ASIGNATURA DE ANÁLISIS MULTIVARIANTE

PRÁCTICA DE MINERÍA DE TEXTO: ANÁLISIS DESCRIPTIVO

13 de diciembre de 2020

Emilio Miguel Gómez Cofrades, Jaime Estebala Márquez

Universidad Complutense de Madrid

Índice general

0.1. INTRODUCCIÓN AL PROBLEMA. PREPROCESADO DE DATOS

Durante esta sección se formulará el problema a resolver durante la práctica y se introducirá brevemente aquello reseñable correspondiente al procesado de la información previo a la aplicación de los modelos.

Se trabajará con el corpus número 3, de 3000 documentos, con el objetivo de determinar de qué temas se habla en los textos a analizar. El objetivo es aplicar las técnicas descriptivas de análisis multivariante aprendidas durante el curso, exponiendo las decisiones de diseño tomadas a lo largo de la práctica en esta memoria. Es posible acceder al código de la práctica mediante el siguiente repositorio de código de la plataforma *GitHub*, que permanecerá público y disponible para todo tipo de usuarios:

`https://github.com/MiguelGomezC/AMUL-PRACTICA`

En cuanto a preprocesado, se trabaja en el fichero *TFIDF.sas*, del directorio principal del repositorio. La cuestión principal a determinar es con qué cantidad de palabras nos quedaremos de todo el vocabulario del corpus para aplicar los modelos descriptivos. Se filtran aquellas palabras que tengan una suma total de *TFIDF* en todo el corpus menor que $2'165$, el valor correspondiente al percentil 95. Esto resulta en una matriz de términos-documentos de 1205 variables, incluida la variable identificadora indicativa del documento al que corresponde cada observación.

A partir de aquí, se hace un análisis inicial en SAS Enterprise Miner para familiarizarse con los posibles temas de los que trata el texto, haciendo un análisis de componentes principales. Lo único a añadir es que se tuvo que cuidar el rol asignado a las variables de la matriz de términos-documentos, asegurándose de que todas ellas fueran variables de entrada o *input*, dado que hubo casos en los que se les asignó uno distinto por defecto, como es el caso de la varable *cost*, que se le puso rol de coste, o *target*, de respuesta.

0.2. ANÁLISIS DE COMPONENTES PRINCIPALES

Los objetivos que se tienen al hacer un análisis inicial de componentes principales es identificar qué palabras pueden tener más peso general dentro del corpus (para lo que se buscará generar una

primera componente de volumen) e intentar identificar si el resto de componentes corresponden a algún tema en concreto, estudiando cómo correlan las variables con cada componente.

0.2.1. Componente de volumen

Dado que la métrica TFIDF es no negativa, se ejecuta el nodo de componentes principales basado en la matriz de varianzas-covarianzas (con fuente de autovalor no corregida). De este modo, está garantizado que la primera componente sea de volumen. Esta componente nos dará información de qué palabras están más presentes a lo largo del corpus, las que más variabilidad explican, aunque no necesariamente las que discriminan más.

Se muestra en la figura 1 un listado con las variables que más correlan con la componente de volumen. Nótese que muchas de ellas son raíces de palabras relacionadas con las tecnologías e internet, como *phone*, *subscrib*, *program*, *server*, *widget*, *system*, *displai*, *comput*... Este tipo de palabras no orientan mucho sobre de qué temas específicos pueden tratar los textos, ya que hablar de internet está a la orden del día y es algo que puede estar presente en multitud de contextos. Sin embargo, es posible que los distintos temas que contiene el corpus tengan en común esta componente tecnológica, es decir, puede que todos ellos tengan algo que ver con las tecnologías e internet, aunque eso de por sí no sea nada demasiado revelador.

