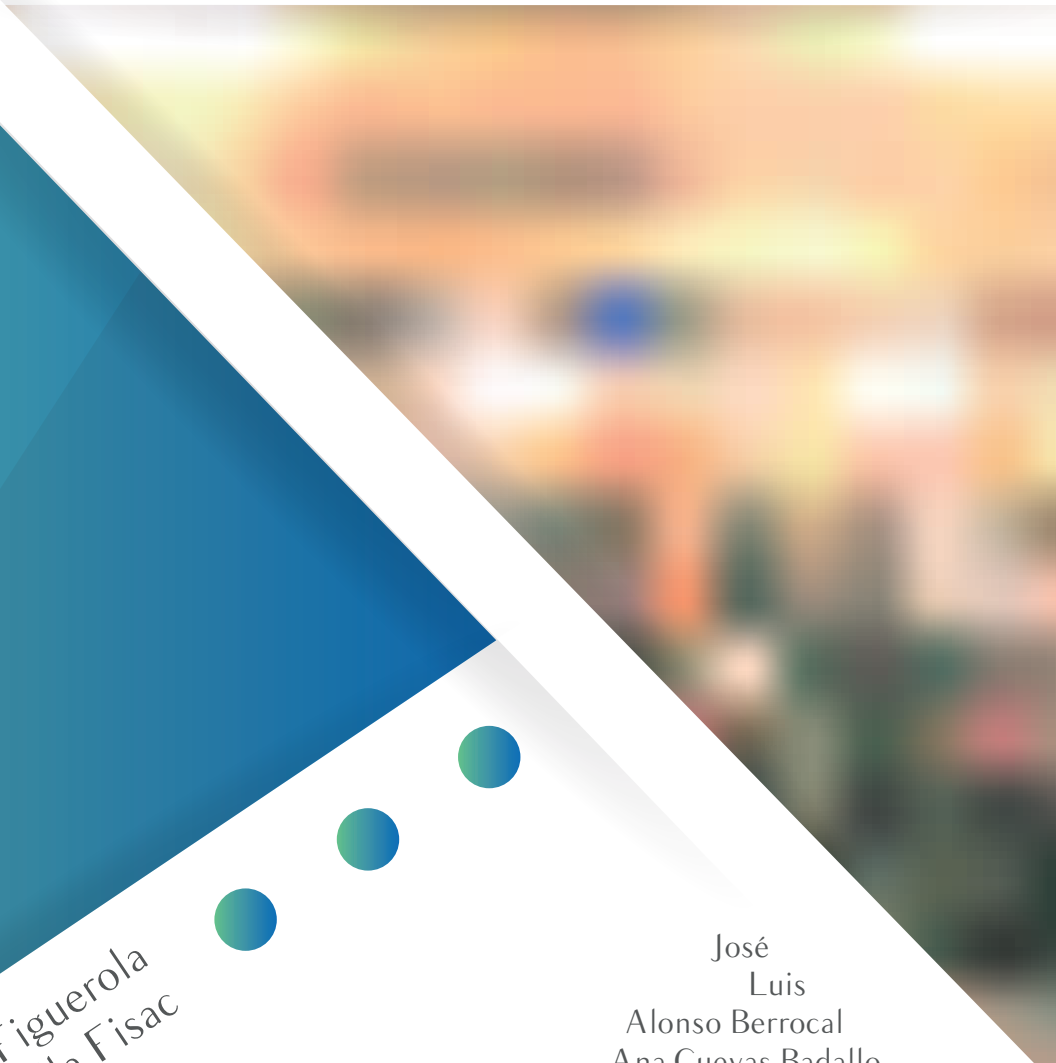
# SISTEMA DE INDICADORES PARA EL SCSC

## (Spanish Corpus of Scientific Culture)




Carlos G. Figuerola
Miguel Angel Quintanilla Fisac

José
Luis
Alonso Berrocal
Ana Cuevas Badallo
Marina Gordaliza Escobar
Tamar Groves
Pilar López Morales
Bruno Maltrás Barba
Esther Palacios Mateos
Libia Santos Requejo
Ángel Francisco Zazo



GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
FECYT FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA
INSTITUTO ECYT

# Proyecto: FCT-15-10287

INVESTIGADORES

**Miguel Angel Quintanilla Fisac**
**Carlos G. Figuerola**

José Luis Alonso Berrocal
Ana Cuevas Badallo
Marina Gordaliza Escobar
Tamar Groves
Pilar López Morales
Bruno Maltrás Barba
Esther Palacios Mateos
Libia Santos Requejo
Ángel Francisco Zazo

Salamanca, 2016

# ÍNDICE


INTRODUCCIÓN
**7**

LA DESCARGA DE NOTICIAS
**9**

EL CASO DEL ABC
**11**

RESULTADOS
**12**

CATEGORIZACIÓN DE LAS NOTICIAS
**14**

INTENSIDAD CIENTÍFICA
**17**

CIENCIA VS. TECNOLOGÍA
**20**

CATEGORÍAS DEL MODELO TEÓRICO DE CULTURA CIENTÍFICA
**27**

DISTRIBUCIÓN TEMÁTICA
**34**

BIBLIOGRAFÍA
**38**

# INTRODUCCIÓN

La realización del proyecto Sistema de indicadores para el SCSC (Spanish Corpus of Scientific Culture) se apoya en los trabajos y datos recopilados en trabajos anteriores, que dieron lugar a la primera versión del Spanish Corpus of Scientific Culture (Quintanilla et al., 2014).
Esta primera versión constaba de aproximadamente 50.000 noticias sobre Ciencia y Tecnología extraídas de las versiones digitales de los periódicos El País, El Mundo y Público durante los años 2002-2011 (para Público solamente 2007-2011).
Esta primera versión permitió diseñar, implementar y poner a prueba una serie de técnicas automáticas capaces de recopilar y analizar cuantitativamente noticias sobre Ciencia y Tecnología, así como calcular una serie de indicadores de cultura científica.

