

YouTube Recommendation System

El objetivo del sistema de recomendación es poder sugerir videos de nuestra base de datos a partir de un nuevo video. Este es extraído mediante la API de YouTube y posteriormente procesado para obtener las variables empleadas por el resto de métodos como pueden ser la extracción del porcentaje de emociones, las palabras de los títulos, comentarios y transcripciones o la probabilidad de que un video sea clickbait.

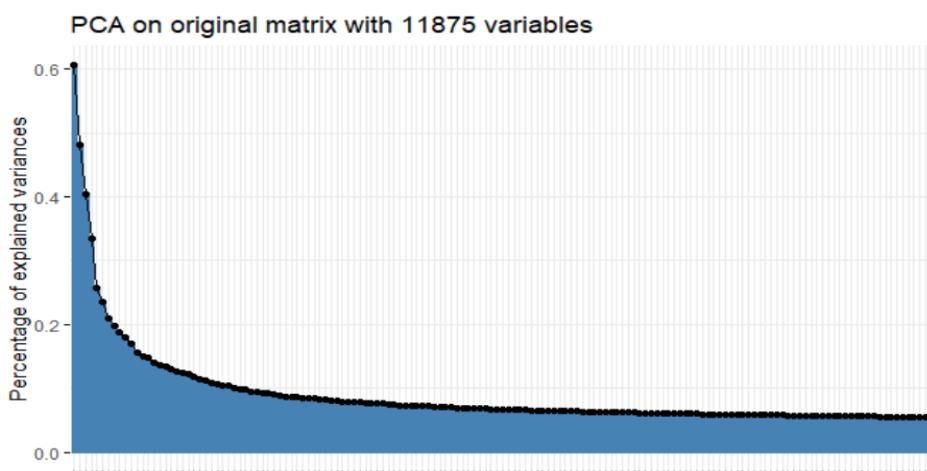
Previamente, se ha realizado un clustering de los videos de la base de datos y se ha entrenado un modelo XG-Boost con el objetivo de obtener la predicción del cluster del nuevo video. De esta manera, podemos reducir la complejidad computacional y centrarnos directamente en los videos de ese cluster. A continuación, sobre estos videos se pueden calcular matrices de distancia como la euclídea o manhattan, pero como justificaremos, las predicciones utilizando la similitud del coseno parecen mucho más estables.

Finalmente, nos quedaremos con aquellos videos que presenten una mayor similitud del coseno con el nuevo video, pero además la ordenación de la recomendación de esos videos se realizará analizando las similitudes de las miniaturas.

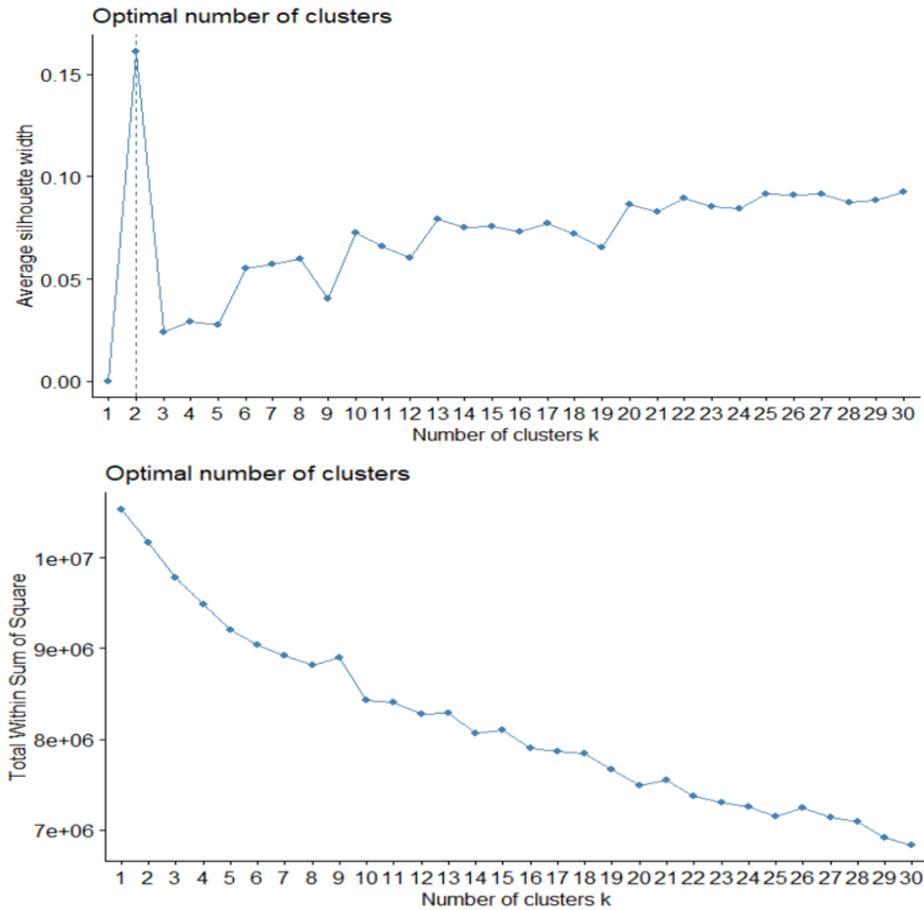
También, se propone la utilización de algoritmos de ranking, de manera que la recomendación puede extenderse a una secuencia de videos visualizados. Además, se van a mostrar técnicas de visualización que nos van a ayudar a entender qué caracteriza a cada uno de los clusters.

1. K-MEANS ON LATENT SPACE FOR CLUSTERING VIDEOS

El primer objetivo ha sido reducir el espacio de variables en un espacio latente, de esta manera conseguimos reducir la disminuir la dimensionalidad quedándonos con las principales fuentes de varianza, eliminando todas las fuentes de ruido que son irrelevantes para la agrupación de videos que queremos realizar, obteniendo así clusters más homogéneos.



Vamos a considerar un espacio latente de 100 componentes principales sobre el cual hemos aplicado k-means:



Los coeficientes de Silhouette nos indican para cada valor de k considerado, la media de cómo es de buena la asignación de los elementos a cada uno de sus clusters. Este valor se maximiza en nuestro caso cuando $k = 2$, aunque si solo consideramos dos clusters, la suma de cuadrados intra-cluster sigue siendo demasiado elevada y además un clustering de los vídeos únicamente en dos grupos se queda algo pobre. Se observa un submáximo de la media de los coeficientes de Silhouette en $k = 20$, donde además ya se reduce bastante la SC intra-cluster. Por lo tanto, vamos a considerar este número de clusters para el análisis.

Los clusters se distribuyen de la siguiente manera:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
103	121	283	203	194	353	664	178	290	628	106	116	1675	146	855	732	185	187	250	997

2. CLUSTERING PREDICTION FOR FEATURE SELECTION

Una vez realizados los clusters, se ha planteado la creación de un modelo de clasificación XG-Boost (*Extreme Gradient Boosting*). Se trata de un modelo paralelizable basado en árboles de decisión que es extremadamente rápido en el entrenamiento de grandes cantidades de datos y ofrece resultados mejores que muchos otros algoritmos mucho más costosos computacionalmente.

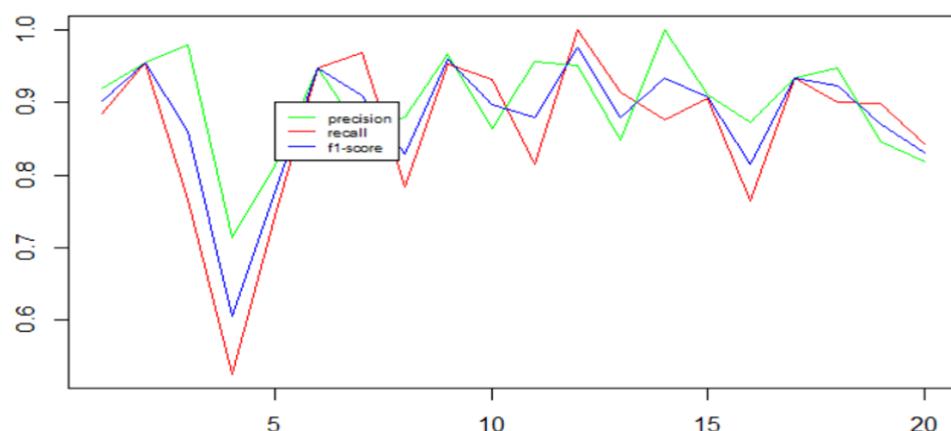
La finalidad de este modelo no es tan solo predecir el cluster para un nuevo video, si no que también nos permite indagar en las variables que más caracterizan a cada cluster, en este caso serán aquellas que provoquen una mayor ganancia de información respecto a la clasificación de cada cluster. Hemos decidido crear 20 árboles de decisión para cada grupo de videos, de manera que el número de rondas del algoritmo será 20.

El accuracy obtenido en el conjunto de test es muy elevado **0.8764** y por lo tanto, es el que emplearemos para predecir el cluster de un nuevo vídeo. La matriz de confusión resultante es la siguiente:

tables_test	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
0	23	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	
1	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
2	0	0	46	0	0	0	0	0	0	0	10	0	0	1	0	0	2	0	0	0	1
3	0	0	0	20	6	0	2	0	0	0	0	0	1	0	0	0	0	0	0	9	
4	0	0	0	0	7	26	0	0	0	0	0	0	0	0	0	0	0	0	0	2	
5	0	0	0	0	0	0	73	0	0	0	0	0	0	0	0	3	0	0	0	1	
6	0	0	0	0	0	0	0	120	0	0	0	0	2	0	0	2	0	0	0	0	
7	0	0	0	0	0	0	0	1	29	0	0	0	6	0	1	0	0	0	0	0	
8	0	0	0	0	0	0	0	0	0	59	1	0	0	2	0	0	0	0	0	0	
9	0	0	0	1	0	0	0	0	1	0	121	0	0	0	0	1	0	0	0	5	
10	0	0	0	0	0	0	0	0	0	0	22	0	1	0	0	1	0	0	0	3	
11	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	
12	0	0	0	0	0	0	0	7	1	0	2	0	1	297	0	4	3	0	0	28	
13	0	0	0	0	0	0	0	0	0	0	0	0	3	21	0	0	0	0	0	0	
14	0	0	0	0	0	0	0	0	1	0	0	0	6	0	153	2	0	0	6	1	
15	2	0	0	0	0	1	7	2	1	0	1	0	11	0	3	117	2	2	0	4	
16	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	28	0	0	0	
17	0	0	0	0	0	0	0	2	0	0	0	0	1	0	0	0	36	0	1	0	
18	0	0	0	0	0	0	0	0	0	0	0	0	1	0	4	0	0	0	44	0	
19	0	1	0	1	0	2	0	0	0	5	0	0	18	0	3	2	0	0	0	172	

[1] 0.8764385

Además se ha calculado la precisión, el recall y el f1-score para cada cluster:



Podemos ver que variables en general están siendo más importante para diferenciar entre los clusters:



Podemos observar palabras que podrían hacer referencia a distintos grupos de vídeos. Por ejemplo, una mayor frecuencia de palabras como recipiente, cocina o gordon (el cocinero), estarían implicando videos de cocina o palabras como web, javascript y tecnología videos orientados a la informática.