Cabe destacar la presencia de la palabra *utexa*. A diferencia de *Mike*, que puede hacer referencia a multitud de individuos con el mismo nombre, el significado de *utexa* es muy concreto. Puede hacer referencia a una empresa de textiles americana (*United Textiles of America*), o bien puede ser un

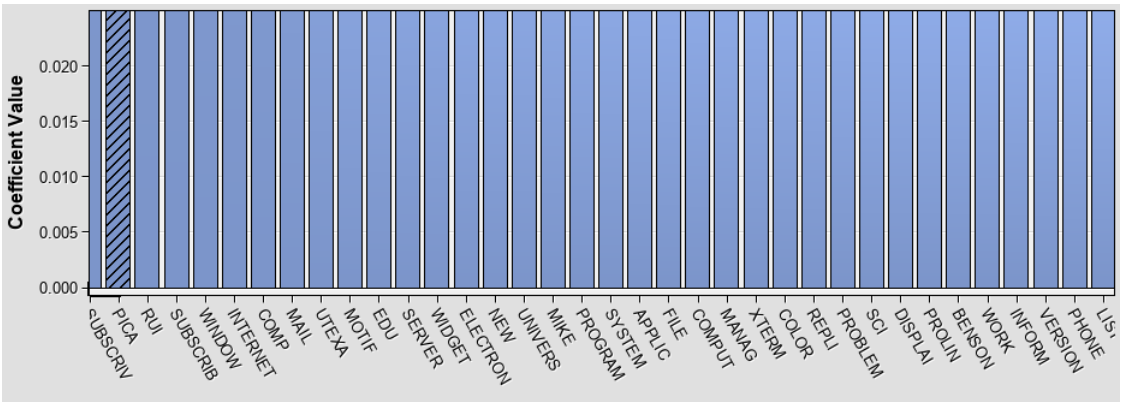

Figura 1: Variables de mayor coeficiente en la componente de volumen.

acortamiento del acrónimo de la universidad de Texas (*UTexas*).

0.2.2. Componentes específicas

En cuanto al resto de componentes que se generan, se debe analizar si es posible asignar temas concretos a cada una de ellas.

La séptima componente que más variabilidad explica se muestra en la figura 2 en forma de un listado con las palabras con las que más correla. En ella, se muestran palabras derivadas de Israel y Palestina, y otros países de oriente medio, y muchas otras altamente relacionadas con conflictos armados; *attack*, *peace*, *kill*, además de *arab* y *jew*, haciendo referencia a las religiones árabe y judía. Parece claro que a esta componente se le podría asignar el tema de conflictos armados y Oriente Medio. Por otro lado, nótese que algunas de las variables que correlan negativamente con esta componente están relacionadas con la tecnología y la ciencia. Esto no solo responde a la lógica, ya que parecen temas que no tienen mucho en común (sobre todo la ciencia con los conflictos armados), sino que indica que un tema independiente a los conflictos armados presente en el corpus podría ser la ciencia.

Esta idea se refuerza aún más si se echa un vistazo a la información de la componente 8, que se muestra en la figura 3. Como se puede ver, correla positivamente con una gran cantidad de palabras relacionadas con tecnología y ciencia, y negativamente con palabras clave de los conflictos armados, como *kill*, *Israel*, y *Palestina*. Podría tratarse de la componente de la tecnología.

Al estudiar las componentes, se ve que existen algunas de interpretación muy parecida a las que ya se han visto, como la 6, que es similar a la 7. Sin embargo, una componente que se interpreta de

