

Práctica final



Amazon fine food reviews

Tratamiento Inteligente de Datos

Felipe Peiró Garrido

José Andrés Bonilla

Juan Carlos Serrano Pérez

Pedro Manuel Gómez-Portillo López

Índice



1. Introducción

2. Dataset

3. Herramientas utilizadas

4. Preprocesamiento de datos

5. Agrupamiento

6. Clasificación

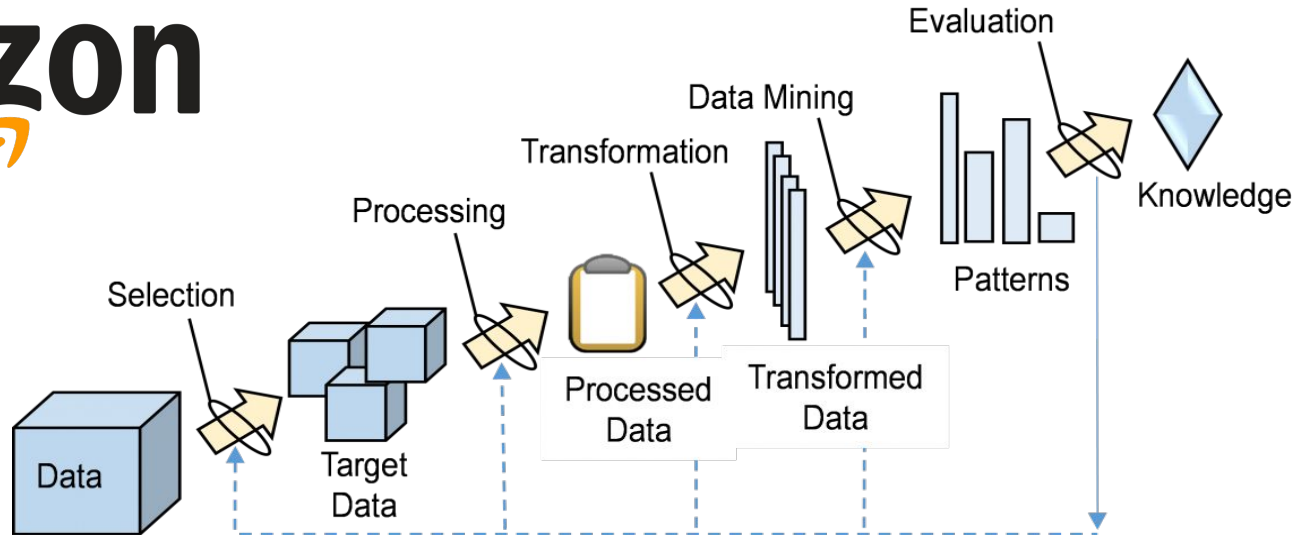
7. Regresión

8. Minería de texto

9. Bibliografía

1. Introducción

Análisis de datos



2. Dataset

Amazon Fine Food Reviews

Descripción

- Realizados entre octubre de 1999 y octubre de 2012
- Hay un total de 568.454
- Realizados por un total de 256.059 usuarios distintos
- Realizados sobre 74.258 alimentos distintos
- 260 usuarios realizaron más de 50 comentarios

The Kaggle logo, featuring the word "kaggle" in a blue, lowercase, sans-serif font. Below the text is a stylized, blue, geometric mountain or iceberg shape composed of many small triangles. The logo is set against a light blue background.

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

Datos



- **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.
- **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.

3. Herramientas utilizadas

Herramientas



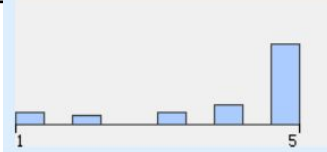
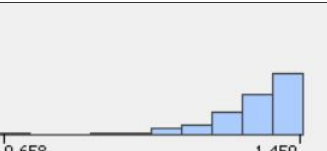


4. Preprocesamiento de datos

Datos estadísticos I

Colums	Min	Mean	Median	Max	Standard dev.
<i>HelpfulnessNumerator</i>	0	1,5737	?	187	5,2309
<i>HelpfulnessDenominator</i>	0	2,015	?	216	5,8078
<i>Score</i>	1	4,1345	?	5	1,3272
<i>Time</i>	9,6E8	1,29E9	?	1,35E9	4,7E7

Datos estadísticos II

Colums	Skewness	Kurtosis	No. Missing	No. + ∞	No. - ∞	Histogram
<i>Helpfulness Numerator</i>	15,8685	391,023	0	0	0	
<i>Helpfulness Denominator</i>	15,7098	364,587	0	0	0	
<i>Score</i>	-1,3562	0,4537	0	0	0	
<i>Time</i>	-1,2314	1,4971	0	0	0	

Datos estadísticos III - Conclusiones



- Los valores *Helpfulness* generalmente toman un valor 0 debido a que normalmente no se puntúan los comentarios de los productos
- *Score* toma valores entre 1 y 5, predominando este último
- Finalmente vemos en el atributo *Time* que predominan los comentarios más antiguos casi de manera exponencial