Sin embargo, se detectaron también numerosos aspectos claramente mejorables, así como diverssos elementos no desarrollados que complementarían notablemente los trabajos desarrollados hasta la fecha. Sobre la mejora de tales aspectos se ha reconstruído el SCSC y se han obtenido diversos indicadores cuantitativos de Cultura Científica que son el objeto principal de este documento.

En la versión final del SCSC se han tomado las siguientes opciones:

- Ampliación de la cobertura temporal, ahora 2002-2015
- Mantener las noticias de los diarios El Mundo y El País
- Eliminar las noticias del diario Público; el cambio de titularidad de esta cabecera supuso la desaparición de Internet (al menos para el acceso público) de las noticias recopiladas. Pero también se pudo constatar que algunos de los presupuestos que habían llevado a la inclusión de este diario en la primera versión del SCSC no se habían hecho efectivos, por lo que su presencia en el SCSC tenía escasa relevancia y planteaba una serie de problemas técnicosimportantes
- Añadir al SCSC las noticias del diario ABC, teniendo en cuenta su representatividad en el ámbito de la prensa nacional (si bien esto ha planteado una serie de problemas técnicos que se comentarán más adelante

De forma sintética, los objetivos del proyecto objeto de este documento han sido:

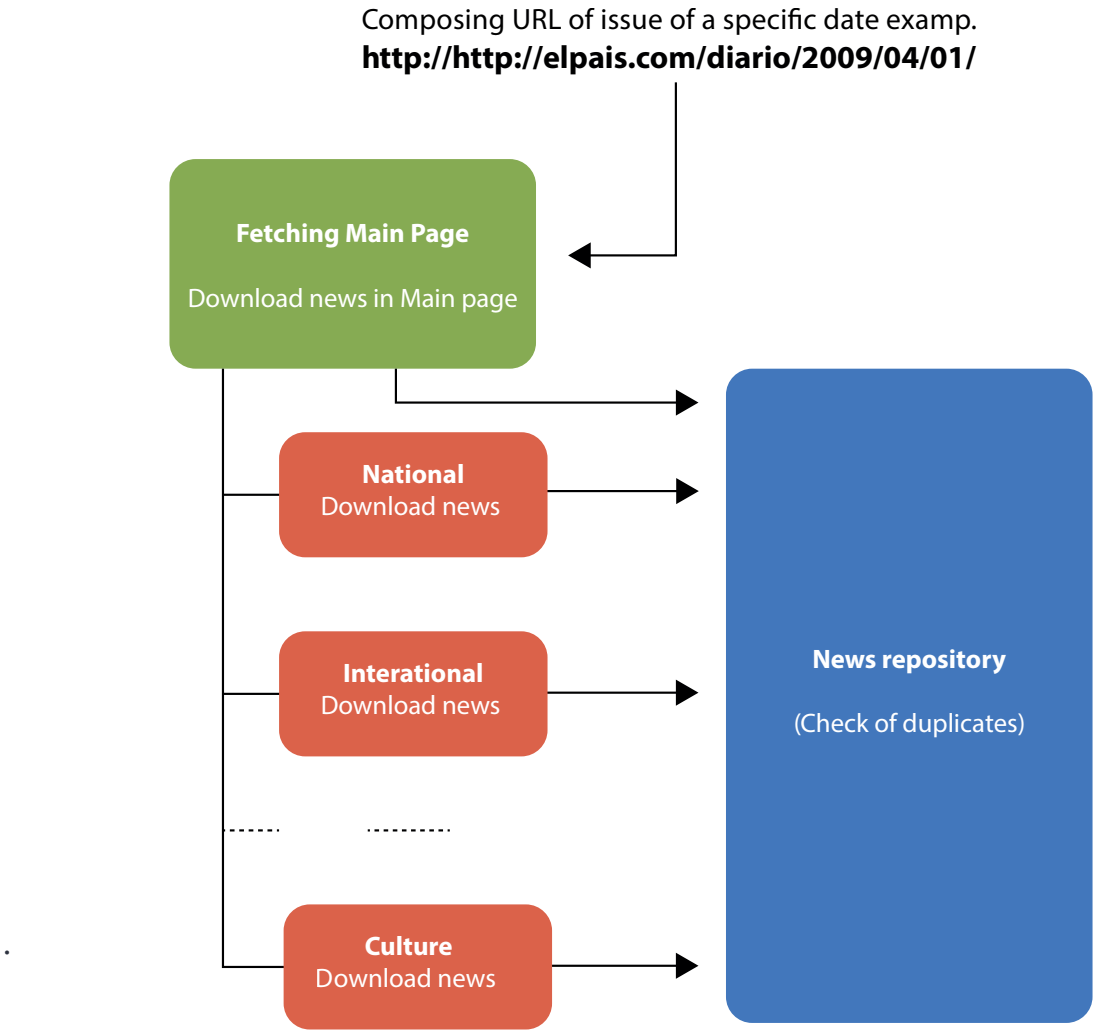
- Evaluar la importancia de las noticias científicas en el contexto español
- Analizar la ratio de componentes intrínsecos y extrínsecos
- Discriminar entre los temas más frecuentes
- Poner a prueba diferentes metodologías capaces de manejar grandes cantidades de datos.