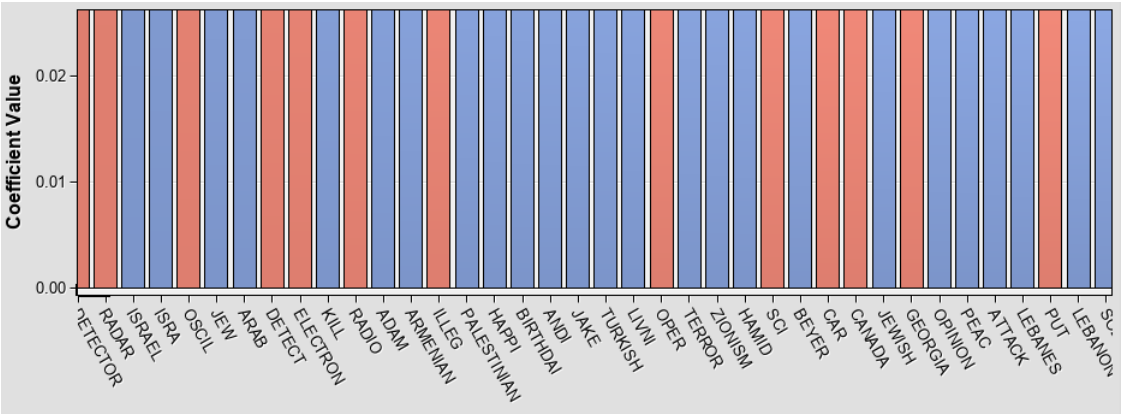

Figura 2: Variables de mayor coeficiente en la componente 7.

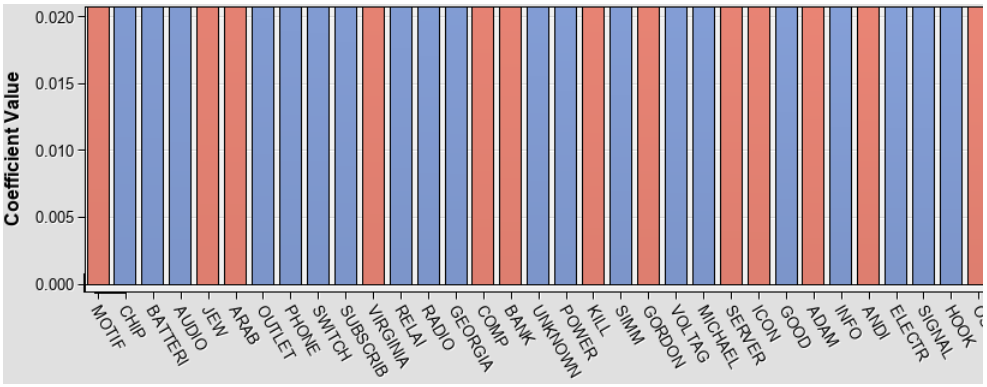

Figura 3: Variables de mayor coeficiente en la componente 8.

forma distinta a ellas y de manera certera, es la componente 4, que se muestra en la figura 4.

Palabras relacionadas con el tema religioso correlan positivamente con esta componente, como *intellect*, *surrend*, *skeptic* o *science* (en efecto, en el paradigma religioso está muy presente el debate de la compatibilidad con la ciencia). Por otro lado, también correlan positivamente palabras como *shame*, *chastiti* (del inglés, castidad), *Pittsburgh* y *Pitt*. Pittsburgh es una ciudad en Pennsylvania, Estados Unidos, en la que a mediados del siglo XX hubo escándalos relacionados con la Iglesia, de sacerdotes que acosaron y violaron a más de 1000 niños [1]. Se podría asignar a esta componente, por tanto, el tema de escándalos de la Iglesia, o algo más generalizado, el de religión cristiana. Además, Pittsburgh tiene una universidad muy conocida en Estados Unidos, que frecuentemente compite contra la universidad de Texas en campeonatos interuniversitarios. Sin embargo, el hecho de que *utexa* correle negativamente con esta componente, es un argumento a favor de que el significado de *utexa* puede estar más ligado a la empresa de textiles que a la universidad.

