considerarla satisfactoria.

## Clustering statistics

### Data Statistics

Statistics	Value
Number of clusters found:	3
Number of objects in clusters:	9999
Number of reference clusters:	5
Total number of patterns:	9999

### Data Statistics

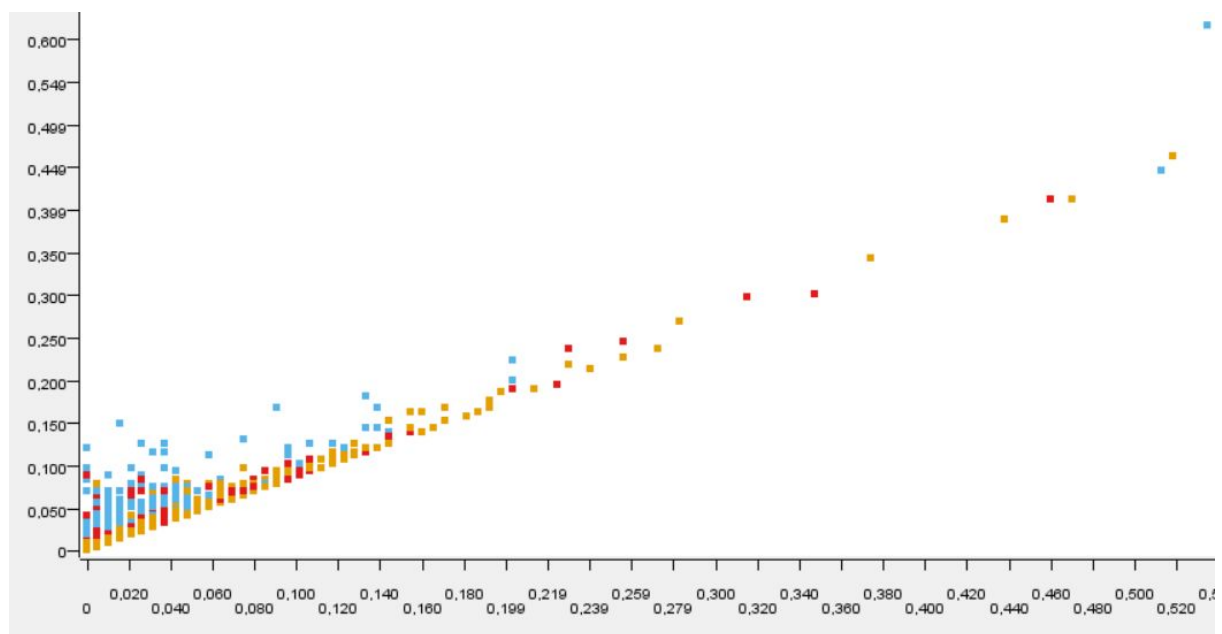
Score	Value
Entropy:	0,3657
Quality:	0,8425

Row ID	I Size	D Entropy	D Normali...	D Quality
cluster_0	6183	0	0	?
cluster_2	2294	0.955	0.411	?
cluster_1	1522	0.963	0.415	?
Overall	9999	0.366	0.158	0.842

El primer grupo está formado por el 61.8% de los datos, el segundo por el 22.9% y finalmente el tercer grupo que tan solo está formado por un 15.2%.

En las dos imágenes posteriores mostramos cómo a cada comentario se le ha asignado un grupo y un color identificativo para representarlo en un gráfico de dispersión en el que podemos ver la gran cantidad de valores agrupados que hay en [0.25, 0.25] y posteriormente aumenta su separación.

