MÁSTER EN SISTEMAS DE
INFORMACIÓN DIGITAL

UNIVERSIDAD DE SALAMANCA
FACULTAD DE TRADUCCIÓN Y DOCUMENTACIÓN
MÁSTER EN SISTEMAS DE INFORMACIÓN DIGITAL

Trabajo Fin de Máster

Organización Automática de Documentos

Técnicas K-Means

Autor: Sergio Pradales Gallego

Tutor: Carlos G. Figuerola

Salamanca,
2017

VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

MÁSTER EN SISTEMAS DE INFORMACIÓN DIGITAL

UNIVERSIDAD DE SALAMANCA
FACULTAD DE TRADUCCIÓN Y DOCUMENTACIÓN
MÁSTER EN SISTEMAS DE INFORMACIÓN DIGITAL

Trabajo Fin de Máster

Organización Automática de Documentos

Técnicas K-Means

Autor: Sergio Pradales Gallego

Tutor: Carlos G. Figuerola

Salamanca,
2017

ASIENTO CATALOGRÁFICO ADAPTADO AL FORMATO DEL REPOSITORIO INSTITUCIONAL GREDOS

- **Título:**

Organización Automática de Documentos. Técnicas K-Means

- **Autor:**

Pradales Gallego, Sergio

- **Director:**

G. Figuerola, Carlos

- **Palabras clave:**

[ES] clasificación automática - prensa digital española - noticias ciencia - K-Means - Scikit-learn - evolución

[EN] automatic clustering - spanish digital press - science news - K-Means - Scikit-learn - evolution

- **Clasificación UNESCO**

5701.02, 5701.04, 1203.04, 1203.23

- **Fecha:**

2017-07-14

- **Resumen:**

[ES] En los últimos años hemos experimentado el crecimiento progresivo de los documentos, sobre todo con la aparición de internet, la globalización y las nuevas tecnologías de la información y comunicación, que han dado lugar a una gran cantidad de documentos en formato digital. Toda esta información era necesario organizarla, y para ello existen diversos sistemas de filtrado de documentos automáticos, siendo uno de los más utilizados el *clustering,* que a través en este caso del algoritmo de *K-Means* y el software de *Scikit-learn*, permite la recuperación y clasificación de documentos afines. Con esta idea se pretendió comprobar la evolución de un conjunto de noticias de ciencia y tecnología extraídas de la prensa digital española, y la utilidad de este sistema de clasificación. Los resultados reflejan que es un buen sistema aunque tiene algunas carencias que dependen sobre todo de factores humanos.

[EN] In recent years we have experienced the progressive growth of documents, especially with the emergence of the internet, globalization and new information and communication technologies, which have resulted in a large number of documents in digital format. All this information was necessary to organize it, and for this purpose there are several automatic document filtering systems, one of the most used being *clustering*, which in this case, the *K-Means* algorithm and the *Scikit-learn* software, Retrieval and classification of related documents. With this idea it was tried to verify the evolution of a set of science and technology news extracted from the Spanish digital press, and the usefulness of this classification system. The results reflect that it is a good system although it has some deficiencies that depend mainly on human factors.

- **Descripción:**

Trabajo de Fin de Máster en Sistemas de Información Digital, 2017.

SUMARIO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

0. Introducción

Durante las últimas décadas del siglo XX se produjo un crecimiento exponencial de los documentos, sobre todo por la aparición de internet, la globalización y la consiguiente irrupción de las nuevas tecnologías de la información y comunicación, incrementándose de manera notoria la cantidad de datos en formato digital.

Toda esta información era necesario organizarla y filtrarla, con el único propósito de permitir y garantizar al usuario el acceso a la información que precisa disponer en cada momento (Bravo, 2008). Este proceso se realizaba de forma manual, un trabajo que requería de mucho tiempo y esfuerzo. Sin embargo, a la par que crecía la información digital, lo hacían las estructuras tecnológicas dedicadas a la gestión, análisis y representación de datos (Vallejo, 2016).

En este sentido, surgieron los diferentes sistemas de clasificación de los documentos, encontrándonos con sistemas de clasificación supervisada, es decir, que requieren en alguna de sus fases de la intervención humana; y por otro lado, sistemas de clasificación no supervisada, que organizan de manera automática los documentos sin necesidad de la participación del hombre, siendo este último caso el que nos ocupa para este Trabajo de Fin de Máster.

Para poder llevar a cabo el proceso de agrupamiento y clasificación documental no supervisada existen diferentes técnicas, nosotros usaremos el *clustering*; consistente en la agrupación y representación de los documentos en forma de vectores, tanto de los documentos a representar como de los documentos que se insertan en los *clusters* o grupos, y mediante la estimación de su similitud obtendremos la referencia de clasificación en un *cluster* u en otro.

Los usos más frecuentes del *clustering* van desde la visualización y navegación en colecciones documentales, pasando por la recuperación de documentos afines, o la desambiguación de resultados de búsquedas convencionales (Figuerola, 2017a).

Los algoritmos de *clustering* se aplican sobre una serie de colecciones dinámicas de documentos, y un aspecto a tener en cuenta sobre ellas es que su organización va cambiando con el tiempo, encontrándonos con colecciones que se mantienen idénticas o colecciones en las que los documentos tienen una vigencia que es finita, es decir, tienen una duración determinada, de modo que sería interesante conocer como una colección de documentos va variando y el tiempo de vida que tiene esa colección.

En esta dirección se pretende el uso de un programa de *clustering*, en este caso a través del algoritmo de *K-Means* aplicado sobre una colección de documentos con fecha de vigencia finita, que abarca noticias de la prensa digital española centrada en la ciencia y tecnología en un sentido amplio desde el año 2002 hasta el año 2015, con una colección total de 60.074 documentos. Mediante la aplicación del *clustering* se ansía la determinación de las ventanas de tiempo de los documentos para conocer los grandes temas que han ocupado a la prensa digital española y su evolución durante estos citados años, así como la evaluación de la utilidad del sistema aplicado.

Por esto los capítulos se han organizado en el siguiente orden: en primer lugar nos encontramos con los antecedentes y el estado de la cuestión, donde hacemos referencia a los diferentes tipos de información y los modelos que nos permiten recuperarla. A continuación, en el segundo capítulo nos centramos en la clasificación automática de los documentos y en el algoritmo de *K-Means*, que será clave en el desarrollo de este Trabajo de Fin de Máster. En un tercer capítulo se hace una descripción del proceso de formación de la colección de noticias que se llevan a análisis. Proseguimos con un cuarto capítulo que hemos denominado *Scikit-learn*, que hace referencia al software utilizado para poder llevar a cabo el análisis de noticias por

medio el algoritmo de *K-Means*. Posteriormente pasamos a la parte práctica del trabajo con el capítulo de Trabajo Experimental, en el que realizamos los diversos análisis sobre la colección de noticias. Continuamos con el penúltimo capítulo de discusión de resultados en el que realizamos las observaciones que hemos considerado oportunas, y finalizamos con las conclusiones que hemos extraído de este Trabajo de Fin de Máster. Así mismo, queda otro capítulo más en el que se incluye la bibliografía utilizada.

1. Antecedentes y estado de la cuestión

En el momento en el que irrumpieron las nuevas tecnologías de la información y comunicación también surgieron los sistemas de recuperación, término que fue utilizado por primera vez en 1959 por el físico y matemático Calvin N. Mooers refiriéndose al ámbito que, "comprende los aspectos intelectuales de la descripción de la información y su especificación para buscar, así como cualquier sistema, técnica o máquina que se emplee para desarrollar la operación" (Piwowarski y Blanco, 2011, p. 34).

Esta y otras muchas más definiciones del ámbito abarcan aspectos muy amplios, pero si por algo se caracterizan todas ellas es porque a grandes rasgos, vienen a determinar el funcionamiento de los sistemas de recuperación de información como sistemas de búsqueda muy simples, que se basan en la recepción de una determinada consulta, véase petición, que realiza un usuario concreto, de modo que el sistema debe ser capaz de recuperar dentro de una gran base de datos, el lugar en el que se alojan aquellos textos que contengan la información que el usuario necesita en función de su búsqueda, devolviéndole el sistema aquellos documentos que son relevantes y que se ajustan a sus criterios de búsqueda, mostrando toda la información en forma de lista ordenada.

1.1. La información y su acceso

Hoy en día, los usuarios están acostumbrados a recuperar información en la web a través de motores de búsqueda tan conocidos como Google o Yahoo!, en los que el usuario escribe una consulta viéndose dispuesta su necesidad en ese momento de una determinada información, de modo que como respuesta obtiene unos documentos que aparecerán mediante una relación sistematizada (Piwowarski y Blanco, 2011).

Sin embargo, estos no son ni mucho menos los únicos sistemas de recuperación de información que existen, ni los únicos modos tampoco de reflejar la información que se precisa, encontrándonos con sistemas que se basan en otros modos de búsqueda como la clasificación; que muestra los documentos por asociación, es decir, se destaca un documento relevante a nuestra búsqueda pero a su vez este está relacionado con otros documentos, de modo que los usuarios pueden navegar por los diferentes documentos hasta conseguir el que realmente necesitan. El filtrado; que lo que hace es "filtrar", es decir, dentro de un conjunto de documentos se escogen unos cuantos relevantes para el usuario. Normalmente, este tipo de sistemas se suelen aplicar para restringir la gran cantidad de documentos que el usuario realiza en su consulta. La recomendación; muestra al usuario aquellos documentos que basados en su perfil y valoraciones se ajustan más a sus preferencias. El resumen; utilizado para disminuir el exceso de información que se presenta al usuario. Por último, tendríamos el agrupamiento; que muestra los documentos en grupos lógicos. Sus usos son varios, pero el más común es la asociación de los resultados de una búsqueda dada en una serie de conjuntos, que vienen a dar respuesta a los términos de esa consulta (Piwowarski y Blanco, 2011).

Este último caso de agrupamiento es en el que nos centraremos nosotros, pero antes hemos de dejar claro los tipos de información y las colecciones que existen.

1.2. Clases de información y colección

En primer lugar, resulta bastante impreciso determinar en la actualidad el término de documento, pues vivimos en una sociedad tecnológica muy cambiante en la que los tipos de información van variando de forma gradual.

Por otro lado, sí que podemos determinar los tipos de información que se buscan con mayor frecuencia y que suelen ser documentos completos, véase como ejemplo de ello cualquier documento albergado en la web, un documento en formato PDF, determinadas porciones de un documento como un capítulo de un libro, anuncios, información multimedia, correos electrónicos etc. (Piwowarski y Blanco, 2011).

Las búsquedas que se realizan son infinitas, y para cada una de ellas se podría pensar en una estrategia de recuperación diferente. Estos son unos aspectos que hay que tener muy presentes en el momento de determinar el uso de uno u otro sistema de recuperación de información.

En segundo lugar, con respecto a las colecciones documentales, básicamente las podemos agrupar en tres grupos (Piwowarski y Blanco, 2011):

Colecciones de uso personal: se refiere a aquellas colecciones que están formadas por documentos que el propio usuario dispone en el disco duro de su ordenador personal. Dado que la cantidad y naturaleza de los documentos puede ser muy diferente, su búsqueda puede resultar un tanto complicada.

Colecciones de uso corporativo: se refiere a aquellas colecciones que están formadas por documentos que han sido producidos por las empresas. La recuperación aquí también presenta una serie de problemas bastante grandes, puesto que el sistema se tiene que adaptar a las necesidades de cada empresa de manera muy precisa, y se podrían así mismo requerir de una serie de limitaciones en relación a las bases de datos.

Colección de uso web: se refiere a aquellas colecciones que están formadas por documentos albergados en la propia web. Para su recuperación se requiere de mayores medios, puesto que las páginas web son muchas y las consultas e informaciones también.

En nuestro caso concreto, nos ocuparemos del análisis de documentos en modo local, es decir, de documentos que han sido extraídos de una colección web, pero que al descargarlos no están sujetos a ningún cambio como lo estarían en la propia web.

1.3. Recuperación de información

Las herramientas de las que disponemos para la recuperación de información son muchas, al igual que los modelos que se utilizan para llevar a cabo tal acción. Cuándo nosotros realizamos una consulta dentro de una colección de documentos, para poder lograr obtener aquellos documentos que realmente nos interesan, se puede hacer a través de diferentes modelos de recuperación de información (Bravo, 2008).

Dado que el interés de este Trabajo de Fin de Máster no son los modelos de recuperación de información, tan solo daremos unas breves pinceladas de los más extendidos.

1.3.1. Modelo Booleano

De todos los modelos que existen de recuperación de información, este es considerado uno de los más sencillos. Se encuentra basado en la *Lógica de Proposiciones* propuesta por Georges Boole en 1854 (Urbano et al, 2010).

Este modelo permite la creación de una expresión, booleana, para así poder formalizar dicha consulta, formada por diferentes operadores booleanos siendo los más conocidos and, or y not (Bravo, 2008).

El sistema asigna un peso determinado, denominado como peso binario, a cada uno de los términos de búsqueda en función de si estos están o no en dicho documento. De modo que no existe una relevancia que pueda ser considerada como parcial, los documentos son relevantes o no (Urbano et al, 2010). Dentro de estas búsquedas se incluyen los operadores booleanos, que dependiendo del que utilicemos en cada caso nos mostrará unos resultados u otros.

1.3.2. Modelo Probabilístico

Este modelo fue propuesto por Robertons y Spacrk-Jones en 1996, conocido también como *Binary Independence Retrieval* (Figuerola, 2017b).

Su base es el cálculo de la probabilidad de que un determinado documento sea considerado como relevante para una consulta dada. De modo que si tenemos una consulta dada, existe un número de documentos concretos que contienen exactamente aquellos documentos que son considerados relevantes y los que no lo son, pudiendo nosotros estimar la probabilidad de que un documento sea pertinente para la consulta realizada. Para realizar tal acción existe una fórmula que es: Probabilidad (relevancia) = n / N, dónde n correspondería con el conjunto de documentos que son considerados relevantes, y N con el conjunto de documentos en sí mismo (Bravo, 2008).

En este modelo se presupone que existe un conjunto de documentos que son relevantes, de manera alguna que para calcular dicha relevancia se utilizan un conjunto de pesos asociados a las morfologías de los documentos. Si queremos conocer la relevancia, es necesario utilizar índices para los términos, también conocidos como descriptores para los pesos que se han fijado previamente (Bravo, 2008*)*.

La distribución de los términos es totalmente independiente para cada documento, sea relevante o no, y la probabilidad de relevancia se basa exclusivamente en la presencia de términos para la consulta realiza en el documento, pero también en la ausencia de ellos (Urbano et al, 2010).

Es un modelo que es un tanto complejo y costoso en relación a su computabilidad, pero es mucho más fuerte. Además como ya hemos explicado necesita de un corpus de entrenamiento, asume la independencia de los términos y les asigna pesos, que pueden ser positivos (relevantes) o negativos (no relevantes) (Urbano et al, 2010*)*.

1.3.3. Modelo Vectorial

Este tercer modelo de recuperación de información, también conocido como *Modelo de Espacio Vectorial*, fue propuesto por Gerard Salton a finales de los años 60, siendo uno de los modelos más extendidos y que presenta mejores resultados en este campo, pues permite la representación consistente de documentos como de las

consultas, permitiendo a su vez la formulación de estas últimas en un lenguaje natural, lo cual convierte a este sistema en base de muchos otros de recuperación (Urbano et al, 2010).

El sistema funciona “modelando” los documentos y las consultas mediante una serie de vectores descriptores o de términos (Urbano et al, 2010). De esta manera se procede a la creación de un espacio vectorial en el que la dimensión va a estar caracterizada por el número correlativo de términos que aparezcan en la colección. Los documentos se vienen a representar como vectores dentro de ese espacio vectorial con unas características propias, y del mismo modo se representan las consultas. Una vez representados ambos conceptos resulta fácil la aplicación de alguna función de similitud para poder determinar la semejanza existente entre los vectores de término de la consulta y los vectores de cada uno de los documentos (Figuerola, 2017b). En función de la consulta que se lleve a término, los grados de similitud varían, de modo que se puede considerar que la obtención de un grado de similitud mayor se ajusta a los criterios de búsqueda y viceversa, mostrando los documentos de una manera organizada, pudiendo también limitar el número de estos si se utiliza un grado de similitud “mínimo” (Bravo, 2008).

Por otro lado, el modelo vectorial también se basa en el “esquema de pesos”, que viene a determinar el modo en el que se reparten los pesos en los términos de los documentos dependiendo claro está, de la importancia de cada uno de ellos con respecto al contenido y que se suele calcular mediante diversos métodos. Para llevar a cabo esta acción se suele utilizar un sistema muy sencillo de asignación de pesos como es el esquema binario, que se basa en la asignación de un término “1” si el término está presente, y de un peso “0” si el término no está presente, pudiendo establecer ponderaciones más complejas si así se desea (Urbano et al, 2010).

A modo de resumen, la forma más simple de calcular la similitud existente entre dos vectores (consulta y documentos), consiste en determinar qué número de términos tienen en común ambos, lo que se conoce como el producto interno de dos vectores (Urbano et al, 2010*)*. De manera alguna, que representado vectorialmente, cuánto más juntos estén los vectores mayor será la similitud entre ambos, y cuánto más alejados estén entre sí menor será la similitud.

Este último modelo de recuperación de información será el que nosotros utilizaremos a través del algoritmo de *K-Means*.

2. K-Means

Todos los documentos, independientemente del formato en el que se encuentren, están conformados por una cantidad de palabras que es finita. De este modo, es posible representarlos como puntos en un vector dentro de un espacio que se denomina t-dimensional y que da nombre al Modelo de Espacio Vectorial.

Siguiendo la idea de este modelo, podemos actuar sobre los documentos empleando sistemas de entrada espaciales, ya sea para hacer búsquedas activas o simplemente para evaluar la similitud entre diferentes documentos. Este último caso es objeto de este trabajo, pues se busca medir la semejanza que existe entre los diferentes documentos que conforman nuestra colección a través de su contenido, ya que al realizar el modelado de dos documentos como vectores, se pueden establecer una serie de frecuencias de término para cada uno de ellos, permitiendo determinar la similitud que hay entre ambos. Esto se puede realizar a través de la clasificación automática de documentos (Álvarez, Vega y Fernández, 2007).

2.1. Clasificación automática de documentos

Cuando hablamos de clasificación, aludimos a un concepto muy simple de organización de los documentos para que estos después puedan ser fácilmente recuperables. En torno a esta idea se han ido fraguando diferentes sistemas con destinos muy dispares.

En el momento que surgen los documentos en formato electrónico, lo hacen también las técnicas que posibilitan su clasificación de forma automática. De modo, que cuando hablamos de clasificación automática de documentos, se distinguen dos atmósferas distintas, como son la clasificación supervisada y la clasificación no supervisada. Esta última será la que nosotros utilizaremos en este estudio.

Ambas tienen la misma base, pues “se parte de una serie de clases o categorías conceptuales prediseñadas a priori, y en la que la labor del clasificador (manual o automático) es asignar cada documento a la clase o categoría que le corresponde” (Figuerola, Alonso y Zazo, 2004, p. 1).

2.1.1. Clasificación supervisada

De manera breve decir, que aunque nosotros vamos a trabajar con la clasificación no supervisada, esta no se entendería bien sin antes hablar de la supervisada y sus algoritmos principales, por ello hemos creído conveniente enunciarla primero.

Como ya hemos explicado anteriormente, este tipo de clasificación supervisada parte de un conjunto de categorías que previamente habrían sido establecidas, debiendo nosotros de ubicar cada documento en cada clase.

Esta clasificación también se le conoce con el nombre de “categorización”, debido en gran parte a que permite la recuperación ad-hoc, restringiendo los sondeos a las categorías que el usuario ha establecido a través de las competencias que tiene sobre esa materia (Figuerola, Alonso y Zazo, 2004).