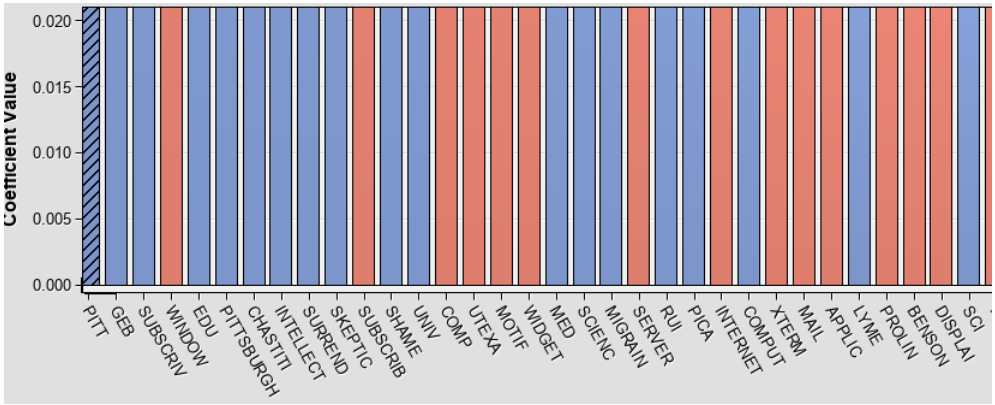

Figura 4: Variables de mayor coeficiente en la componente 4.

Por tanto, las conclusiones que se sacan a raíz de este primer análisis de componentes principales, es que el corpus consta de dos temas claramente definidos e independientes: uno de conflictos en Oriente Medio y otro de relacionado con la tecnología y la ciencia, pudiendo haber un tercero relativo a la religión cristiana o escándalos religiosos en Pittsburgh. Además, se sospecha que pueda haber más temas subyacentes que no se caracterizan de forma tan clara como los anteriormente mencionados, que puedan tener que ver con tecnología, internet y ciencia.

0.3. ANÁLISIS CLUSTERING

Tras el análisis inicial de componentes principales, se procede a ejecutar un análisis clustering usando Enterprise Miner, esperando obtener resultados parecidos a lo ya intuido a través de las componentes principales. Por desgracia, como veremos, esto será insuficiente, por lo que posteriormente se procede a realizar un análisis cluster con código de SAS en el que se diseñan los centriodes de partida a través de un algoritmo jerárquico y a través de la intuición y de lo aprendido sobre el corpus a lo largo de la práctica. Además, se probará a hacer un análisis clustering tras reducir la dimensionalidad mediante componentes principales y cambiando el percentil representativo de las variables con las que nos quedamos.

0.3.1. Clustering a través del Miner

El primer clustering que se lleva a cabo en Miner (en el diagrama *ANÁLSIS CLUSTER*, en el proyecto de Miner del repositorio) se ejecuta sobre la matriz de términos documentos, sin realizar ninguna transformación de la dimensionalidad, y arroja resultados positivos; pero no definitivos, puesto que aparecen varios clúster residuales que contienen palabras que deberían ser reclasificadas.

Se observa claramente la presencia de grupos de palabras relacionadas con la tecnología en el segmento 3 (ver figura 5) pues aparecen palabras como *window*, *comp* (de *computer*), *server*, *sci* (de *science* o derivadas), *display*... con gran poder discriminante.

Asimismo, surge una nueva temática que en las componentes no fue posible identificar: la medicina. Como representante de ella, destaca el cluster número 5 (ver figura 6) , constituido en su mayor parte por la palabras del campo médico como *med*, *sci*, *patient*, *medic*, *food*, *doctor*,...