Row ID	S Cluster	S ProductId	S UserId	S ProfileN...	D Helpful...	D Helpful...	D Score	D Time
1	cluster_0	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	0.005	0.005	1	0.878
2	cluster_1	B00813GRG4	A1D87F6ZCVE5NK	dli pa	0	0	0	0.989
3	cluster_2	B000LQOCHO	ABXLMWJDXAIN	Natalia Corr...	0.005	0.005	0.75	0.661
4	cluster_1	B000UA0QIQ	A395BORC6FGVXV	Karl	0.016	0.014	0.25	0.889
5	cluster_0	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bi...	0	0	1	0.999
6	cluster_2	B006K2ZZ7K	ADT0SRK1MGOEU	Twoapenny...	0	0	0.75	0.976
7	cluster_0	B006K2ZZ7K	A1SP2KVKFXXRU1	David C. Sulli...	0	0	1	0.972
8	cluster_0	B006K2ZZ7K	A3JRGQVEQN31IQ	Pamela G. W...	0	0	1	0.961
9	cluster_0	B000E7L2R4	A1MZY09TZK08BI	R. James	0.005	0.005	1	0.925
10	cluster_0	B00171APVA	A21BT40VZCCYT4	Carol A. Reed	0	0	1	1
11	cluster_0	B0001PB9FE	A3HDKO7OW0QNK4	Canadian Fan	0.005	0.005	1	0.375
12	cluster_0	B0009XLVG0	A2725IB4YY9JEB	A Poeng Spa...	0.021	0.019	1	0.825
13	cluster_1	B0009XLVG0	A327PCT23YH90	LT	0.005	0.005	0	0.97
14	cluster_2	B001GVISJM	A18ECVX2RJ7HUE	willie roadie	0.011	0.009	0.75	0.84
15	cluster_0	B001GVISJM	A2MUGFVZTDQ47K	Lynrie Oh HE...	0.021	0.023	1	0.787
16	cluster_0	B001GVISJM	A1CZX3CP8IKQIJ	Brian A. Lee	0.021	0.023	1	0.771
17	cluster_1	B001GVISJM	A3KLWF6WQ5BNYO	Erica Neathery	0	0	0.25	0.992
18	cluster_0	B001GVISJM	AFKW14U97Z6QO	Becca	0	0	1	0.984
19	cluster_0	B001GVISJM	A2A9X58G2GTBLP	Wolfee1	0	0	1	0.932
20	cluster_0	B001GVISJM	A3IV7CL2C13K2U	Greg	0	0	1	0.915
21	cluster_0	B001GVISJM	A1W00KGLPR5PV6	mom2emma	0	0	1	0.903
22	cluster_0	B001GVISJM	AZOF9E17RGZH8	Tammy Ande...	0	0	1	0.892
23	cluster_0	B001GVISJM	ARYVQL4N737A1	Charles Brown	0	0	1	0.881
24	cluster_0	B001GVISJM	AJ613OLZZUG7V	Mare's	0	0	1	0.88
25	cluster_0	B001GVISJM	A22P2J09NJ9HKE	S. Cabanaug...	0	0	1	0.857
26	cluster_0	B001GVISJM	A3FONPR03H3PJS	Deborah S. L...	0	0	1	0.839



## 5.2. Clustering jerárquico

El segundo método de agrupamiento que vamos a utilizar es el método de clustering jerárquico usando el *nodo hierarchical clustering* de KNIME.

En primer lugar hemos usado una función de distancia de Manhattan en el nodo hierarchical clustering con el que hemos obtenido tan solo un 27% de calidad y vemos que el agrupamiento no es aceptable debido a que el primer y segundo grupo tienen tan solo 1 elemento cada uno, mientras que el tercero tiene 9997 elementos.

# Clustering statistics

Data Statistics	
Statistics	Value
Number of clusters found:	3
Number of objects in clusters:	9999
Number of reference clusters:	5
Total number of patterns:	9999
Data Statistics	
Score	Value
Entropy:	1,6944
Quality:	0,2702

Row ID	I Size	D Entropy	D Normali...	D Quality
cluster_0	1	0	0	?
cluster_1	1	0	0	?
cluster_2	9997	1.695	0.73	?
Overall	9999	1.694	0.73	0.27

Nuevamente hemos tratado de usar el nodo hierarchical clustering, pero en este caso con una de distancia de Euclídea con el que hemos obtenido nuevamente tan solo un 27% de calidad y vemos que el agrupamiento no es aceptable nuevamente debido a que el primer y segundo grupo tienen tan solo 1 y 2 elementos cada uno respectivamente, mientras que el tercero tiene 9996 elementos.