3. EXAMPLE OF RECOMMENDATION

Para mostrar el trabajo realizado se va a mostrar un ejemplo de recomendación con un nuevo video. Consideraremos el siguiente video de un unboxing del Samsung Galaxy S22, del cual se extrae mediante la API sus visitas, la duración, el título, una muestra de comentarios y la transcripción:

Views <int>	duracion <dbl>	Likes <int>	Title <chr>
79907	9.566667	1660	Samsung Galaxy S22 Unboxing, First Impressions!



comentarios

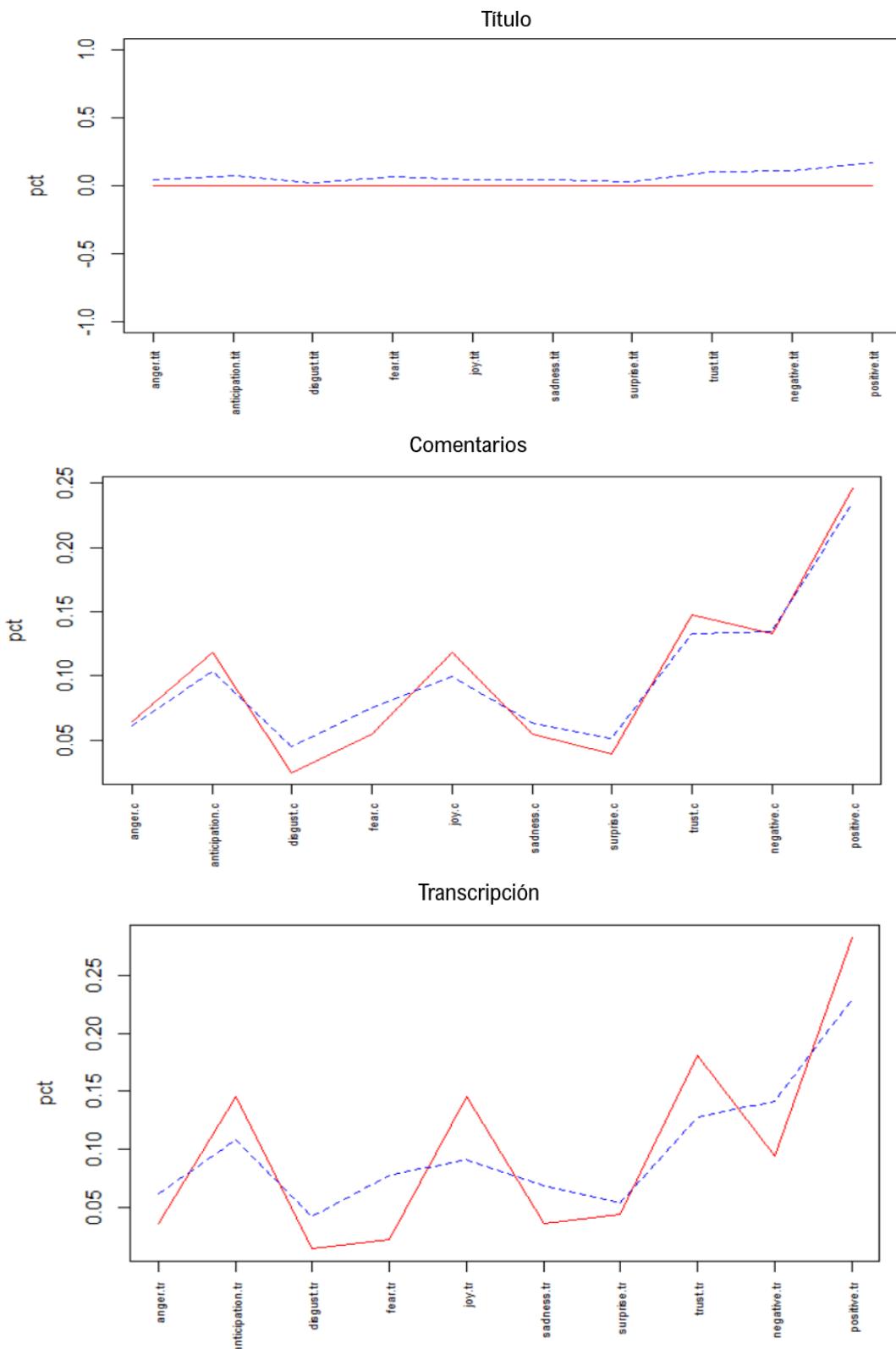
Samsung Galaxy S22 Unboxing, First Impressions! In this video, I share with you my first impressions and thoughts on the Samsung Galaxy S22, share what you think, want to see

transcripcion
<chr>

so what is up guys nick here helping you to master your technology and welcome to my galaxy s22 unboxing and first impressions i did decide to go with the phantom black edition you can s

3.1. SENTIMENT ANALYSIS FROM TITLE, COMMENTS AND TRANSCRIPTION

El primer paso es obtener el porcentaje de sentimientos del título, de los comentarios y de la transcripción mediante la librería *syuzhet*. Podemos comparar gráficamente el porcentaje de emociones de este nuevo video con la media de nuestra base de datos:

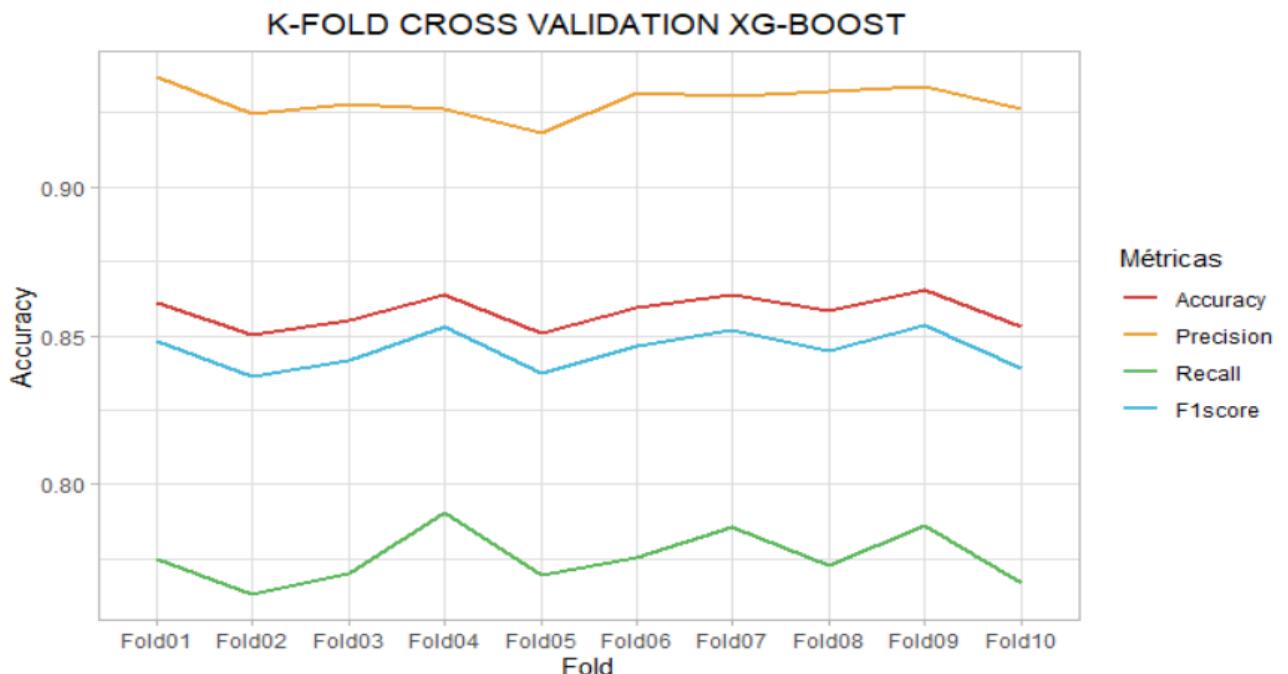


3.2. EXTRACTING WORDS FROM TITLE AND CLICKBAIT PREDICTION

Se extraen del video las palabras del título que estén contempladas en nuestras variables y se predice la probabilidad de ser clickbait a través de estas. El modelo XG-Boost para predecir el clickbait se ha creado a partir de una gran muestra de títulos etiquetados de distintas fuentes:

	titulos	clickbait
13 Firms That Received Bailout Money Owe Back Taxes		0
34 Pictures That'll Make You Want To Pack Your Bags And Move To Wales		1
Actor Jerry Orbach dead at age 69		0
18 Of The Greatest Photos To Have Taken Place Inside A Photo Booth		1
Israel Says Actions in Gaza Not War Crimes		0
21 Photos That Are Too Real For Indecisive People		1
Judge Puts Halt to ID Theft Inquiry Focusing on Immigrants		0
Icelandic volcanic eruption prompts evacuation, flight diversions		0
Grandparents Give Love Advice		1
21 Photos Of Flight Attendants Living The Life In Overhead Bins		1

El resultado obtenido en el conjunto de test es satisfactorio (87,8% accuracy), aunque tenemos bastantes falsos negativos de clickbait, aún así hemos considerado que podemos generalizar el modelo a nuevos datos:



Resultados en el conjunto de test:

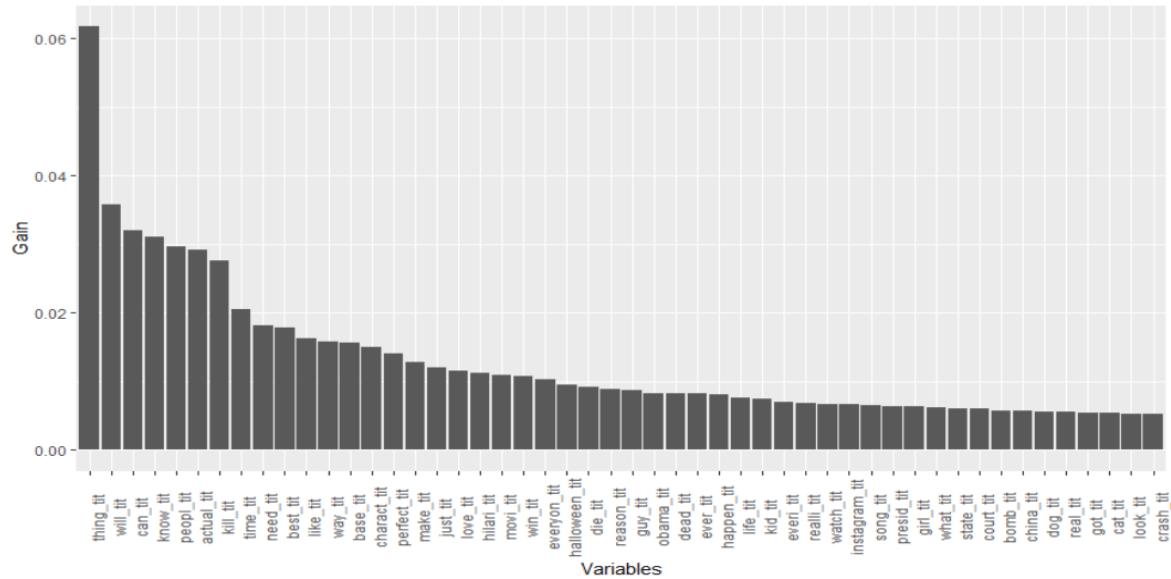
pred		
	FALSE	TRUE
0	6039	381
1	1186	5231

Predicción de clickbait sobre el nuevo video

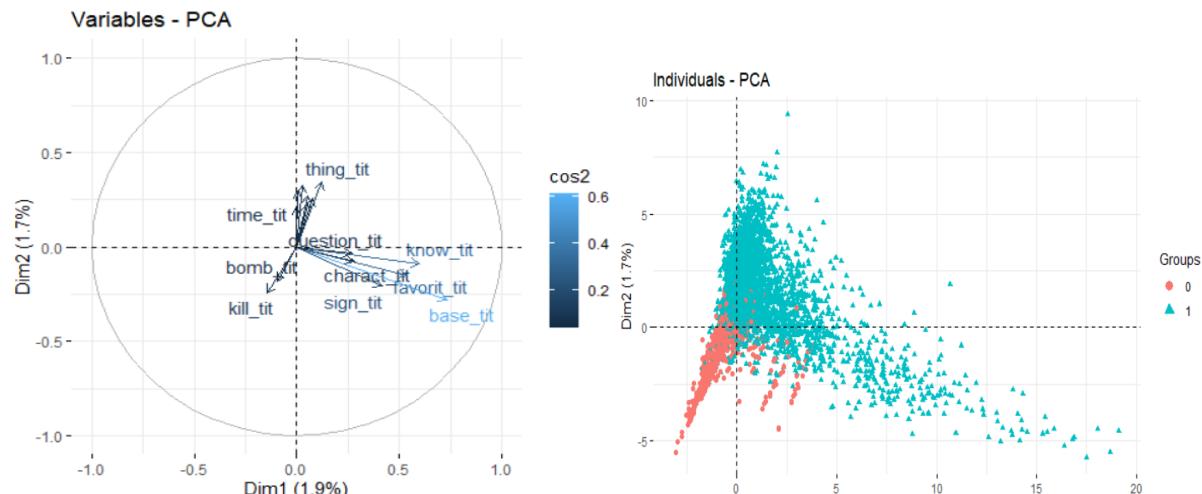
Creamos un vector con las palabras del título del nuevo video que estén como variables predictoras en el modelo:

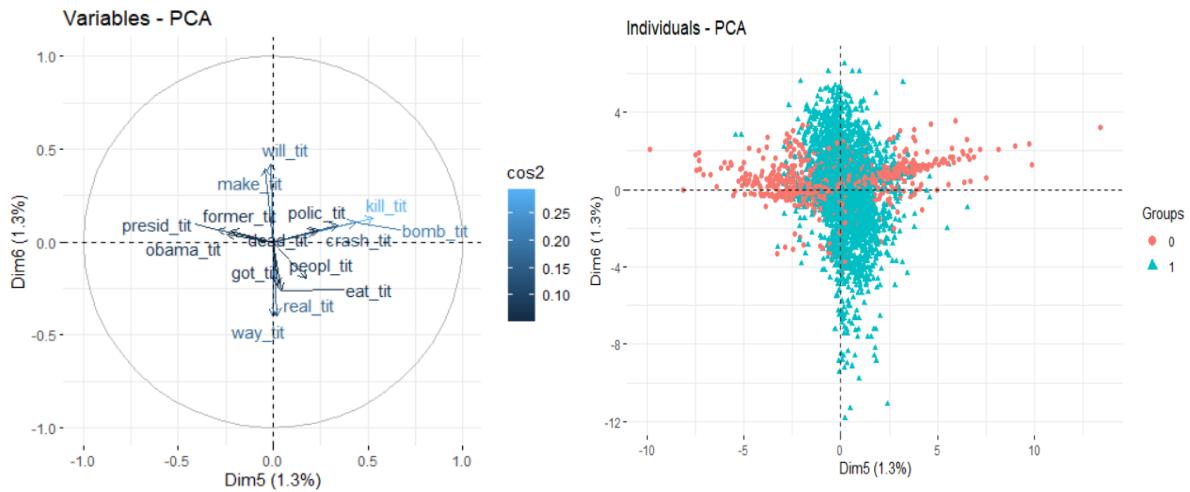


La predicción es menor a 0.5, de modo que en ese caso consideraríamos que el video no es clickbait. Podemos analizar las variables que más influyen a determinar si un video es o no clickbait analizando la ganancia de información del modelo:



Podemos visualizar un PCA de estas variables en algunas dimensiones:





Podemos observar que la primera componente estaría explicada principalmente por palabras como know, favourite, base... El hecho de que un título presente estas palabras está relacionado con que el video sea clickbait. Por otro lado, la segunda componente está explicada por un lado por las palabras como thing y time, las cuales también conllevarían a clasificar un video como clickbait. En dirección contraria estarían las palabras bomb y kill que no implicarían clickbait. En el segundo gráfico, se observa que los videos que tomen scores tanto positivos como negativos en la dimensión 6, tenderán a no ser clickbait, es decir aquellos que contengan en sus títulos palabras como presidente o obama. Por otro lado, la dimensión 6 está explicada por otras palabras que también explican clickbait como will, make, real...