Sus aplicaciones son varias, desde el filtrado de documentos pasando por el *routing*, recuperación mediante *browsing* o la asignación automática de descriptores, encabezamientos de materias y similares (Figuerola, 2017a).

2.1.1.1. Algoritmos

Existen un gran variedad de algoritmos que se muestran capaces de realizar una clasificación supervisada teniendo un planteamiento similar, ya que de lo que se trata es de construir un patrón que sea representativo de las categorías que se van a formar, y a continuación aplicar una determinada función que logre establecer la similitud existente, entre el documento que se pretende clasificar y los patrones establecidos para cada una de las categorías. De esta manera, se determina la categoría a la que pertenece el documento que queremos clasificar, ya que el patrón de similitud así lo indica. Sin embargo, en sistemas con asignación de clases múltiples, es el umbral de similitud el que indica a qué clase hay que asignar cada documento (Figuerola, 2017a*)*.

Para poder construir patrones en las categorías, previamente los documentos se han de clasificar de forma manual, y utilizar estas series como ejemplos. Todo el proceso de formación de los patrones se conocerá con el nombre de "entrenamiento o aprendizaje", mientras que la colección de documentos que se han clasificado previamente se conoce como "colección de entrenamiento" (Figuerola, Alonso y Zazo, 2004). Por lo tanto, la intervención humana se va a precisar, tanto para la creación de los patrones como para la revisión y posterior refinamiento de los resultados.

Los algoritmos más utilizados en la clasificación supervisada son el Algoritmo de Naive Bayes; basado en el principio de la probabilidad, se encarga de evaluar la probabilidad de que un determinado documento coincida en una clase mediante la posibilidad de que, determinados documentos encierren ciertos términos que pertenezcan a esa clase. Estas probabilidades se estimarían mediante los términos que se muestran en los documentos de entrenamiento. Sin embargo, en la práctica solo se van a estimar los pesos binarios de esos términos. Este es un método que es eficaz, y resulta fácil y rápido de implementar (Figuerola, 2017a).

Por otro lado, tenemos el Algoritmo de Rocchio que se encuentra basado en los mismos conceptos que la realimentación por relevancia, es decir, que trata de construir una serie de vectores que representen a cada una de las clases por medio de los documentos de la colección de entrenamiento. De modo que para el vector de cada clase, los documentos de entrenamiento de una clase concreta se utilizarían como elementos positivos, mientras que los documentos de entrenamiento del resto de clases se utilizarían como elementos negativos. Los vectores que son representativos de cada clase se construyen sumando los pesos de los términos que se han establecido como elementos positivos. De estos se restarían los pesos de los términos de los elementos que son negativos. Mediante la aplicación de una serie de coeficientes multiplicadores, sería posible establecer el mayor o menor grado de importancia de los elementos positivos y negativos. Resultado de ellos, surge un vector de términos con una serie de pesos como el que se aplica en el modelo vectorial. Entonces, si queremos ordenar un documento nuevo, no habría más que estimar el grado de similitud que existe entre, el vector del documento nuevo y los vectores que tienen las clases dónde se va a clasificar.

También nos encontramos con el algoritmo del Vecino más próximo y Knn, que parte del hecho de que una colección de entrenamiento se puede indizar a través de cualquier motor de recuperación. De modo que, cuando se necesita clasificar un documento nuevo, este documento se va a utilizar como una consulta contra ese motor de recuperación, es decir, la consulta se va a realizar contra la colección de entrenamiento. Resultado de esa búsqueda aparecerá un documento que es considerado como el más relevante, indicándonos la clase a la que debe pertenecer el documento que necesitamos clasificar. La variante K-nn, básicamente radica en

estimar los K primeros documentos más notables, en vez de estimar solo el primer documento.

Por último, nos encontraríamos con el algoritmo de Máquinas de Vectores Soporte (SVM), que mediante un plano con miles de dimensiones pretende ubicar puntos de dos clases. El problema radica en separar ambas clases, es decir, que necesitamos de "un hiperplano que maximice la separación existente entre las muestras de entrenamiento. En ocasiones no es posible aplicar una solución lineal. Este algoritmo utiliza funciones kernel para remapear las muestras de forma que un hiperplano pueda separarlas" (Figuerola, 2017a, p. 19).

2.1.2. Clasificación no supervisada

La clasificación no supervisada, más conocida como *clustering* o agrupamiento documental, no tiene un conjunto de clases preestablecidas, sino que es el propio sistema quién va a establecer una serie de clases o *clusters* (elementos o registros de un conjunto de datos semejantes entre sí) de forma totalmente automática (Álvarez, Vega y Fernández, 2007).

Es importante hacer bien esta distinción, ya que lo que se pretende con el agrupamiento documental, es particionar una colección de documentos en grupos o *clusters* que estén a la vez lo más cohesionados internamente y lo más separados entre ellos (Ares, Parapar y Barreiro, 2011). Con ello se descubrirían grupos de documentos que describen fenómenos similares dentro de la heterogeneidad de la colección, permitiendo y facilitando así el estudio de las estructuras intrínsecas de la misma.

Así, el principio fundamental que rige el *clustering* se basa en garantizar "que los grupos sean lo más heterogéneos entre sí, pero que los elementos del grupo sean lo más homogéneos posibles, basados en un criterio de optimización" (Vallejo, 2016, p. 29), buscando con ello minimizar la distancia y similitud dentro de cada clúster consiguiendo con ello cohesión, y maximizar la distancia y similitud entre *clusters* para garantizar la separación.

Orientado en este sentido, hemos de distinguir los diferentes sistemas de *clustering* o como hemos mencionado anteriormente, aprendizaje no supervisado, de las técnicas de clasificación o aprendizaje supervisado.

Se puede afirmar, que las técnicas de clasificación se caracterizan por el empleo de modelos extraídos de una serie por lo general etiquetada de forma manual, conocida como colección de entrenamiento, que yéndonos a un lenguaje más técnico se conoce como "training", para así poder entrenar los clasificadores de las diferentes categorías en las que se quieren reunir los documentos de otras series, conocidos como colección de prueba o "test". Sin embargo, en los sistemas de *clustering*, esta acción de reunión es llevada a término sin necesidad de conocimiento del dominio, no hay por tanto ninguna necesidad de efectuar un entrenamiento del sistema e incluso sin tener constancia de las clases que ya existían (Ares, Parapar y Barreiro, 2011).

Tradicionalmente, los sistemas de agrupamiento documental se han venido utilizando en distintas áreas del marco de la computación en función de los documentos y los datos, siendo los usos más frecuentes los relacionados con la navegación y visualización de las colecciones documentales, recuperación de documentos afines o la desambiguación de resultados de búsquedas convencionales (Figuerola, 2017a).

2.1.2.1. Representación de documentos y medidas de similitud

Existen una gran variedad de algoritmos relacionados con el agrupamiento documental o *clustering* teniendo como principio común, la existencia de una representación documental de todo el conjunto de datos que forman parte de la colección, así como la utilización de una serie de medidas que facilitan la medición de la similitud y distancia que existe entre dos instancias de datos dadas (Ares, Parapar y Barreiro, 2011).

La forma de representar los documentos estriba en gran medida en el origen de la información, que comprende las instancias de datos que forman parte de una colección de datos con una serie de valores cuantitativos y cualitativos.

Por lo general, la representación de los documentos textuales se realiza siguiendo un patrón que es considerado común y muy parecido al que se utiliza en el Modelo de Espacio Vectorial, es decir, que "cada documento es un vector donde cada componente corresponde al valor asociado en la representación elegida a un determinado término" (Ares, Parapar y Barreiro, 2011, p. 397).

Ahora bien, el modo en el que se contabilizan los pesos de cada uno de los términos en cada documento es distinto, pues está ligado al modo de representación que se determine, encontrándonos con la Frecuencia Relativa de los Términos; que vendría a ser un sistema para la representación de documentos textuales, que solo considera los datos locales de cada documento para determinar su representación. En esta representación cada documento estaría conformado por un vector con distintas posiciones, identificadas cada una de ellas con la frecuencia relativa de un término concreto dentro del documento. La Frecuencia de Término en el Documento Ponderada por su Frecuencia en el Documento; su utilización ha aportado grandes logros en los modelos de recuperación de información, pues se basa en la utilización de estadísticas de la serie documental, permitiendo "ponderar el peso de cada término en el documento (información local), por su especificidad (información global) en la colección" (Ares, Parapar y Barreiro*,* 2011, p. 398*)*. El método más habitual de conjugar la información local con la información global para su posterior agrupamiento documental es la utilización de la frecuencia inversa logarítmica. Por último tendríamos la Información Mutua; partiendo del hecho de que tendríamos un vector de información mutua para la representación de los documentos textuales, se determinaría un peso a cada término de nuestra colección, utilizando para ello la información local del documento como la información global de la colección (Ares, Parapar y Barreiro, 2011).

2.1.2.2. Medidas de distancia

En el momento que hemos elegido nuestro método para la representación de los documentos textuales, los algoritmos utilizados para el agrupamiento necesitan de una medida que logre cuantificar la similitud que hay entre los documentos. Según los autores Ares, Parapar y Barreiro afirman que:

> En el caso de que el valor de la medida aumente cuanto más diferentes sean, hablaremos de una medida de distancia, mientras que si disminuye hablaremos de una medida de similitud (es posible convertir fácilmente de un tipo de medida a otro, por ejemplo restándole el valor de la misma al máximo valor posible). (p. 401).

La selección de una medida de distancia es una cuestión decisiva, ya que si eligiéramos una medida que no es la adecuada, esta influiría de una manera incipiente en la calidad de la salida en el proceso de *clustering*.

En función del modelo de representación escogido, dependiendo si son vectores con valores discretos, continuos o representaciones conformadas en elementos cualitativos, obtendremos diferentes alternativas en el momento de seleccionar las medidas de distancia. En nuestro caso, pondremos la mirada en los documentos textuales, que por otro lado es el caso que es más común.

Las dos distancias que más fama tienen son la Distancia Euclídea; tradicionalmente utilizada en tareas de agrupamiento cuando las representaciones de valores continuos con las que se trabajaban estaban en espacios de baja dimensionalidad. Dada la inexistencia de normalización exige obrar sobre representaciones que sí que están normalizadas, pues de no ser así, los vectores que tengan magnitudes elevadas desarrollarán valores mucho más altos.

Por otro lado, tenemos la Distancia del Coseno; de mayor uso en la recuperación de información y minería de datos, al contrario que la anterior, esta se aplica sobre información con una alta dimensionalidad, véase las colecciones de texto. Su aplicación es simple, se fundamenta en calcular la cercanía entre dos documentos por medio del coseno del ángulo entre las representaciones (Ares, Parapar y Barreiro, 2011).

2.1.2.3. Algoritmos de Agrupamiento

Son muchos los autores que han mostrado interés en los trabajos concernientes al agrupamiento o *clustering*, no pocos proponiendo diferentes algoritmos para llevar a término tal acción.

En función del método utilizado en la creación de los *clusters*, se pueden determinar dos clases de algoritmos; jerárquicos y particionales (Ares, Parapar y Barreiro, 2011). En este trabajo se hará más hincapié en los particionales pues será el que describamos más adelante.

Los algoritmos jerárquicos realizan la división de los documentos creando *clusters*, que a su vez forman una jerarquía entre ellos. Los algoritmos aglomerativos son los representantes más populares de este tipo de algoritmos.

Los algoritmos que realizan *clustering* jerárquico, poseen la misión de reunir *clusters*, ya sea para formar un nuevo o en su defecto dividir alguno de los que ya existen dando lugar a otros dos nuevos. De esta manera, si el proceso se repitiera de forma progresiva se maximizarían o minimizarían algunas medidas de similitud (Vallejo, 2016). Es de destacar, que en este tipo de algoritmos no es necesario prefijar el número de K *clusters*. En consecuencia, obtenemos una especie de árbol, un árbol de clúster que también se suele denominar “dendrograma”, en el que se ven reflejadas las relaciones que se producen entre los *clusters*. En el caso de que nosotros seccionemos el dendrograma en un determinado nivel (umbral), lo que obtendríamos sería una agrupación de elementos, pero estos estarían en grupos desiguales. Por último, hay que destacar que los algoritmos de *clustering* jerárquicos siguiendo la metodología usada para formar *clusters*, se pueden dividir a su vez en divisivos y aglomerativos, con una gran variedad de vertientes en cada una de ellas (Vallejo, 2016).

Por otro lado, nos encontramos con el agrupamiento particional más conocido como “*flat clustering*”, encargado de adquirir “una partición única de los datos en vez de una estructura de agrupación” (Vallejo, 2016, p. 32), a diferencia de como lo llevaba a cabo el dendrograma por medio del método jerárquico. La principal ventaja de esta técnica en contraposición con el agrupamiento jerárquico radica en las aplicaciones que

conllevan el control de ingentes conjuntos de datos, ya que en estos casos no se puede realizar un dendrograma pues computacionalmente está prohibido (Vallejo, 2016*)*.

Los algoritmos que utilizan métodos particionales, realizan particiones de los datos en k conjuntos por separado de manera simultánea, es decir, los k conjuntos obtenidos no tienen nada que ver con otros elementos de esos mismos k conjuntos, por lo tanto se producen k clases que carecen de relación entre ellas mismas. En este sentido Vallejo (2016) afirma que:

> Las técnicas particionales suelen producir clusters optimizando alguna función objetivo (criterio) definida en forma local (usando parte de los patrones) o global (usando la totalidad de los datos). El criterio más comúnmente usado es el error cuadrático y uno de los problemas asociados al uso de un algoritmo particional es la elección del k número de grupos y puntos iniciales. La búsqueda combinatoria del conjunto de posibles clases para el valor óptimo de la función objetivo es computacionalmente prohibitiva. Por esta razón, en la práctica, el algoritmo se ejecuta típicamente múltiples veces con diferentes estados iniciales, y la mejor configuración obtenida de todas las ejecuciones es la que se utiliza como clustering de salida, es decir, que la optimización de la función objetivo se realiza mediante un proceso iterativo. (p. 33).

En nuestro caso concreto solo nos vamos a centrar en el algoritmo de *K-Means*, objeto de este Trabajo de Fin de Máster.

2.1.2.4. K-Means

El algoritmo particional de *K-Means* asociado al *clustering* y a la estadística, ha sido descrito en la literatura bajo diversos nombres. Comúnmente es conocido como K-Medias, debido a la representación que lleva a cabo de los grupos “por la media” de sus puntos, que reciben el nombre de centroides (Vallejo, 2016*)*.

La primera vez que hay constancia de su utilización procede del año 1967 bajo la mano de McQueen, aunque si bien es cierto la idea del algoritmo procede de Steinhaus, quien formuló la teoría en 1956, siendo redescubierto en 1957 por Lloyd como “una técnica de cuantificación para la modulación por impulsos codificados” (Vallejo, 2016, p. 33). En 1965, Forgy se decidió por publicar el mismo método. De modo que nos encontramos con muchas variantes del mismo método, pero el más popular debido en gran parte a su complejidad, y en teoría bajo coste computacional que le proporcionan una facilidad en relación a la implementación, velocidad y eficacia es el algoritmo de McQueen.

Este algoritmo es iterativo pues de lo que trata es de “distribuir los documentos en k *clusters* de un modo tal que la suma residual de cuadrados (RSS, por las iniciales en inglés *Residual Sum of Squares*) de la solución alcanzada al final del proceso sea mínima” (Ares, Parapar y Barreiro, 2011, p. 403), es decir, que trata de reducir a la mínima expresión el cálculo de la distancia existente entre los puntos de cada uno de los grupos y sus respectivos centroides (Vallejo, 2016).

Formalmente se pueden establecer varias fases de consecución algorítmica:

En la primera fase, que hemos denominado como fase previa, el usuario debe de especificar el número de K grupos que desea formar para que el proceso se inicie (Vallejo, 2016). Después, se han de seleccionar k puntos iniciales de entre todos los datos que forman parte del conjunto. Una vez disponemos de los k puntos iniciales es el momento de realizar los cálculos de las distancias existentes entre los centroides y

los demás puntos restantes, reasignándose posteriormente hacia aquellos puntos que se encuentren más cercanos.

En la segunda fase, una vez terminada la fase de reasignación, se llevará a cabo de nuevo un recálculo de los centroides que forman parte de cada *cluster*, de modo que se volverá a repetir este proceso hasta que no aparezca cambios en los k centroides.

Se puede decir que el algoritmo lo que viene a realizar es una minimización de la función que es objetivo, sin embargo, en ningún caso esto puede ser sinónimo de garantía para que se alcance el mínimo global. De modo, que llegará un momento en el que *K-Means* llegue a "converger en un mínimo local del vector de cuantificación del error" (Vallejo, 2016, p. 34), y todo el proceso ha de ser reiniciado de nuevo muchas veces, proceso que por otra parte es muy costoso en términos computacionales, y más cuándo estamos hablando de conjuntos de datos extensos. Todo este proceso lo podemos observar en la Figura 1.

Por otro lado, el principal problema que tiene *K-Means* es su gran dependencia de la situación inicial, es decir, todo el proceso va a estar ligado a la elección inicial de los k documentos (Ares, Parapar y Barreiro, 2011). Esta situación influirá de manera decisiva en la calidad de la solución que obtengamos, de modo que la mejor solución aunque no sea la más óptima, pasa por repetir el algoritmo, escogiendo distintos puntos iniciales asociado también a la elección de la solución que contenga el valor mínimo de la función que es objetivo.

En comparación con los métodos jerárquicos, este algoritmo tiene una eficiencia mucho mayor, debido en gran parte a que los tiempos para el cómputo que necesita son lineales en relación con el número de documentos que se van a asociar, aunque como hemos mencionado anteriormente es muy dependiente de la situación inicial (Bravo, 2008). Por ello, los resultados son muy heterogéneos y todo vuelve a quedar en manos del azar.

Figura 1. Proceso del algoritmo K-Means

3. Descripción de la Colección

La colección de documentos que es objeto de análisis para este Trabajo de Fin de Máster está compuesta por 60.074 noticias de la prensa digital española, en concreto de los periódicos El Mundo y El País. Estas noticias no han sido seleccionadas al azar, sino que solo se han extraído noticias desde el año 2002 hasta el año 2015 que trataban temas de Ciencia y Tecnología en sentido amplio.

Esta colección de noticias procede de un proyecto denominado como "Sistema de indicadores para el SCSC (*Spanish Corpus of Scientific Culture)"* que permitió la realización del diseño, implementación y puesta a prueba de un conjunto de técnicas automáticas que fueran "capaces de recopilar y analizar cuantitativamente noticias sobre Ciencia y Tecnología, así como calcular una serie de indicadores de cultura científica" (Figuerola y Quintanilla, 2016, p. 7). En nuestro caso nos centraremos en la última versión realizada en el año 2016.

El interés suscitado de la ciencia en los diferentes medios de comunicación es un campo en expansión dentro de la investigación académica, pues muchas veces los medios han sido tildados de un tratamiento de la información con poco rigor, fiabilidad e inadecuados en relación a la ciencia. Su estudio nos permite reflejar el posicionamiento que ocupa la ciencia dentro de la sociedad, por ello nos encontramos muchas veces que tanto Ciencia como Tecnología aparecen expuestos de formas muy distintas dependiendo del medio de comunicación que se analice (Groves, Figuerola y Quintanilla, 2016).

En nuestro caso particular tenemos por objetivo el proporcionar una visión que de manera general muestre la evolución de las noticias de Ciencia y Tecnología a lo largo de catorce años, cuestión que se puede realizar partiendo de la base de que la mayoría de las noticias van a tratar temas específicos dentro de sus publicaciones, en este caso de Ciencia y Tecnología. Una cuestión importante a tener en cuenta en el momento de recolectar las noticias es lo que se entiende por Ciencia y por Tecnología.