Diagrama de cajas I

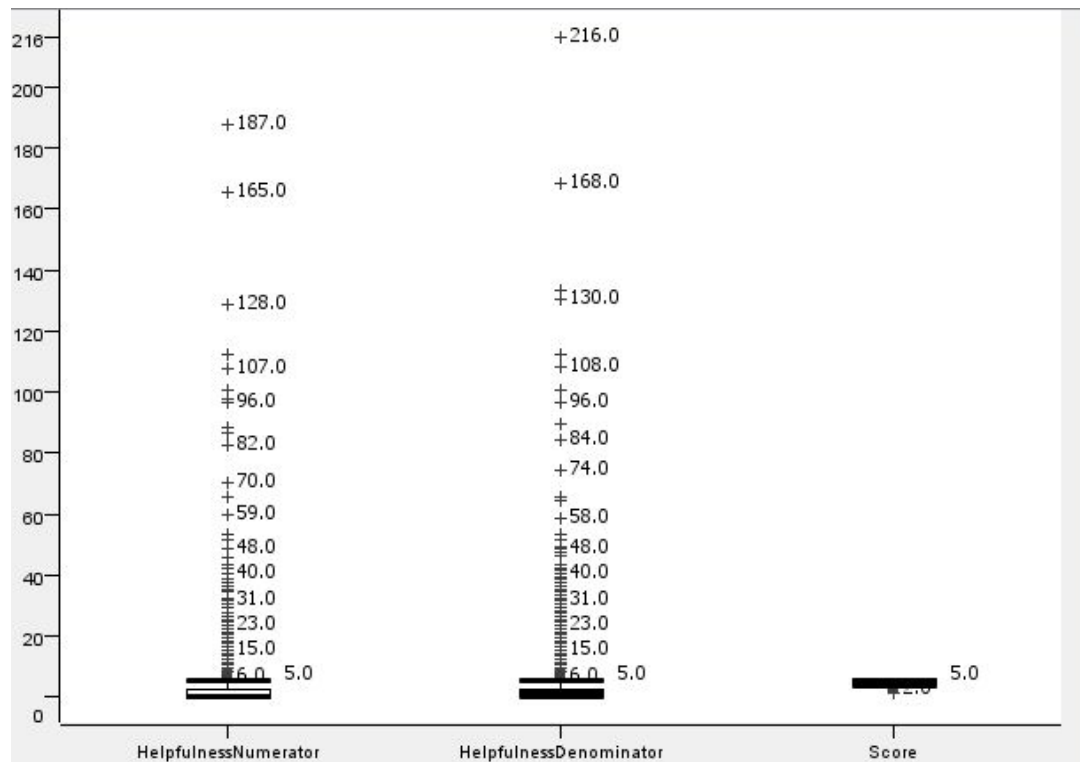


Diagrama de cajas II

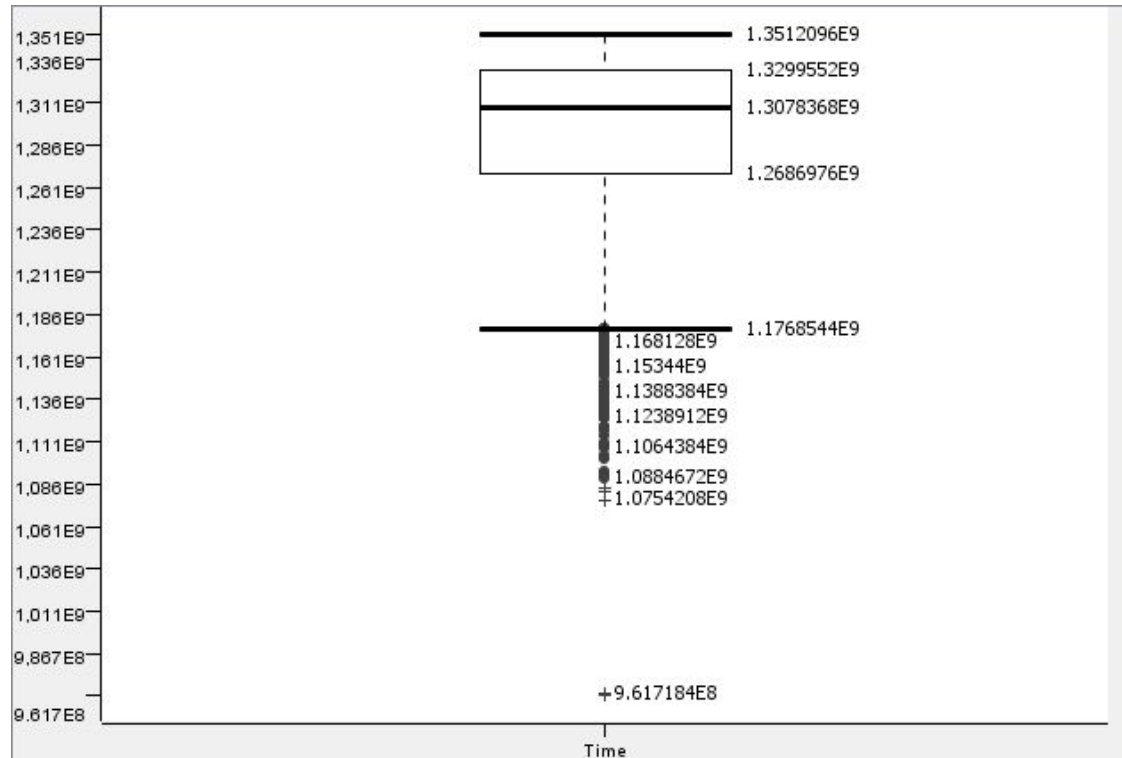


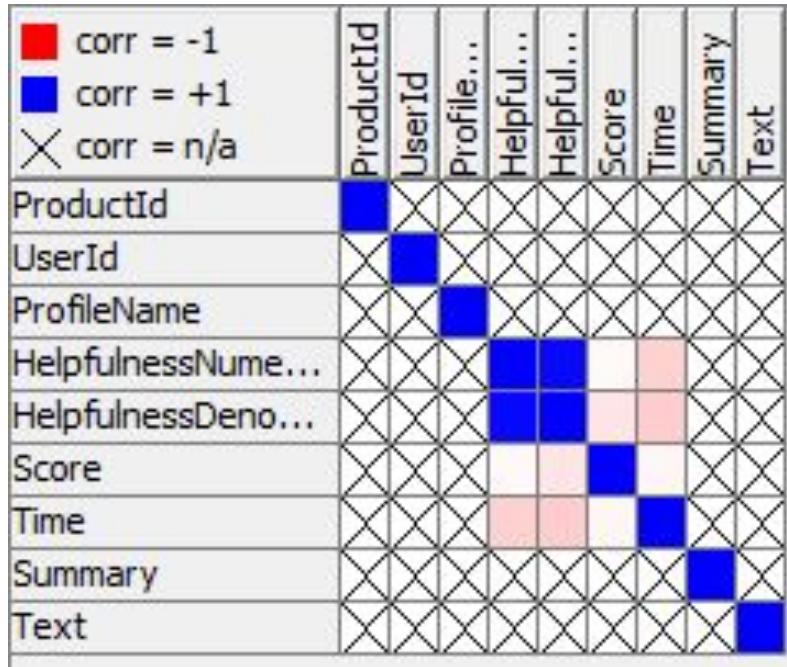
Diagrama de cajas III

Valor	<i>HelpfulnessNum</i>	<i>HelpfulnessDenom</i>	Score	<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)

Diagrama de cajas - Conclusiones

- *HelpfulnessNumerator* normalmente toma valores de 0 a 5
- *HelpfulnessDenominator* normalmente toma valores de 0 a 5
- *Score* normalmente toma valores de 3 a 5 por lo que no tiene valores anómalos
- *Time* normalmente toma valores de 1176854400 (18/04/2007) a 1351209600 (26/10/2012) por lo que el valores mayores podrían considerarse anómalos