El desarrollo de estos objetivos, así como los resultados obtenidos , extán expuestos en las siguientes páginas.

# LA DESCARGA DE NOTICIAS

En una primera fase se procedió a descargar todas las noticias posible (de cualquier temática) de los periódicos seleccionados. Para esto se recurrió a las hemerotecas digitales de esos periódicos; en el caso de El Mundo y El País dichas hemerotecas remiten a las noticias en formato web. En el caso del ABC remiten a las páginas en papel digitalizadas, lo cual, como se comentará, produce algunos problemas.

En general, las hemerotecas digitales están organizadas por fechas, permitiendo acceder a un navegador humano a la página principal del día o fecha elegidos. Por supuesto, las hemerotecas permiten también efectuar búsquedas de las noticias que contienen determinadas palabras, pero este tipo de recuperación no se ha considerado útil para nuestros propósitos (Figuerola et al., 2014)



Una vez en la página principal o portada de la fecha deseada, ésta contiene algunas o todas las noticias de ese día, completas o sólo en titulares y entradillas (leer más, etc.); enlaces a las secciones del periódico, cada una de las cuales contiene las noticias de esa sección de ese día (muchas veces también de otros días), nuevamente completas o en forma resumida y con enlaces a la noticia completa.

Las hemerotecas digitales y los ejemplares de cada fecha están diseñados para ser navegados por personas, pero una descarga sistemática de todas las fechas del período analizado (2002-2015) y, para cada fecha, de todas las noticias de todas las secciones es inabordable, obviamente, de forma manual. La cuestión se ha resuelto mediante el diseño de crawlers específicos, es decir, programas capaces de navegar de forma autónoma y automática las hemerotecas descargando las páginas web navegadas (Castillo, 2004; Olston et al., 2010).
Las noticias aparecen en la misma página con otros elementos que deben ser depurados, tales como publicidad, noticias relacionadas (en ocasiones de fechas posteriores), comentarios de los lectores, etc.. Además, la estructura interna de las páginas y de la codificación de las propias noticias no sólo difiere de un periódico a otro, sino incluso entre secciones del mismo periódico; y, además, varían con el tiempo. En efecto, los periódicos experimentan en diversos momentos cambios de formato, estructura, proveedores de servicios web, etc., lo cual supone cambios en su codificación. Es preciso rastrear y detectar dichos cambios para poder individualizar y limpiar cada noticia. Dado el elevado número de éstas, esa detección ha de conseguirse por medios automáticos.

# EL CASO DEL ABC

En una primera fase se procedió a descargar todas las noticias posible (de cualquier temática) Aunque el diario ABC tiene su hemeroteca digital que permite seleccionar fechas o días concretos (además de las consabidas búsquedas por palabras), ésta remite a los ejemplares del periódico en papel, escaneados y convertidos a formato PDF. En la hemeroteca de ABC cada página física de papel es un archivo PDF en el web. Es posible extraer el texto de esos PDF, naturalmente; pero dado que cada página suele contener más de una noticia, y que el maquetado de las páginas es diverso, no es posible individualizar el texto de esas noticias de forma automática. Dicho de otra forma, en el caso del ABC el elemento individual más básico que podemos conseguir es cada página del periódico.

Para sistemas de análisis automático (clasificadores, detección de topics, etc.) basados en palabras esto no es un problema, si admitimos como unidad de análisis la página (en lugar de la noticia, como ocurre en los otros periódicos). Afortunadamente, aunque cada página contiene varias noticias, éstas no suelen ser heterogéneas temáticamente sino que versan sobre objetos cercanos.

Así pues, hemos adoptado este enfoque con el ABC; de manera que las cifras brutas que se darán para este diario deben entenderse como páginas y no como noticias, lo cual, siempre en el terreno de las cifras brutas, hace difícil la comparación con los otros diarios. Sin embargo, si operamos con cifras relativas al total de noticias -páginas, en el caso del ABC- creemos que los datos pueden resultar significativos. De hecho, como se verá, las series de datos así obtenidas del ABC son consistentes con las obtenidas a partir de noticias individuales de los otros diarios.