En determinados estudios utilizan una visión amplia de Ciencia en el que se incluyen las ciencias sociales y las humanidades, mientras que en otros son mucho más exigentes y lo restringen únicamente a las consideradas como las ciencias puras así como a la biología y la medicina (Groves, Figuerola y Quintanilla, 2016). Se define a la Ciencia como la agrupación de los conocimientos que se han logrado obtener por medio de la observación y del razonamiento, que aparecen de manera ordenada y de los que se pueden extraer una serie de principios comprobables de forma experimental. Por otro lado, la Tecnología sería la puesta a punto del conocimiento científico para dar respuesta a las necesidades que tienen los humanos, y aquí radica la diferencia con la Ciencia. A pesar de que los análisis se podían haber hecho por separado, nosotros los hemos hecho en conjunto.

3.1. Procedimiento

La formación de la colección fue un proceso laborioso cuanto menos, pues lo primero que se hizo fue proceder a su descarga, y hablamos de descarga porque la revolución digital también ha llegado a los medios contemporáneos. Es un hecho que la mayoría de los periódicos que son importantes tienen sus ediciones también en formato digital, pues los usuarios tienen cada vez más a consultar sus noticias de forma remota.

Por esta razón, el corpus fundamental de noticias que forma parte de este trabajo tiene su origen en las ediciones digitales de los periódicos EL Mundo y El País.

En un primer momento se descargaron de las hemerotecas digitales de estos periódicos todas las noticias que la página web les permitía independientemente de la temática. Las hemerotecas estaban en formato digital, y por lo tanto las noticias también lo estaban. Partiendo de esta base fue posible la búsqueda de noticias en la página principal de las mismas mediante las fechas que se necesitaban, en este caso desde el año 2002 hasta el año 2015. La descarga de todas las noticias fue un proceso que no se podía realizar de otra manera sino automáticamente, ni que pensar queda que este proceso se hubiera podido realizar de forma manual pues sería inabarcable. Este proceso automático se pudo realizar a través de *crawlers*, que son un conjunto de programas que pueden navegar de una forma libre y automática, que aplicado a las hemerotecas permite la descarga de las páginas web por las que se ha estado navegando (Figuerola y Quintanilla, 2016).

Las noticias no aparecen "solas", es decir, muchas de ellas contienen publicidad o diversos comentarios, e incluso si atendemos al paso del tiempo muchos periódicos sufren variaciones en relación a su formato o estructura lo que modifica su codificación, de modo que hay que hacer un expurgo de todo lo que no nos interesa, detectar las variaciones de las noticias para dejarlas en definitiva "limpias", y esto solo se pudo realizar a través de técnicas automáticas. Una vez estaban descargados los artículos, estos se guardaron en formato HTML, y posteriormente se guardaron en texto pero sin formato (Groves, Figuerola y Quintanilla, 2016).

Dado que el objetivo era trabajar con noticias de Ciencia y Tecnología, no se necesitaban todas las noticias pues se llegaron a recopilar más de 1.5 millones de noticias entre los tres periódicos. Por otro lado, se decidió prescindir de las noticias del ABC, ya que inicialmente se descargaron noticias de este periódico junto con las de El Mundo y El País, pero debido a una serie de problemas de formato que hacían imposible individualizar noticias a partir de las páginas escaneadas, se eliminaron sus noticias de este análisis. Una vez extraídas las noticias del ABC, se puede observar la evolución del número total de noticias de los periódicos EL Mundo y El País de forma anual en la Figura 2 y la Tabla I respectivamente.

Noticias generales El Mundo y El País

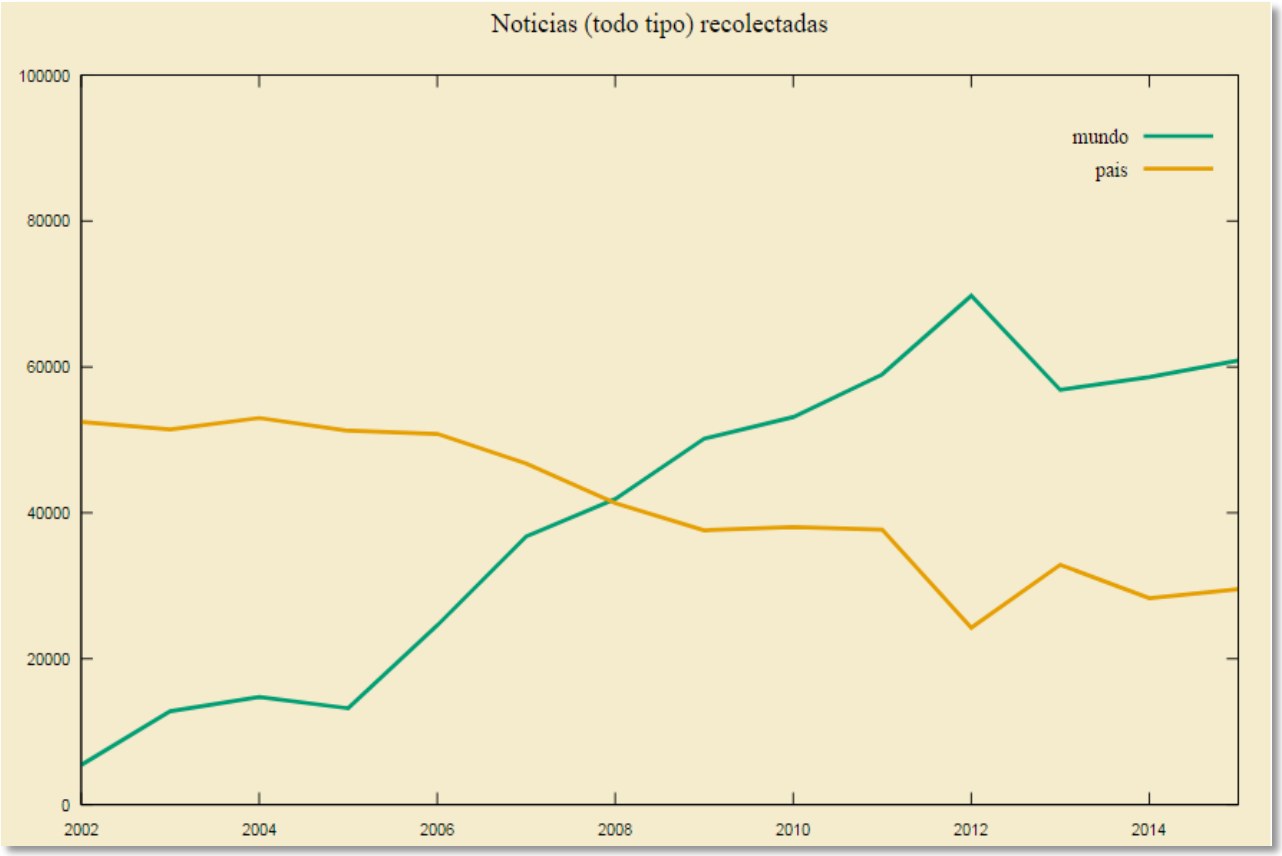


Figura 2. Gráfico noticias El Mundo - El País

	EL MUNDO	EL PAÍS
2002	**5441**	**52483**
2003	**12817**	**51450**
2004	**14745**	**53004**
2005	**13209**	**51249**
2006	**24572**	**50824**
2007	**36772**	**46755**
2008	**41887**	**41329**
2009	**50171**	**37601**
2010	**53133**	**38028**
2011	**58965**	**37696**
2012	**69749**	**24266**
2013	**56860**	**32889**
2014	**58618**	**28312**
2015	**60874**	**29544**

Tabla I. Noticias totales periódicos

Sin embargo, nosotros no vamos a trabajar con tal cantidad de noticias, de modo que reducimos la dimensionalidad únicamente a las noticias de Ciencia y Tecnología obteniendo la cantidad de 60.074 noticias. Este proceso es realmente difícil llevarlo a cabo de forma manual pues resulta inviable, por ello se recurrió a categorizadores automáticos para que seleccionen solo aquellas noticias que se precisan. En este caso se utilizó un categorizador SVM, propuesto por V. Vapnik en 1995, el cual realiza un escalamiento que es bastante bueno cuando nos encontramos con un número de características elevado, su entrenamiento es muy rápido y más eficiente, luego los resultados de clasificación son mucho mejores (Figuerola, 2017a).

“Los categorizadores necesitan de una fase de entrenamiento, mediante técnicas de *machine learning*, durante la cual establecen las características que definen las clases o categorías con las que hemos de trabajar” (Figuerola y Quintanilla, 2016, p. 27).

Estos categorizadores necesitan de una colección de documentos que previamente han sido etiquetados de forma manual para poder extraer de ellos sus principales características. Esta colección que se ha clasificado de forma manual recibe el nombre de colección de entrenamiento y es vital que cumpla con una serie de parámetros. En primer lugar, la colección ha de ser "representativa de los casos que se pueden representar" (Figuerola y Quintanilla, 2016, p. 27), algo que no es ni mucho menos fácil de vaticinar. Por otro lado, debe de tener un tamaño que sea adecuado, es decir, se van a precisar de un buen porcentaje de noticias, pero tampoco hay que excederse pues los categorizadores podrían sufrir como es habitual un "sobreentrenamiento" por el tamaño de la colección, y esto produciría ruido en la misma y restaría efectividad al proceso.

Una vez realizado este proceso se logró desarrollar colecciones de entrenamiento con una calidad suficiente, que unido a las posteriores revisiones de forma manual realizadas a distintas muestras de noticias de Ciencia y Tecnología de forma totalmente aleatoria, mostraron un porcentaje de aciertos de hasta el 95 %, acción que se vio superada cuándo se realizó un posterior análisis automático de los temas (Figuerola y Quintanilla, 2016).

Las opciones que se mostraron en el momento de clasificar las noticias fueron varias, pero en este caso concreto que nos ocupa se decidió por realizar una detección de noticias que tratasen a la vez sobre temas de Ciencia y Tecnología por medio de un categorizador automático. Esta solución tiene una gran ventaja, que es la detección de las noticias que tratan los temas al mismo tiempo, por el contrario, puede producir mucho ruido en las clasificaciones ya que también clasifica las noticias como no pertenecientes a Ciencia y Tecnología a la vez (Figuerola y Quintanilla, 2016).

En otro sentido, lo que realmente nos interesa a nosotros son los temas tratados en las noticias, es decir, que dentro del conjunto de las noticias que versan sobre ciencia y tecnología determinar qué temas concretos y en qué proporción hay dentro de esos grandes conjuntos, que muchas veces y como es el caso no siempre van a coincidir mucho con los temas matrices (Figuerola y Quintanilla, 2016).

El principal escollo con el que nos encontramos, es la elección de una tecnología automática que determine los temas tratados en las noticias dentro de ese gran conjunto de 60.074 documentos. Las tecnologías disponibles son muchas, siendo una de las más comunes la aplicación de *Topic Modeling* por medio de alguna de las implementaciones conocidas como *Latent Dichlet Allocation* (Figuerola y Quintanilla, 2016). El concepto es muy simple, pues se parte de la base que cada documento va a contener distintos temas en una proporción variable, de modo que algunas técnicas como el LDA van a determinar los temas en conjunto y el porcentaje de cada uno de ellos.

Pero esta técnica no es la única, ni mucho menos, y como alternativa podemos encontrar como en nuestro caso las técnicas de *clustering* de documentos, teniendo una cierta certeza de que las noticias que posean un mayor parecido en cuento a contenido temático se refiere van a ser agrupadas de manera conjunta. Esta técnica también requiere una revisión de los *clusters* o temas que se han obtenido. El etiquetado de sus temas debería arrojarnos una lista con los principales y su intensidad dentro de cada *cluster* y con respecto al total (Figuerola y Quintanilla, 2016).

Las técnicas de *clustering* como ya hemos mencionado en apartados anteriores son muchas y se caracterizan por su exigencia en el proceso. Nosotros utilizaremos una de las técnicas que se suelen utilizar con mayor frecuencia como es *K-Means*, que nos va a exigir determinar antes de iniciar el proceso el número de *clusters* que queremos

obtener, lo que va a llevar a temas cuya proporción va a ser más creciente en unas noticias que en otras, o todo lo contrario si establecemos un número muy elevado de *clusters*.

3.2. Datos

Hecho el *clustering,* podemos llevar los datos a un gráfico y observar mejor la disposición de las noticias de los periódicos El Mundo y El País.

Cómo se puede visualizar a continuación en la Figura 3, de entre el total de noticias que son 60.074, la proporción que existe en relación a los periódicos que hemos determinado de forma totalmente manual, es ligeramente superior en El Mundo, con 33.435 noticias que suponen el 56 % del total, mientras que en El País nos encontramos con 26.639 noticas que suponen el 44 % restante. Estos datos vienen a concluir que El Mundo ha tratado más temas de Ciencia y Tecnología que El País.

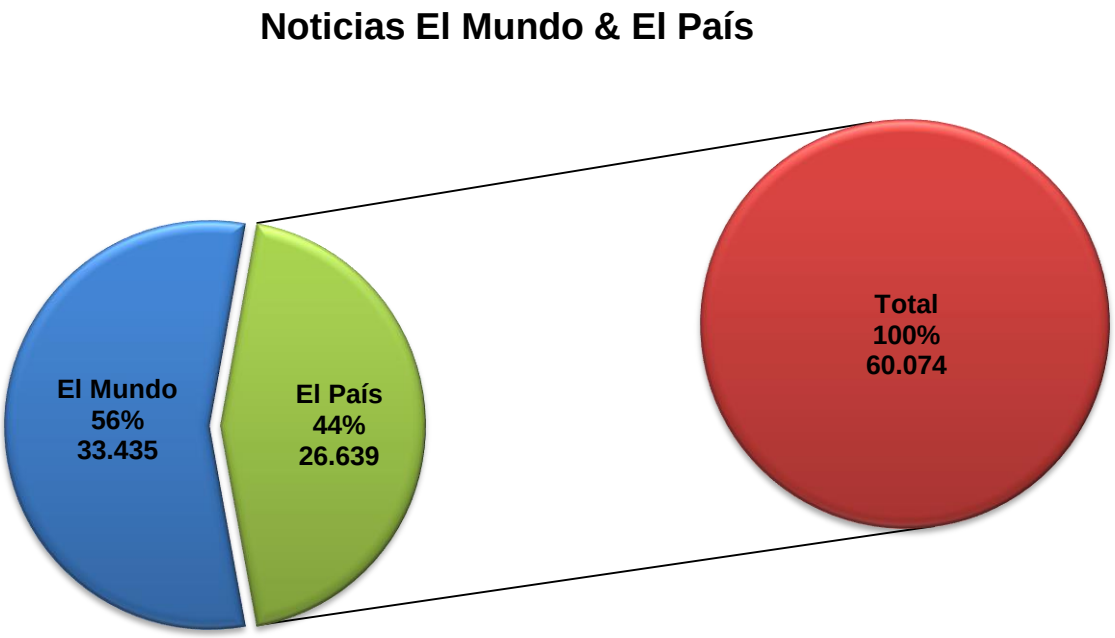


Figura 3. Gráfico porcentaje noticias

En otro término, en la Figura 4 aparece reflejado el análisis de las noticias por años, pudiéndose observar cómo los *clusters* no son para nada equitativos, es decir, no existe la misma proporción de noticias en ambos periódicos ni se mantiene lineal en el tiempo, pues nos encontramos con un mínimo de noticias en el año 2.002 con 2.822 noticias y un máximo en el año 2.011 con 5.756 noticias.

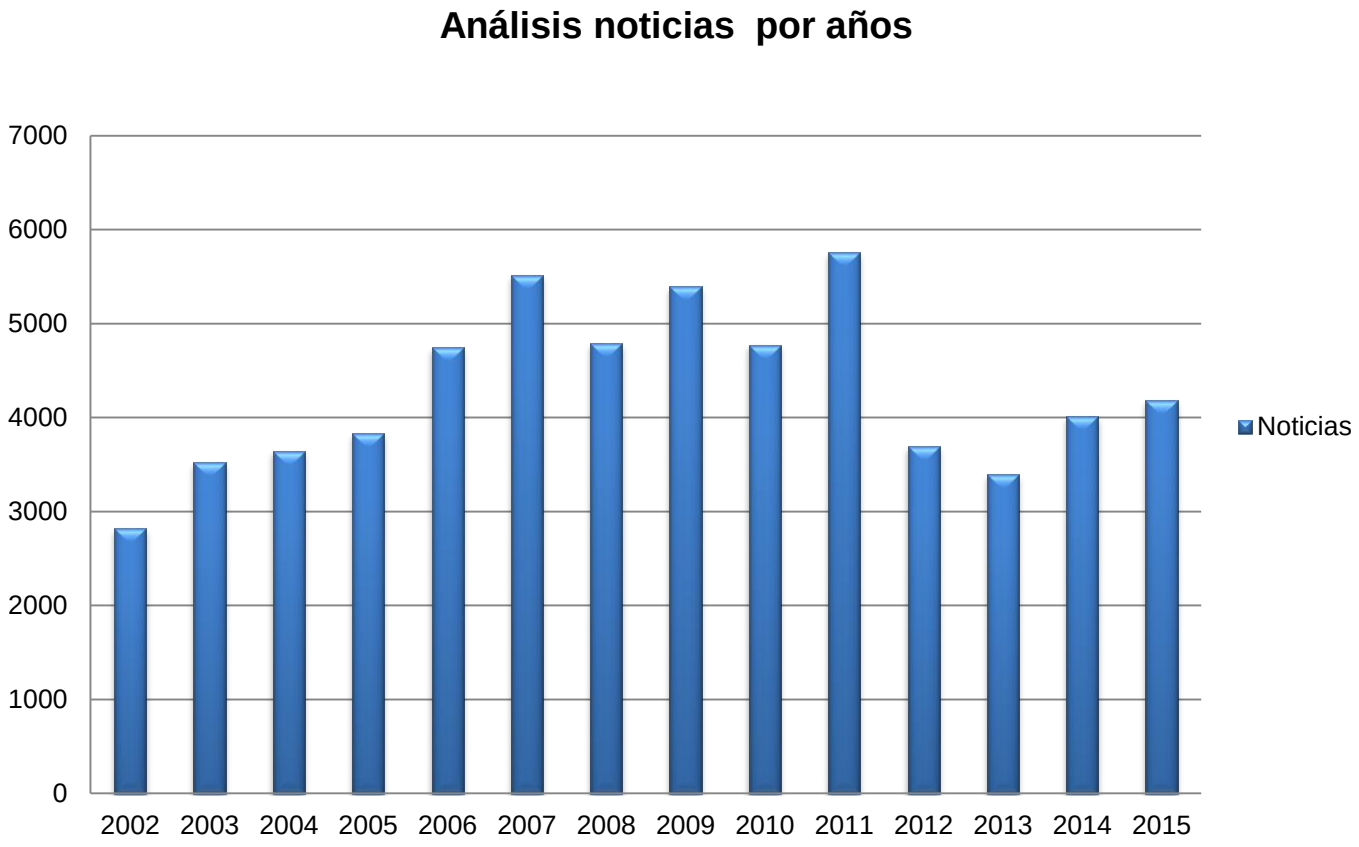


Figura 4. Gráfico noticias por años

Por último, en la Figura 5 se puede observar la evolución de los periódicos a lo largo del periodo comprendido entre en el año 2.002 y 2.015.

Empezando con el periódico EL Mundo, este tiene una evolución *in crescendo* desde el año 2.002 comenzando con apenas 440 noticias, aumentado ligeramente desde el año 2.003 hasta el año 2.005, pasando de 952 noticias hasta las 1.003 y 1.020 respectivamente. Ya en el año 2.006 se produce un gran ascenso con más del doble de noticias que se incrementaron hasta las 2.137, cifra que siguió aumentando en el año 2.007 con 3.056 noticias. Sin embargo en el año 2.008 hay un breve repunte hacia abajo con 3.005 noticias que se recuperaron en 2.009 con 3.654 noticias, y que volvieron a descender en 2.010 con 3.468. En el año 2.011 se llega al pico máximo con 4.379 noticias, y a partir de aquí comienza el descenso en los años 2.012 y 2.013 con 2.829 y 2.241 noticias respectivamente. En el año 2014 se vuelve a recuperar con 2.582, y finalmente termina en el año 2.015 con 2.669 noticias.