Correlación lineal I



- Los datos están muy débilmente correlacionados
- Los únicos datos que están correlacionados son *HelpfulnessNumerator* y *HelpfulnessDenominator*

Correlación lineal II

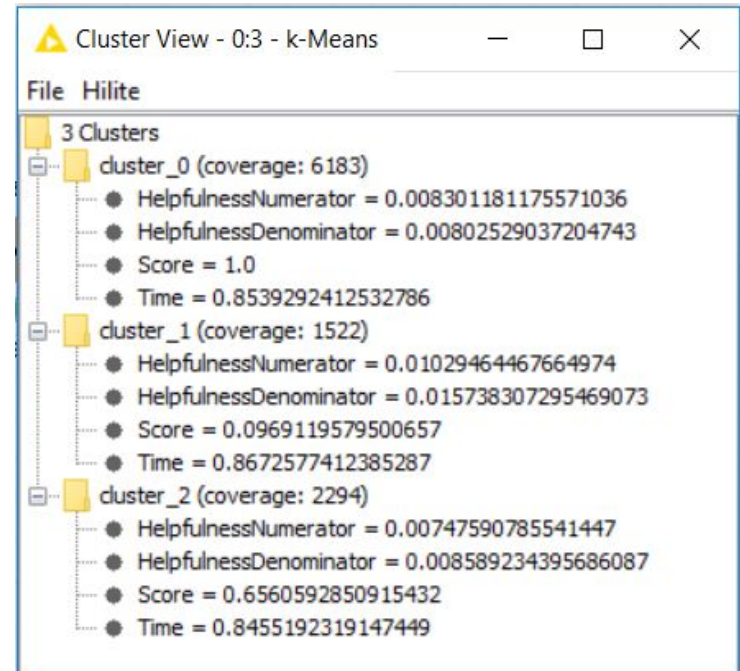


- Eliminar los atributos **ProductId**, **UserId** y **ProfileName**.
 - Son valores de identificación de otros atributos de cuyo conjuntos de datos no disponemos
- Eliminar el atributo **Summary** ya que es un resumen de **Text**
 - Para aplicar la minería de texto será más útil realizarla con el comentario original