3.3. EXTRACTING WORDS FROM COMMENTS AND TRANSCRIPTION

Ahora vamos a obtener las palabras de los comentarios y de la transcripción del video. Nos quedaremos solo con aquellas que hemos considerado para predecir los clusters y además observaremos cuáles de ellas son más relevantes para la predicción.

Comentarios

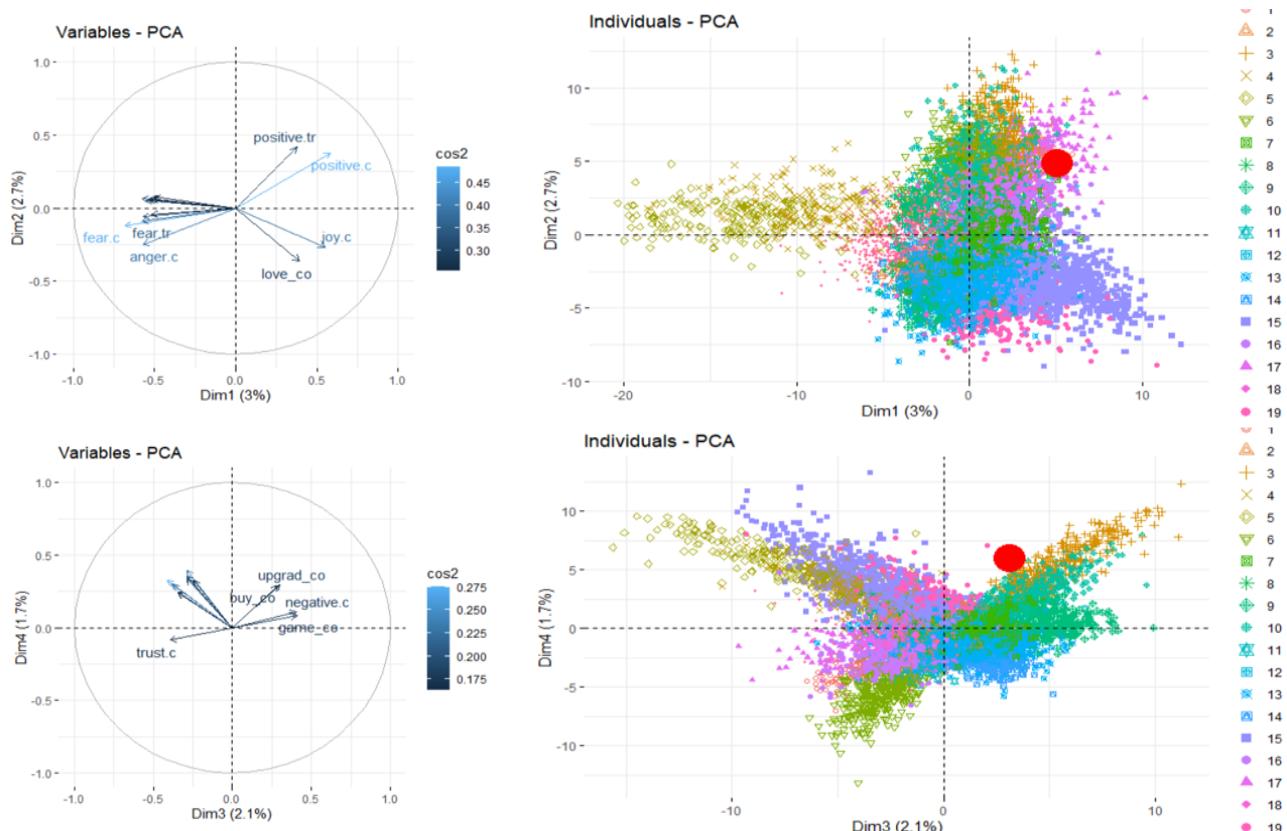


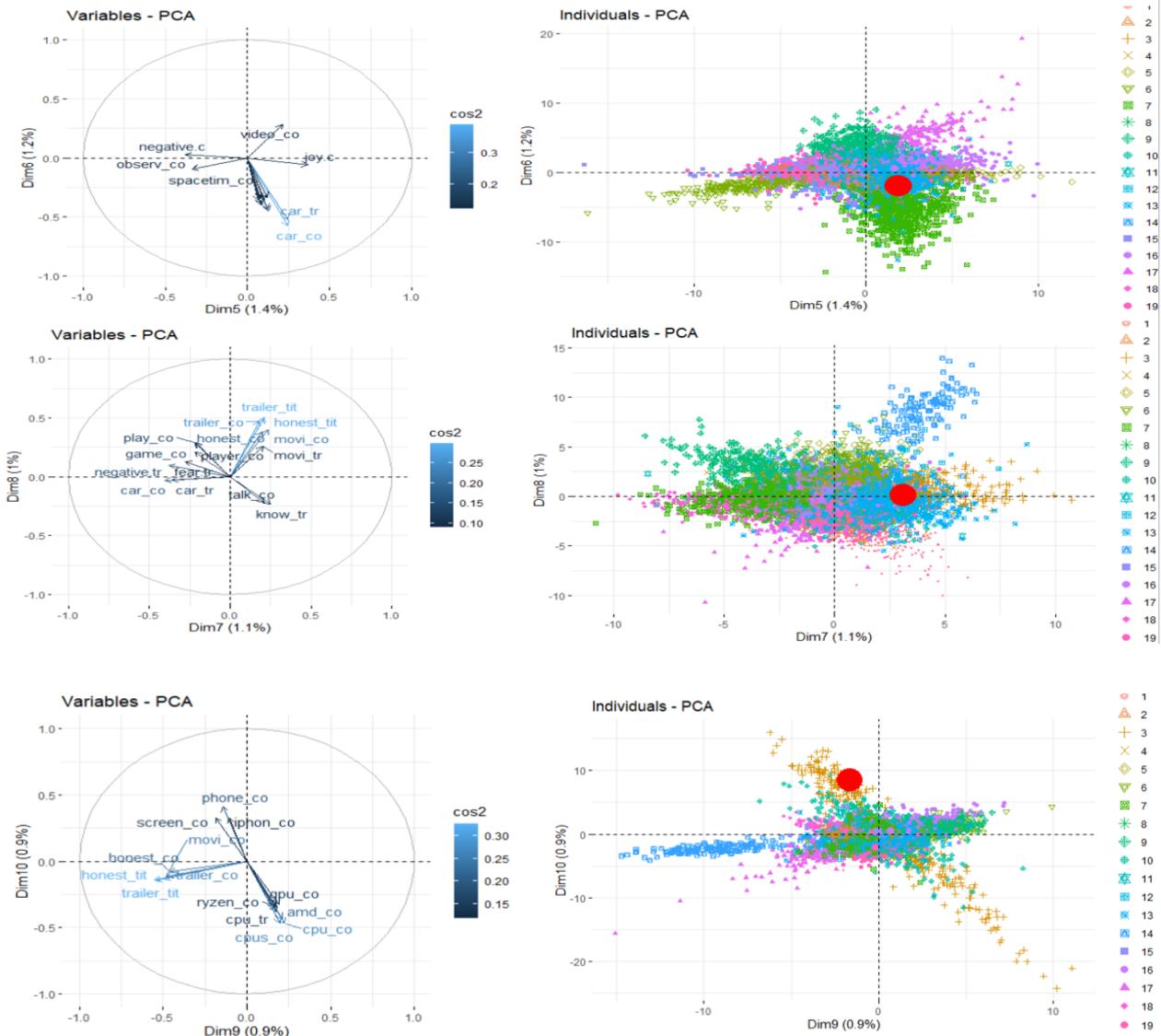
Transcripción



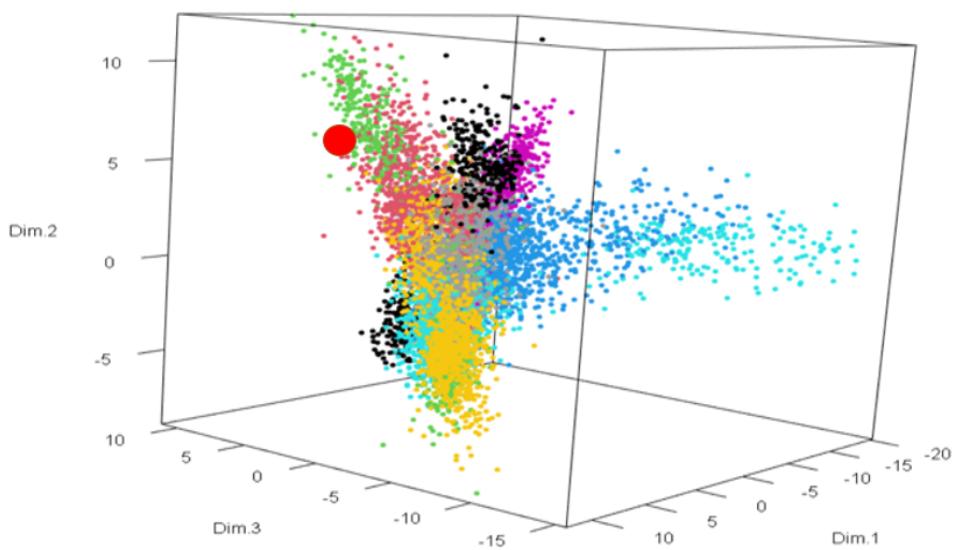
3.4. PCA WITH TOP 500 VARIABLES FOR VISUALIZATION

Vamos a visualizar mediante un PCA las frecuencias relativas de cada una de las palabras que más influyen a las predicciones de los clusters, de esta manera observamos qué palabras caracterizan a los clusters y qué relación hay entre ellas. En los gráficos se remarcá donde se ubicaría el nuevo dato tras recalcular el PCA.





Visualización 3D de las 3 primeras dimensiones

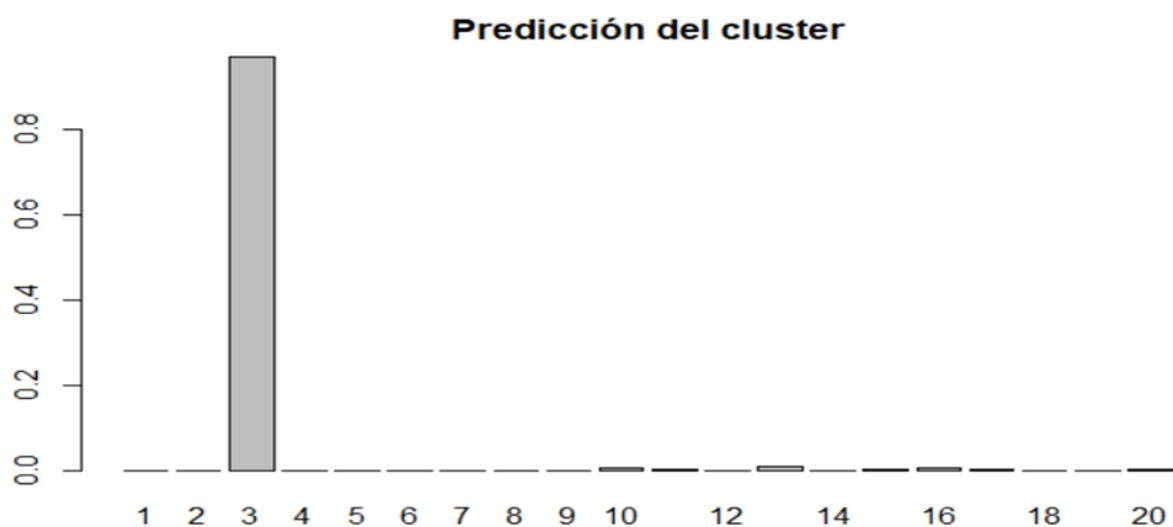


El nuevo dato toma scores positivos en la primera y segunda dimensión indicando que tendrá un porcentaje elevado de sentimientos positivos. Observamos que también tiene scores positivos en la cuarta dimensión lo que implicará la aparición de las palabras upgrade y buy, las cuales están correlacionadas entre ellas. Por último, destaca que el la décima dimensión posee scores elevados lo que conlleva la aparición de palabras como phone y screen, las cuales están incorrelacionadas con otro tipo de palabras más orientadas al sector de los ordenadores de sobremesa y no de móviles como por ejemplo gpu, ryzen, amd...