Por otro lado, la evolución de noticias de El País, es prácticamente contraria a la de El Mundo, pues comienza en el año 2.002 con 2.382 noticias, continua aumentando ligeramente durante los años 2.003, 2.004 y 2.005 desde las 2.570 noticias, pasando a 2.642 y finalizando con 2.814 siendo estas su punto máximo. En el año 2.006 comienza a descender con 2.609 noticias, pasando en el año 2.007 a 2.459 noticias. Ya en los años siguientes que van abarcan los años 2.008, 2.009 y 2.010 la caída es progresiva con 1.782, 1.740 y 1.301 noticias, aumentando en 2.011 con 1.377, y teniendo su mínimo en el año 2.012 con 864 noticias. En el año 2.013 se comienza a recuperar con 1.154 noticias, aumentando en 2.014 a 1.429 y finalizando en 2.015 con 1.516.

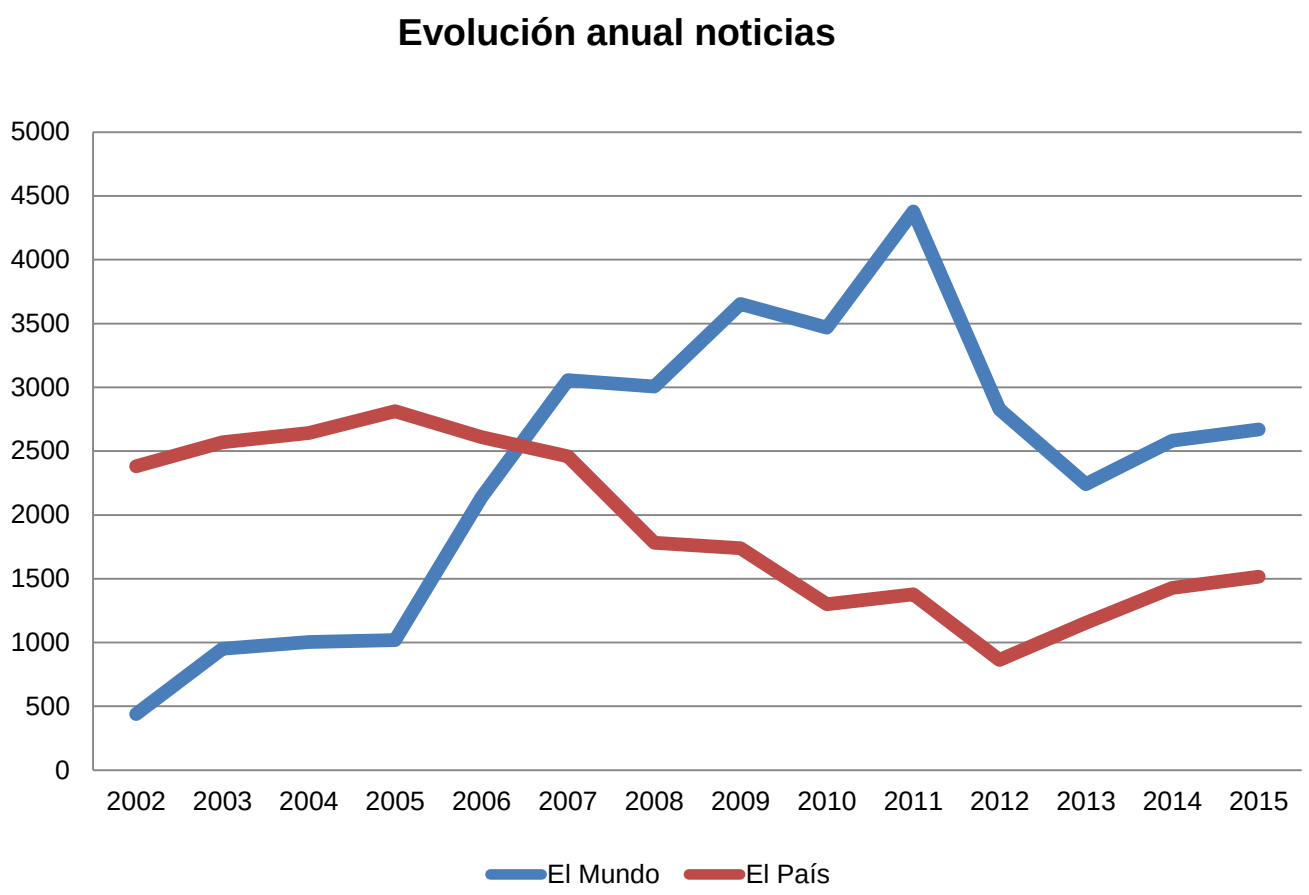


Figura 5. Gráfico evolución noticias

4. Scikit-learn

Una vez hemos determinado el algoritmo que vamos a utilizar para este proceso, véase *K-Means*, y descrito por otro lado todo el proceso de formación de nuestra colección de documentos, es el momento de hablar del software que hemos utilizado para la implementación del algoritmo.

En este caso hemos empleado un software denominado como *scikit-learn* utilizado para el lenguaje de programación de *Python*. Desarrollado en 1991 por Guido Van Rossum con licencia de código abierto, se encuentra actualmente administrado por Python Software Foundation License. Posee diversas implementaciones y entornos de desarrollo integrado, contando además con distintos servidores para sus aplicaciones (Challenger, Díaz y Becerra, 2014).

Este proyecto de *scikit-learn* nace en el año 2007 de la "*Google Summer of Code",* una convención en la que se muestran una serie de proyectos que se encuentran en su fase inicial, y allí se presentó el proyecto que estaba siendo desarrollado por David Cournapeau, y que sería apoyado por Google y otras empresas.

Posteriormente, se unió al estudio Matthieu Brucher que empezó a trabajar en él como parte de su propia tesis, pero no será hasta el año 2010 cuando Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort y Vicent Michel procedentes del centro de investigación francés INRIA (Instituto Nacional de Investigación en Informática y Automática), quienes asuman el mando del proyecto y decidieron hacer pública la primera versión a finales del año 2010. Desde ese momento se han sucedido diversas publicaciones más que lo han convertido en líder dentro de la comunidad internacional en su ámbito (Pedregosa et al, 2011).

Scikit-learn se nutre del lenguaje de programación de *Python* para el desarrollo de no pocos algoritmos de aprendizaje automático con una interfaz sencilla y fácil de utilizar (Müller y Guido, 2016). Su creación responde a la creciente demanda por personas no especialistas no solo del ámbito de la industria sino también de la ciencia, la física, la biología o la informática entre otros, para llevar a cabo el análisis de diversos datos de uso estadístico dentro del campo conocido como la minería de datos (Müller y Guido, 2016).

Este software está accesible para cualquier persona independientemente del lugar en el que se encuentre, pues está concebido como *Open Source*, es decir, como un software de código abierto o libre, distribuible comercialmente mediante la licencia BSD que permite la utilización de su código fuente en otros softwares que no son libres (VanderPlas, 2016).

Además, este software fue diseñado para poder interoperar con las librerías científicas NumPy, SciPy y matplotlib. Por otro lado, también se encuentra disponible para diversos sistemas operativos como Windows, macOS y Linux (Pedregosa et al, 2011).

En la Figura 6 que representa la imagen de un diagrama de flujo, podemos observar cómo actúa el software de *scikit-learn* con la información que le proporcionamos. Nosotros como ya hemos explicado en capítulos anteriores hemos hecho *clustering* de documentos, en este caso noticias, y aquí se encontraría reflejado el camino que realiza este software hasta realizar el proceso completo.

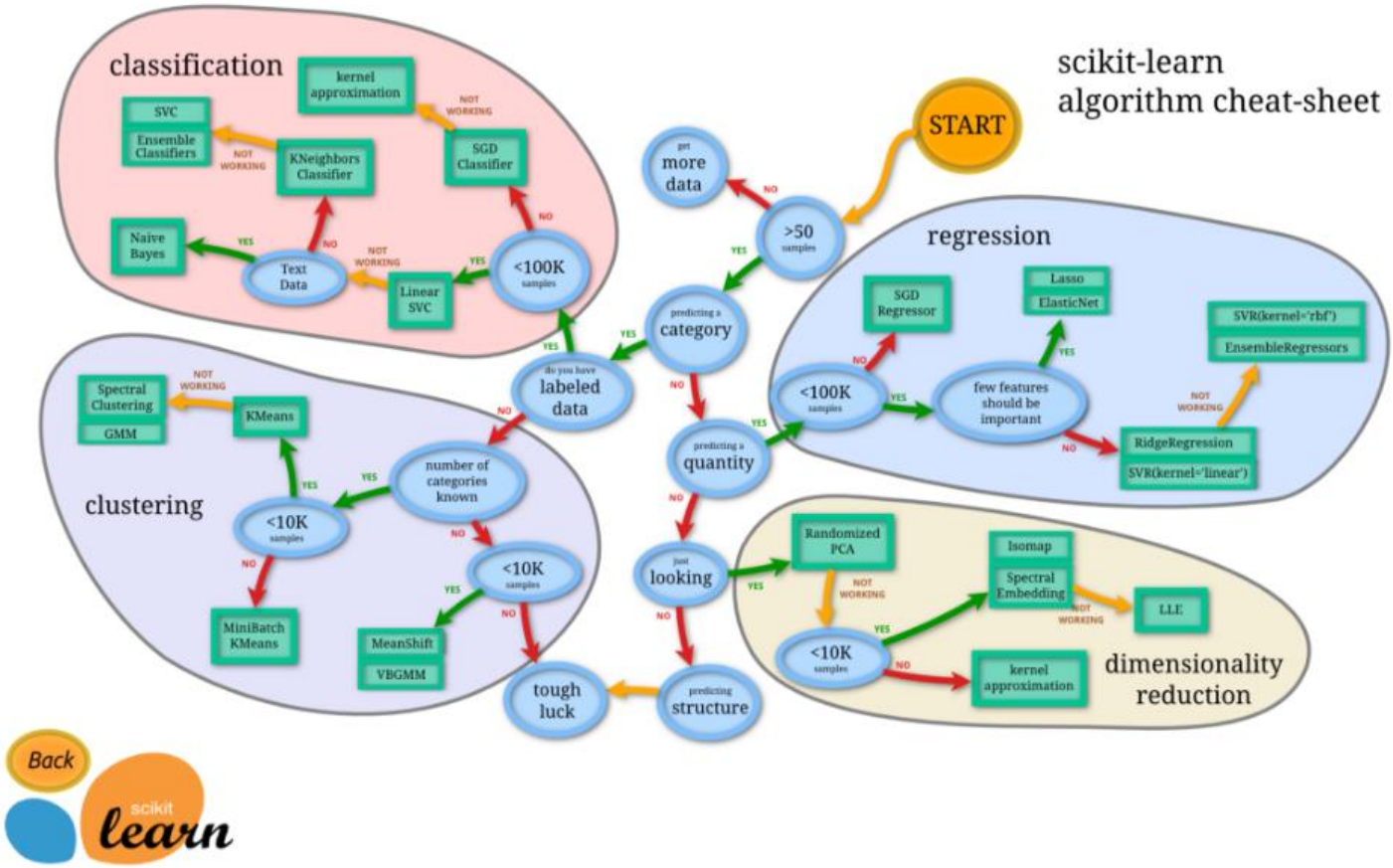


Figura 6. Imagen proceso de actuación del software Scikit-learn

5. Trabajo experimental – Gráficos

Una vez que dispusimos tanto del algoritmo de *K-Means* como del software de *Scikit-learn* implementado a través de *Python*, es el momento de poner en práctica todos los conocimientos teóricos.

En primer lugar, como ya hemos explicado en los capítulos anteriores, a pesar de que el algoritmo de *K-Means* actúa de forma automática, es necesario que introduzcamos de forma manual el número de *clusters* o temas que queremos obtener. Dada la extensa colección de documentos, hicimos varias pruebas hasta determinar que el número de *clusters* más adecuado eran 10.

Introducidos el número de *clusters* de forma manual, debemos ahora también especificar que queremos hacer un *clustering* por años, desde el año 2.002 hasta el año 2.015 de forma independiente.

Determinados los *clusters* y los años, el algoritmo nos va a proporcionar los diez temas que hemos seleccionado previamente pero no los cataloga, es decir, que el algoritmo agrupa las noticias en *clusters* pero no establece un tema general para cada uno de ellos, esta cuestión se debe determinar de forma manual y no es nada sencillo, pues establecer un tema general que englobe a todas las noticias es complicado cuanto menos.

Un ejemplo de ello lo encontramos en la Figura 7, en ella aparece reflejado un *cluster*, en este caso el número cuatro, con una serie de noticias del año 2012. Para poder clasificar las noticias tan solo disponemos de una breve descripción del titular de las mismas. El algoritmo de *K-Means* no dispone los nombres a los *clusters*, sino que por el contrario se encarga de agrupar los documentos en grupos y asignarles un número. De modo que nosotros debemos de revisar las noticias de forma manual y determinar con sus títulos un nombre general que englobe a todos los documentos de ese *cluster*. En este caso, como podemos ver en la imagen todas las noticias están relacionadas con la astronomía y el espacio, de modo que aquí podríamos hacer una primera división por ejemplo de Astronomía, pero habría que homogeneizar este *cluster* con el resto que sean parecidos, de modo que tras realizar el análisis de todos los grupos, este lo englobaríamos dentro del *cluster* de Aeronáutica/Aeroespacial, explicando el tipo de noticias que conforman el grupo.

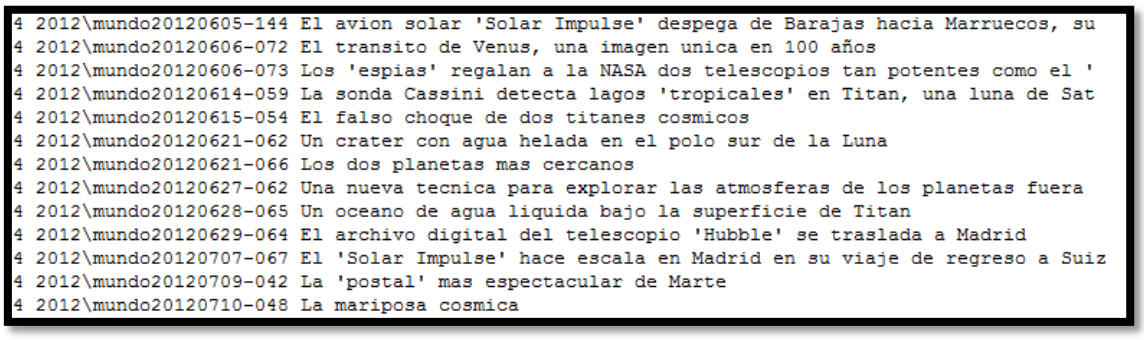


Figura 7. Imagen ejemplo de cluster

5.1. Detección de Temas

En un principio tras realizar varias pruebas, estimamos que el número de *clusters* o temas que debíamos de establecer eran 10. A continuación se encuentran dispuestos en la Tabla II la lista de *topics* o temas obtenidos tras el análisis.

Aeronáutica/Aeroespacial	Temas relacionados con la Astronomía y el Espacio.
Automóviles	Temas relacionados con el mundo del motor y sus novedades.
Cambio Climático y Medio Ambiente	Temas relacionados con el calentamiento global y sus consecuencias que afectan al medio ambiente.
Ecología	Temas relacionados con los seres vivos y su hábitat.
Fuentes de Energía	Temas relacionados con los diferentes combustibles para la obtención de energía.
Investigaciones	Temas relacionados con investigaciones amplias sobre diferentes materias relativas a la ciencia y la tecnología.
Portugués	Temas relacionados por estar escritos en portugués.
Recursos Hídricos	Temas relacionados con el agua.
Sanidad	Temas relacionados con la sanidad que incluirían todo tipo de enfermedades y tratamientos.
Tecnológicas de la Información	Temas relacionados con las grandes empresas internacionales del mundo de la tecnología y sus innovaciones.
Varios	Temas que no se han conseguido agrupar.

Tabla II. Lista de temas detectados

En la Tabla III que se muestra a continuación, se refleja una relación completa de los temas que se han obtenido en cada año. A simple vista podemos observar como en ningún año se establecieron diez temas, ya que una vez hecha la relación manual de los mismos aunque sí que se podrían haber establecido diez temas para muchos casos, eran temas demasiado específicos que se podían juntar en otros más generales y así aparece en la siguiente relación. De modo que no todos los temas aparecen en todos los años, de hecho el mínimo de temas está en 5 y el máximo en 8, oscilando la mayoría entorno a los 7.

Como se ha mencionado anteriormente estos temas son generales, pues en origen teníamos una clasificación más específica como por ejemplo la de Sanidad, que se encontraba dividida previamente en enfermedades diferenciando entre Sida, Ébola o Gripe Aviar entre otras, temas que cómo es evidente se podían y de hecho están agrupados en un tema más general que es Sanidad. En otros casos ocurre exactamente lo mismo como en el caso de las Investigaciones, divididas inicialmente en Estudios Genético-Celulares, Investigaciones Ciencia e Investigaciones Tecnología. En el caso de las Tecnologías de la Información nos encontramos con el mismo paradigma, con subdivisiones en diferentes empresas como Microsoft, Google o Apple. En el *cluster* de Aeronáutica/Aeroespacial tenemos el mismo caso, aquí había temas que eran exclusivos de la Nasa y otros que solo hablaban de Astronomía, por ello se englobaron en un único tema todos juntos.

Por último, nos encontramos con otros casos particulares como el tema de los Automóviles, que solo ha aparecido en un *cluster* bastante claro, o el de Portugués en otros dos. Esto no significa que no haya más *clusters* que contengan noticias relacionadas con estos temas, sino que muchas veces la gran cantidad de noticias que aparecían en un solo *cluster* eran tan variadas y diferentes que resultaba imposible establecer un tema general para todas, por ello hemos establecido un tema que es Varios, dónde se encuentran englobadas noticias que podrían pertenecer a muchos de los temas generales previamente establecidos.

Año	Nº de temas	Temas
2002	**7**	**Tecnologías de la Información, Varios, Aeronáutica/Aeroespacial, Fuentes de Energía, Sanidad, Investigaciones y Ecología.**
2003	**5**	**Aeronáutica/Aeroespacial, Investigaciones, Ecología, Tecnologías de la Información Tecnológicas, Sanidad.**
2004	**6**	**Tecnologías de la Información, Investigaciones, Aeronáutica/Aeroespacial, Cambio Climático y Medio Ambiente, Recursos Hídricos y Sanidad.**
2005	**6**	**Recursos Hídricos, Investigaciones, Sanidad, Cambio Climático y Medio Ambiente, Fuentes de Energía y Aeronáutica/Aeroespacial.**

2006	**8**	**Cambio Climático y Medio Ambiente, Sanidad, Ecología, Tecnologías de la Información, Varios, Investigaciones, Recursos Hídricos y Aeronáutica/Aeroespacial.**
2007	**7**	**Investigaciones, Cambio Climático y Medio Ambiente, Aeronáutica/Aeroespacial, Fuentes de Energía, Tecnologías de la Información, Varios y Sanidad.**
2008	**7**	**Varios, Investigaciones, Aeronáutica/Aeroespacial, Tecnologías de la Información, Cambio Climático y Medio Ambiente, Fuentes de Energía y Sanidad.**
2009	**7**	**Varios, Cambio Climático y Medio Ambiente, Aeronáutica/Aeroespacial, Fuentes de Energía, Investigaciones, Sanidad y Tecnologías de la Información.**
2010	**7**	**Tecnologías de la Información, Ecología, Varios, Fuentes de Energía, Sanidad, Automóviles y Aeronáutica/Aeroespacial.**
2011	**6**	**Varios, Investigaciones, Aeronáutica/Aeroespacial, Tecnologías de la Información, Sanidad y Fuentes de Energía.**
2012	**7**	**Aeronáutica/Aeroespacial, Investigaciones, Sanidad, Varios, Ecología, Tecnologías de la Información y Fuentes de Energía.**
2013	**7**	**Tecnologías de la Información, Investigaciones, Fuentes de Energía, Ecología, Sanidad, Varios y Aeronáutica/Aeroespacial.**
2014	**6**	**Portugués, Sanidad, Cambio Climático y Medio Ambiente, Tecnologías de la Información, Aeronáutica/Aeroespacial y Varios.**
2015	**7**	**Aeronáutica/Aeroespacial, Tecnologías de la Información, Portugués, Varios, Cambio Climático y Medio Ambiente, Sanidad y Ecología.**

Tabla III. Relación completa de temas por año

5.1.1. Porcentaje de temas

En la siguiente Figura 8 se muestran las relaciones de temas y su proporción global.