## Clustering statistics

### Data Statistics

Statistics	Value
Number of clusters found:	3
Number of objects in clusters:	9999
Number of reference clusters:	5
Total number of patterns:	9999

### Data Statistics

Score	Value
Entropy:	1,6943
Quality:	0,2703

Row ID	I Size	D Entropy	D Normali...	D Quality
cluster_0	1	0	0	?
cluster_2	2	1	0.431	?
cluster_1	9996	1.695	0.73	?
Overall	9999	1.694	0.73	0.27



## 6. Clasificación

En este apartado trataremos de inferir la puntuación de los productos a través de los comentarios que recibe. Para ello se hará uso de las columnas *ProductID*, *HelpfulnessNumerator* y *HelpfulnessDenominator*.

### 6.1. Árbol de decisión

En primer lugar haremos uso de un árbol de decisión. Este hace uso de un aprendizaje lógico para tratar de realizar la clasificación. Se utilizará en este caso un conjunto de entrenamiento de 9000 datos y un conjunto de prueba de 1000 datos. Los resultados son los siguientes:

Row ID	I 5	I 1	I 2	I 3	I 4
5	256	6	0	4	0
1	50	13	0	0	0
2	39	6	0	1	0
3	52	2	1	2	0
4	78	1	0	0	0

Donde se puede concluir que el árbol de decisión ha clasificado correctamente el 53,03% de los datos.

La mayor cantidad de datos acertados son aquellos valorados con cinco estrellas. Son estos además los más comunes, por lo que el aprendizaje ha sido más efectivo.

### 6.2. Naive-Bayes

A continuación se va a clasificar los datos haciendo uso del algoritmo Naive-Bayes. Este expresa la probabilidad de que un dato se presente en el conjunto a lo largo del tiempo. La partición de datos es la misma que en el anterior apartado. Para este los resultados son los siguientes:

Row ID	I 5	I 3	I 1	I 4	I 2
5	268	5	0	0	0
3	59	0	0	0	0
1	66	0	0	0	0
4	60	3	0	0	0
2	50	0	0	0	0

En este caso el porcentaje de datos bien clasificados es del 52,45%, similar al resultado del árbol de decisión aunque mejora en las clases predominantes y empeora en el resto.

### 6.3. C4.5 de Quinlann

Para utilizar el algoritmo C4.5 debemos instalar el plugin KNIME *Weka Data Mining Integration* (3.7) donde se usarán los nodos *J48Graft* y *Weka Predictor*. Este algoritmo forma un árbol de decisión haciendo uso de la entropía de la información, es decir, eligiendo los atributos que dividen mejor el conjunto de datos.

Los resultados obtenidos son los siguientes.

Row ID	I 5	I 1	I 2	I 3	I 4
5	247	2	3	0	0
1	59	11	3	2	0
2	44	5	3	0	0
3	45	6	2	0	0
4	78	1	0	0	0

Este algoritmo arroja un porcentaje de datos bien clasificados del 51,08%, muy similar a los anteriores aunque mejorando en las clases menos predominantes.

### 6.4. Random forest

El algoritmo Random forest combina árboles de clasificación por el cual cada uno de ellos depende de los valores de un vector aleatorio de igual distribución proporcionando así una discriminación estocástica. Los resultados obtenidos son los siguientes.

Row ID	I 5	I 1	I 4	I 2	I 3
5	263	5	1	0	1
1	57	18	0	0	1
4	69	4	0	0	0
2	35	3	0	1	0
3	50	3	0	0	0

Este algoritmo ya mejora algo al resto proporcionando un 55,19% de acierto. Aún así las clases menos predominantes siguen siendo las que peor resultado tienen.