3.5. PREDICTION OF THE CLUSTER

Ya tenemos todos los datos con toda la extracción de los sentimientos y de las palabras de los títulos, comentarios y transcripciones. Ahora el objetivo es predecir el cluster de ese video para centrarnos únicamente en esos videos. En este caso, la predicción ha sido muy robusta, ya que modelo está con 97,23% de probabilidad seguro de que el nuevo video pertenece al cluster 3.

```
[1] 0.0003281126 0.0003994162 0.9721336961 0.0003368916 0.0003268810 0.0003322857 0.0004029597 0.0003281985 0.0003314444 0.0054778270 0.0006601877
[12] 0.0003262071 0.0083222399 0.0003281865 0.0006857063 0.0066933958 0.0009226804 0.0003421729 0.0003280649 0.0009934552
```



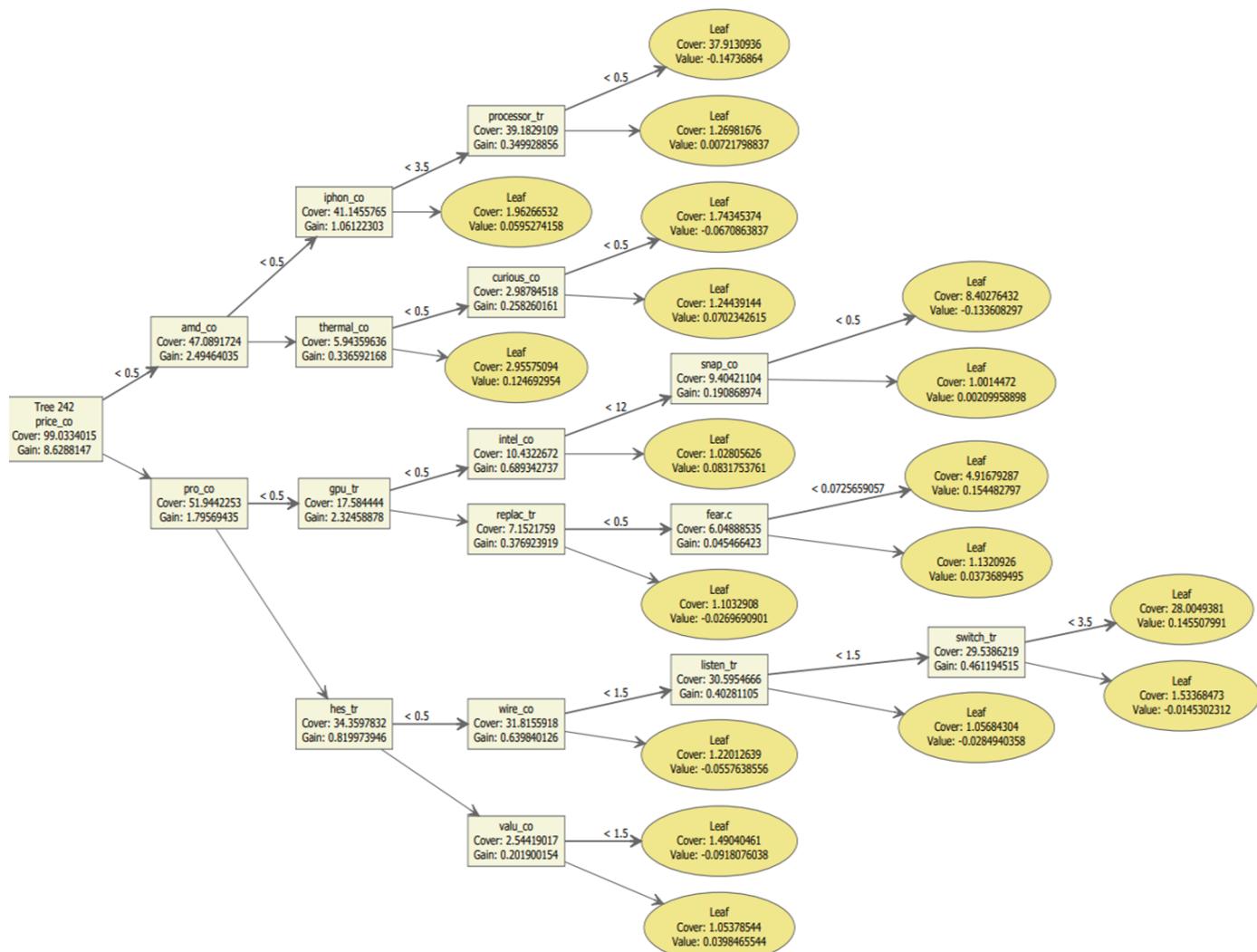
Vamos a observar los videos de nuestra base de datos que pertenecen a ese cluster:

Channel <chr>	Subscribers <dbl>	Title
TechLinked	1690000	Is This REALLY the RTX 3080??
TechLinked	1690000	It's over, Intel.
Techquickie	4000000	What is a Core i3, Core i5, or Core i7 as Fast As Possible
Hardware Canucks	1710000	My first 4K Curved Smart TV Is the Curve worth it?
Austin Evans	5240000	iPhone 6 vs Samsung Galaxy S5 Speed Test!
Hardware Canucks	1710000	TOP CASES OF 2016
Techquickie	4000000	Why Intel is STRUGGLING Against AMD
CNET	3350000	iPhone 5: The final rumors
Techquickie	4000000	32-bit vs 64-bit Computers & Phones as Fast As Possible
Linus Tech Tips	14300000	How many Chrome tabs can you open with 2TB RAM?

Las variables con más ganancia de información para predecir el cluster 3 son phone, gpu, price, ryzen, porcentaje alto de sentimientos positivos en los comentarios...



Podemos observar uno de los árboles de decisión asociados a determinar ese cluster:



Vemos un ejemplo, de si este árbol ofrecería una probabilidad positiva de que el nuevo video pertenezca a este cluster:

price_co <dbl>	amd_co <dbl>	iphon_co <dbl>
0	0	24

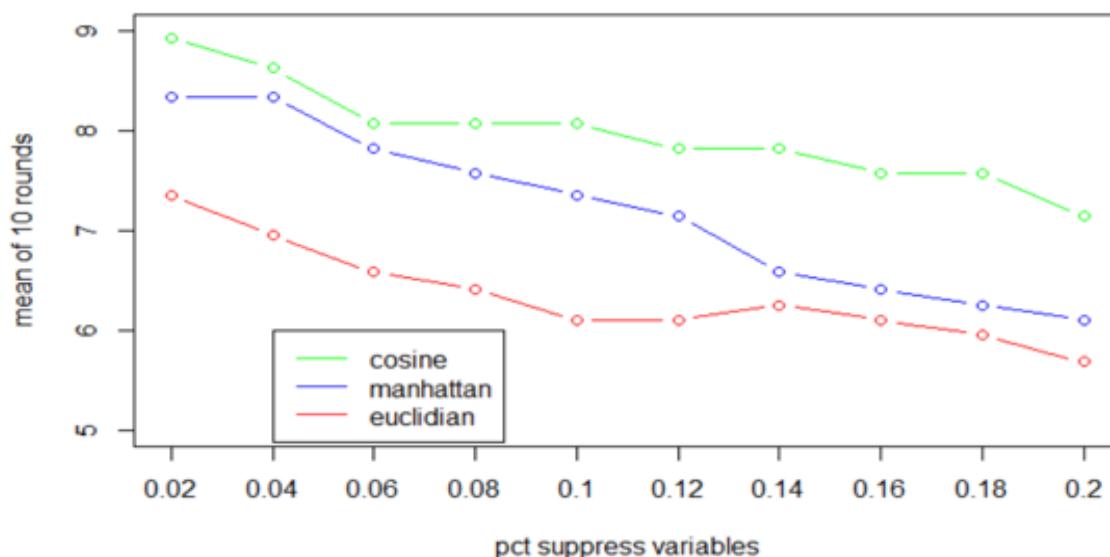
Como no aparece ni la palabra price, ni amd, pero aparece iphone 4 o más veces, la probabilidad que ofrece este árbol es mayor a 0.5 de pertenecer a este cluster (value = 0.05953)

3.6. EUCLIDIAN, MANHATTAN DISTANCE AND COSINE SIMILITUDE

La mejor forma de evaluar los distintos métodos realizados para “predecir” las recomendaciones sería que los propios usuarios nos indiquen si el contenido recomendado está siendo de su agrado. A pequeña escala con las pruebas realizadas subjetivamente nos ha gustado más la similitud del coseno, además se sabe que este tipo de similitud funciona mejor con el lenguaje natural y matrices de altas dimensiones. Un método que hemos empleado para ver la estabilidad de las predicciones es el siguiente:

1. Realizar N iteraciones aleatorias en la que se eliminarán k variables antes de calcular la matriz de similitud o distancias.
2. Para cada iteración se hacen 10 rondas y se recomiendan un top N de videos a partir de un nuevo dato arbitrario (se eliminan cada vez k pct de variables).
3. Calcular el número de ítems que se han recomendado en total y compararlo con el top N seleccionado. Interesa que el número de ítems sea exactamente el top N seleccionado, es decir, que en todas las rondas se recomiendan siempre los mismos top N artículos, de manera que el método sea muy sólido y estable al recomendar y no dependa de la influencia de la varianza asociada a muy pocas variables.

En el siguiente ejemplo, se ha realizado este método con diversos videos con recomendaciones de 25 videos en cada ronda, eliminando entre 0.02 y 0.2 por ciento de las muchas variables originales. Se observa como la función del coseno parece recomendar videos de forma más estable (el modelo más estable sería aquel que en cada ronda recomendase siempre los mismos 25 videos, lo que implicaría un valor de 10)



Por lo tanto, utilizaremos la distancia del coseno. Podemos comparar las recomendaciones que ofrecen las distintas matrices ya sea usando todas las variables originales o solo aquellas más importantes para predecir el cluster según la ganancia de información.

COSINE SIMILITUDE MATRIX OF THE CLUSTER

Channel <chr>	Subscribers <dbl>	Title <chr>
447 Marques Brownlee	15200000	iPhone 7 Unboxing: Jet Black vs Matte Black!
3349 CNET	3410000	Samsung Galaxy S21s Galaxy S22 Plus is just the right size (full review)
3332 CNET	3410000	Samsung Galaxy S22 review: For people who love smaller phones
6057 Marques Brownlee	15500000	Galaxy S22 Impressions: 1 Real Downgrade!
2001 CNET	3350000	Galaxy S9+ unboxing: Everything you get
3392 CNET	3410000	Samsung Galaxy S21 FE review: A solid \$700 phone that comes
6054 Marques Brownlee	15500000	Galaxy S22 Review: The iPhone of Android!
6069 Marques Brownlee	15500000	Oppo Find N Impressions: The Best Folding Phone?!
6097 Marques Brownlee	15500000	Samsung Galaxy Z Fold 3 Impressions: 3 New Features!
6096 Marques Brownlee	15500000	Samsung Galaxy Z Flip 3 Impressions: Design Refresh!
Channel <chr>	Subscribers <dbl>	Title <chr>
3332 CNET	3410000	Samsung Galaxy S22 review: For people who love smaller phones
6057 Marques Brownlee	15500000	Galaxy S22 Impressions: 1 Real Downgrade!
6069 Marques Brownlee	15500000	Oppo Find N Impressions: The Best Folding Phone?!
6097 Marques Brownlee	15500000	Samsung Galaxy Z Fold 3 Impressions: 3 New Features!
3330 CNET	3410000	Samsung Galaxy Book 2 Pro and Pro 360 are better for working anywhere
447 Marques Brownlee	15200000	iPhone 7 Unboxing: Jet Black vs Matte Black!
708 Austin Evans	5240000	iPhone 6 vs Samsung Galaxy S5 Speed Test!
3349 CNET	3410000	Samsung Galaxy S21s Galaxy S22 Plus is just the right size (full review)
3319 CNET	3410000	Galaxy S22 Ultra vs. Note 20 Ultra
6096 Marques Brownlee	15500000	Samsung Galaxy Z Flip 3 Impressions: Design Refresh!