Podemos observar que los temas menos tratados son los de Automóviles con apenas 113 noticias, Portugués con 168 y Recursos Hídricos con 451. En el primer caso, ha resultado totalmente imposible englobar este pequeño *cluster* que estaba bastante definido en otro más grande y por eso aparece solo, aunque bien es cierto que en otros *clusters* como alguno de los Varios, hay muchas noticias que también son de automóviles pero debido a la gran variedad temática no se ha podido establecer un tema único. En el caso de Portugués ocurre prácticamente lo mismo, solo que en este *cluster* es bastante probable que el sistema haya agrupado todas las noticias por el idioma, en este caso portugués, de ahí que el *cluster* general lleve ese nombre. Por último aparecen los Recursos Hídricos, que ha suscitado un interés menor para los periódicos pero aun así ha sido un tema que se ha tratado.

Luego tenemos temas como las Fuentes de Energía, que sí que han suscitado el interés de la prensa con 1.829 noticias, las Tecnologías de la Información con 2.472 noticias, y el Cambio Climático y el Medio Ambiente, un tema muy recurrido con 2.883 noticias.

También tenemos otros temas intermedios como Aeronáutica/Aeroespacial y Ecología, con 3.967 y 4.511 noticias respectivamente, temas recurrentes por la prensa a lo largo de estos 14 años.

Sin embargo, los temas que más se han tratado han sido los de Investigaciones con 9.395 noticias, y los de Sanidad con 15.743 noticas. Esto es debido, como ya hemos mencionado anteriormente, a que son temas generales que tratan otros más específicos y que se han englobado dentro de estos otros.

Por último nos encontramos con Varios con 18.542 noticias, que no significa que hablen de un solo tema sino que debido al gran volumen y variedad de noticias que acumulan en su interior, ha resultado totalmente imposible determinar un tema general que logre englobar a todas las noticias, y por ello han recibido el nombre de Varios.

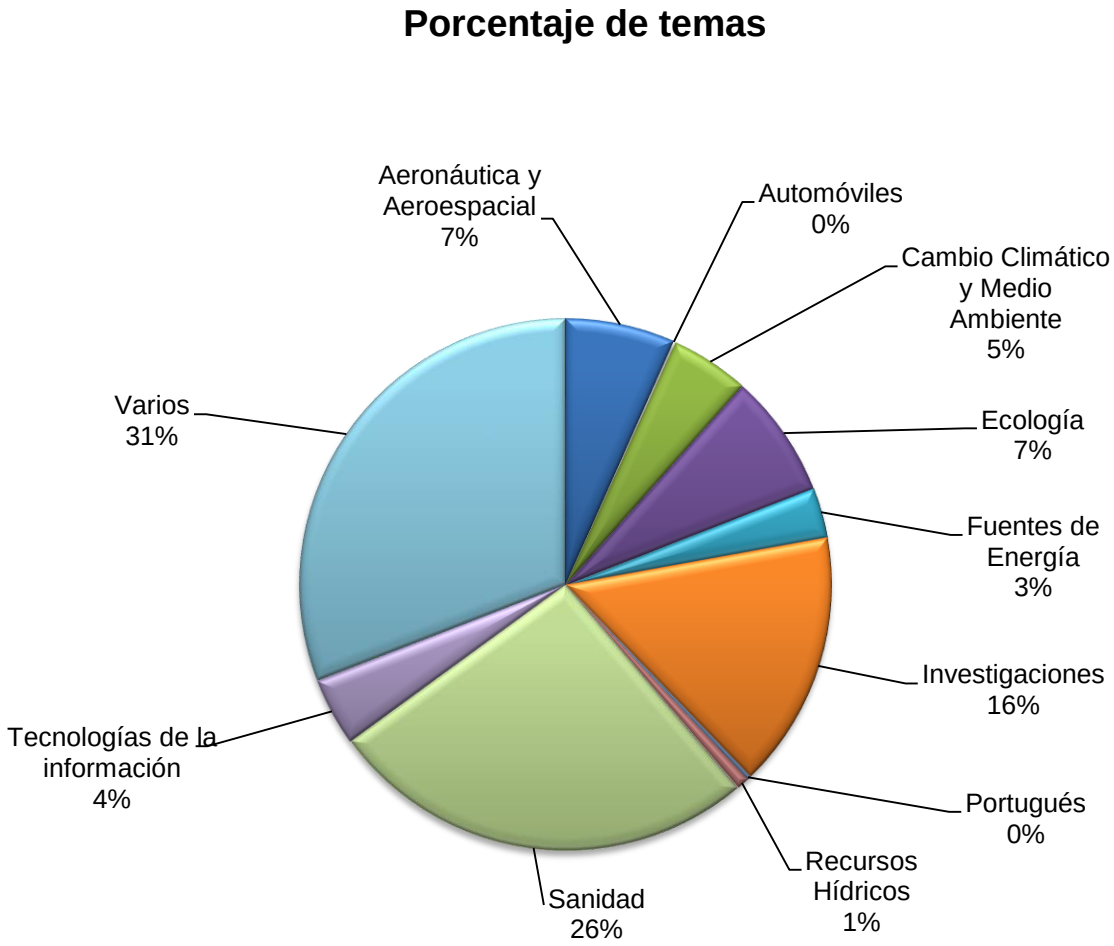


Figura 8. Gráfico porcentaje temas

5.2. Análisis temático

En las siguientes figuras aparecen reflejadas unas descripciones breves de cada *cluster*, el tipo de noticias que alberga a través de un texto en formato plano, con las noticias de El Mundo y El País elegidas de forma aleatoria dentro del *cluster* seleccionado, así como la evolución de los temas desde el año 2.002 hasta el año 2.015 incluido.

Empezando por el *cluster* Aeronáutica/Aeroespacial, podemos observar a través de la Figura 9 cómo es un tema que ha sido fácil de detectar pues la mayoría de las noticias están relacionadas con la exploración espacial, misiones, investigaciones relacionadas y la astronomía.

Aunque esta y el resto de imágenes que aparecen para cada tema solo reflejan una pequeña muestra del contenido de cada año, sí que permite visualizar en este caso los *clusters* que van con números correlativos representados en la primera columna. Por otro lado nos encontramos con el año analizado (2003) y con las noticias de los periódicos (EL Mundo y El País). Por último nos encontraríamos con los títulos de las noticias.

Como se puede observar, determinar temas generales para cada *cluster* atendiendo tan solo a los breves títulos, que en muchos casos son muy confusos, es una tarea que no es nada fácil.

```
1 2003\mundo20031105-035 La NASA asegura que la sonda 'Spirit' funcionara en Marte a pesar de
1 2003\mundo20031108-009 La nave 'Voyager I', muy proxima al fin del Sistema Solar
1 2003\mundo20031203-028 La NASA advierte de que la proxima mision a Marte sera complicada y
1 2003\mundo20031209-011 Japon esta a punto de abandonar su mision espacial a Marte
1 2003\mundo20031218-034 La sonda europea Mars Express inicia su fase mas critica
1 2003\mundo20031221-016 Rusia enviara al tercer 'turista espacial' a la ISS en octubre
1 2003\mundo20031222-009 El 'Beagle2' se separa con exito de la sonda 'Mars Express'
1 2003\mundo20031222-016 Luna de miel en la ISS
1 2003\mundo20031229-012 El jefe de mision de la NASA asegura que el descenso a Marte 'son sei
1 2003\mundo20031231-024 Lanzan con exito la 'Sonda N-1', la primera mision espacial eurochin
1 2003\mundo20031231-027 La Mars Express cambia de orbita para facilitar un posible contacto c
1 2003\pais20030115-040 Se retrasa el lanzamiento de 'Rosetta' hacia un cometa
1 2003\pais20030129-068 Nuevos paseantes para Marte
1 2003\pais20030203-090 Despues de la tragedia
1 2003\pais20030204-104 China mantiene su plan de hacer una mision espacial tripulada
1 2003\pais20030210-045 La aventura espacial
1 2003\pais20030213-046 La NASA insiste en que no detecto anomalias previas en el 'Columbia'
1 2003\pais20030223-081 El ala del 'Columbia' fue alcanzada por tres fragmentos
1 2003\pais20030301-120 La tripulacion del 'Columbia' estaba tranquila, segun un video de l
1 2003\pais20030402-077 La investigacion del 'Columbia' apunta a un problema del ala
```

Figura 9. Imagen cluster Aeronáutica/Aeroespacial

Pasando al análisis de la evolución del *cluster*, podemos ver reflejado en la Figura 10 como su recorrido a lo largo de estos años se ha mantenido de forma constante en ligero ascenso. Comienza en el año 2.002 con 133 noticias ascendiendo hasta el año 2.004 con 175 noticias, desciende después hasta 2.006 con 161, vuelve a ascender en 2.007 descendiendo ligeramente los siguientes dos años, teniendo un pico en 2.010 con 366 noticias, descendiendo el siguiente año y aumentando de forma progresiva hasta 2.013 con 441 noticias. Finalmente en 2.010 comienza ligeramente a descender quedándose en 2.015 con 408 noticias.

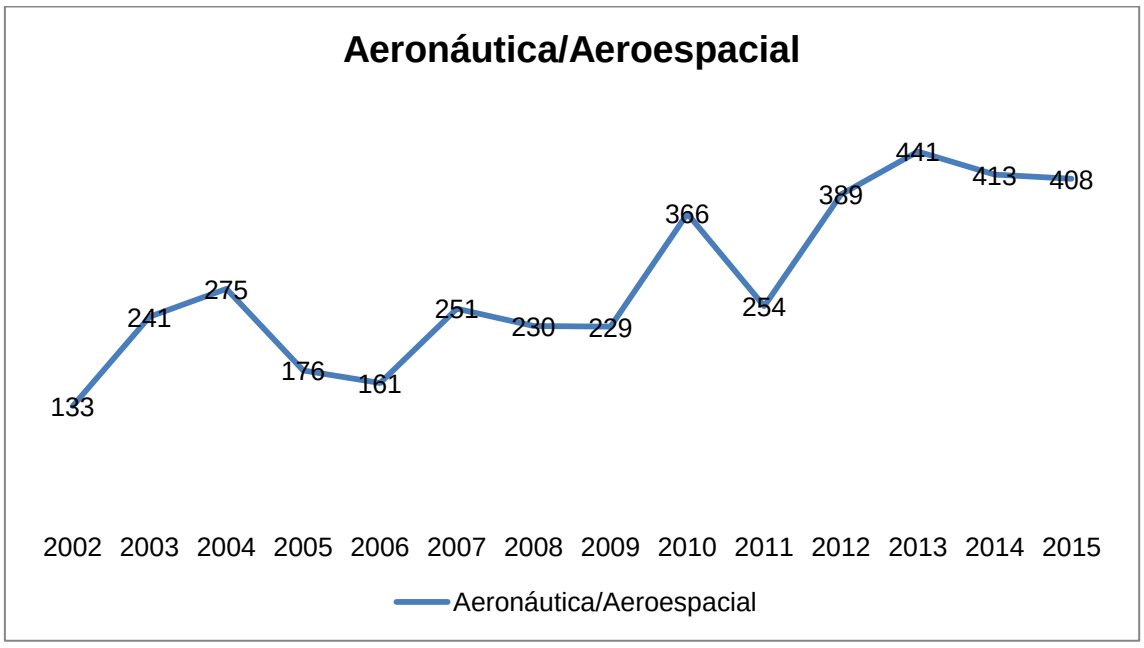


Figura 10. Gráfico noticias Aeronáutica/Aeroespacial

En segundo lugar nos encontramos con el *cluster* de Automóviles, que podemos ver reflejado su composición en la Figura 11. En ella vemos cómo la totalidad de las noticias hablan sobre las diferentes marcas de automóviles y sus respectivas novedades en relación a sus nuevos modelos de vehículos.

```
6 2010\mundo20100107-151 Delhi: el salon de los coches 'low cost'
6 2010\mundo20100111-147 Peugeot: futuro hibrido y deportivo
6 2010\mundo20100118-152 Nissan Cube: inspirado en Jennifer Lopez
6 2010\mundo20100119-140 Mini 2010: mas emocion, menos emisiones
6 2010\mundo20100120-140 BMW afina el Coupe y el Cabrio de la Serie 3
6 2010\mundo20100121-121 Countryman, el Mini campero
6 2010\mundo20100122-106 Citroen explora su lado chic con el DS3
6 2010\mundo20100128-141 Renault Master: mas oferta y economia
6 2010\mundo20100130-086 Montero 2010: una 'bestia' en tierra y asfalto
6 2010\mundo20100208-140 ¿Su automovil mira por el bolsillo?
6 2010\mundo20100210-111 Los Volvo V70 y S80 mas ecologicos
6 2010\mundo20100210-112 Volvo S60: la nueva cara de Volvo
6 2010\mundo20100218-140 Toyota Auris 2010: mejorado y mas juvenil
6 2010\mundo20100220-090 GTI: El Polo mas deportivo
6 2010\mundo20100222-134 Otro Opel electrico
6 2010\mundo20100314-109 Fiat Panda 4x4 y Sedici: eficacia y ahorro
6 2010\mundo20100315-149 Ford S-Max y Galaxy: apuesta por el cambio automatico
6 2010\mundo20100316-144 Audi RS5: deportividad sostenible
6 2010\mundo20100317-135 Peugeot RCZ: el coupe accesible
6 2010\mundo20100317-136 Peugeot RCZ: fuera de serie
```

Figura 11. Imagen cluster Atomóviles

Por otro lado, es de destacar que en el *cluster* de Automóviles no hemos podido establecer una consecución en el tiempo como se refleja en la Figura 12, solo aparece en el año 2.010 con 113 noticias, no volviendo a hacer acto de presencia ni anterior ni posteriormente a esa fecha.

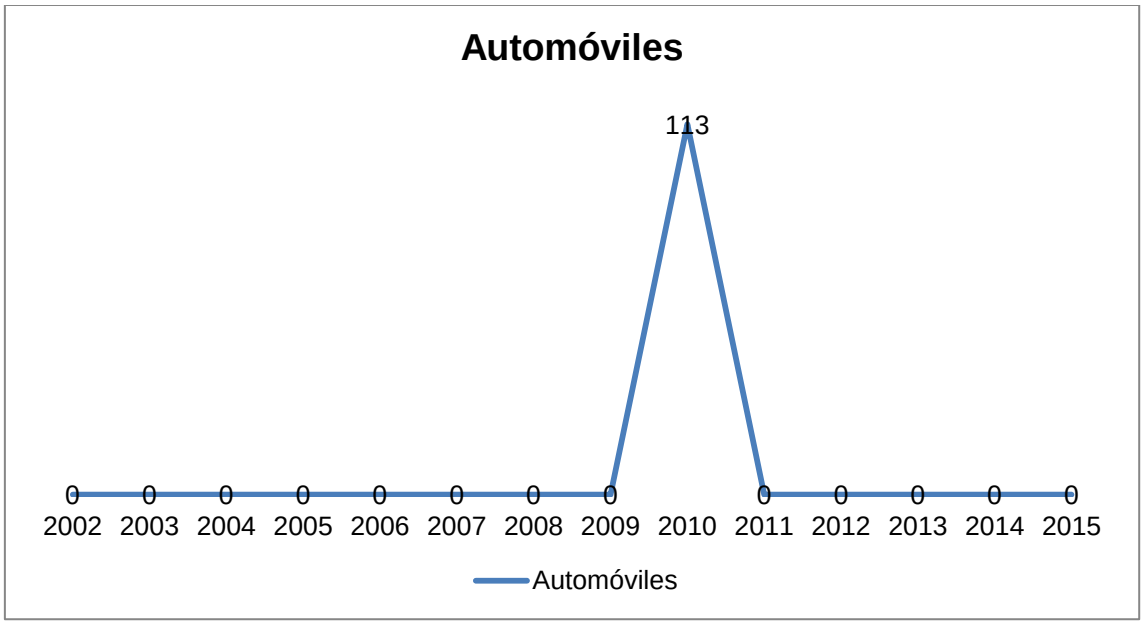


Figura 12. Gráfico noticias Automóviles

En tercer lugar hemos detectado el *cluster* de Cambio Climático y Medio Ambiente que se ve reflejado en la Figura 13. La mayoría de las noticias informan sobre el calentamiento global y el cambio climático, sus efectos y las medidas que se están llevando a cabo para tratar de frenarlo.

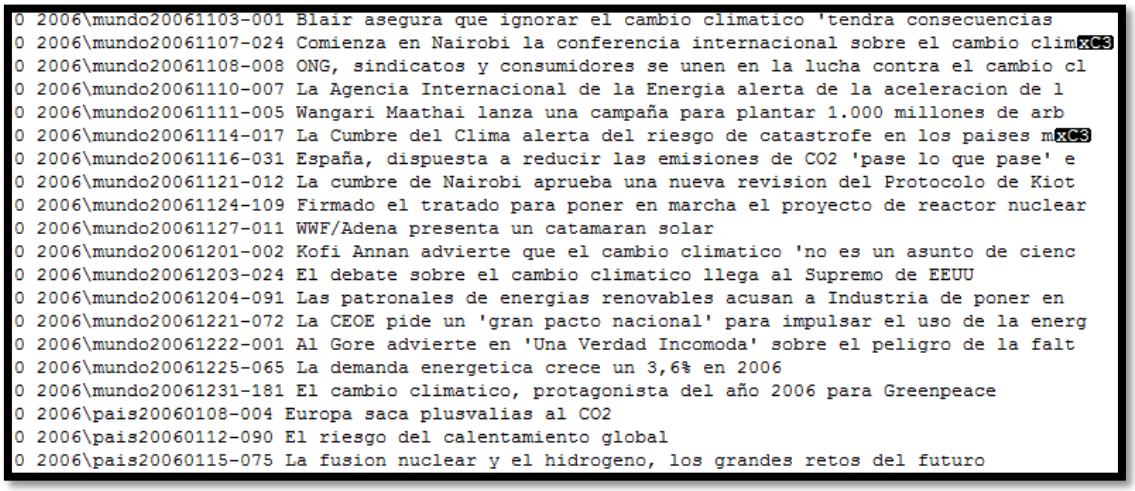


Figura 13. Imagen cluster Cambio Climático y Medio Ambiente

Este *cluster de* Cambio Climático y Medio Ambiente, aparece reflejado en la Figura 14 de manera intermitente comenzando en el año 2.004 con 126 noticias, continuando en ascenso hasta el año 2.008 con 988 noticias encontrando aquí su máximo pico, ya que el siguiente año desciende hasta las 376 noticias. Después, no será hasta el año 2.014 cuando vuelva a aparecer en escena con 157 noticias, finalizando el año 2.015 con 262 noticias.