5. Agrupamiento

K-means I

- Usaremos el nodo *k-Means* de KNIME
- Dividiremos el conjunto de 9999 elementos en 3 grupos usando el valor medio que obtenga cada una



K-means II

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

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

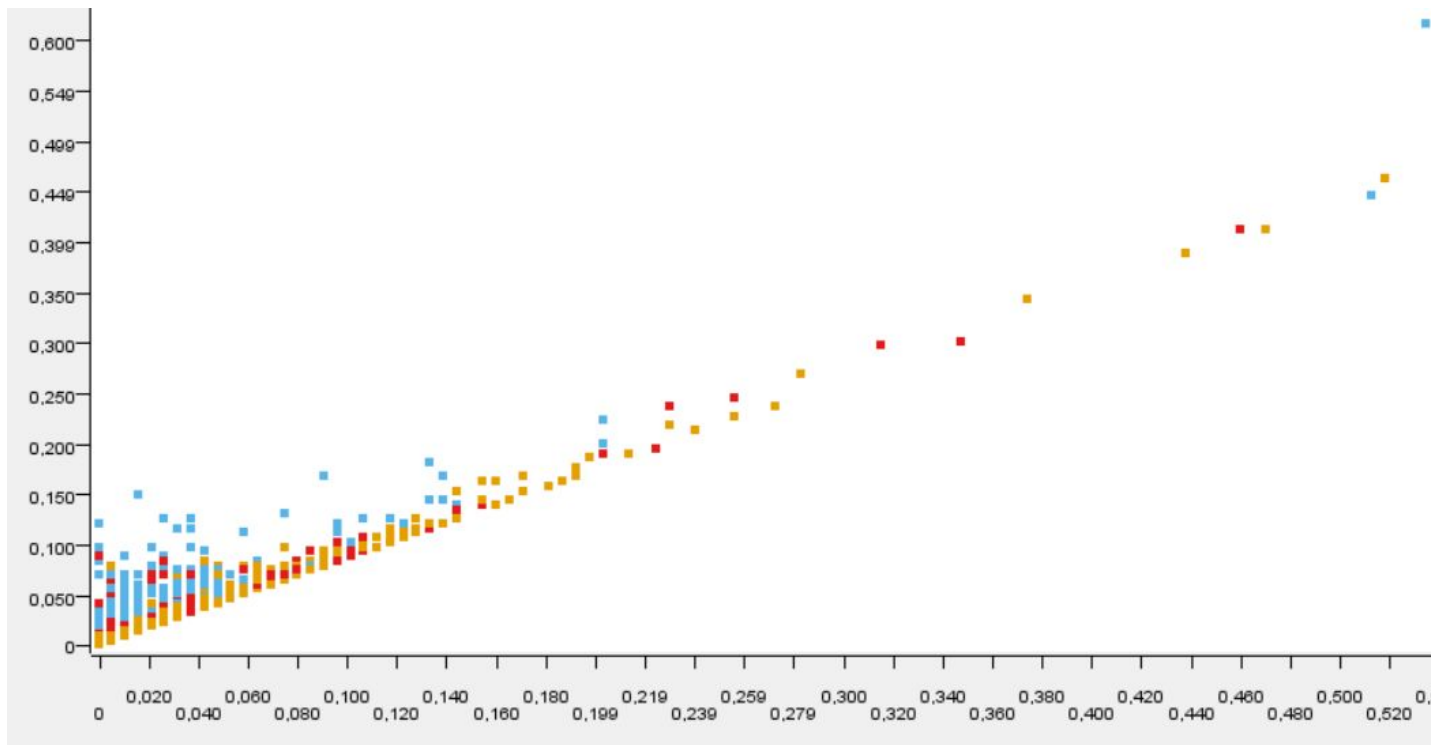
Score	Value
Entropy:	0,3657
Quality:	0,8425

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

K-means - Agrupación de comentarios

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	dll pa	0	0	0	0.989
3	cluster_2	B000LQOCH0	ABXLMWJIXXAIN	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	Twoapennyt...	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	A1MZY09TZK0BBI	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	A2MUGFV2TDQ47K	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	mom2emmas	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	A22P2J09N9HKE	S. Cabanaug...	0	0	1	0.857
26	cluster_0	B001GVISJM	A3FONPR03H3PJ5	Deborah S. L...	0	0	1	0.839

K-means - Gráfico de dispersión



Clustering jerárquico I

- Nodo *Hierarchical clustering* de KNIME
- Función de distancia de Manhattan
 - 27% de calidad
 - Primer y segundo grupo tienen tan solo 1 elemento cada uno
 - Tercero 9997

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

Clustering jerárquico II

- Nodo *Hierarchical clustering* de KNIME
- Función de distancia de Euclídea
 - 27% de calidad
 - Primer grupo: 1 elemento
 - Segundo 2
 - Tercero 9997