## 6.5. Random forest con clústeres

Por último clasificamos haciendo uso de los clústeres previamente definidos. Los resultados son los siguientes:

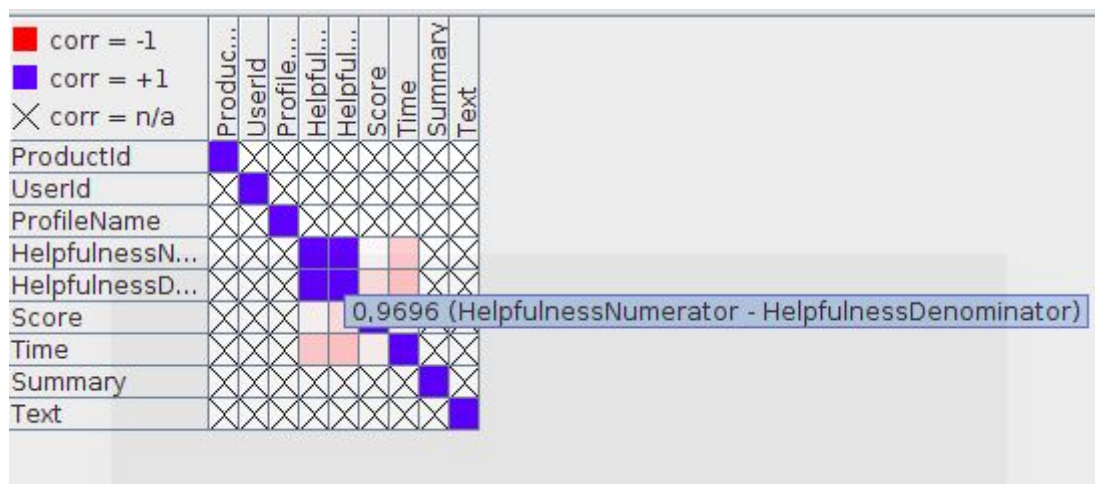
Row ID	I cluste...	I cluste...	I cluste...
cluster_0	173	0	3
cluster_1	0	174	1
cluster_2	0	0	161

En este caso sí mejora bastante debido a que se ha pasado de 5 a 3 clústeres y a que la similitud entre estos es menor. El porcentaje de acierto es del 99,22%.

## 7. Regresión

En esta sección vamos a intentar predecir los valores de algunos datos a partir de otros. Para ello se hará uso de la matriz de correlación (nodo Linear Correlation) que ya se analizó por encima en la sección de pre-procesamiento y de la matriz de gráficos de dispersión (nodo Scatter Matrix) entre cada variable.