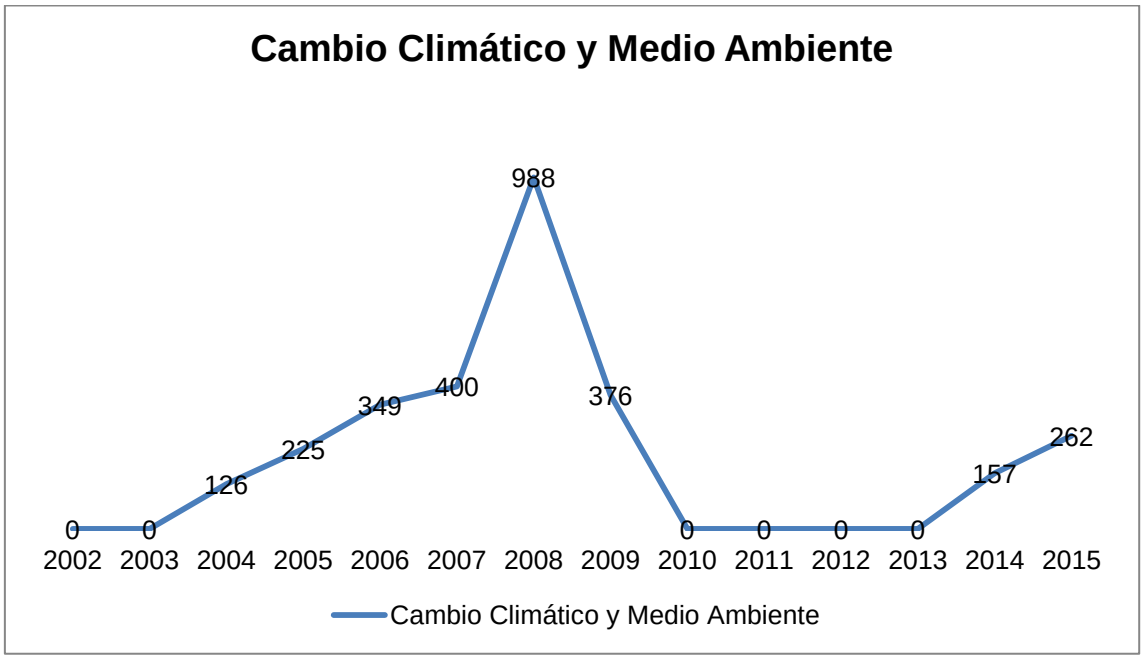


Figura 14. Gráfico noticias Cambio Climático y Medio Ambiente

En cuarto lugar hayamos el *cluster* de Ecología. Cómo vemos en la Figura 15, las noticias están relacionadas con los seres vivos y sus relaciones con el medio ambiente. Por otro lado, también hay noticias de ecología que se vinculan a la prehistoria y muchas más relacionadas con la biodiversidad, pero que hemos considerado oportuno englobar en el mismo *cluster* al estar muy relacionadas entre sí.

```
1 2010\mundo20100726-031 Una ruta nocturna divulgara los valores ambientales del parque natura
1 2010\mundo20100727-024 Bruselas pide rebajar las medidas contra el mal de las 'vacas locas' a
1 2010\mundo20100727-037 Nace en Fuengirola una cria de talapoin norteño, el mas pequeño p
1 2010\mundo20100727-040 Localizan herramientas de hominidos anteriores al 'Homo Antecessor'
1 2010\mundo20100727-057 Vuelven la nutrias al Delta del Ebro
1 2010\mundo20100727-172 La enfermedad de las 'vacas locas' ha afectado a 217 personas en Europ
1 2010\mundo20100728-035 'Mini-vacas' para proteger el medio ambiente
1 2010\mundo20100801-025 Una rata gigantesca mas grande que un gato
1 2010\mundo20100801-026 Canarias, esperanza para un guacamayo extinto en Brasil
1 2010\mundo20100801-030 El mar se queda sin plancton
1 2010\mundo20100801-047 Alerta en el Cantabrico por las picaduras de la peligrosa 'carabela p
1 2010\mundo20100802-020 La 'epidemia' de las ballenas varadas en la Patagonia
1 2010\mundo20100802-030 La desproteccion de los tiburones en Canarias
1 2010\mundo20100803-025 La joya amenazada del mar Mediterraneo
1 2010\mundo20100803-047 Encuentran las huellas de reptil mas antiguas
1 2010\mundo20100803-048 La unica ballena con olfato
1 2010\mundo20100804-031 Arañas que 'tienden puentes'
1 2010\mundo20100804-033 Las nuevas joyas de la naturaleza Patrimonio de la Humanidad
1 2010\mundo20100805-038 Los ecologistas denuncian ante el Seprona el estado de las pinadas de
```

Figura 15. Imagen cluster Ecología

En relación al *cluster* de Ecología reflejado en la Figura 16, abre el año 2.002 con 288 noticias, y curiosamente su máximo lo tiene al año siguiente con 1.070 noticias. Los dos siguientes años no hay constancia de noticias de ecología, reapareciendo en 2.006 con 828 noticias, desapareciendo de nuevo hasta 2.010 con 636. En 2.011 tampoco aparece, y reaparece en 2.012 con 739, descendiendo en 2.013 hasta 638 y desapareciendo en 2.014. Finaliza el año 2.015 con 312 noticias.

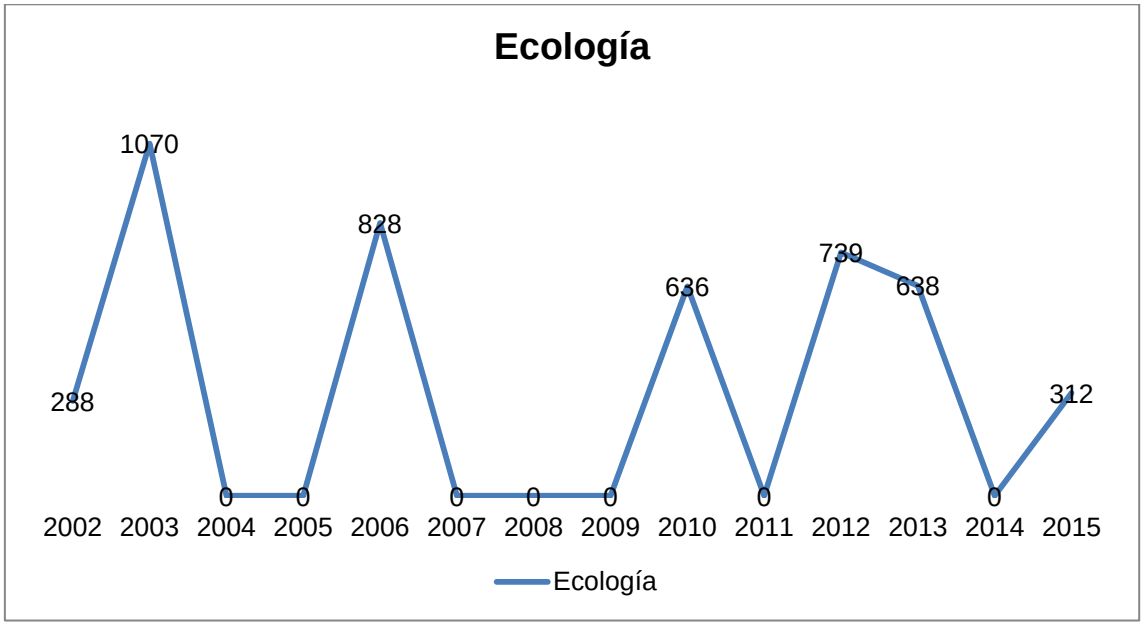


Figura 16. Gráfico noticias Ecología

En quinto lugar tenemos el *cluster* de Fuentes de Energía, y cómo podemos ver reflejado en la Figura 17 los temas que se tratan en él están todos relacionados con la energía eléctrica, el carbón, el viento o la energía solar entre muchas otras más.

```
3 2010\pais20100912-028 Iberdrola Renovables crece en el exterior
3 2010\pais20100913-062 La agonia del carbon
3 2010\pais20100918-107 "Zarra, 304 puntos; Asco 300"
3 2010\pais20100919-030 Renovables hasta en el gallinero
3 2010\pais20100926-122 Una oportunidad para todos
3 2010\pais20100928-069 La eolica en Canarias ya es mas barata que la electrica convenciona
3 2010\pais20101002-009 ¿Y si aliamos agua y energia?
3 2010\pais20101004-006 Bruselas comprende
3 2010\pais20101010-041 Siete propuestas al Gobierno sobre energia electrica
3 2010\pais20101027-062 EE UU refuerza su apuesta solar con la mayor planta del mundo
3 2010\pais20101029-020 Ayudar al carbon desviste a otros
3 2010\pais20101031-044 Redes electricas para un desarrollo sostenible
3 2010\pais20101101-038 Un molino de viento casi tan alto como Torrespaña
3 2010\pais20101104-038 Bruselas abre la via para exportar residuos nucleares
3 2010\pais20101110-019 La subvencion a energia fosil es cinco veces mayor que la de renova
3 2010\pais20101110-054 El PP trunca el pacto sobre el impuesto verde a los carburantes
3 2010\pais20101110-075 "La fusion ha dejado de ser un sueño"
3 2010\pais20101117-004 "Con el mayor laser del mundo estudiaremos la fisica en miniatura"
3 2010\pais20101118-003 La AIE advierte del efecto "devastador" de ahorrar en renovables
3 2010\pais20101119-074 El Gobierno examina hoy los recortes a la energia solar
```

Figura 17. Imagen noticias Fuentes de Energía

Las Fuentes de Energía es un tema también intermitente que aparece reflejado en la Figura 18 comenzando en el año 2.002 con apenas 86 noticias, no volviendo a aparecer hasta 2.005 con 45. Al año siguiente tampoco aparece y reaparece en 2.007 con 223, desciende en 2.008 hasta 136, y aumenta progresivamente en 2.009 desde 233 noticias hasta las 496 del año 2011, su máximo. El siguiente año desciende hasta las 112, aumentando en 2.013 hasta las 253 noticias no volviendo a aparecer los dos últimos años.

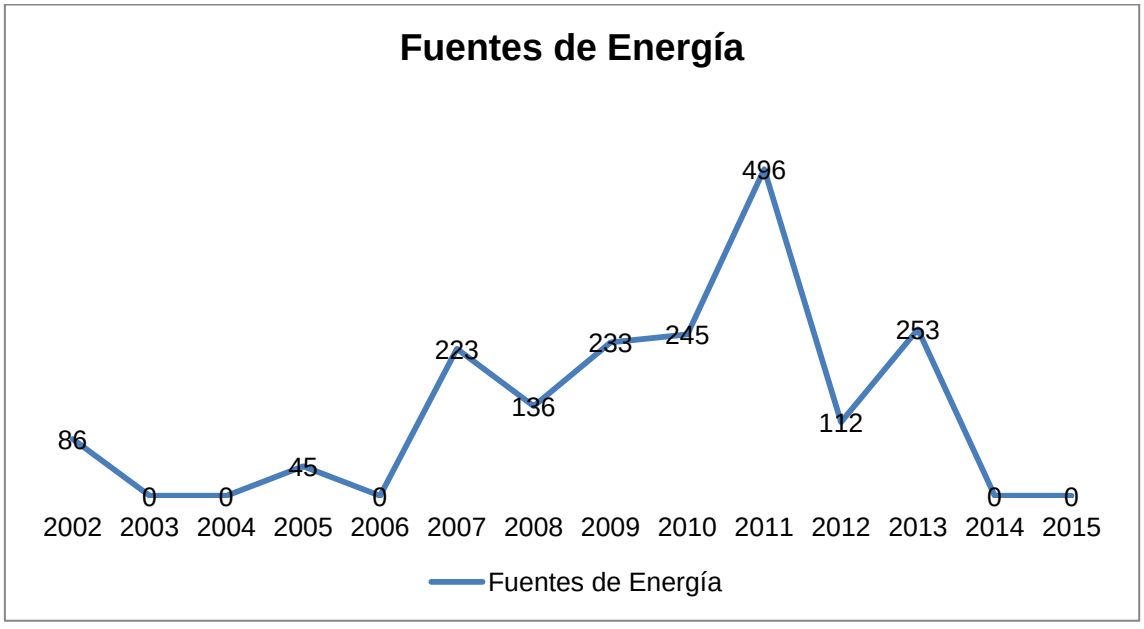
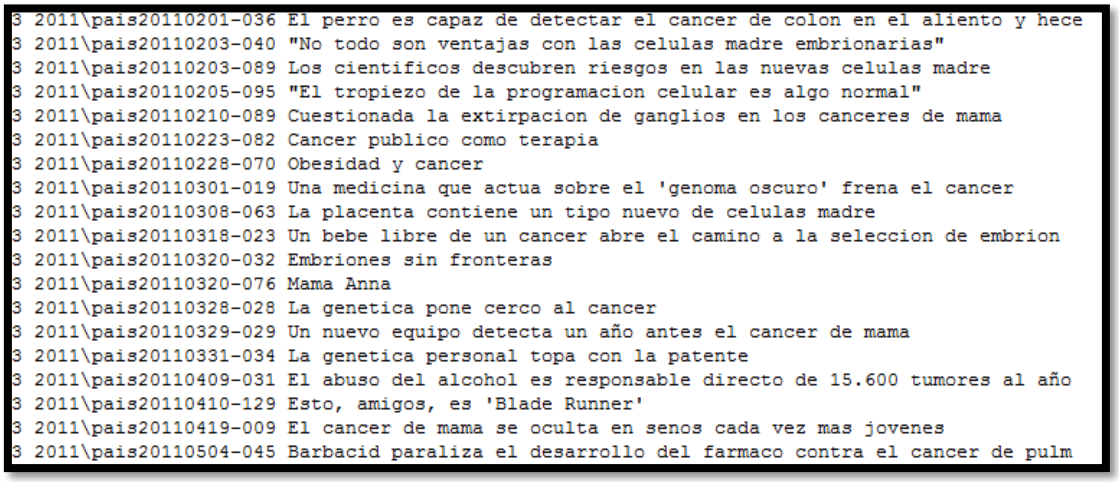


Figura 18. Gráfico noticias Fuentes de Energía

En sexto lugar está el *cluster* de Investigaciones reflejado en la Figura 19, aquí podemos ver como la mayoría de estudios en este ejemplo concreto tratan sobre el cáncer, sin embargo hay muchos más temas como los estudios genético-celulares, investigaciones sobre la ciencia o en materia tecnológica, todos ellos agrupados bajo este mismo *cluster*.



```
3 2011\pais20110201-036 El perro es capaz de detectar el cancer de colon en el aliento y hece
3 2011\pais20110203-040 "No todo son ventajas con las celulas madre embrionarias"
3 2011\pais20110203-089 Los cientificos descubren riesgos en las nuevas celulas madre
3 2011\pais20110205-095 "El tropiezo de la programacion celular es algo normal"
3 2011\pais20110210-089 Cuestionada la extirpacion de ganglios en los canceres de mama
3 2011\pais20110223-082 Cancer publico como terapia
3 2011\pais20110228-070 Obesidad y cancer
3 2011\pais20110301-019 Una medicina que actua sobre el 'genoma oscuro' frena el cancer
3 2011\pais20110308-063 La placenta contiene un tipo nuevo de celulas madre
3 2011\pais20110318-023 Un bebe libre de un cancer abre el camino a la seleccion de embrion
3 2011\pais20110320-032 Embriones sin fronteras
3 2011\pais20110320-076 Mama Anna
3 2011\pais20110328-028 La genetica pone cerco al cancer
3 2011\pais20110329-029 Un nuevo equipo detecta un año antes el cancer de mama
3 2011\pais20110331-034 La genetica personal topa con la patente
3 2011\pais20110409-031 El abuso del alcohol es responsable directo de 15.600 tumores al año
3 2011\pais20110410-129 Esto, amigos, es 'Blade Runner'
3 2011\pais20110419-009 El cancer de mama se oculta en senos cada vez mas jovenes
3 2011\pais20110504-045 Barbacid paraliza el desarrollo del farmaco contra el cancer de pulm
```

Figura 19. Imagen cluster Investigaciones

El tema de las Investigaciones reflejado en la Figura 20, es un tema que ha sido bastante tratado ya que al agrupar diferentes categorías los resultados son mucho mayores, por ello comienza en el año 2.002 con 488 noticias aumentado hasta el año 2.005 con 2.422, su máximo. A partir de aquí desciende hasta el año 2.007 con 138, aumentando el siguiente año hasta 1.290, y descendiendo en 2.008 hasta las 123 noticias, desapareciendo en 2.010. Reaparece en 2.011 con 352 y desciende progresivamente hasta el año 2.013 con 152, no mostrándose más en 2.014 y 2.015.

Figura 20. Gráfico noticias investigaciones

En séptimo lugar tenemos el *cluster* de Portugués, el cual ha resultado bastante sencillo de detectar, en el sentido de que como se puede ver en la Figura 21 todas las noticias de este *cluster* están escritas en portugués, aunque bien es cierto que las noticias no tienen nada que ver unas con otras. La totalidad de las noticias proceden del periódico El País, y estas noticias probablemente aparezcan porque desde hace algunos años este periódico tiene una sección digital para Brasil y las noticias están escritas en portugués.

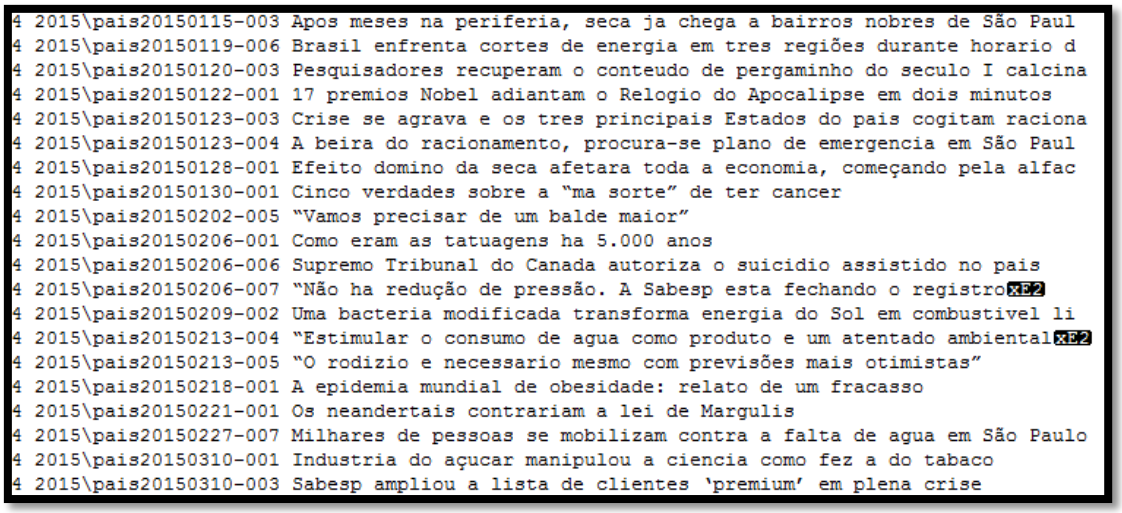


Figura 21. Imagen cluster Portugués

En el caso del tema de Portugués ocurre algo parecido que con el tema de Automóviles, y es que como se puede ver en la Figura 22 aparece en 2.014 con 74 noticias y aumenta en 2.015 hasta las 94, no apareciendo antes de estas fechas.