MANHATTAN DISTANCE MATRIX OF THE CLUSTER

Channel <chr>	Subscribers <dbl>	Title <chr>
6056 Marques Brownlee	15500000	Galaxy S22 Ultra Impressions: It's a Note!
6069 Marques Brownlee	15500000	Oppo Find N Impressions: The Best Folding Phone?!
3352 CNET	3410000	Samsung Galaxy S22 Ultra Review: An Upgrade for Galaxy Note Fans
3332 CNET	3410000	Samsung Galaxy S22 review: For people who love smaller phones
6097 Marques Brownlee	15500000	Samsung Galaxy Z Fold 3 Impressions: 3 New Features!
6096 Marques Brownlee	15500000	Samsung Galaxy Z Flip 3 Impressions: Design Refresh!
3361 CNET	3410000	Galaxy S22 Ultra vs. S21 Ultra spec comparison
6057 Marques Brownlee	15500000	Galaxy S22 Impressions: 1 Real Downgrade!
299 Hardware Canucks	1710000	Samsung Galaxy Note 8 - A True User Review
6093 Marques Brownlee	15500000	Samsung Z Flip 3 Review: The First Big Step!
Channel <chr>	Subscribers <dbl>	Title <chr>
299 Hardware Canucks	1710000	Samsung Galaxy Note 8 - A True User Review
6082 Marques Brownlee	15500000	NEW M1 Max MacBook Pro Reaction: The Ports are Back!
6052 Marques Brownlee	15500000	Galaxy Tab S8 Ultra: A Monster Tablet!
6077 Marques Brownlee	15500000	M1 Max MacBook Pro Review: Truly Next Level!
4841 Hardware Canucks	1730000	Lenovo Legion 7 AMD Review - Worth it vs the Legion 5 PRO?
4848 Hardware Canucks	1730000	Its FINALLY Here - ASUS ROG Zephyrus G14 (2021) Review
4835 Hardware Canucks	1730000	Lenovo ThinkPad X1 Carbon Gen 9 Review - PERFECTION
6097 Marques Brownlee	15500000	Samsung Galaxy Z Fold 3 Impressions: 3 New Features!
4820 Hardware Canucks	1730000	Dell Inspiron 14 2-in-1 Review - An Affordable AMD POWERHOUSE
5877 Linus Tech Tips	14400000	2 extra inches of Mac is a BIG difference! - 16 inch M1 MacBook Pro

EUCLIDIAN DISTANCE MATRIX OF THE CLUSTER

	Channel <chr>	Subscribers <dbl>	Title
2793	Austin Evans	5280000	Before you upgrade to iPhone 13...
6097	Marques Brownlee	15500000	Samsung Galaxy Z Fold 3 Impressions: 3 New Features!
4789	Hardware Canucks	1730000	Surface Pro 8 Review - The Microsoft Tax
3352	CNET	3410000	Samsung Galaxy S22 Ultra Review: An Upgrade for Galaxy Note Fans
6101	Marques Brownlee	15500000	The iPhone 13 Models!
6081	Marques Brownlee	15500000	Surface Duo 2: Can This Be Saved?
6052	Marques Brownlee	15500000	Galaxy Tab S8 Ultra: A Monster Tablet!
3319	CNET	3410000	Galaxy S22 Ultra vs. Note 20 Ultra
6057	Marques Brownlee	15500000	Galaxy S22 Impressions: 1 Real Downgrade!
2850	Austin Evans	5280000	iPhone 12 Mini - I'm switching!
	Channel <chr>	Subscribers <dbl>	Title
6097	Marques Brownlee	15500000	Samsung Galaxy Z Fold 3 Impressions: 3 New Features!
3332	CNET	3410000	Samsung Galaxy S22 review: For people who love smaller phones
4820	Hardware Canucks	1730000	Dell Inspiron 14 2-in-1 Review - An Affordable AMD POWERHOUSE
3319	CNET	3410000	Galaxy S22 Ultra vs. Note 20 Ultra
6054	Marques Brownlee	15500000	Galaxy S22 Review: The iPhone of Android!
6056	Marques Brownlee	15500000	Galaxy S22 Ultra Impressions: It's a Note!
2822	Austin Evans	5280000	Apple was SO close... New iMac 2021
299	Hardware Canucks	1710000	Samsung Galaxy Note 8 - A True User Review
1476	Hardware Canucks	1710000	OnePlus 7T vs 7 Pro - We Have A WINNER!
6096	Marques Brownlee	15500000	Samsung Galaxy Z Flip 3 Impressions: Design Refresh!

3.7. ORDERING THE TOP N BY MINIATURE SIMILARITY

La predicción de los top 10 videos más similares se ordenan en función de la similitud de la miniatura, así obtenemos patrones de forma y colores similares de las imágenes que captarán la atención de los usuarios. Se muestra el ejemplo de los más similares de el video de ejemplo:

ORDERING THE TOP N BY MINIATURE SIMILARITY

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	x
447	1	1	1	4	141	139	134	175	147	144	162	190	210	221	227	229	194	175	
3349	1	1	1	2	16	18	11	21	24	30	31	29	29	24	13	9	6	2	
3332	1	1	1	1	207	204	199	205	212	214	203	207	209	208	208	208	207	199	
6057	1	1	1	1	210	216	218	212	209	212	212	174	191	215	214	218	217	204	
2001	1	1	1	0	14	18	20	20	20	22	17	15	16	12	16	15	17	18	
3392	1	1	1	1	111	78	121	90	101	139	134	117	117	112	112	102	113	114	
6054	1	1	1	1	232	226	223	225	224	221	223	224	225	224	225	226	225	220	
6069	1	1	1	4	225	220	119	53	41	55	63	47	43	116	128	134	133	131	
6097	1	1	1	0	54	107	62	59	98	93	200	157	98	173	138	100	32	25	
6096	1	1	1	0	219	221	221	217	194	186	198	220	182	244	252	238	121	84	
nuevo_dato	1	1	1	1	1	3	6	20	20	12	18	94	67	15	6	8	4	3	

6069 447 3392 3332 6054 6096 6057 6097 3349 2001
0.7690037 0.7338240 0.7328232 0.7304587 0.7286724 0.6871569 0.6771738 0.6521816 0.5081922 0.5006916



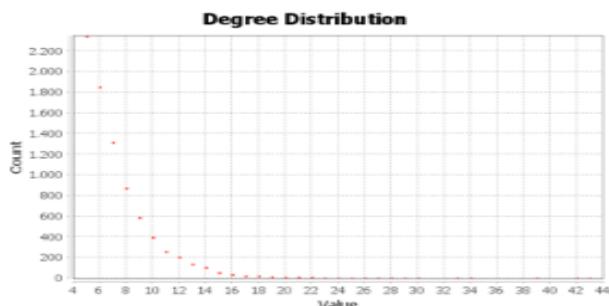
Channel <chr>	Subscribers <dbl>	Title <chr>
6069 Marques Brownlee	15500000	Oppo Find N Impressions: The Best Folding Phone?!
447 Marques Brownlee	15200000	iPhone 7 Unboxing: Jet Black vs Matte Black!
3392 CNET	3410000	Samsungâ€“200â€“231s Galaxy S21 FE review: A solid \$700 phone that cor
3332 CNET	3410000	Samsung Galaxy S22 review: For people who love smaller phones
6054 Marques Brownlee	15500000	Galaxy S22 Review: The iPhone of Android!
6096 Marques Brownlee	15500000	Samsung Galaxy Z Flip 3 Impressions: Design Refresh!
6057 Marques Brownlee	15500000	Galaxy S22 Impressions: 1 Real Downgrade!
6097 Marques Brownlee	15500000	Samsung Galaxy Z Fold 3 Impressions: 3 New Features!
3349 CNET	3410000	Samsungâ€“200â€“231s Galaxy S22 Plus is just the right size (full review)
2001 CNET	3350000	Galaxy S9+ unboxing: Everything you get



En este caso tras ver la similitud de imágenes, el video que se recomienda primero es el que se recomendaba el octavo solo atendiendo a la similitud, esto se debe al parecido en los colores de la pantalla y se trata de las primeras impresiones del móvil Oppo Find N. Es un video que perfectamente puede interesar al usuario que ha visto el unboxing del Samsung Galaxy S22, ya que se puede tratar de un usuario que quiera comprarse un móvil y quiera ver opiniones de distintos modelos.

3.8. K-NEIGHBORS SIMILARITY GRAPH (k=5)

Para entender la estructura de recomendación de los videos se ha realizado el siguiente grafo en el cual cada video está unido potencialmente a sus vecinos más cercanos, es decir, con los que tiene mayor similitud del coseno. Además, cada video está coloreado en función de su cluster y se ve perfectamente esta estructura entre los videos.