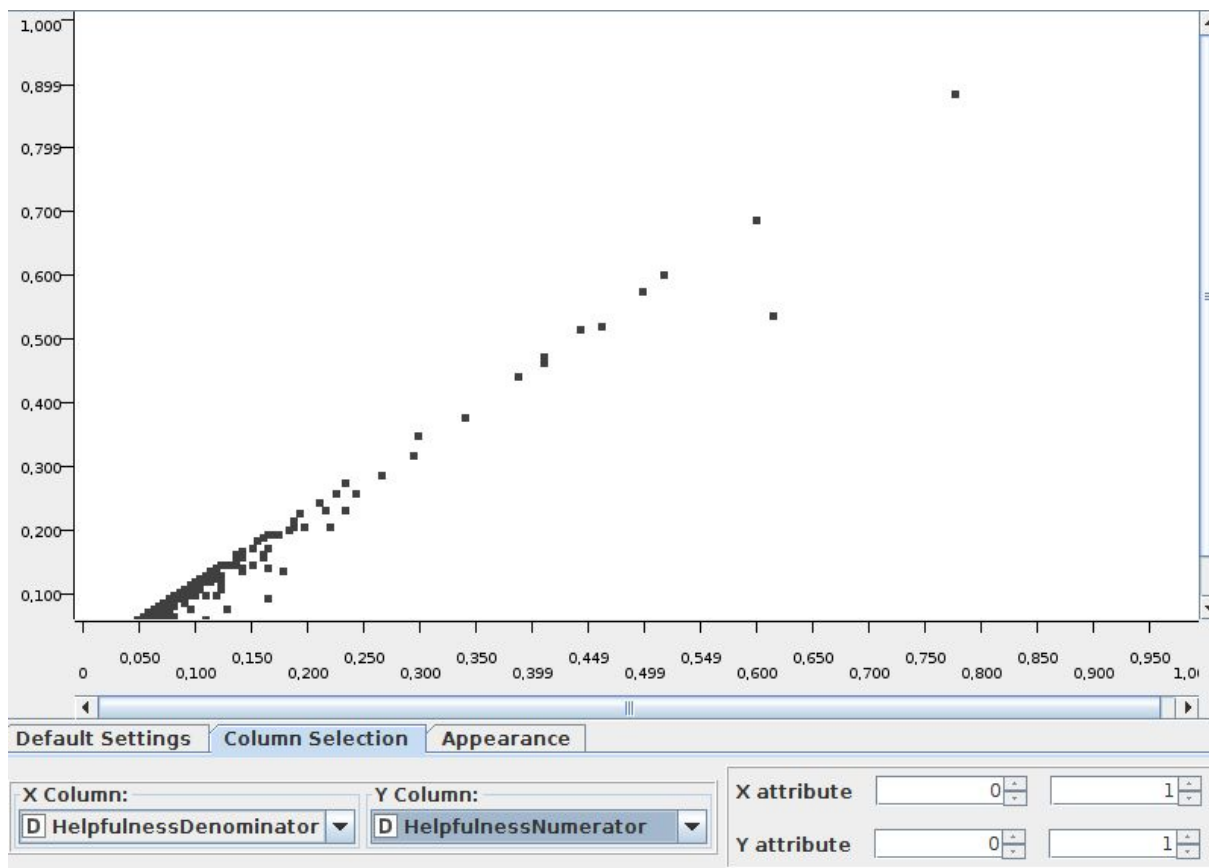
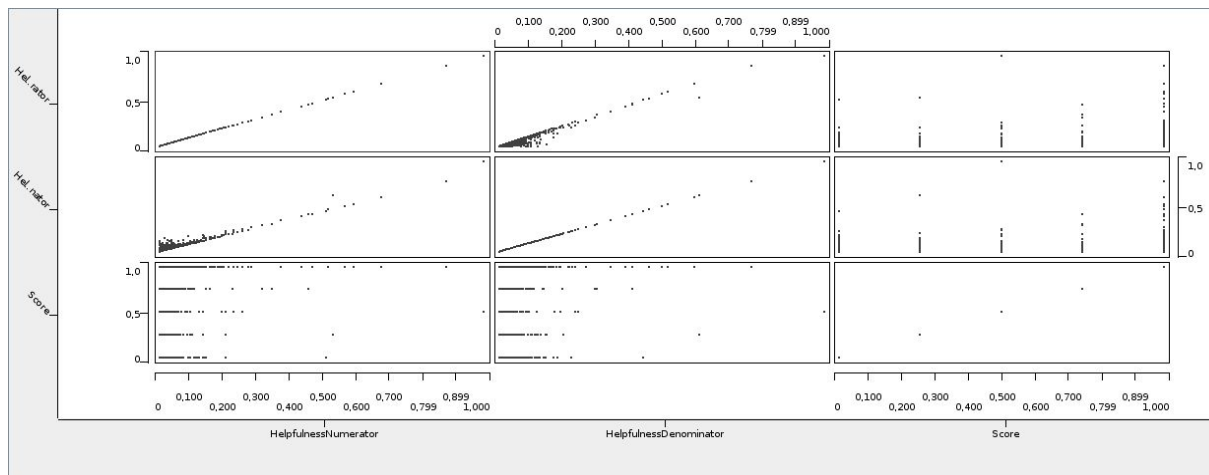
### 7.1. Correlación



Como podemos ver los únicos datos que están correlacionados son *HelpfulnessNumerator* y *HelpfulnessDenominator*. lo que nos indica que en esta parte del análisis vamos a procurar hacer la regresión lineal basado en estas dos variables, igualmente se procede hacer la Scatter Matrix para mostrar la correlación gráficamente.