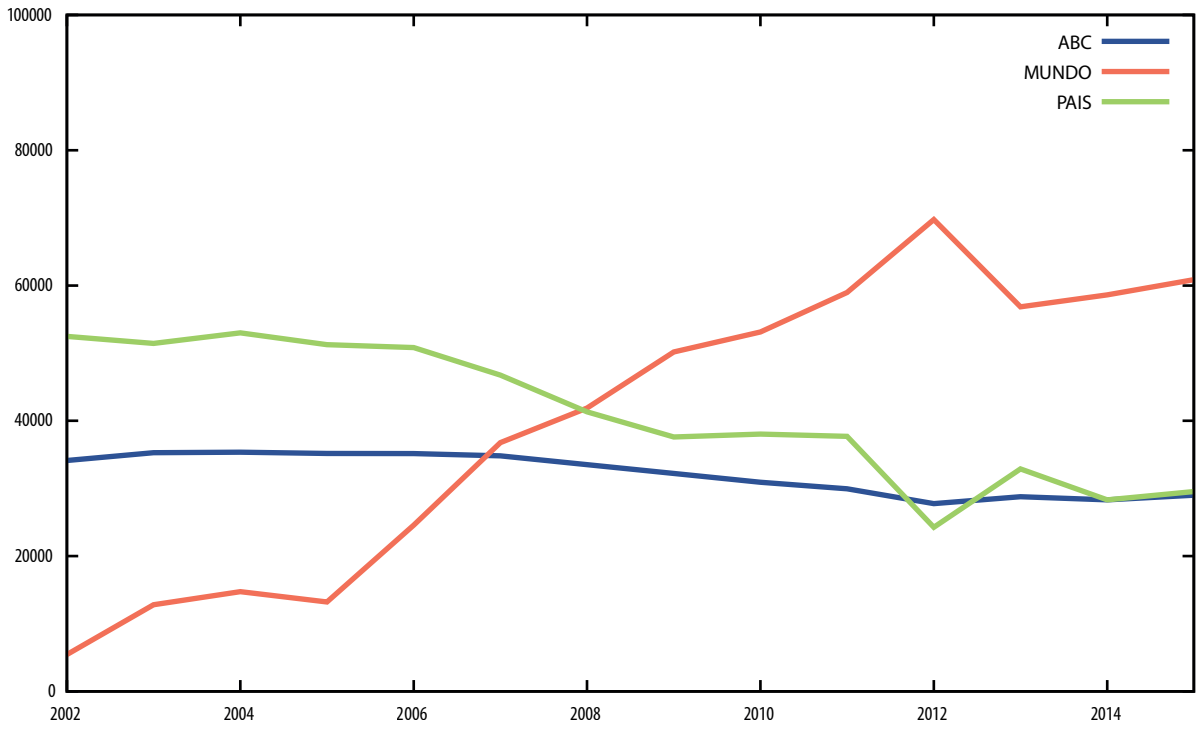
# RESULTADOS

Con estos planteamientos metodológicos, el total de noticias (páginas, en el caso del ABC) obtenidas queda reflejada en la tabla [todas.xlsx] y en el gráfico [todas.svg]. Se trata de un conjunto de 1,583,607 noticias, lo que nos permite afirmar que el espacio de análisis de este trabajo es realmente exhaustivo.

**Tabla / Noticias (todo tipo) recolectadas**

| YEAR | ABC | MUNDO | PAIS |
|---|---|---|---|
| 2002 | 34126 | 5441 | 52483 |
| 2003 | 35280 | 12817 | 51450 |
| 2004 | 35354 | 14745 | 53004 |
| 2005 | 35170 | 13209 | 51249 |
| 2006 | 35151 | 24572 | 50824 |
| 2007 | 34829 | 36772 | 46755 |
| 2008 | 33522 | 41887 | 41329 |
| 2009 | 32215 | 50171 | 37601 |
| 2010 | 30924 | 53133 | 38028 |
| 2011 | 29944 | 58965 | 37696 |
| 2012 | 27762 | 69749 | 24266 |
| 2013 | 28782 | 56860 | 32889 |
| 2014 | 28305 | 58618 | 28312 |
| 2015 | 29000 | 60874 | 29544 |

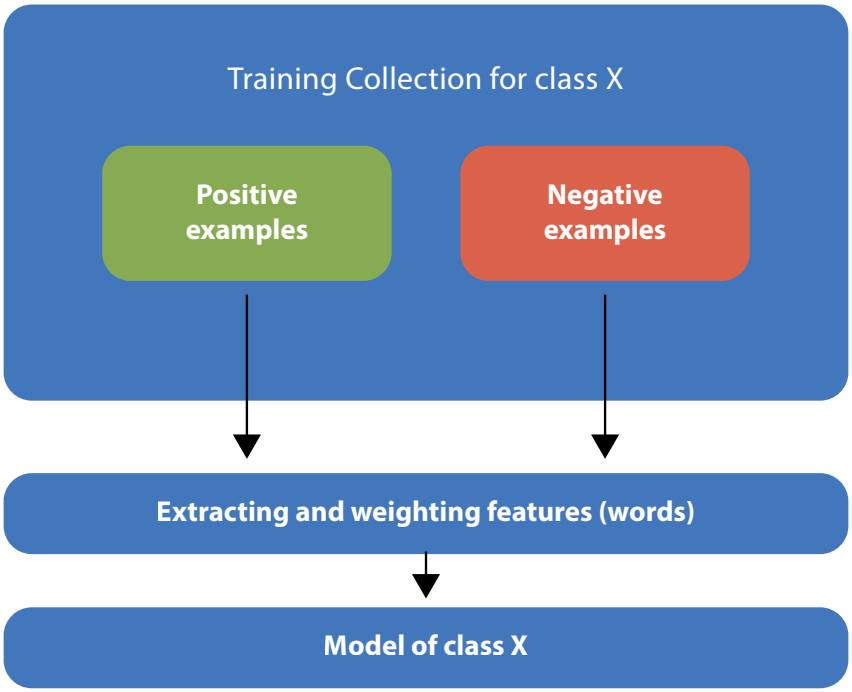
**Gráfico / Noticias (todo tipo) recolectadas**



# CATEGORIZACIÓN DE LAS NOTICIAS

A los efectos de constitución del SCSC, análisis y obtención de indicadores es obvio que no se necesita trabajar con todas las noticias, sino con aquéllas que tratan de Ciencia y/o Tecnología. Nuevamente, elelevado número de documentos con que trabajamos hace inviable una aproximación manual, de manera que es preciso recurrir a categorizadores automáticos que analicen el millón y medio largo de noticias recolectadas y selecciones o aíslen aquéllas que tratan de Ciencia y/o Tecnología. Dado que, posteriormente, será preciso clasificar tales noticias según el modelo teórico de cultura científica aplicado, el uso de categorizadores será recurrente.