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

6.1. Árbol de decisión

Row ID	5	1	2	3	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

6.2. Naive-Bayes

Row ID	5	3	1	4	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

6.3. C4.5 de Quinlann

Row ID	5	1	2	3	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

6.4. Random forest

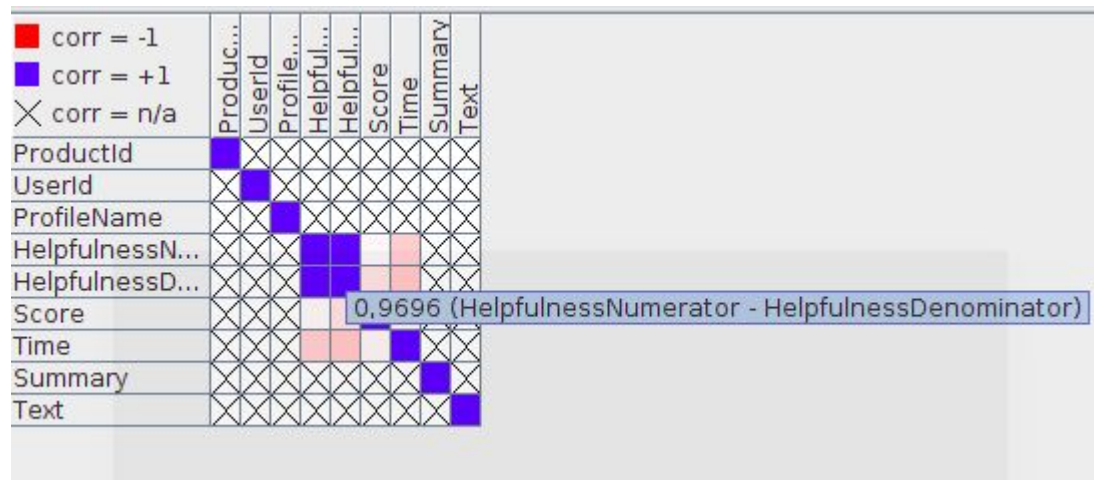
Row ID	5	1	4	2	3
5	263	5	1	0	1
1	57	18	0	0	1
4	69	4	0	0	0
2	35	3	0	1	0
3	50	3	0	0	0

6.5. Random forest con clústeres

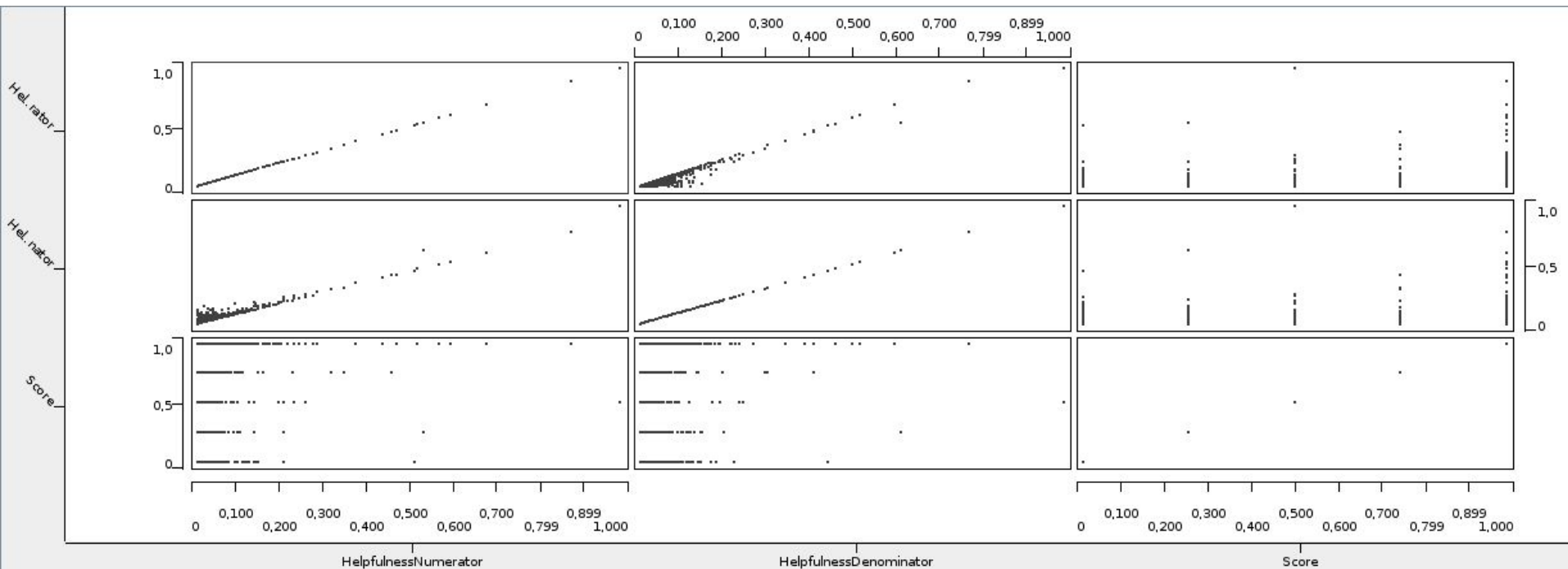
Row ID	Cluster 1	Cluster 2	Cluster 3
Cluster 1	173	0	3
Cluster 2	0	174	1
Cluster 3	0	0	161