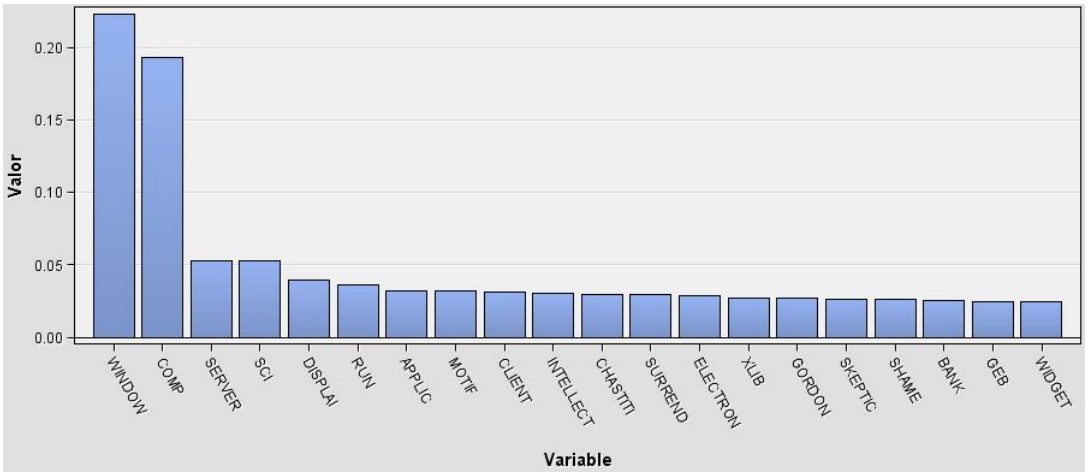

Figura 5: Variables de mayor poder discriminante en el cluster 3.

Además, como se había notado en el análisis de componentes principales, destaca la temática relativa a Israel y Palestina, como demuestra la aparición del cluster número 1 (ver figura 7), que contiene palabras como *arab*, *Israel*, *jew*, *jewish*, *policy*, *peace*...

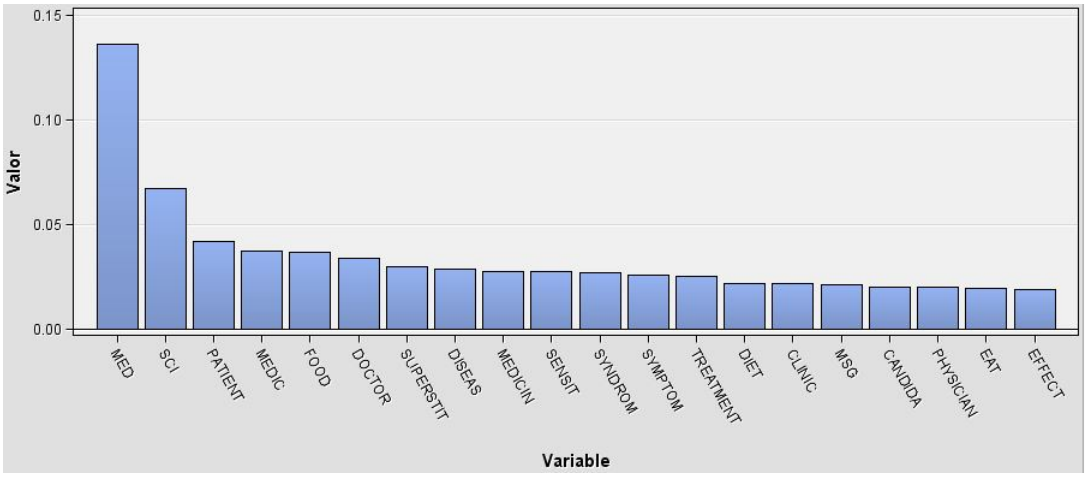

Figura 6: Variables de mayor poder discriminante en el cluster 5.

Los clusters de mayor tamaño del análisis inicial que se han comentado sí parecen contener palabras de la misma temática, aunque permanecen aún algunos de menor tamaño que, *a priori*, no aportan información muy relevante. Sin embargo, uno de ellos sí resultará de importancia posteriormente: el cluster número 8 (ver figura 8), de temática relacionada con un conflicto turco-armenio.

De este modo, se toman los centroides de los tres o cuatro segmentos que mayor cantidad de palabras agrupan y que lo hacen con mayor acierto para emplearlos como semilla de un nuevo análisis