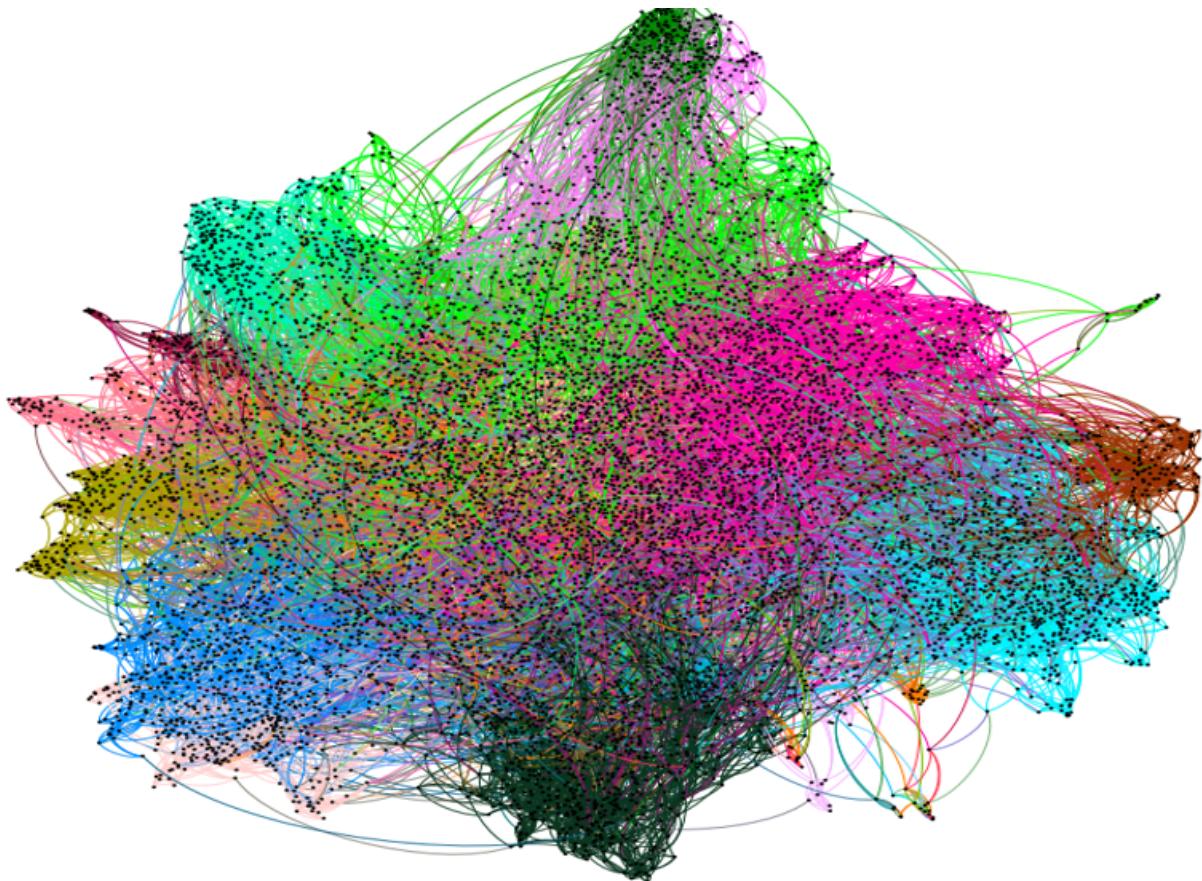


$$G = (V, E) \text{ GND}$$

$$V = \{\text{Videos de YouTube}\}$$

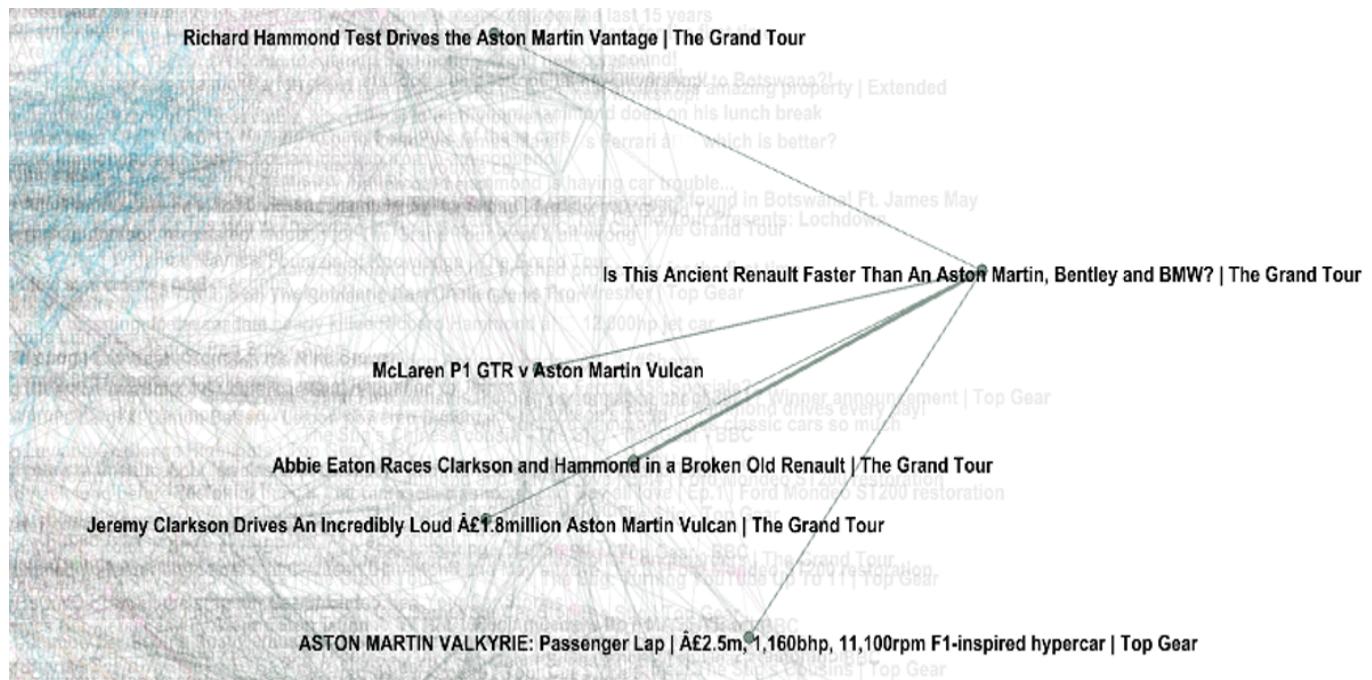
$$E = \{(x, y) \in V / x \text{ e } y \text{ son similares}\}$$

$$p(x, y) = \text{"Similitud coseno entre } x \text{ e } y\text{"}$$



Vamos a visualizar algún ejemplo de los videos:

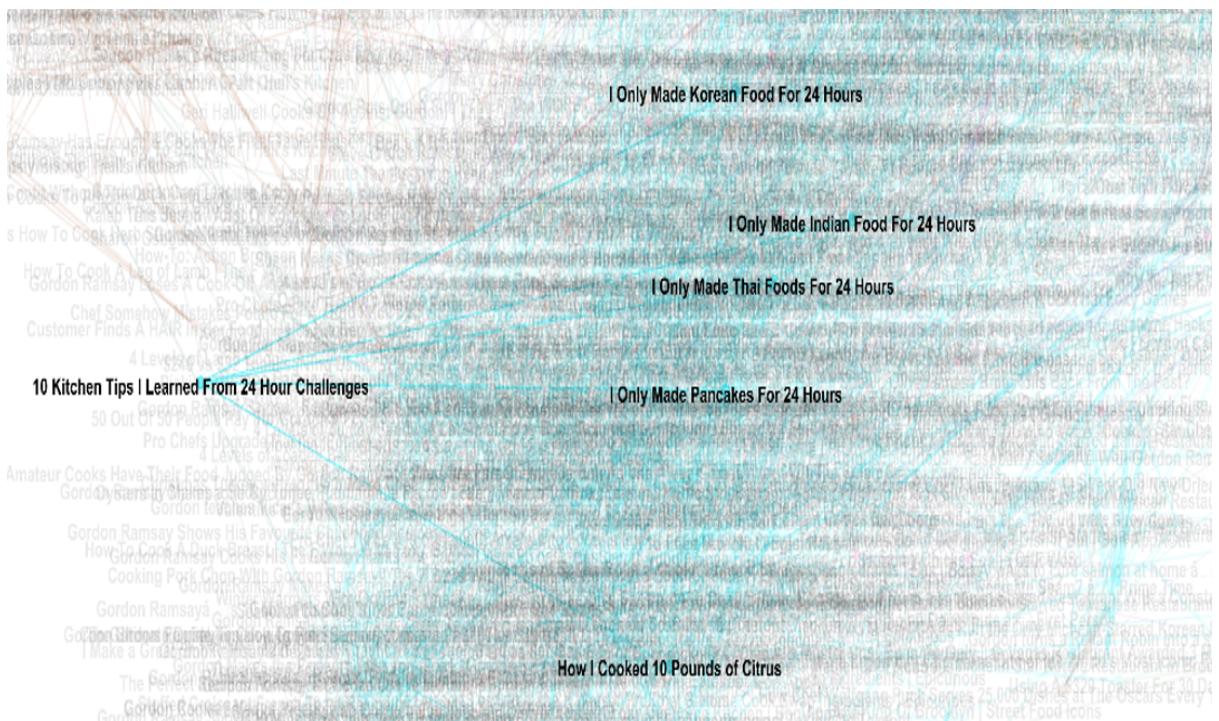
Top 5 videos más similares a *Is This Ancient Renault Faster Than An Aston Martin y sus incidentes* (cluster de coches)



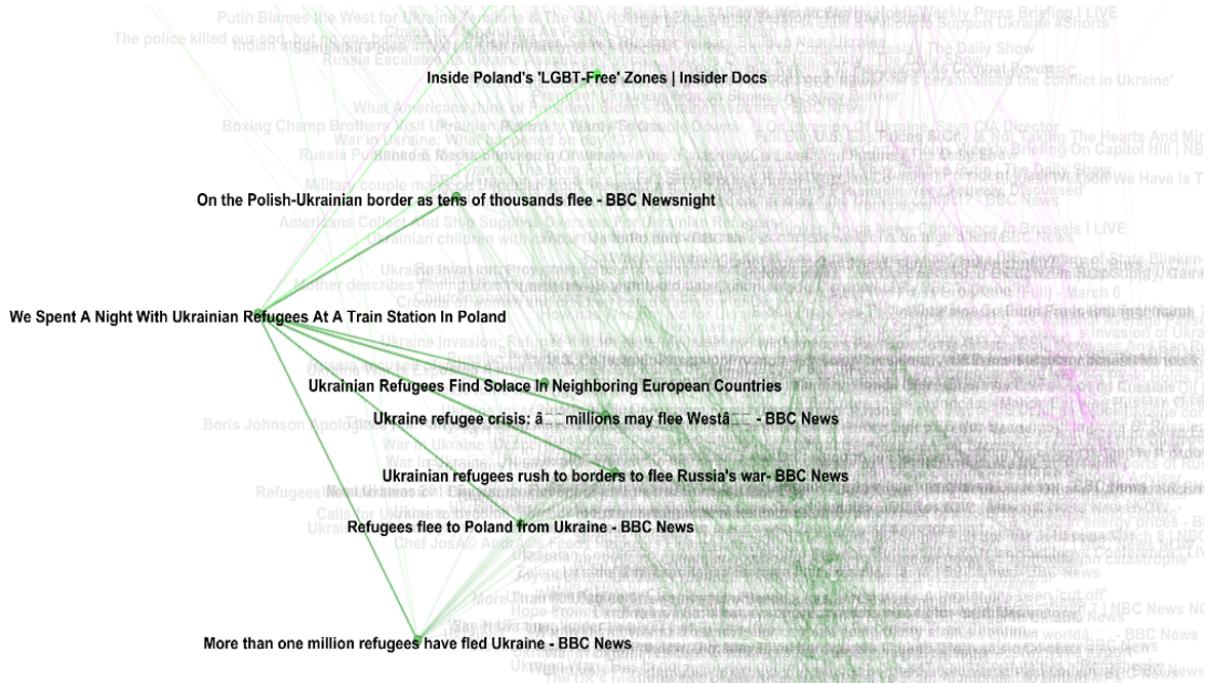
Top 5 videos más similares a *Is These Chips Are Better Than CPUs y sus incidentes (cluster de coches)*



Top 5 videos más similares a *10 Kitchen Tips | Learned From 24 Hour Challenges y sus incidentes (cluster de cocina)*



Top 5 videos más similares a *We Spent A Night With Ukrainian Refugees At A Train Station in Poland y sus incidentes* (cluster de guerra de Ucrania)



3.9. RANKING ALGORITHM OF THE VIDEOS IN THE CLUSTER

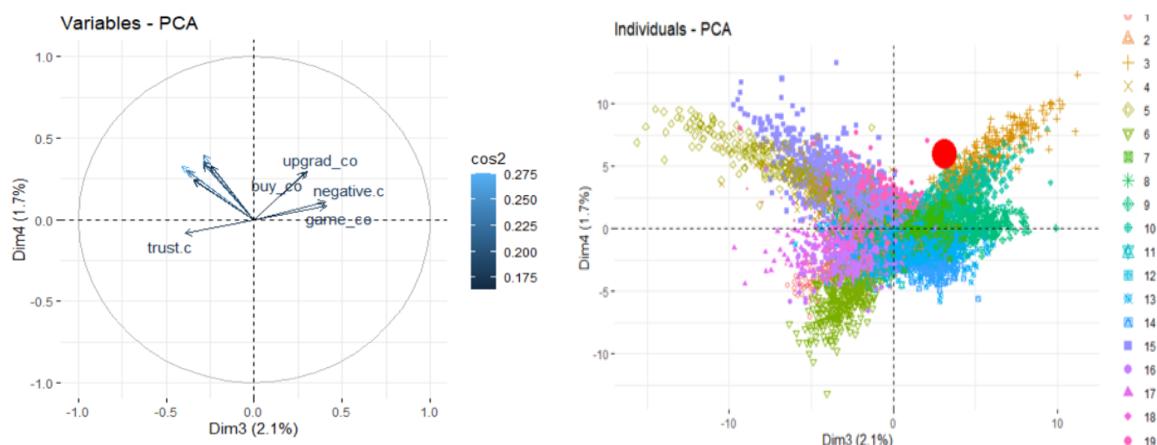
Para generalizar la idea de la recomendación a una secuencia de videos se ha planteado usar un algoritmo universal de ranking que asigna a cada nodo o video una puntuación de ser recomendado a partir de una o más queries. Se ha empleado la función de kernel de Laplacian sobre toda nuestra matriz de datos. Según la documentación, el método se basa en crear una red ponderada donde los nodos de las queries bombean sus puntuaciones a sus vecinos hasta converger:

Un algoritmo de clasificación universal simple que explota la estructura geométrica global intrínseca de los datos. En muchas aplicaciones del mundo real, esto debería ser superior a un método local en el que los datos simplemente se clasifican por distancias euclidianas por pares. En primer lugar, se define una red ponderada sobre los datos y se asigna una puntuación autorizada a cada consulta. Los puntos de consulta actúan como nodos de origen que bombean continuamente sus puntuajes autorizados a los puntos restantes a través de la red ponderada y los puntos restantes distribuyen aún más los puntuajes que recibieron a sus vecinos. Este proceso de dispersión se repite hasta la convergencia y los puntos se clasifican según su puntuación al final de las iteraciones.