7. Regresión

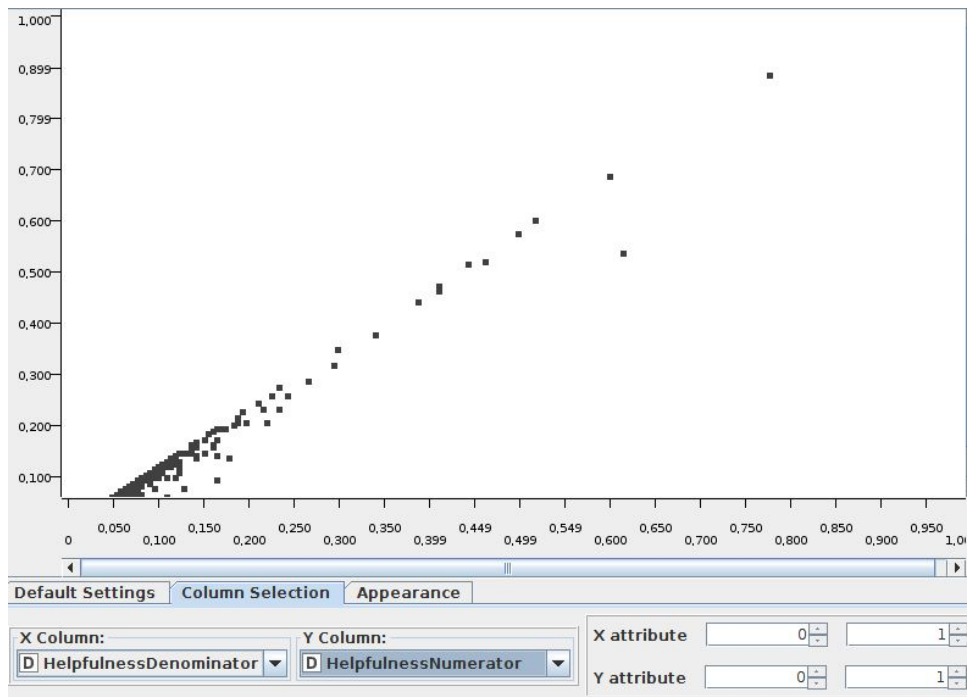
Correlación



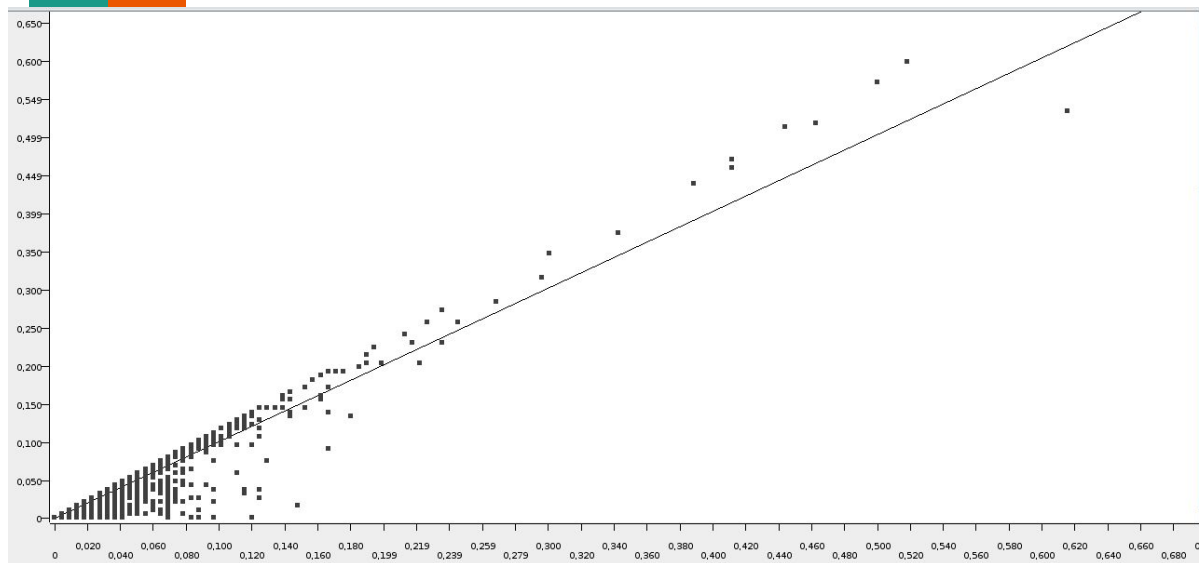
Matriz de dispersión



HelpDenominator VS HelpNumerator



Función lineal



Coeficiente de
Determinación
0,9401

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

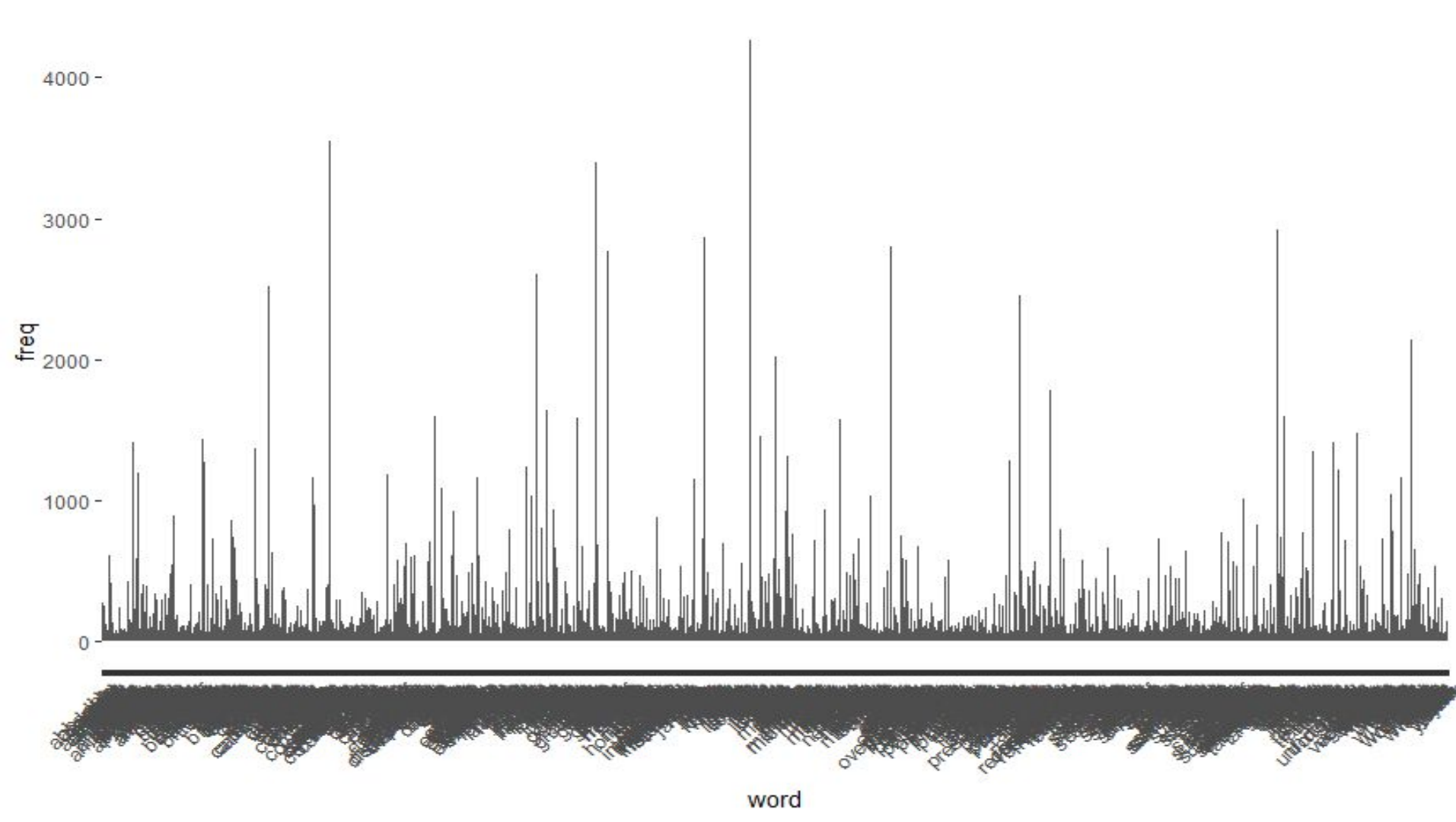