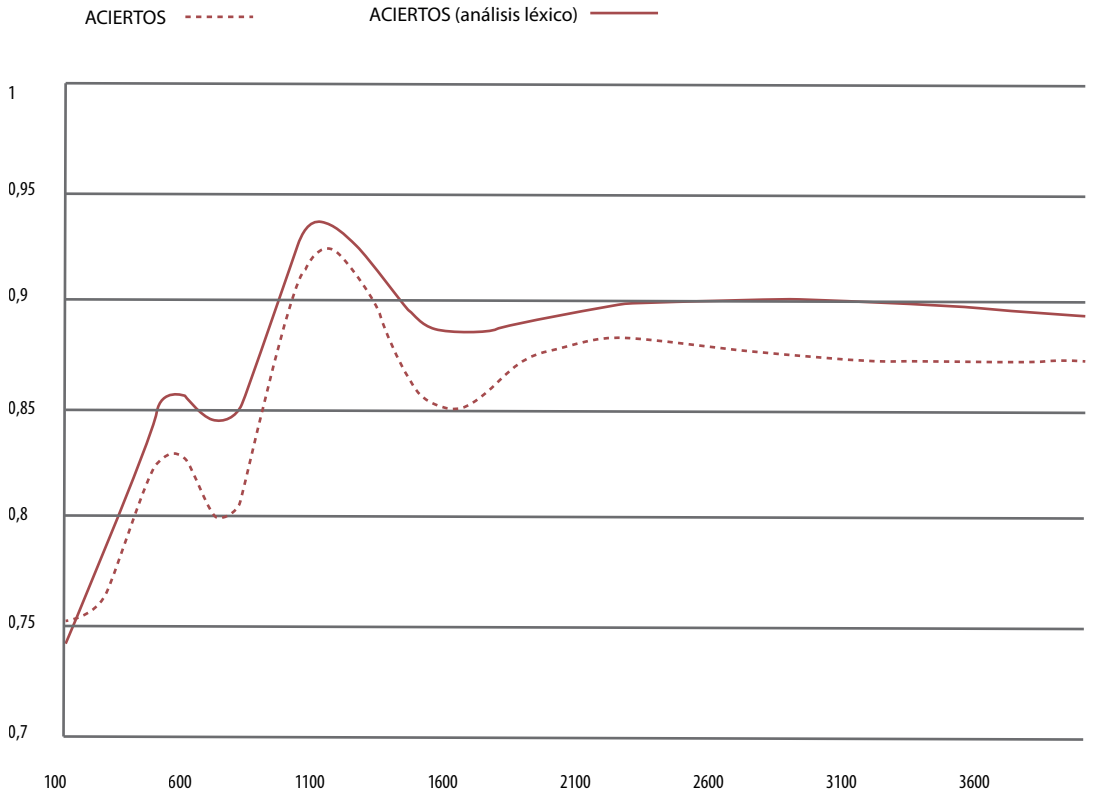
En la primera fase de la constitución del SCSC, anterior al proyecto actual, se aplicó un categorizador bayesiano para extraer las noticias de Ciencia y/o Tecnología. Aunque los categorizadores bayesianos son reconocidos por la literatura científica como muy eficaces a la hora de categorizar documentos o texto (McCallum y Nigam, 1998; Eyheramendy et al., 2003; Kim et al. 2006), parece que los categorizadores basados en SVM, una tecnología más moderna, producen resultados ligeramente mejores (Vapnik, 1995; Joachims, 1998; Hu et al., 2003; Steinwart y Christman, 2008). Así pues, en la presente fase del SCSC hemos utilizado categorizadores SVM tanto para la obtención de las noticias de Ciencia y Tecnología como para la aplicación del modelo teórico de cultura científica.



Sin embargo, la experiencia acumulada en esa primera fase de categorizadores bayesianos nos permitió abordar con facilidad la aplicación de este tipo de técnicas optimizando su rendimiento; por ejemplo, con la utilización de técnicas avanzadas de análisis léxico de las noticias (stemming, bigramas …) o con la formación de buenas colecciones de entrenamiento.

En general, los categorizadores (bayesianos, SVM o de otro tipo) necesitan de una fase de entrenamiento, mediante técnicas de machine learning, durante la cual establecen las características que definene las clases o categoriás con las que hemos de trabajar. Para ello precisan de una colección de documentos etiquetados manualmente, de los cuales captar esas características. Dicha colección clasificada manualmente, conocida como colección de entrenamiento es crítica, y debe cumplir una serie de requisitos. Naturalmente, debe ser representativa de los casos que se puedan presentar, cosa que no es fácil de prever; también debe tener un tamaño adecuado. Es preciso contar con un número suficiente de casos (documentos, noticias, etc.), pero los categorizadores suelen sufrir de sobreentrenamiento si el tamaño de esa colección es demasidao grande, introduciendo ruido y perdiendo eficacia.

**Gráfico / ruido**



El gráfico muestra los resultados de algunas pruebas destinadas a determinar el tamaño óptimo de la colección de entrenamiento. A partir de las operaciones efectuadas en la fase preliminar del SCSC, aplicando un método iterativo de prueba, revisión manual, incorporación de noticias a colección de entrenamiento, etc. se consiguió disponer de colecciones de entrenamiento de una calidad razonable.