### 7.2 Scatter Matrix

En la matriz de dispersión se comparan todas las variables con el objetivo de detectar cuáles podrían ser aptas para el modelamiento en la regresión lineal.



Las variables comparadas son Score, *HelpfulnessNumerator* (HN) y *HelpfulnessDenominator* (HD), como muestra la gráfica la nube de puntos que tiene una forma lineal es en la que están relacionadas HN y HD las cuales se usarán para el modelamiento de Regresión lineal.

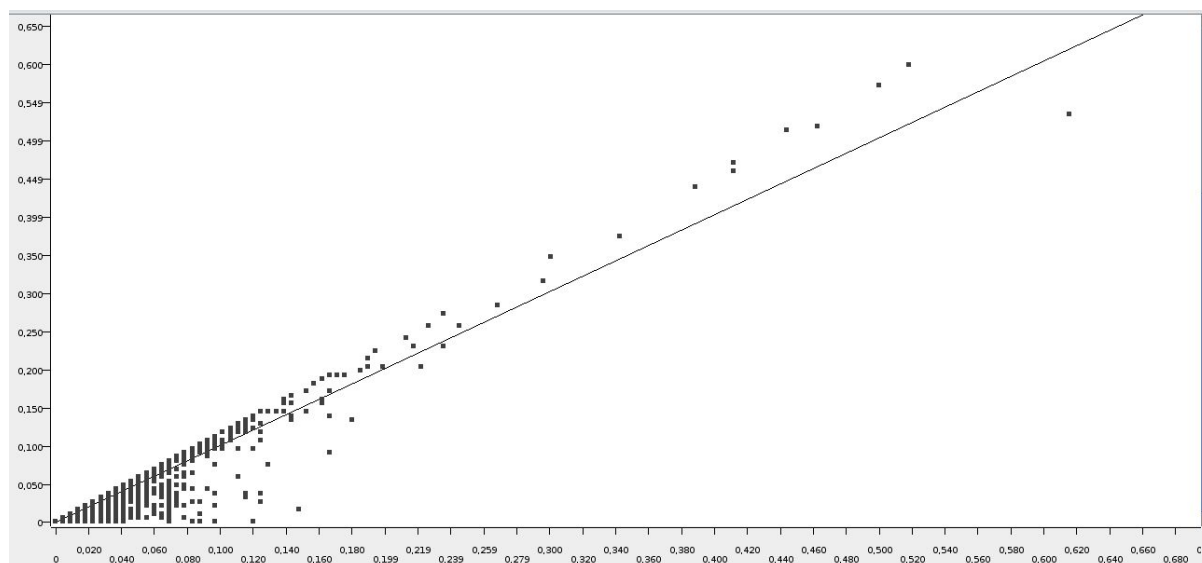
La función generada comparada con los datos de *HelpfulnessNumerator* es la siguiente:

## Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P> t
HelpfulnessDenominator	1,0087	0,0025	396,0556	0.0
Intercept	-0,001	7,25E-5	-13,7204	0.0

Multiple R-Squared: 0,9401

Adjusted R-Squared: 0,9401



En esta sección se puede concluir que las únicas variables que se pueden predecir linealmente y están fuertemente correlacionadas son (*HN*) y (*HD*) un coeficiente de determinación = 0,9401.

## 8. Minería de texto

### 8.1. Frecuencia de las palabras

Para obtener los resultados a través de la minería de texto se ha comenzado extrayendo la frecuencia de aparición de las 10 palabras más comunes y el número de veces que aparece nada una, que han sido *like*, *tast* y *flavor*.