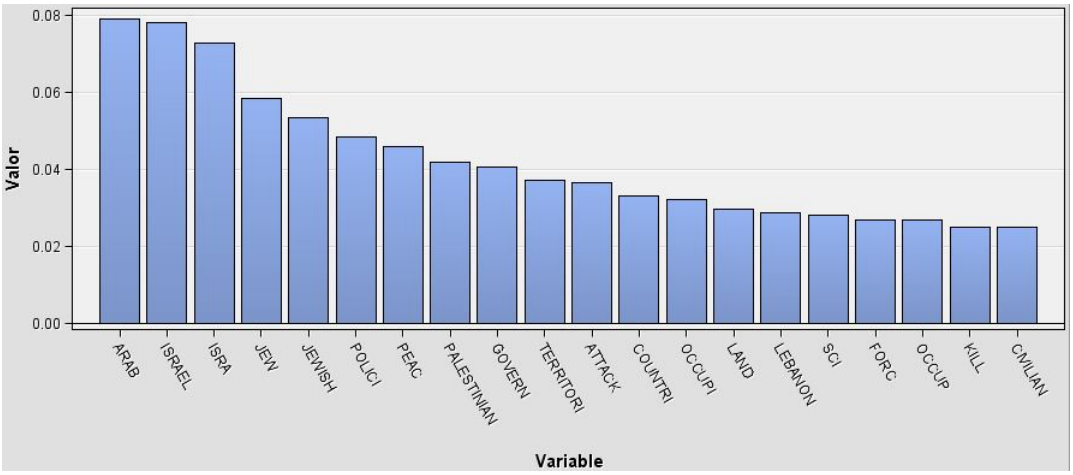

Figura 7: Variables de mayor poder discriminante en el cluster 1.

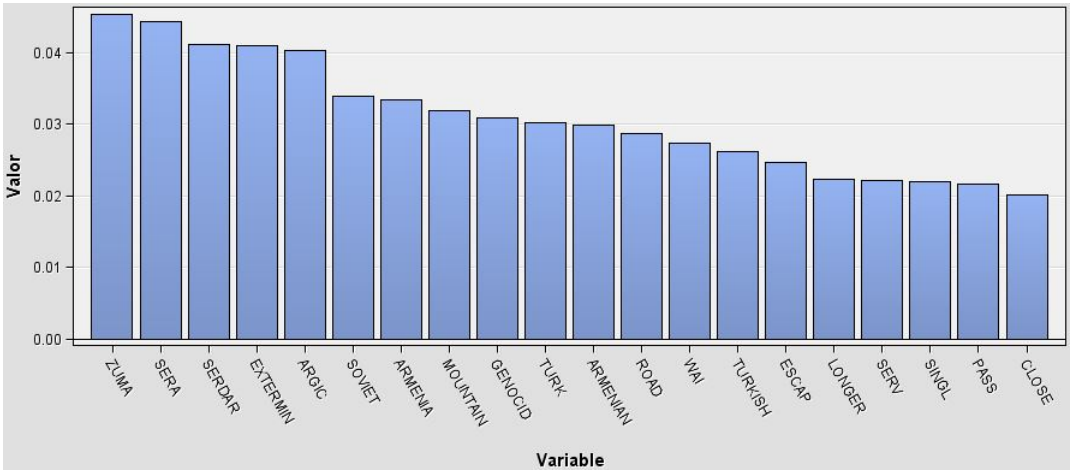

Figura 8: Variables de mayor poder discriminante en el cluster 8.

de grupos.

0.3.2. Clustering con centroides de partida personalizados

En un primer momento, la idea que surgió fue seleccionar aleatoriamente una plabra de cada uno de los grupos descritos en 0.3.1 como centroide de un segundo análisis cluster, aplicando técnicas de clustering de optimización (*k-means*). Se probó tambien a tomar como centroides de inicio los determinados a través de la aplicación de un clustering jerárquico (*proc cluster*).Todo esto se hizo con la motivación de orientar las agrupaciones mediante la intuición y el sentido que le dimos a los datos durante el estudio.

La puesta en práctica demostró, sin embargo, que no era eficiente ninguno de los métodos. De hecho, solamente se consiguió que los términos se agrupasen principalmente en un cluster y se generasen dos o tres residuales (claramente, resultado insatisfactorio, como se ve en la figura 9).