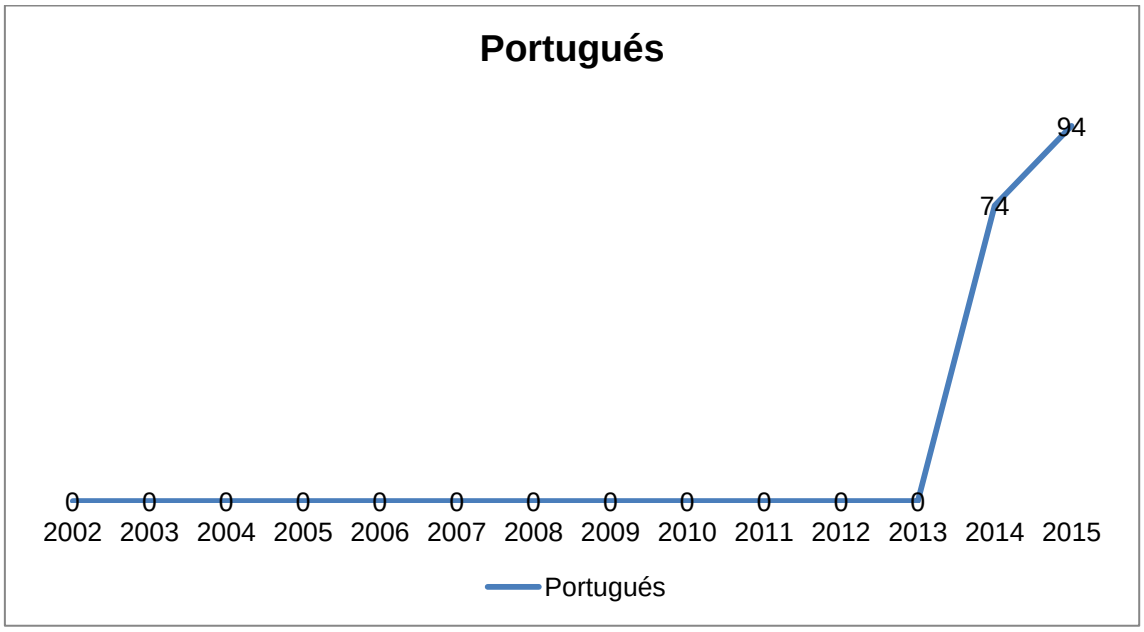


Figura 22. Gráfico noticias Portugués

En octavo lugar podemos visualizar en la Figura 23 el *cluster* de Recursos Hídricos. Cómo se puede ver reflejado todas las noticias están relacionadas con el agua, ya sea cuándo se habla de las reservas o de los trasvases.

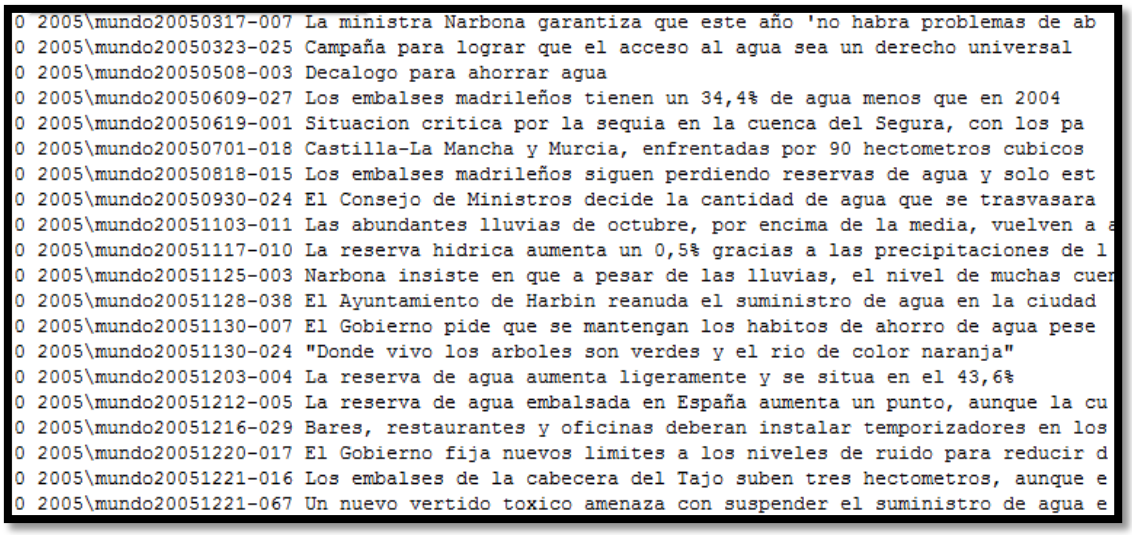


Figura 23. Imagen cluster Recursos Hídricos

Los Recursos Hídricos reflejados en Figura 24 apenas aparecen en tres años de forma consecutiva aunque no progresiva. Comienzan en el año 2.004 con 114 noticias, aumentando el año siguiente con 207 y finalizando en 2.006 con 130 noticias.

Figura 24. Gráfico noticias Recursos Hídricos

En noveno lugar tenemos el *cluster* de Sanidad, reflejado en la Figura 25 dónde podemos ver en este caso que sirve de ejemplo para los demás, que la mayoría de noticias tratan sobre enfermedades y los posibles tratamientos a llevar a cabo.

```
0 2003\pais20030127-066 Un grupo español logra una vacuna que protege a los perros contra la
0 2003\pais20030127-106 La cosecha mundial de transgenicos crecio un 12% en 2002
0 2003\pais20030127-108 Despilfarro farmaceutico
0 2003\pais20030128-020 Laboratorios y OMC acercan posturas sobre farmacos baratos para pais
0 2003\pais20030128-101 Comienza la dispensacion compasiva de un nuevo grupo de farmacos con
0 2003\pais20030129-031 La OMS elige al coreano Lee como director
0 2003\pais20030129-093 Las violaciones disparan el sida entre las niñas africanas
0 2003\pais20030202-067 El gasto en farmacos genericos ha pasado en seis años de 378.000 eu
0 2003\pais20030203-035 El alto tribunal gallego ordena devolver la pension a un enfermo de
0 2003\pais20030204-025 "La industria farmaceutica es el enemigo numero uno de los farmaceu
0 2003\pais20030210-073 El comite etico europeo pide limitar los ensayos clinicos en paise
0 2003\pais20030214-017 El Reino Unido quiere imponer pruebas de sida a los inmigrantes
0 2003\pais20030214-066 Fronteras con seguro
0 2003\pais20030215-102 Un virus inocuo para los humanos frena el desarrollo del sida
0 2003\pais20030218-051 La OMS propone subir el precio y limitar la publicidad en la primera
0 2003\pais20030220-141 EE UU impide el acuerdo sobre la venta de farmacos baratos a paises
0 2003\pais20030221-125 Sida en casa del ministro
0 2003\pais20030224-058 Un informe preve que las medidas antitabaco no destruiran empleo
0 2003\pais20030225-099 Fracasa el primer ensayo a gran escala de una vacuna contra el sida
```

Figura 25. Imagen cluster Sanidad

El tema más recurrido cómo se pude ver reflejado en la Figura 26 es el de sanidad, que con diversas fluctuaciones es el tema de mayor magnitud de entre los clasificados. Comienza en 2.002 con 551 noticias, aumentando hasta 1.203 en 2.003, descendiendo los dos siguientes años hasta las 759 noticias, remontando en 2.005 dónde comienza ascendiendo hasta el año 2.006 con 1.319 noticias, desciende ligeramente en 2.007, y en 2.008 tiene su máximo con 2.458 noticias. A partir de aquí comienza su declive hasta el año 2.013 dónde encontramos su mínimo en 545 noticias. El siguiente año aumenta hasta las 1.172, y finaliza en el año 2.015 con 1.162 noticias.

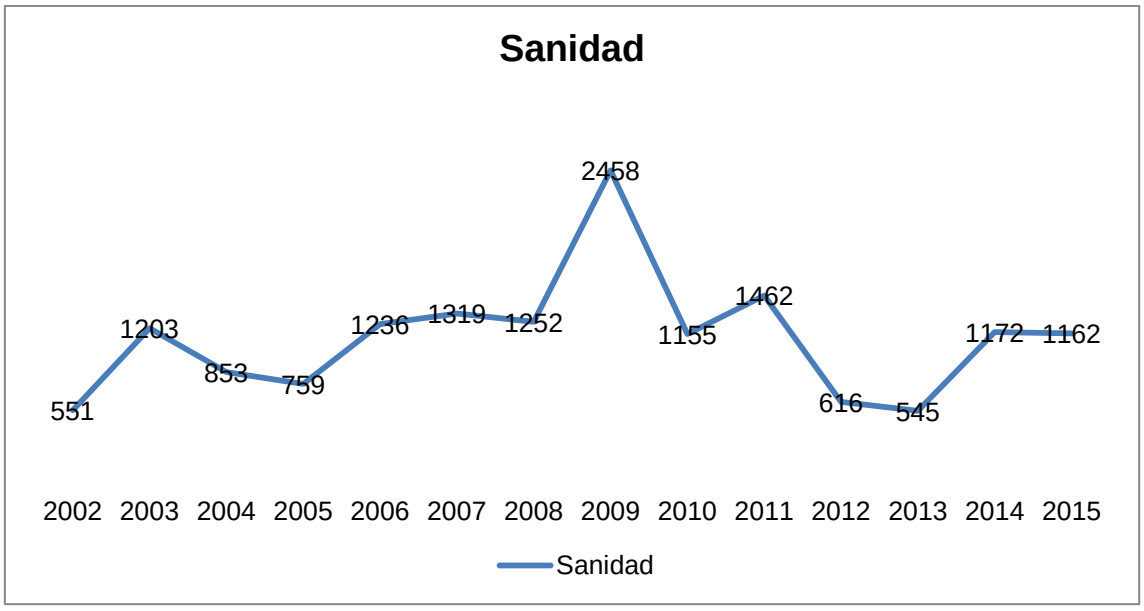


Figura 26. Gráfico noticias Sanidad

En décimo lugar hemos observado el *cluster* de Tecnologías de la Información. En la Figura 27 se puede ver como los grandes temas los ocupan Microsoft, Google, Apple o Sony para este ejemplo concreto, pero se habla en general de las innovaciones en el sector de las tecnologías de la información y comunicación.

```
3 2014\mundo20140512-055 Unir todas las cosas, un negocio mundial de 13,5 billones de euros
3 2014\mundo20140514-065 Sony recupera el mitico cassette y lo llena de 60 millones de cancion
3 2014\mundo20140518-069 Los bandazos de la XBox One de Microsoft
3 2014\mundo20140520-182 China prohibe Windows 8 en los ordenadores gubernamentales
3 2014\mundo20140520-184 Microsoft presenta la tercera generacion de su tableta Surface
3 2014\mundo20140521-066 Intel presenta su revolucion 3D, una fabrica de ideas en cada casa
3 2014\mundo20140530-081 La taza con pantalla personalizada y la tetera que se activa por wifi
3 2014\mundo20140601-106 De Internet al robot equilibrista
3 2014\mundo20140602-170 Apple acerca el iPhone y el iPad al Mac
3 2014\mundo20140603-060 Lo mas in en 2014: Tabletas mini, gafas inteligentes... y los mayores
3 2014\mundo20140603-061 Lo que no invente Asia... Raton con ventilacion y cepillo dental con
3 2014\mundo20140610-184 Google adquiere la empresa de satelites Skybox por 369 millones
3 2014\mundo20140616-183 El futuro en blanco de Sony
3 2014\mundo20140619-123 Moviles contaminados de fabrica
3 2014\mundo20140625-169 Google renueva Android para llevarlo a nuevos mercados
3 2014\mundo20140627-185 Google tienta a la empresa con almacenamiento ilimitado
3 2014\mundo20140704-188 ¿Movil o navegador GPS?
3 2014\mundo20140711-097 Las aplicaciones de tecnologia cloud y Big Data, en auge
3 2014\mundo20140711-178 Microsoft afirma haber liberado 4,7 millones de PC infectados
3 2014\mundo20140711-179 Sony encuentra por fin la formula de la portatil
```

Figura 27. Imagen cluster Tecnologías de la Información

En la Figura 28 de las Tecnologías de la Información vemos como su ascenso es muy moderado, comienza en el año 2.002 con 68 noticias, asciende ligeramente en 2.003 con 93, desciende en 2.004 con 58 y en 2.005 no aparece ninguna noticia. En 2.006 comienza fuertemente con 177 y asciende en 2007 con 283 para volver a descender el año siguiente con 148, aumentando progresivamente hasta 2.011 dónde tiene su máximo con 433. A partir de aquí comienza a descender hasta el año 2.013 con 135, aumenta en 2014 hasta 163 y se estabiliza con 147 noticias en el año 2.015.

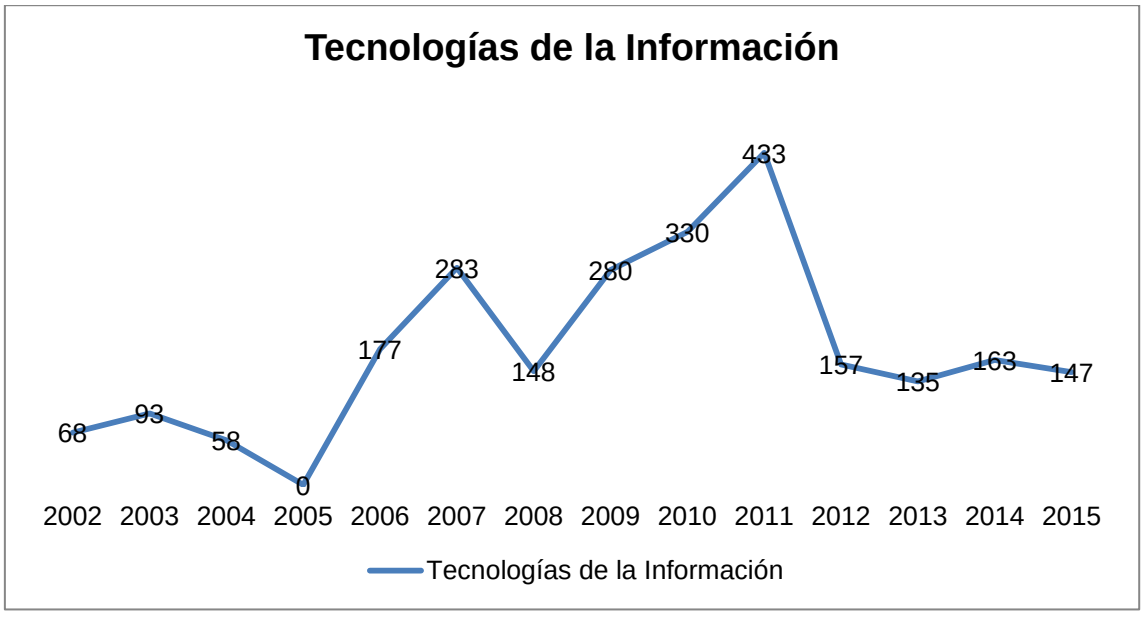


Figura 28. Gráfico noticias Tecnologías de la Información

En último lugar nos encontramos con el *cluster* de Varios, en el que se engloban noticias de todo tipo que debido a su variedad temática ha resultado imposible englobar en un único tema. Como ejemplo podemos observar en la Figura 29 que la heterogeneidad de las noticias es bastante grande.

```
0 2009\mundo20090207-022 'Walking house': la vivienda caracol
0 2009\mundo20090208-107 Repsol descubre un nuevo yacimiento en el Golfo de Mexico
0 2009\mundo20090209-002 Un craneo fosil desvela las claves del origen de los animales terres
0 2009\mundo20090209-004 Los monos capuchinos eligen su cascanueces
0 2009\mundo20090209-008 Las ballenas parian en tierra hace 47 millones de años
0 2009\mundo20090209-009 Descubren el planeta extrasolar mas pequeño, algo mayor que la Tierr
0 2009\mundo20090209-010 Descubren el fosil de la serpiente prehistorica mas grande hallada
0 2009\mundo20090209-014 Bañarse en acido
0 2009\mundo20090209-016 La Comision Europea adopta un plan de accion para la conservacion d
0 2009\mundo20090211-002 El origen de las uñas, en un fosil de 390 millones de años
0 2009\mundo20090211-003 El laser que rastrea el CO2 de los bosques
0 2009\mundo20090211-012 Descubren en Peru cientos de huellas de dinosaurios y otros animales
0 2009\mundo20090211-013 'SandBot': el nuevo robot que se mueve sobre la arena
0 2009\mundo20090212-002 Google en el contador de la luz
0 2009\mundo20090212-015 'Bejar': una bola de fuego de un cometa de Jupiter que cayo en Sala
0 2009\mundo20090215-006 Multiculturalismo evolutivo
0 2009\mundo20090215-009 La linterna mas pequeña del mundo
0 2009\mundo20090215-012 Una galaxia espiral en homenaje a Galileo
0 2009\mundo20090215-015 Presentado el primer borrador del genoma del neandertal
```

Figura 29. Imagen cluster Varios

Este último grupo de Varios reflejado en la Figura 30 comienza en el año 2.002 con un *cluster* de 1.208 noticias que no se han podido clasificar. Los tres siguientes años son considerados como “buenos”, es decir, todas las noticias se han logrado clasificar en algún *cluster*. En 2.006 nos encontramos con 753 noticias, aumentando en 2.007 con 2.901 noticias, descendiendo en 2.008 hasta las 743. Después aumenta progresivamente hasta el año 2.009 con 2.759 noticias. Desciende hasta 2.013 con 1.231 noticias, aumenta en 2.014 con 2.032 y finaliza el año 2.015 con 1.800 noticias.

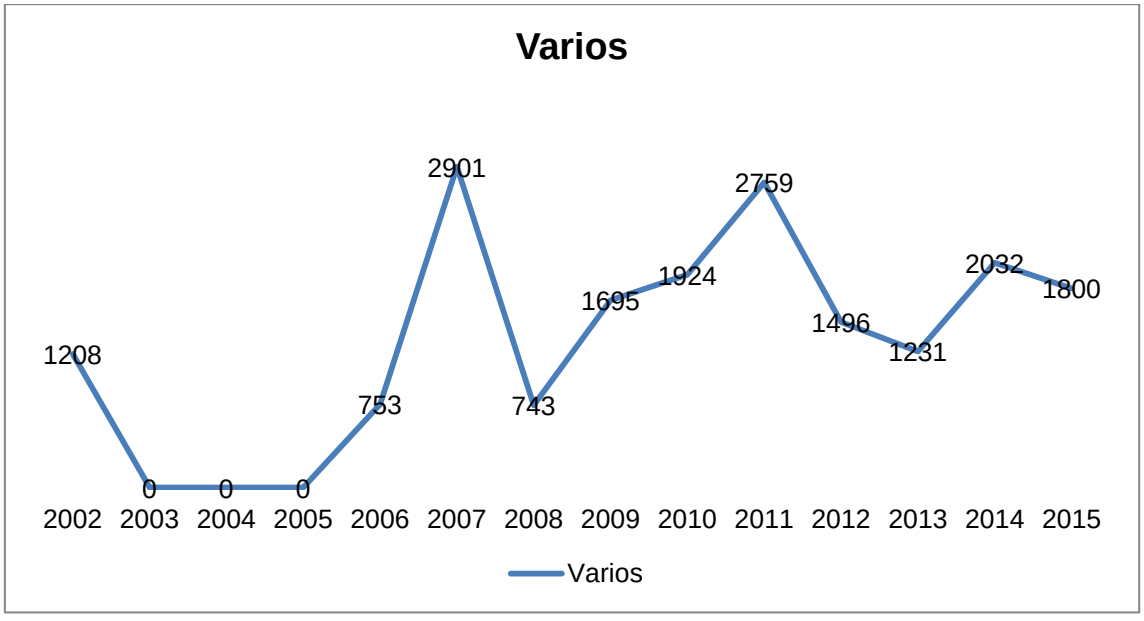


Figura 30. Gráfico noticias Varios

6. Discusión de los resultados

Pasando al análisis de los resultados, la distribución temática obtenida con el algoritmo de *K-Means* no difiere tanto de la obtenida en el *Spanish Corpus of Scientific Culture* (SCSC).