Revisiones manuales sobre muestras aleatorias efectuadas sobre la categorización final para extraer las noticias de Ciencia y/o Tecnología arrojaron un resultado de un **95 %** de aciertos. Como se indicará más adelante, esta selección se ha visto todavía mejorada a través del análisis (automático) de temas.

Este tipo de operaciones puede aplicarse, circunscribiéndose sólo a las noticias de Ciencia y/o Tecnología, a fin de determinar cuáles se refieren a Ciencia y cuáles a Tecnología (admitiendo que hay noticias que pueden referirse a ambos campos). Y también para determinar qué noticias pueden adscribirse a las diferentes categorías del modelo teórico de cultura científica, admitiendo también que esas categorías son compatibles entre sí y que unamisma noticia puede tener, por ejemplo, elementos de Ciencia Intrínseca pero también de Ciencia Extrínseca.

# INTENSIDAD CIENTÍFICA

Los gráficos y tablas pertinentes muestran las noticias de Ciencia y/o Tecnología obtenidas para cada periódico y su evolución a lo largo del período contemplado. Dado que la cantidad de noticias recolectadas es muy diferente para cada diario, es tal vez más interesante observar los datos relativos, es decir, los porcentajes referidos a la cantidad global de noticias para cada diario y año.

Ese porcentaje es lo que llamamos **intensidad científica** (Bauer, 2009), que sitúa, de forma global en **5.22 %**.

**Tabla / Noticias Ciencia y Tecnología recolectadas**

| YEAR | ABC | MUNDO | PAIS |
|---|---|---|---|
| 2002 | 1639.0 | 441.0 | 2381.0 |
| 2003 | 1967.0 | 952.0 | 2570.0 |
| 2004 | 1814.0 | 1004.0 | 2641.0 |
| 2005 | 1824.0 | 1018.0 | 2816.0 |
| 2006 | 1864.0 | 2137.0 | 2609.0 |
| 2007 | 2020.0 | 3056.0 | 2459.0 |
| 2008 | 1882.0 | 3005.0 | 1782.0 |
| 2009 | 1541.0 | 3654.0 | 1740.0 |
| 2010 | 1374.0 | 3468.0 | 1301.0 |
| 2011 | 1383.0 | 4379.0 | 1377.0 |
| 2012 | 1264.0 | 2829.0 | 864.0 |
| 2013 | 1304.0 | 2241.0 | 1154.0 |
| 2014 | 1421.0 | 2582.0 | 1429.0 |
| 2015 | 1348.0 | 2669.0 | 1516.0 |

**Tabla / Noticias Ciencia y Tecnología recolectadas (%)**

| YEAR | ABC | MUNDO | PAIS |
|---|---|---|---|
| 2002 | 4.80278966184 | 8.10512773387 | 4.53670712421 |
| 2003 | 5.5753968254 | 7.42763517204 | 4.99514091351 |
| 2004 | 5.13096113594 | 6.80908782638 | 4.98264281941 |
| 2005 | 5.18623827125 | 7.7068665304 | 5.49474136081 |
| 2006 | 5.30283633467 | 8.69689076998 | 5.13340154258 |
| 2007 | 5.79976456401 | 8.31067116284 | 5.25933055288 |
| 2008 | 5.61422349502 | 7.17406355194 | 4.31174236009 |
| 2009 | 4.78348595375 | 7.28309182596 | 4.62753650169 |
| 2010 | 4.44315095072 | 6.52701710801 | 3.42116335332 |
| 2011 | 4.61862142666 | 7.42643941321 | 3.65290747029 |
| 2012 | 4.5529860961 | 4.05597212863 | 3.5605373774 |
| 2013 | 4.53060940866 | 3.9412592332 | 3.50877192982 |
| 2014 | 5.02031443208 | 4.40479033744 | 5.04732975417 |
| 2015 | 4.64827586207 | 4.3844662746 | 5.13132954238 |