Adjusted R-Squared: 0,9401

8. Minería de texto

Frecuencia de las palabras

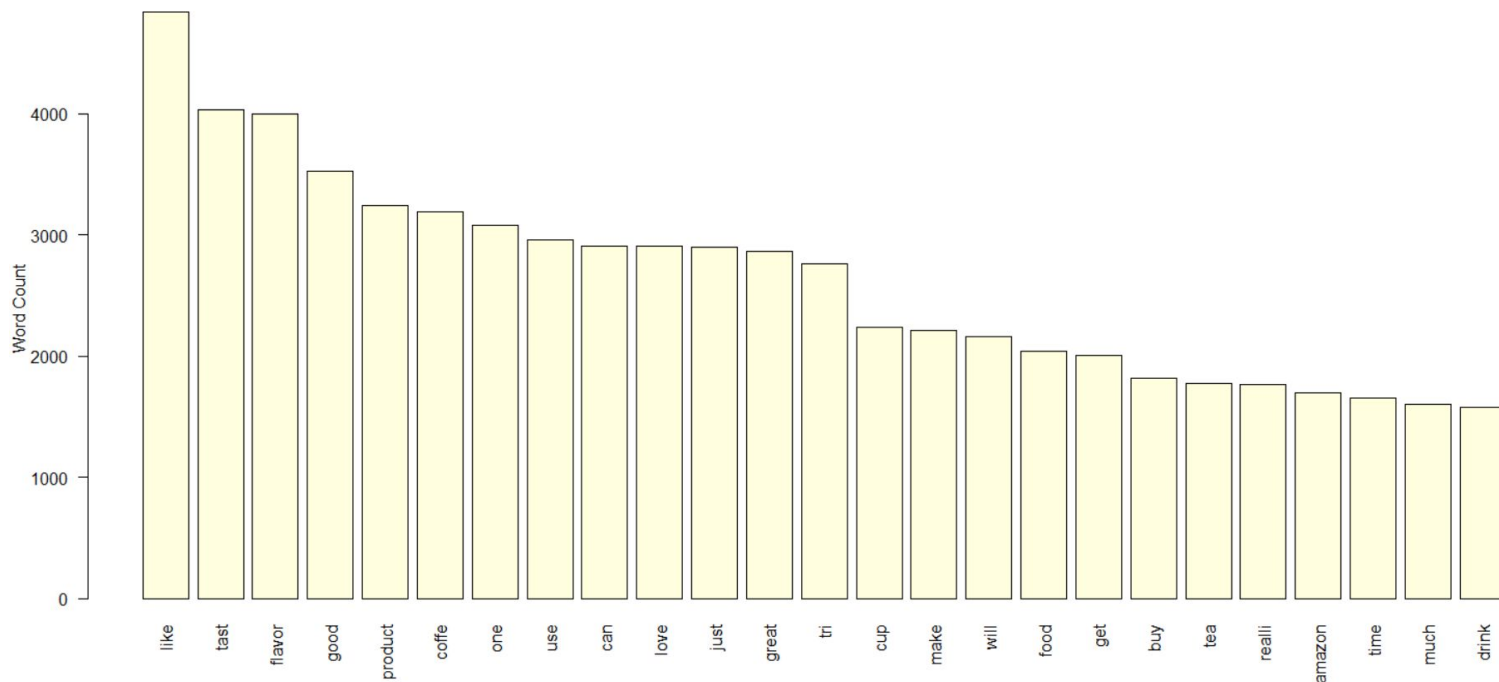


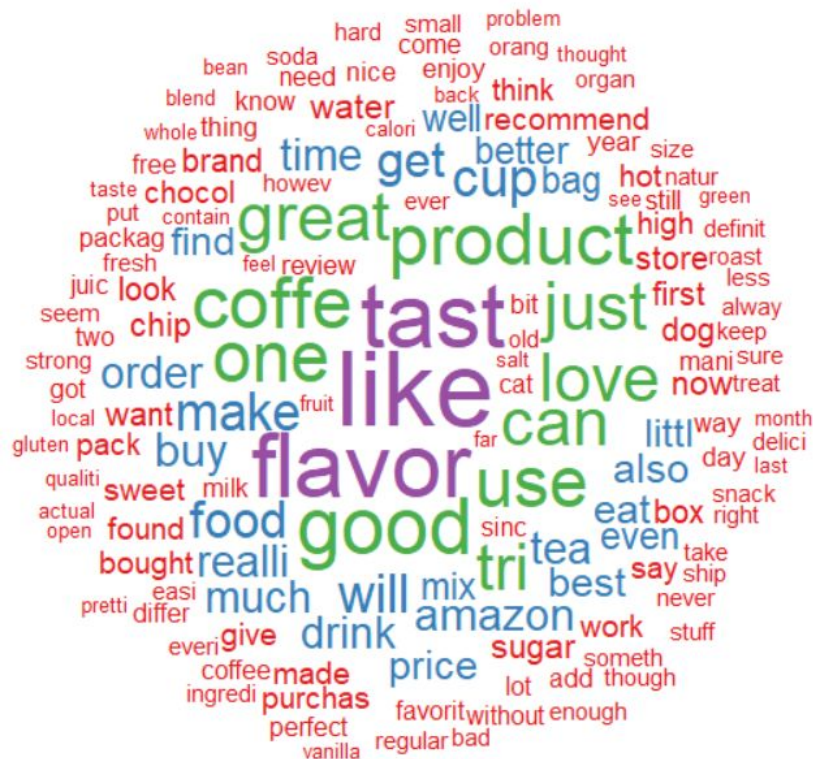
Palabra	Frecuencia
like	4837
tast	4035
flavor	4000
good	3527
product	3249
coffe	3193
one	3081
use	2963
can	2911
love	2911



Frecuencia de las palabras

Most Frequent Words From Amazon Reviews





Asociación de palabras: 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

Asociación de palabras: 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

9. 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].

Gracias por su atención
¿Preguntas?