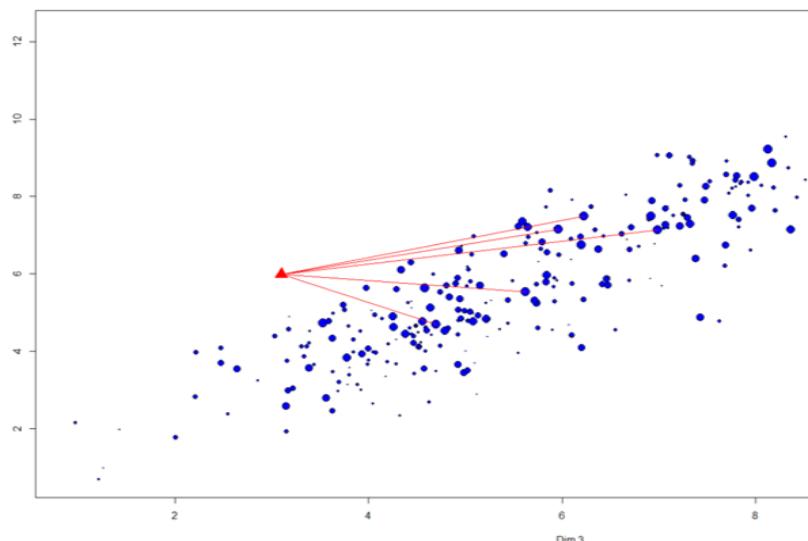
Siguiendo con el mismo ejemplo, estas son las recomendaciones que ofrece este método con el video del unboxing del Samsung Galaxy S22:

	Channel <chr>	Subscribers <dbl>	Title <chr>
	4789 Hardware Canucks	1730000	Surface Pro 8 Review - The Microsoft Tax
	6052 Marques Brownlee	15500000	Galaxy Tab S8 Ultra: A Monster Tablet!
	4840 Hardware Canucks	1730000	Apple M1 MacBook Air - Long Term User Review
	4784 Hardware Canucks	1730000	My Life with the MacBook Pro M1
	6101 Marques Brownlee	15500000	The iPhone 13 Models!
	4772 Hardware Canucks	1730000	Absolutely LOVE These Cases - Fractal TORRENT Compact & Nano!
	2850 Austin Evans	5280000	iPhone 12 Mini - I'm switching!
	3349 CNET	3410000	SamsungÃ¢Â€Â“s Galaxy S22 Plus is just the right size (full review)
	3332 CNET	3410000	Samsung Galaxy S22 review: For people who love smaller phones
	4768 Hardware Canucks	1730000	It Actually Got BETTER Ã¢Â€Â“ Zephyrus G14 2022 vs G14 2021

Veamos las puntuaciones asignadas a los videos en alguna dimensión del PCA:



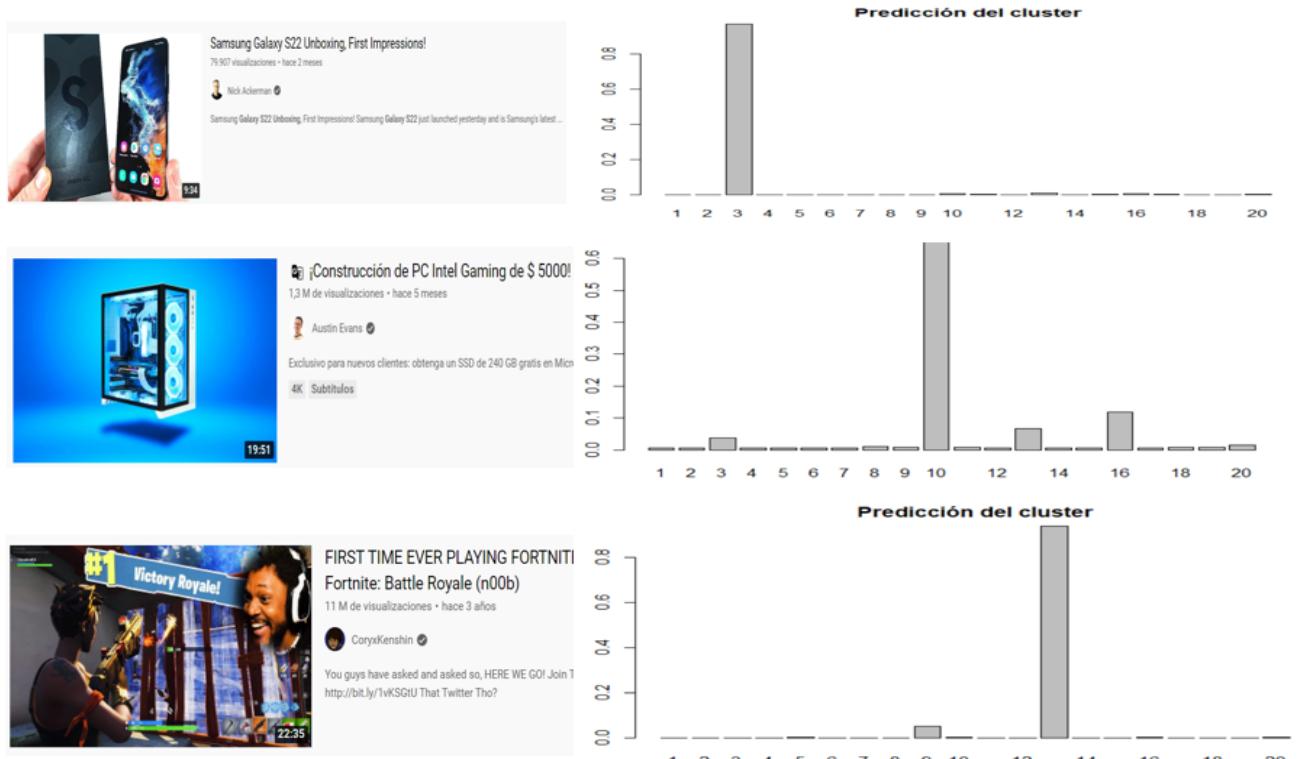
En el siguiente gráfico se representan los videos del cluster donde el tamaño hace referencia a la puntuación obtenida. De esta forma, el video está unido a los top 5 videos más recomendables a partir de él.



La potencia de este método es que podemos añadir más videos a parte de este a la consulta. En este caso vamos a considerar otros dos videos, de los cuales predecimos el cluster. A la hora de recomendar consideramos la unión de los videos de todos los clusters predecidos de los distintos videos. En este ejemplo, usaríamos a parte de los videos del cluster 3, los del 10 y el 13.

Queries

- Samsung Galaxy S22 Unboxing (cluster 3, orientado más a tecnología de móviles)
- Construcción de PC Intel Gaming (cluster 10, orientado a ordenadores)
- First Time Ever Playing Fortnite (cluster 13, videojuegos variados)



En este caso la recomendación de videos ha sido capaz de entender y combinar las consultas y ha recomendado tanto videos de ordenadores gaming, como de móviles y también de algún videojuego:

Channel <chr>	Subscribers <dbl>	Title
843 Linus Tech Tips	14300000	I LOVE <u>BUILDING COMPUTERS!!</u>
1626 LGR	1560000	LGR - Building a 486 DOS PC!
935 Linus Tech Tips	14300000	Building a \$500 AMD Gaming PC
5842 Linus Tech Tips	14400000	Our First Budget Gaming Build in Over a Year!
6133 Marques Brownlee	15500000	What I REALLY Think of the iPhone!
6084 Marques Brownlee	15500000	Reviewing EVERY iPhone Ever!
2824 Austin Evans	5280000	Building a Gaming PC...at Best Buy??
725 CDawgVA	2600000	I Spent \$2000 On Making A YÃŽÃ±oi
5983 Markiplier	32500000	Inscription - Part 12 (ENDING) juego

3.10. PREDICT CLUSTER ONLY WITH MINIATURE

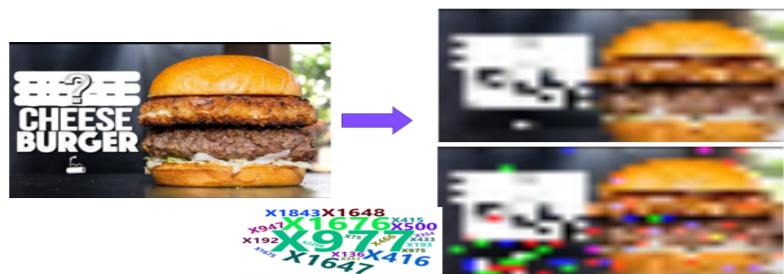
En este apartado, se propone la idea de predecir el cluster solo con la miniatura utilizando también el modelo XG-Boost, aunque en la práctica este método no ha sido utilizado ya que utilizando todas las palabras de los títulos, comentarios y transcripciones el resultado de clasificación es mucho mejor.

En el conjunto de test se obtiene una accuracy de 0.43 y esta es la matriz de confusión. Vemos por ejemplo que hay muchos falsos positivos del cluster 12 o del 19

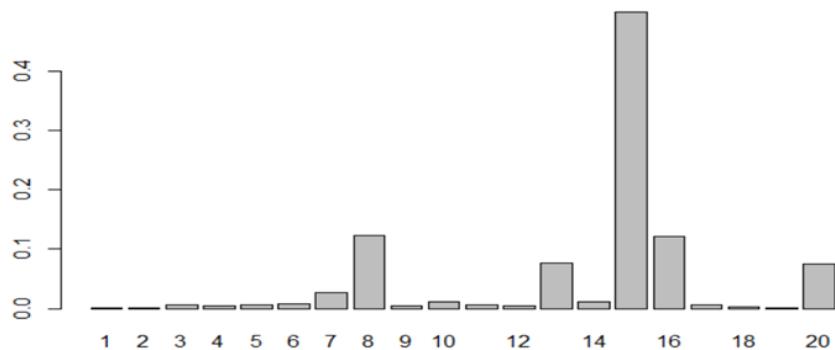
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0	16	0	0	0	0	0	0	0	1	0	0	0	5	0	1	0	0	1	0	0
1	0	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	0	2	0	0	0	0	1	1	0	0	0	25	0	0	2	8	0	1	2
3	0	0	1	5	12	0	0	0	1	4	0	0	12	0	1	4	0	0	0	6
4	0	0	0	4	26	0	1	0	0	0	0	0	3	0	0	2	0	0	0	6
5	0	0	0	0	0	26	1	1	0	1	0	1	18	0	1	3	0	0	1	14
6	0	0	0	0	0	0	60	0	0	5	0	0	34	0	4	13	0	0	0	11
7	0	0	0	0	0	0	0	0	0	0	0	0	17	0	7	4	0	0	0	4
8	0	0	1	0	0	0	0	2	0	24	1	0	0	19	0	3	8	0	1	6
9	0	1	2	0	0	0	0	6	1	0	40	0	0	41	0	14	13	0	0	11
10	0	0	0	0	0	0	0	1	0	0	0	1	0	16	0	1	1	0	0	2
11	0	0	0	0	0	0	0	0	0	0	0	0	2	14	0	2	2	0	0	2
12	1	2	1	0	0	0	6	6	0	0	11	0	1	216	0	21	17	0	1	42
13	0	0	1	0	0	0	0	1	0	0	2	0	0	13	5	1	5	0	0	3
14	0	0	0	0	0	0	0	8	0	0	3	0	0	39	0	97	11	0	0	8
15	1	0	0	0	0	0	2	12	0	1	8	1	0	59	0	16	18	2	2	22
16	0	1	0	0	0	0	3	0	0	0	3	0	0	6	0	2	3	20	0	5
17	0	0	0	0	0	0	1	8	0	0	0	1	0	6	1	5	4	0	5	8
18	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	8	0	0	0	37
19	0	0	1	1	1	10	8	0	0	7	1	1	49	1	3	20	0	0	5	0

Nota: Aquí los clusters comienzan en el 0 hasta el 19

Veamos un ejemplo de predicción del cluster a través de la miniatura de un nuevo video que no está en nuestra base de datos. La imagen se transforma a 28*28*3 píxeles, además se representan los píxeles que más ayudan a la predicción según la ganancia de información.



Predicción en base a la miniatura



Vemos que a pesar de que el modelo puede dudar con algún cluster predice claramente el 15. Los videos de este cluster son justamente de cocina:

	Channel <chr>	Subscribers <dbl>	Title <chr>
2	Mythical Kitchen	1900000	\$420 Pizza Hut Stuffed Crust Pizza Fancy Fast Food Mythical Kitchen
11	Epicurious	4010000	How To Slice Every Fruit Method Mastery Epicurious
24	FoodTribe	379000	James May shocks electricians with vegan hot dog prank
30	Bon AppÃ©tit	5930000	Every Way to Cook a Potato (63 Methods) Bon AppÃ©tit
67	Epicurious	4010000	4 Levels of Onion Rings: Amateur to Food Scientist Epicurious
71	Epicurious	4010000	How To Fillet Every Fish Method Mastery Epicurious
78	Epicurious	4010000	\$98 vs \$9 Burger: Pro Chef & Home Cook Swap Ingredients Epicurious
108	About To Eat	870000	I Only Made Indian Food For 24 Hours
160	Gordon Ramsay	18700000	Gordon Ramsay Demonstrates Key Cooking Skills
195	Munchies	4590000	How to Cook From Tokyo's Vending Machines

3.11. COMPLEX NETWORK CORRELATION WORDS

Para entender la estructura de correlación entre las palabras se propone representar el subgrafo formado por las 1000 palabras que más ayudan a predecir los clusters y de cada una de ellas buscar con qué otras palabras están correlacionadas ($p \geq 0.5$).

$$G = (V, E) \text{ GND}$$

$$V = \{\text{Palabras de los títulos, comentarios y transcripciones}\}$$

$$E = \{(x, y) \in V / x \text{ e } y \text{ están correlacionados con } p \geq 0.5\}$$