**Gráfico / Noticias Ciencia y Tecnología recolectadas**

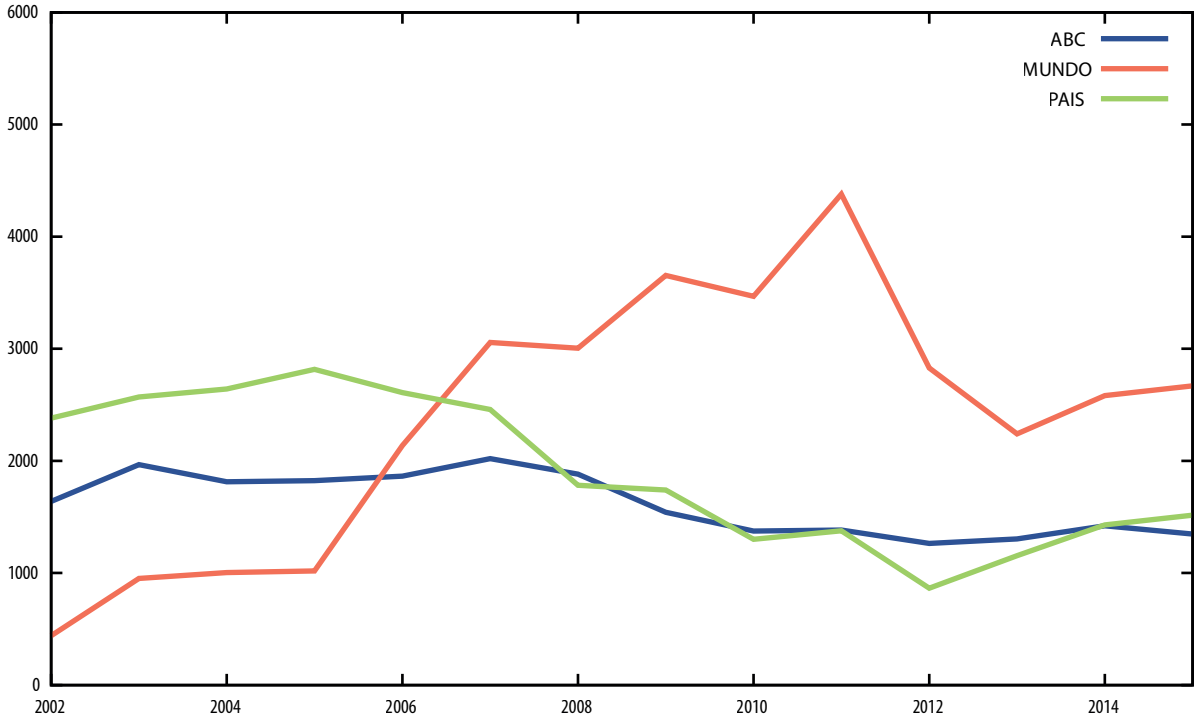
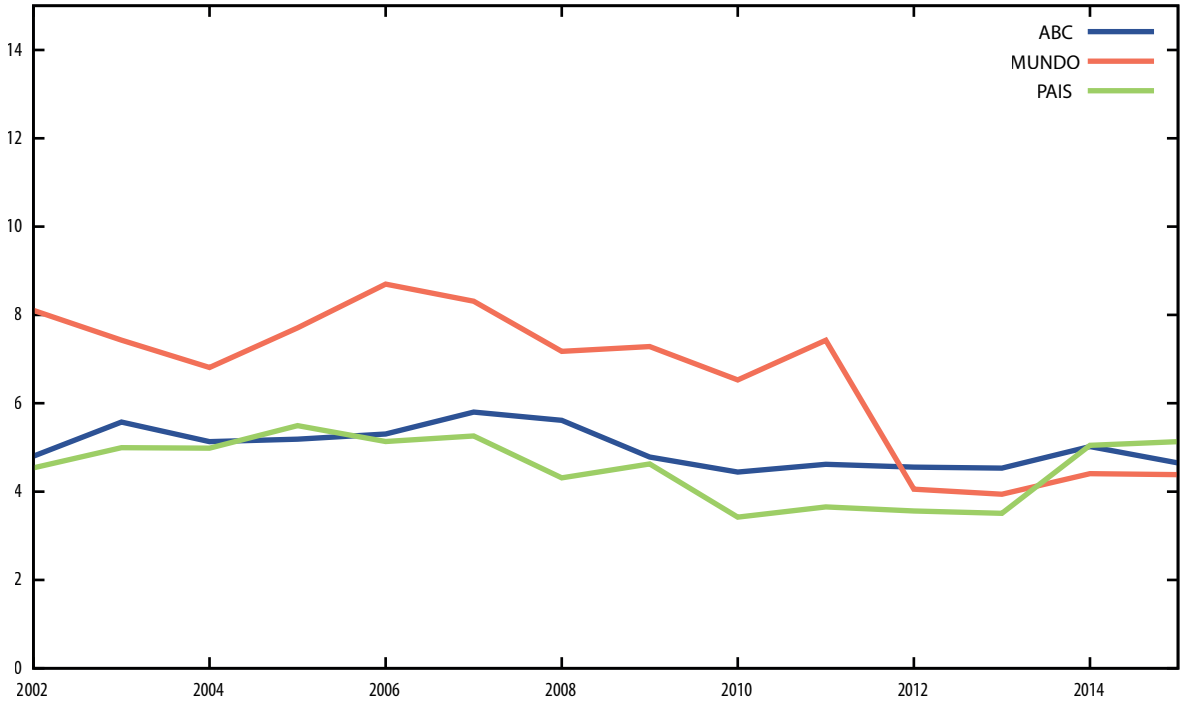


**Gráfico / Noticias Ciencia y Tecnología recolectadas (%)**



# CIENCIA VS. TECNOLOGÍA

Podemos preguntarnos acerca de la proporción de noticias sobre Ciencia y noticias sobre Tecnología, incluso noticias sobre ambas cosas. Si abordamos esta cuestión mediante un categorizador automático tenemos dos opciones: una, considerar ambas categorías (Ciencia, Tecnología) como categorías autónomas y categorizar cada noticia como Ciencia / No Ciencia y como Tecnología / No Tecnología. Este planteamiento tiene la ventaja de detectar noticias que tratan de ambas cosas, Ciencia y Tecnología, al mismo tiempo. Pero también el inconveniente de que pueda haber noticias clasificadas como No Ciencia y No Tecnología al mismo tiempo.

Este último es el caso de 1.595 (=1.93 %) noticias, las cuales fueron seleccionadas por el categorizador como de Ciencia y/o Tecnología pero para las cuales el categorizador no ha sido capaz de decidir si tratan de Ciencia o de Tecnología.

En cualquier caso, según este planteamiento, 61.360 noticias trataban sobre Ciencia, cifra equivalente al 74.19 %. Mientras que 23.780 trataban sobre Tecnología (= 28.75 %) y 4.025 eran de ambas cosas (=4.87 %), solapándose con las cifras apuntadas.

El otro enfoque es considerar ambas categorías como excluyentes y obligar al categorizador a clasificar obligatoriamente cada noticia enuna de las dos categorías. Esta opción no permite noticias ambivalentes, pero tampo deja noticias inclasificadas. Según este punto de vista, 50.052 noticias (el 60.51 %) son de Ciencia, mientras que 32.667 noticias (39.49 %) son de Tecnología.