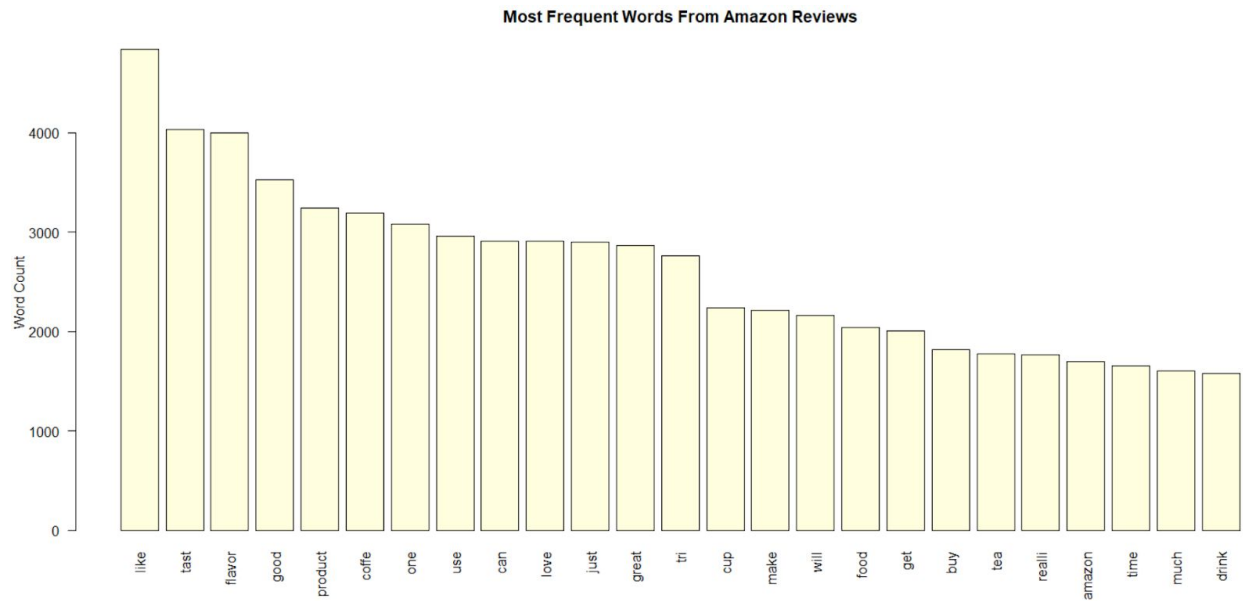
Al tratarse de comentarios en una web realizados sobre alimentos es normal la aparición de palabras que indican el agrado sobre el producto como *like*, *good* o *love*, así como nombres de productos, unidades o términos como probar o sabor.

Palabra	Frecuencia
like	4837
tast	4035
flavor	4000
good	3527
product	3249
coffe	3193
one	3081
use	2963
can	2911
love	2911

A partir de la tabla anterior se han representado las 25 palabras con mayor número de apariciones en una gráfica en la que poder estudiar de forma más visual cuáles son las palabras con mayor frecuencia de aparición.

Podemos observar cómo la palabra *like* tiene cerca de 800 repeticiones más que la segunda. Posteriormente, las sucesivas palabras no tienen una diferencia tan significativa entre ellas.

Ya que en esta gráfica aparecen más palabras que anteriormente, podemos observar que se cumple lo comentado anteriormente y generalmente las palabras que encontramos son referentes al agrado sobre alimentos.





A través de una nube de palabras podemos hacer una representación visual de las palabras que conforman los comentarios, en donde el tamaño es mayor para las palabras que aparecen con más frecuencia.

[illegible]

## 8.3. Asociación de palabras

Con esta técnica podemos ver qué índice de relación tiene cada palabra con el resto.