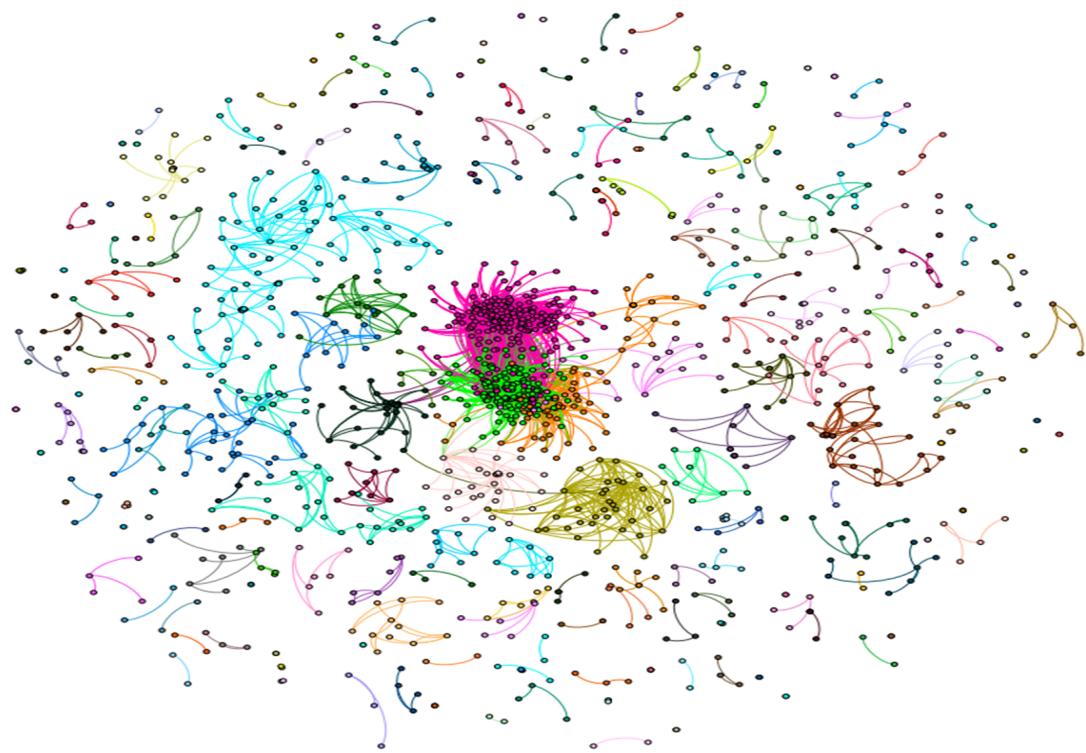
$$p(x, y) = \text{"Correlación entre las palabras } x \text{ e } y"$$

Proceso y consideraciones

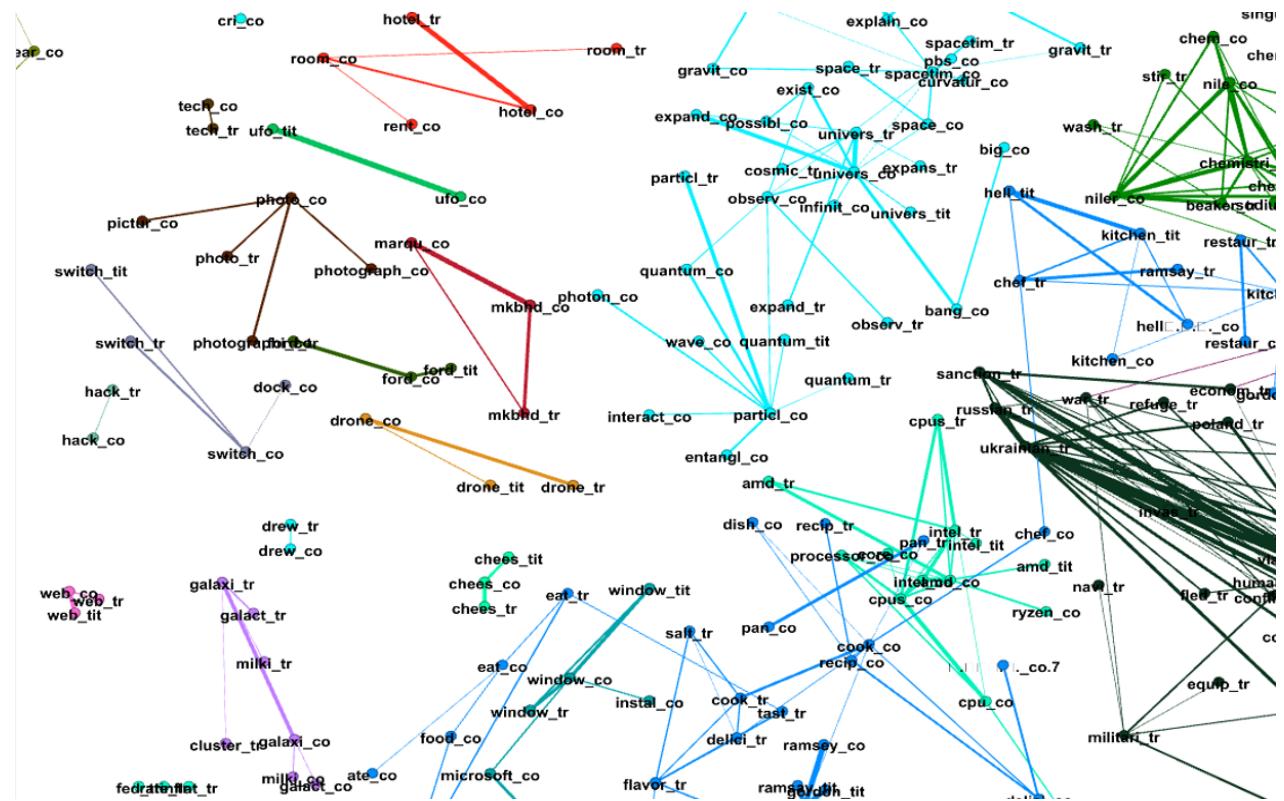
Al no estar representada toda la red:

- Se buscan 1000 palabras potenciales (las que más ayudan a predecir los clusters) y las correlaciones de estas con las demás donde $p \geq 0.5$
- Habrá palabras que no aparezcan por no estar relacionadas (de esas 1000) con el resto, pero tampoco estarán aquellas que no hayan sido consideradas, y de las consideradas ninguna se relaciona con ellas.
- Las palabras que aparezcan y no sean de las 1000 consideradas, pueden no estar bien representadas al no haber considerado sus correlaciones directas con el resto.
- Lo ideal es generar la red completa, lo cual puede ser computacionalmente costoso, pero sería una representación REAL de la estructura de las correlaciones.

Resultado de la red de correlación



Explorando una parte de la red



Observamos grupos de palabras correlacionadas, cada grupo tenderá a pertenecer a un cluster. Por ejemplo:

- Amd está relacionada con Intel, Intel con cpus...
- Salt está relacionada con cook, cook con delicious, delicious con flavor...
- Space está relacionado con univers, univers con observation, observation con particle, particle con quantum...
- Ukrainian con russian y refugees, russian con war y con sanction...

3.12. ASSOCIATION RULES WITH CLUSTER

Para entender la influencia de las palabras y sentimientos en los clusters se ha propuesto crear reglas de asociación. Para ello se han discretizado las palabras en intervalos de frecuencias de aparición {[0,1), [1,3), [3,10), [10,100]}, también se han discretizado en intervalos las frecuencias de sentimientos (cada uno de ellos en tres bloques, que podemos considerar como bajo, medio o alto porcentaje de ese tipo de emoción). Mostramos algún ejemplo asociado al cluster 3 (móviles y tecnología), cluster 10 (ordenadores, consolas...) y del 15 (comida). Para realizar las reglas se ha establecido un soporte de 0.002, es decir, los items de una regla deben aparecer en al menos $0.002 * \text{num_videos}$, en nuestro caso $0.002 * 8266 = 17$ aproximadamente y la confianza mínima de estas es del 70%

Cluster 3

lhs <chr>	rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
{ryzen_co=[3,10), perform_co=[10,100]}	=> {cluster=3}	0.002298572	1.0000000	0.002298572	29.20848	19
{cpu_tr=[3,10), perform_co=[10,100]}	=> {cluster=3}	0.003024437	1.0000000	0.003024437	29.20848	25
{bought_co=[3,10), perform_co=[10,100]}	=> {cluster=3}	0.002903460	1.0000000	0.002903460	29.20848	24
{game_tr=[3,10), perform_co=[10,100]}	=> {cluster=3}	0.003024437	1.0000000	0.003024437	29.20848	25
{univers_co=[0,1), galaxi_tr=[10,100]}	=> {cluster=3}	0.002056617	1.0000000	0.002056617	29.20848	17
{laptop_co=[10,100], ryzen_co=[3,10)}	=> {cluster=3}	0.003145415	0.9629630	0.003266392	28.12668	26
{camera_tr=[10,100], display_co=[3,10)}	=> {cluster=3}	0.003024437	0.9615385	0.003145415	28.08508	25
{cpu_co=[3,10), perform_co=[10,100]}	=> {cluster=3}	0.002903460	0.9600000	0.003024437	28.04014	24
{laptop_co=[10,100], perform_co=[10,100]}	=> {cluster=3}	0.002782482	0.9583333	0.002903460	27.99146	23
{gpu_co=[10,100], littl_tr=[3,10)}	=> {cluster=3}	0.002661505	0.9565217	0.002782482	27.93855	22

Si aparece la palabra ryzen entre 3 y 9 veces, y la palabra perform entre 10 y 100 veces en los comentarios entonces ese video pertenecerá al cluster 3 con un 100% de confianza. Una regla también interesante es la quinta, la cual dice que si no aparece la palabra universo pero si galaxi entre 10 y 100 veces la confianza de pertenecer al cluster 3 también es del 100%. El hecho de que no aparezca la palabra universo está forzando a que la palabra galaxi esté haciendo referencia al Samsung Galaxy y no a la palabra galaxia, ya que si no este video pertenecería al cluster de ciencia.

Cluster 10

{lgr_co=[10,100]}	=>	{cluster=10}	0.004597145	1	0.004597145	13.16242	38
{razer_co=[10,100], temp_co=[0,1]}	=>	{cluster=10}	0.002177595	1	0.002177595	13.16242	18
{car_co=[0,1], hardwar_co=[10,100]}	=>	{cluster=10}	0.002056617	1	0.002056617	13.16242	17
{razer_co=[3,10], cpu_co=[0,1]}	=>	{cluster=10}	0.002056617	1	0.002056617	13.16242	17
{xbox_co=[10,100], didnt_tr=[1,3]}	=>	{cluster=10}	0.002056617	1	0.002056617	13.16242	17
{xbox_co=[10,100], negative.c=[0.122,0.149]}	=>	{cluster=10}	0.002298572	1	0.002298572	13.16242	19
{xbox_co=[10,100], anger.c=[0.0521,0.069]}	=>	{cluster=10}	0.002177595	1	0.002177595	13.16242	18
{xbox_co=[10,100], positive.c=[0.216,0.252]}	=>	{cluster=10}	0.002419550	1	0.002419550	13.16242	20
{xbox_co=[10,100], anticipation.tr=[0.0968,0.115]}	=>	{cluster=10}	0.002419550	1	0.002419550	13.16242	20
{xbox_co=[10,100], use_tr=[3,10]}	=>	{cluster=10}	0.002056617	1	0.002056617	13.16242	17

Por ejemplo, la regla 3 dice que si no aparece la palabra coche pero sí hardware entre 10 y 100 veces, el video pertenece al cluster 10 con 100% de probabilidad. El lift es de 13, lo que quiere decir que la probabilidad de que el video pertenezca al cluster 10 es 13 veces más probable si sabemos a priori la información del antecedente.

Cluster 15

{game_co=[0,1], delici_tr=[10,100]}	=>	{cluster=15}	0.002661505	1	0.002661505	9.667836	22
{pleas_tr=[0,1], delici_tr=[10,100]}	=>	{cluster=15}	0.002056617	1	0.002056617	9.667836	17
{delici_tr=[10,100], howev_tr=[0,1]}	=>	{cluster=15}	0.002419550	1	0.002419550	9.667836	20
{recip_co=[10,100], delici_co=[10,100]}	=>	{cluster=15}	0.002540527	1	0.002540527	9.667836	21
{delici_co=[10,100], peopl_tr=[0,1]}	=>	{cluster=15}	0.003145415	1	0.003145415	9.667836	26
{delici_co=[10,100], sadness.c=[0,0.0556]}	=>	{cluster=15}	0.003992257	1	0.003992257	9.667836	33
{delici_co=[10,100], fear.c=[0,0.0637]}	=>	{cluster=15}	0.004113235	1	0.004113235	9.667836	34
{delici_co=[10,100], disgust.c=[0,0.0372]}	=>	{cluster=15}	0.003266392	1	0.003266392	9.667836	27
{delici_co=[10,100], anticipation.c=[0.11,1]}	=>	{cluster=15}	0.003629325	1	0.003629325	9.667836	30
{delici_co=[10,100], positive.c=[0.252,0.75]}	=>	{cluster=15}	0.003629325	1	0.003629325	9.667836	30

Aquí, vamos a ver la importancia de los sentimientos. Por ejemplo, la regla 6 nos dice que si la palabra delicioso aparece entre 10 y 100 veces, y además el porcentaje de tristeza en los comentarios es bajo (primer intervalo) entonces ese video pertenece al cluster 15 con un 100% de confianza.