La principal diferencia que encontramos entre los resultados obtenidos en una tabla y otra reside principalmente en el método utilizado para detectar los *topics* o temas para cada caso. Nosotros hemos utilizado técnicas de *clustering* para la clasificación automática de las noticias a través del algoritmo de *K-Means*, el cual ha precisado de nuestra intervención para el establecimiento de los temas que hemos considerado oportunos y la posterior revisión manual para determinar los temas de cada *cluster.*

En el caso del SCSC se utilizó una metodología que se basa en las Tecnologías de Análisis de Redes, que sin entrar demasiado en detalles específicos, básicamente una red se compone de una serie de nodos que se encuentran unidos por unos arcos, los cuales vienen a representar las relaciones existentes entre cada uno de estos nodos. Los arcos pueden tener un peso específico que vendría caracterizado por la intensidad de su relación con el nodo, de modo que estos nodos se pueden representar en una red. Los nodos con mayor fuerza se situarán más próximos y los que tienen menos fuerza se ubicarán en posiciones más alejadas. Una vez representados estos nodos nos encontraremos con una serie de comunidades, es decir, grupos de nodos que se encuentran muy juntos en relación a otros. Llevado a la práctica los nodos serían las noticias y los arcos la similitud, en este caso semántica que hay entre las noticias. Para detectar estas similitudes se pueden utilizar diferentes algoritmos, pero el SCSC se decantó por utilizar InfoMap (Figuerola et al, 2016).

Como podemos ver en la Tabla IV con el algoritmo de InfoMap se obtuvieron siete comunidades, que vendrían a ser siete temas, frente a los once que obtuvimos con el algoritmo de *K-Means* que podemos ver en la Tabla V.

En primer lugar, en el caso de las noticias del SCSC son 82.680 en las que vienen incluidas las noticias del periódico ABC. En nuestro caso, tenemos 60.074 noticias y no hay ninguna noticia del ABC.

Si hacemos un breve análisis de los temas podemos observar como muchos de los temas que hemos obtenido nosotros los podemos agrupar en otros más amplios, de modo que si juntamos los temas de Ecología, Recursos Hídricos, Fuentes de Energía, Cambio Climático y Medio Ambiente en un único *cluster* como Medio Ambiente en sentido general, y añadimos los temas de Sanidad, Tecnologías de la Información, Aeronáutica/Aeroespacial, Investigaciones y Varios tenemos los 6 grandes temas que han ocupado a la prensa digital española a lo largo de los años 2002 hasta 2015, teniendo una visión global de conjunto de todos estos miles de noticias como podemos observar en la tabla.

En esta clasificación hemos obviado los temas de Automóviles, que tras un análisis de los *clusters* más cercanos hemos visto que la evolución de estos no es progresiva porque las noticias que deberían de incluirse en este *cluster* son tan pocas que se han mezclado con otros *clusters* cercanos y por eso no hay evolución de los mismos. Por otro lado, en el caso del *cluster* Portugués, el algoritmo clasificó las noticias atendiendo al idioma, de modo que también hemos obviado este *cluster* al considerar como un error cometido en la descarga de noticias, pues nosotros estamos trabajando con noticias de la prensa digital española, y se sobreentiende que todas las noticias deberían de estar escritas en español, y no en portugués, por este motivo hemos excluido este *cluster.* Por ello, aunque excluimos los *clusters,* sí que incluimos sus

noticias dentro de la categoría de Varios pues apenas tienen evolución en uno y dos años respectivamente, no proporcionándonos una visión de conjunto.

TOPIC	NEWS
Aerospace and Astronomy	8.385
Energy and Environment	28.569
Health Science	28.725
Information Technology	7.805
Others	3.115
Paleontology Evolution	2.845
Science Policy	3.236
Total	82.680

Tabla IV. Topics algoritmo InfoMap

TOPIC	NEWS
Aeronáutica / Aeroespacial	3.967
Automóviles	113
Cambio Climático y Medio Ambiente	2.883
Ecología	4.511
Fuentes de Energía	1.829
Investigaciones	9.395
Portugués	168
Recursos Hídricos	451
Sanidad	15.743
Tecnologías de la Información	2.472
Varios	18.542
Total	60.074

Tabla V. Topics algoritmo K-Means

Si comparamos los temas de la Tabla IV con los de esta Tabla VI vemos como hemos obtenido temas muy parecidos con una evolución lineal en el tiempo. El tema de Aeronáutica/Aeroespacial está muy relacionado con el tema de *Aeroespace and Astronomy*, luego tenemos el de Medio Ambiente que es el mismo que el de *Energy and Environment*. El tema de Investigaciones es un *cluster* nuevo que no aparece en el SCSC, y continuando con el de Sanidad sí que aparece reflejado en el SCSC como *Health Science*. El tema de Tecnologías de la Información es el mismo que *Information Technology*, y el de *Others* se corresponde con el nuestro de Varios. Por otro lado, nosotros no tenemos los temas de *Paleontology Evolution*, ni el de *Science Polity*.

TOPIC	NEWS
Aeronáutica / Aeroespacial	3.967
Medio Ambiente	9.654
Investigaciones	9.395
Sanidad	15.743
Tecnologías de la Información	2.472
Varios	18.863
Total	60.074

Tabla VI. Topics generales agrupados con K-Means

En las siguientes Figuras 31 y 32 se refleja la evolución lineal de cada uno de los temas obtenidos por el SCSC y nosotros. Si realizamos un análisis comparativo, a simple vista podemos observar cómo en la Figura 31 los temas que más destacan son los de *Health Science* y *Energy and Environment*, mientras que en nuestra Figura 32 obviando al tema de Varios, los *topics* más sobresalientes son Sanidad, Medio Ambiente e Investigaciones. Por el contrario los temas que menos repuntan en la Figura 31 son *Others, Paleontology Evolution* y *Science Polity*, mientras que en la Figura 32 son Tecnologías de la Información y Aeronáutica/Aeroespacial.

En general los temas de la Figura 31 son muy constantes, mientras que los temas de la Figura 32 tienen muchas más fluctuaciones y repuntes, debido principalmente a algún acontecimiento que sucedió ese año reflejado en el índice de noticias, sirva de ejemplo el año 2009 en el *cluster* de Sanidad, tiene su máximo histórico con 2.458 noticias. Analizando este año en concreto nos damos cuenta de que la prensa puso toda su atención en la sanidad debido al origen de la pandemia ocasionada por la Gripe Aviar, centro de las noticias de ambos periódicos.

Temas InfoMap

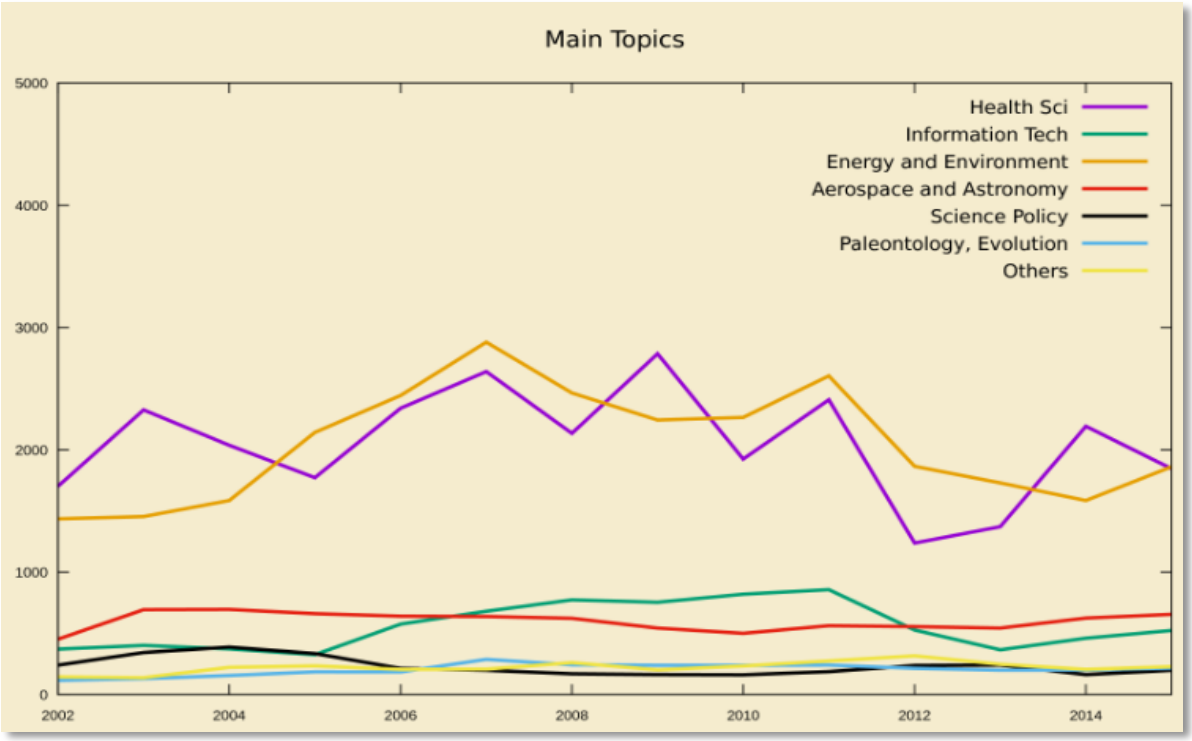


Figura 31. Gráfico evolución topics SCSC

Temas K-Means

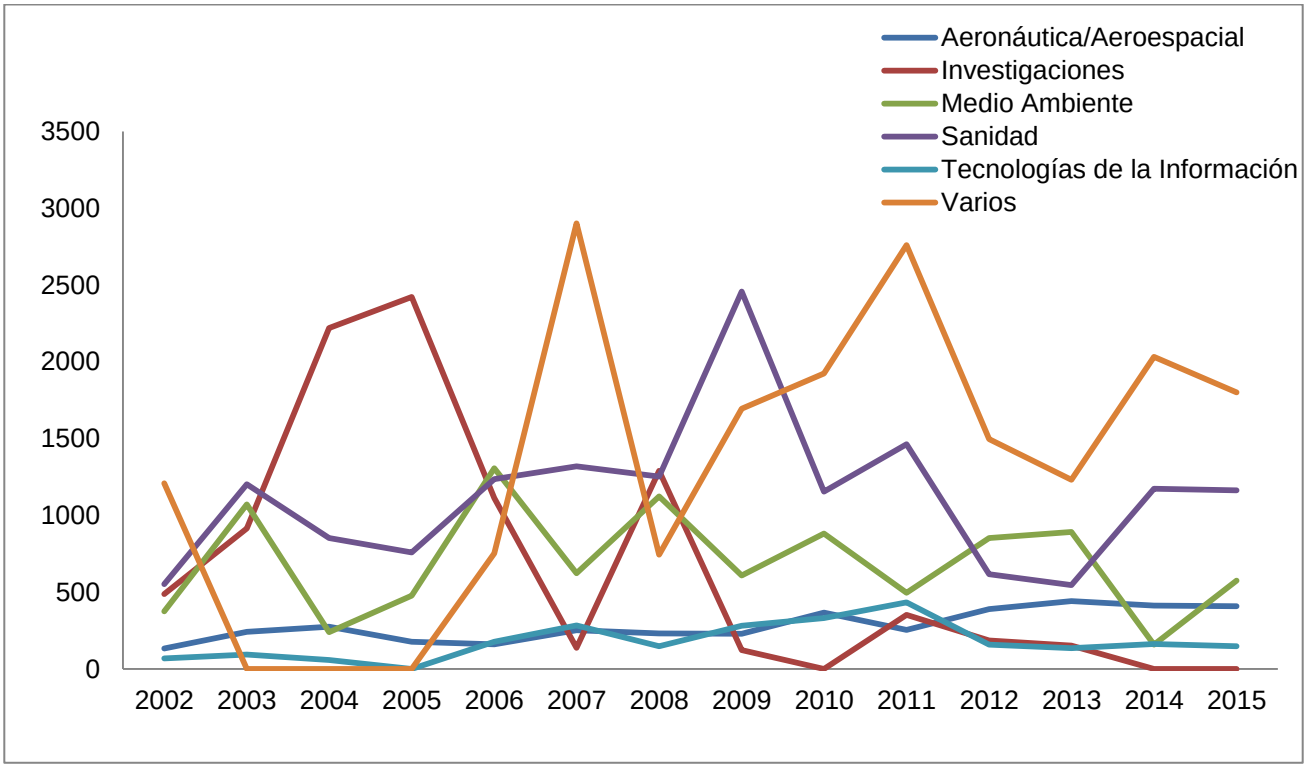


Figura 32. Gráfico evolución topics con K-Means

7. Conclusión

La primera de las conclusiones a las que me referiré se plasma en la Figura 33. Aquí observamos el porcentaje de aciertos y fallos que ha tenido el algoritmo de *K-Means*. Estos porcentajes no son fidedignos en el sentido de que el algoritmo en verdad ha clasificado todas las noticias en los *clusters* que nosotros le hemos especificado, pero tras el análisis manual ha resultado totalmente imposible determinar un tema general para determinados *clusters*, de modo que si hacemos balance de las noticias clasificadas correctamente estas se contabilizan en 41.532, que suponen un 70 % de aciertos con respecto al total. Por otro lado, las noticias que no se han logrado clasificar son 18.542, el 30 % restante.

Este porcentaje de fallos es muy alto si lo comparamos el algoritmo de InfoMap de Análisis de Redes, donde apenas era del 3.76 %.

Figura 33. Gráfico resultados K-Means

Originalmente teníamos una clasificación temática de 11 temas. En algunos de estos temas tan sólo hay noticias para uno o dos años determinados, probablemente porque los *clusters* con estas noticias concretas eran tan reducidos que se han mezclado con otros *clusters* cercanos temáticamente como sucedió con el tema de los Automóviles.

Los temas no son continuos a través de los años, y no tienen por qué serlos, cada año produce temas que no siempre son los mismos para los años siguientes. Sin embargo, si agrupamos los temas cercanos temáticamente podemos distinguir líneas de continuidad en algunos temas, no en todos.

Esto nos ha permitido clasificar las 60.074 noticias originales en 6 temas generales que nos proporcionan una visión del interés de la prensa digital española a lo largo de los años comprendidos entre 2002 y 2015.

La importancia que da la prensa a las noticias de Ciencia y Tecnología en el contexto de España, es creciente, sin embargo se centran mucho más en la ciencia que en la

propia tecnología, pues tras el análisis temático de los grandes *topics* que han ocupado a estos periódicos, la mayoría de las noticias se centran en la Sanidad y el Medio Ambiente, relegando a un segundo lugar los avances en materia tecnológica.

Por otra parte, hemos de destacar que las técnicas de *clustering* permiten agrupar documentos, véase noticias, con contenidos temáticos que son cercanos. Sin embargo, ningún sistema es perfecto, y el algoritmo de *K-Means* tampoco lo es, ya que su principal problema reside en prefijar de antemano los temas que queremos obtener, exigiendo una revisión manual de los mismos una vez ha finalizado el proceso de *clustering*. De modo que en definitiva los resultados dependen en gran medida de las capacidades que tenga la persona física en el momento de determinar un tema general para cada *cluster*, pues no todas utilizan los mismos criterios ni tienen los mismos conocimientos que les permitan establecer unos temas más adecuados a los contenidos que se expresan en los documentos.

8. Bibliografía

Álvarez, P. A., Vega, I. F. y Fernández, E. (2007). Análisis Comparativo de las Medidas de Semejanza Aplicadas al Contenido de Documentos Web. Universidad Autónoma de Sinaloa. Recuperado el 10 de julio de 2017, de https://researchgate.net/publication/261027922_Analisis_Comparativo_de_las_Medidas_de_Semejanza_Aplicadas_al_Contenido_de_Documentos_Web

Ares, M. E., Parapar, J. y Barreiro, Á. (2011). Agrupamiento Documental. En Cacheda, F., Fernández, J. M. y Huete, J. F. *Recuperación de Información. Un enfoque práctico y multidisciplinar* (pp. 393-416). Madrid: Ra-Ma.

Bravo, D. (2008). Clasificación no supervisada de documentos (proyecto fin de carrera). Universidad CEU San Pablo. Recuperado el 11 de julio de 2017, de http://biocomp.cnb.csic.es/~coss/Articulos/Bravo2008.pdf

Cacheda, F. y Martínez, J. A. (2011). Modelos de Recuperación de Información Clásicos. En Cacheda, F., Fernández, J. M. y Huete, J. F. *Recuperación de Información. Un enfoque práctico y multidisciplinar* (pp. 95-103). Madrid: Ra-Ma.

Campos, L. M. y Romero, A. E. (2011). Clasificación Documental. En Cacheda, F., Fernández, J. M. y Huete, J. F. *Recuperación de Información. Un enfoque práctico y multidisciplinar* (pp. 359-389). Madrid: Ra-Ma.

Challenger, I., Díaz, J. y Becerra, R. A. (2014). El lenguaje de programación Python/The programming language Python, *Ciencias Holguín,* abril-junio, 3-4. Recuperado el 9 de julio de 2017, de http://ciencias.holguin.cu/index.php/cienciasholguin/article/view/826/872

Figuerola, C. G. (2017a) Clasificación automática de documentos. Salamanca: Universidad de Salamanca. Recuperado el 7 de julio de 2017, de https://moodle2.usal.es/mod/resource/view.php?id=259140

Figuerola, C. G. (2017b). Modelos Teóricos de Recuperación de Información. Salamanca; Universidad de Salamanca. Recuperado el 7 de julio de 2017, de https://moodle2.usal.es/mod/resource/view.php?id=259114

Figuerola, C. G., Alonso, J. L. y Zazo, A. F. (2004). Algunas Técnicas de Clasificación Automática de Documentos. *Cuadernos de Documentación Multimedia*, 15. Recuperado el 8 de julio de 2017, de http://multidoc.rediris.es/cdm/viewarticle.php?id=28&layout=html

Figuerola, C. G. y Quintanilla, M. A. (2016). *Sistema de Indicadores para el SCSC (Spanish Corpus of Scientific Culture).* Salamanca: Instituto EcyT. Recuperado el 8 de julio de 2017, de http://grulla.usal.es/figuerola2017sistema.pdf

Groves, T., Figuerola, C. G., y Quintanilla, M. A. (2016) Ten years of science news: A longitudinal analysis of scientific culture in the Spanish digital press. *Public Understanding of Science*, *25* (6), 691-705. Recuperado el 10 de julio de 2017, de http://journals.sagepub.com/doi/pdf/10.1177/0963662515576864

Müller, A. C. y Guido, S. (2016). Introduction to Machine Learning with Python. A Guide for Data Scientists. California: O'Reilly Media.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. y Duchesnay, E. (2011). Scikit-learn: Máquina de aprendizaje en Python, *Journal of Machine Learning Research*, 12,

2825-2830. Recuperado el 11 de julio de 2017, de http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Piwowarski, B. y Blanco, R. (2011). Introducción a la Recuperación de Información. En Cacheda, F., Fernández, J. M. y Huete, J. F. *Recuperación de Información. Un enfoque práctico y multidisciplinar* (pp. 33-43). Madrid: Ra-Ma.

Urbano, J., Morato, J., Marrero, M. y Sánchez-Cuadrado, S. (2010). Recuperación y Acceso a la Información. Madrid: Universidad Carlos III de Madrid. http://ocw.uc3m.es/ingenieria-informatica/recuperacion-y-acceso-a-la-informacion

Vallejo, D. F. (2016). Clustering de documentos con restricciones de tamaño (trabajo fin de máster). Universidad Politécnica de Valencia. Recuperado el 10 de julio de 2017, de http://mugi.webs.upv.es/wp-content/uploads/2016/11/TFM-Diego-Vallejo-MUGI.pdf

Vanderplas, J. (2016). Python Data Science Handbook. Essential Tools for Working with Data. California: O'Reilly Media.