### 8.3.1. Like

tast	realli	flavor	just	one	much	think	can
0.31	0.23	0.20	0.19	0.19	0.18	0.18	0.16
tri	seem	carbon	also	first	review	get	thing
0.16	0.16	0.16	0.15	0.15	0.15	0.15	0.15
drink	better	good	make	give	someth	kind	sweet
0.15	0.14	0.14	0.14	0.14	0.14	0.13	0.13
say	though	probabl	might	orang	feel	juic	look
0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.12
sugar	bit	eat	know	will	pretti	littl	even
0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12
see	strong	soda	well	want	still	water	decid
0.12	0.12	0.11	0.11	0.11	0.11	0.11	0.11
way	peopl	fruit	tangerin	expect	100	appl	natur
0.11	0.11	0.11	0.11	0.10	0.10	0.10	0.10
bad	said						
0.10	0.10						

### 8.3.2. Tast

like	orang	drink	soda	nirvana	tangerin	good
0.31	0.24	0.20	0.18	0.18	0.18	0.17
tri	carbon	ami	brian	coconut	coconuts	flavor
0.17	0.17	0.17	0.17	0.17	0.17	0.16
sweet	juic	bud	sugar	just	realli	artifici
0.16	0.16	0.16	0.15	0.15	0.15	0.15
vitacoco	water	bad	lemon	better	can	buds
0.15	0.14	0.14	0.14	0.13	0.13	0.13
slight	bitter	switch	weird	want	first	expect
0.13	0.13	0.13	0.13	0.12	0.12	0.12
even	natur	probable	sour	zevia	juice	pulp
0.12	0.12	0.12	0.12	0.12	0.12	0.12
unalt	actual	much	review	appl	pretti	coffe
0.12	0.11	0.11	0.11	0.11	0.11	0.11
thing	differ	calori	odd	burnt	vitamin	lorann
0.11	0.11	0.11	0.11	0.11	0.11	0.11
strong	ounc	though	regular	taste	added	
0.11	0.10	0.10	0.10	0.10	0.10	

### 8.3.3. Flavor

like	coffe	tangerin	tast	chip	sweet
0.20	0.18	0.17	0.16	0.16	0.15
orang	strong	drink	juic	acerola	van
0.15	0.14	0.14	0.14	0.14	0.14
houutt	soda	enjoy	tri	favorit	natur
0.14	0.13	0.13	0.13	0.13	0.13
hazelnut	artifici	carbon	light	grape	rich
0.13	0.13	0.13	0.12	0.12	0.12
extra	pack	href	http	www	medium
0.12	0.12	0.12	0.12	0.12	0.12
roast	blend	prefer	overpow	vinegar	bbq
0.12	0.12	0.12	0.12	0.12	0.12
puck	barely	nom	salt	appl	com
0.12	0.12	0.12	0.11	0.11	0.11
gum	cup	flavors	nice	aroma	
0.11	0.11	0.11	0.11	0.11	

# Bibliografía

[1] Zurutuza, U. (2019). Minería de Textos para la Clasificación de Documentos en Español con R | Investigación en TICs. [online] Investigación en TICs. Available at: <https://mukom.mondragon.edu/ict/mineria-de-textos-para-la-clasificacion-de-documentos-en-espanol-con-r/> [Accessed 23 Jan. 2019].

[2] Knime.com. (2019). Simple Example with Statistics | KNIME. [online] Available at: <https://www.knime.com/nodeguide/analytics/statistics/simple-example-with-statistics> [Accessed 23 Jan. 2019].

[3] Knime.com. (2019). Performing a k-Means Clustering | KNIME. [online] Available at: <https://www.knime.com/nodeguide/analytics/clustering/performing-a-k-means-clustering> [Accessed 23 Jan. 2019].

[4] Knime.com. (2019). Classification and Predictive Modelling | KNIME. [online] Available at: <https://www.knime.com/nodeguide/analytics/classification-and-predictive-modelling> [Accessed 23 Jan. 2019].

[5] Knime.com. (2019). Learning a Simple Regression Tree | KNIME. [online] Available at: <https://www.knime.com/nodeguide/analytics/regressions/learning-a-simple-regression-tree> [Accessed 23 Jan. 2019].