Resumen de conglomerados						
Cluster	Frecuencia	Desviación estándar RMS	Distancia máxima de la semilla a la observación	Radio sobrepasado	Conglomerado más próximo	Distancia entre Centroides del cluster
1	2984	0.0131	2.3128		3	0.3262
2	6	0.0131	0.5138		1	0.4197
3	10	0.0112	0.4700		1	0.3262

Figura 9: Clusters generados mediante centroides iniciales del análisis jerárquico.

Llegados a este punto, como se habían llevado a cabo diferentes pruebas y, por el momento, no se obtenían resultados todo lo satisfactorios que se deseaba, se llegó a la conclusión de que el motivo podía ser el disponer de datos con demasiada variabilidad, haciendo más probable que se mezclasen las observaciones de cada temática en los grupos generados. Por ello, se decidió disminuirla, para lo cual fue preciso repetir los pasos que se habían dado, seleccionando un percentil de corte mayor al ya establecido: se pasó del 95 % al 99 %. Al conseguir una menor variabilidad se esperaba poder obtener agrupamientos de términos más definidos.

0.3.3. Análisis con un cambio en el percentil de corte de los datos

Se decide, por tanto, hacer un análisis descriptivo filtrando las variables que acumulen una suma de *TF-IDF* menor a 6′0623, el valor correspondiente al percentil 99, con la esperanza de poder así obtener agrupaciones más puras al disminuir la cantidad de variabilidad analizada del corpus orignial.

Se ejecuta primero un análisis de componentes principales con la misma configuración que la explicada en 0.2.1, y los resultados refuerzan la idea de que existen de forma clara temas en el corpus relacionados con la tecnología. Para ilustrar esto, se muestra en la figura 10 las palabras que más correlan con la tercera componente, donde se distinguen multitud de elementos tecnológicos que correlan positivamente, como *server*, *detector*, *widget*, *window*, *radar*, *program*, *displai*, *electron*, *version*,...

Además, componentes como la 6 tienen un alto nivel de relación con el tema de conflictos de

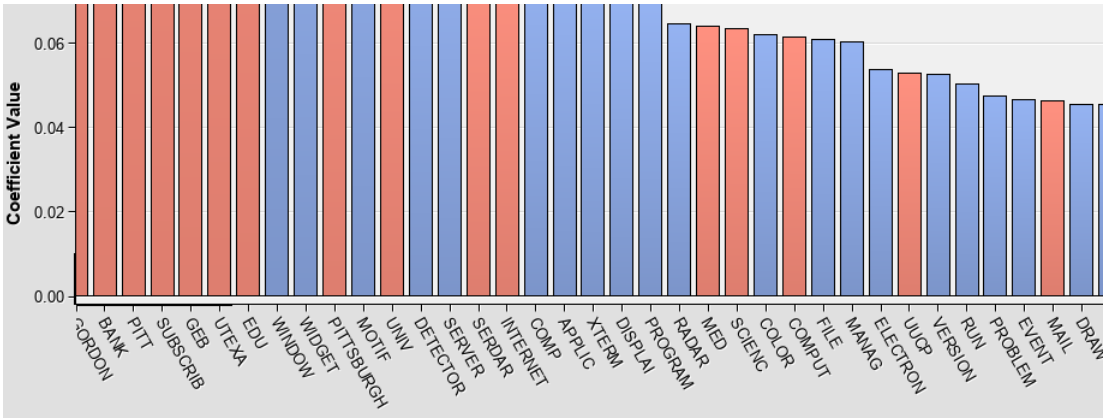

Figura 10: Variables de mayor coeficiente en la componente 3 usando el percentil 99.

oriente Medio inferido anteriormente, y la componente 5 correla positivamente con variables como *Pittsburgh*, *Med* y *Univ*, y otras relacionadas con los estudios universitarios y con la medicina, por lo que se deduce que el concepto de Pittsburgh en el corpus puede estar más frecuentemente asociado a su universidad que al tema de los escándalos de la Iglesia mencionado en 0.2.1.

Se procede entonces a hacer un análisis cluster directamente sobre la nueva matriz de términos-documentos (nodo *resultados finales* en el diagrama *ANÁLISIS CLUSTERING*), de donde se obtienen resultados verdaderamente reveladores. Tanto es así, que a pesar de haber realizado también un análisis clustering reduciendo la dimensionalidad con componentes principales para este percentil 99, se elige este método directo como referencia para reproducir los resultados finales de la práctica. Esto se debe a que en este caso no se gana tanto en cuanto a homogeneidad de cardinal de cada cluster como para renunciar a la fácil interpretabilidad que ofrece el hacer el análisis cluster de forma directa.

La distribución de las observaciones después de este análisis se acumula en 9 clusters, con uno residual de 19 registros. Se pueden identificar 4 temas claros a partir de los clusters de mayor tamaño: tecnología, medicina, conflicto israel-palestina y conflicto turco-armenio. Resulta especialmente interesante mostrar las variables que más valor discriminativo tienen con respecto al cluster bautizado con este último tema, ya que figuran las palabras *Argic* y *Serdar*. Argic Serdar fue un alias utilizado en un incidente digital cuyo objetivo era negar el Genocidio Armenio. Hay que señalar que, al estar centrado en el cero y ser el *TF-IDF* una medida no negativa, uno se puede esperar que una variable es buena discriminando si toma valores alejados de cero, es decir, tiene una gran presencia en el grupo formado.

	NOMBRE DE LA VARIABLE ANTERIOR	ARMENIA	NOMBRE DE LA VARIABLE ANTERIOR	ISRAELPALESTINA	NOMBRE DE LA VARIABLE ANTERIOR	MEDICINA	NOMBRE DE LA VARIABLE ANTERIOR	TECNOLOGIA
2	ARMENIAN	12.865156055	ISRAEL	18.604290723	MED	16.152319566	WINDOW	37.130572721
3	TURKISH	7.9369807342	ISRA	15.795426547	GORDON	12.020987877	ELECTRON	23.720961143
4	UTEXA	6.7514229901	ARAB	12.000698159	BANK	11.597282075	COMP	22.477106951
5	SERDAR	6.6269318098	JEW	11.298522617	PITT	11.106601286	UNIVERS	14.978796247
6	ARMENIA	6.0135570858	KILL	7.3397499913	SCI	9.2667346655	MOTIF	14.59326755
7	ARGIC	5.8024219898	PALESTINIAN	6.7721058749	GEB	8.9090263265	PROGRAM	14.338715348
8	UUCP	3.4149749394	ADAM	6.1790986106	EDU	8.5683037081	SERVER	14.317446184
9	GOVERN	1.7409696996	POLICI	5.2824493756	DOCTOR	8.4881264363	WIDGET	14.041886031
10	DAVID	1.7023807385	JEWISH	5.2694973176	MEDIC	7.518504706	SYSTEM	13.501649137

Figura 11: Tabla con las palabras que más suma de *TF-IDF* tiene en cada agrupación de la clasificación final

Los clusters bautizados con los temas mencionados en el anterior párrafo tienen 1239 (Tecnología), 510 (Medicina), 324 (Israel-Palestina) y 142 (Turquía-Armenia) observaciones, pero se puede asignar alguno de los temas ya mencionados al resto de clusters de manera bastante limpia, dado que pueden ser caracterizados fácilmente en función del valor discrimintativo de las variables más importantes en cada uno de ellos. De esta manera es como se conforma la clasificación final, asignando los clusters 8, 3, 6, 4 y 1 a Tecnología, 9 y 7 a Medicina, 5 y 10 a conflicto Israel-Palestina y 2 a conflicto Armenia-Turquía.

Se muestra en la figura 11 las diez palabras que mayor suma de *TF-IDF* tienen a lo largo de cada grupo creado en la clasificación final.

Bibliografía

[1] Marta Torres. Los templos del horror en pensilvania: "los sacerdotes violaban a niños y niñas pequeños". *El mundo*, 2018. https://www.elmundo.es/internacional/2018/08/15/5b746dac46163f6ea68b4637.html.