El detalle de datos por periódico y año estudiado puede observarse en las tablas y gráficos siguientes, pero, independientemente del punto d evista adoptado, pero parece haber un claro predominio de la Ciencia sobre la Tecnología en las noticias del SCSC.

**Tabla / Noticias solo Ciencia recolectadas**

| YEAR | ABC | MUNDO | PAIS |
|---|---|---|---|
| 2002 | 1217.0 | 334.0 | 1778.0 |
| 2003 | 1471.0 | 749.0 | 1861.0 |
| 2004 | 1401.0 | 741.0 | 1948.0 |
| 2005 | 1387.0 | 768.0 | 2141.0 |
| 2006 | 1433.0 | 1647.0 | 1933.0 |
| 2007 | 1492.0 | 2452.0 | 1738.0 |
| 2008 | 1376.0 | 2340.0 | 1190.0 |
| 2009 | 1141.0 | 2860.0 | 1266.0 |
| 2010 | 1002.0 | 2565.0 | 885.0 |
| 2011 | 884.0 | 3157.0 | 783.0 |
| 2012 | 901.0 | 2017.0 | 635.0 |
| 2013 | 973.0 | 1699.0 | 853.0 |
| 2014 | 1101.0 | 1948.0 | 1152.0 |
| 2015 | 1020.0 | 1891.0 | 1239.0 |

**Tabla / Noticias solo Tecnología recolectadas**

| YEAR | ABC | MUNDO | PAIS |
|---|---|---|---|
| 2002 | 441.0 | 131.0 | 696.0 |
| 2003 | 529.0 | 262.0 | 786.0 |
| 2004 | 426.0 | 312.0 | 768.0 |
| 2005 | 464.0 | 286.0 | 781.0 |
| 2006 | 454.0 | 513.0 | 729.0 |
| 2007 | 567.0 | 653.0 | 778.0 |
| 2008 | 582.0 | 731.0 | 629.0 |
| 2009 | 508.0 | 844.0 | 503.0 |
| 2010 | 477.0 | 993.0 | 454.0 |
| 2011 | 600.0 | 1336.0 | 623.0 |
| 2012 | 464.0 | 926.0 | 247.0 |
| 2013 | 408.0 | 635.0 | 322.0 |
| 2014 | 369.0 | 709.0 | 303.0 |
| 2015 | 368.0 | 857.0 | 316.0 |

**Tabla / Noticias solo Ciencia recolectadas (% por periódico y año)**

| YEAR | ABC | MUNDO | PAIS |
|---|---|---|---|
| 2002 | 74.2525930445 | 75.7369614512 | 74.6745065099 |
| 2003 | 74.7839349263 | 78.6764705882 | 72.4124513619 |
| 2004 | 77.2326350606 | 73.8047808765 | 73.7599394169 |
| 2005 | 76.0416666667 | 75.442043222 | 76.0298295455 |
| 2006 | 76.8776824034 | 77.0706598035 | 74.0896895362 |
| 2007 | 73.8613861386 | 80.2356020942 | 70.6791378609 |
| 2008 | 73.1137088204 | 77.8702163062 | 66.7789001122 |
| 2009 | 74.0428293316 | 78.2703886152 | 72.7586206897 |
| 2010 | 72.9257641921 | 73.9619377163 | 68.0245964643 |
| 2011 | 63.9190166305 | 72.0940854076 | 56.862745098 |
| 2012 | 71.2816455696 | 71.2972781902 | 73.4953703704 |
| 2013 | 74.6165644172 | 75.8143685855 | 73.9168110919 |
| 2014 | 77.4806474314 | 75.4453911696 | 80.6158152554 |
| 2015 | 75.6676557864 | 70.8505058074 | 81.72823219 |

**Tabla / Noticias solo Tecnología recolectadas (% % Ciencia por periódico y año)**

| YEAR | ABC | MUNDO | PAIS |
|---|---|---|---|
| 2002 | 26.9066503966 | 29.7052154195 | 29.2314153717 |
| 2003 | 26.8937468226 | 27.5210084034 | 30.5836575875 |
| 2004 | 23.4840132304 | 31.0756972112 | 29.0798939796 |
| 2005 | 25.4385964912 | 28.094302554 | 27.734375 |
| 2006 | 24.356223176 | 24.0056153486 | 27.9417401303 |
| 2007 | 28.0693069307 | 21.3678010471 | 31.6388775925 |
| 2008 | 30.9245483528 | 24.3261231281 | 35.2974186308 |
| 2009 | 32.9656067489 | 23.0979748221 | 28.908045977 |
| 2010 | 34.7161572052 | 28.6332179931 | 34.8962336664 |
| 2011 | 43.3839479393 | 30.5092486869 | 45.2432824982 |
| 2012 | 36.7088607595 | 32.7324142807 | 28.587962963 |
| 2013 | 31.2883435583 | 28.3355644801 | 27.9029462738 |
| 2014 | 25.9676284307 | 27.4593338497 | 21.2036389083 |
| 2015 | 27.2997032641 | 32.1094042713 | 20.8443271768 |

**Tabla / Noticias Ciencia y Tecnología recolectadas**

| YEAR | CIENCIA | TECNOLOGÍA |
|---|---|---|
| 2002 | 3329.0 | 1268.0 |
| 2003 | 4081.0 | 1577.0 |
| 2004 | 4090.0 | 1506.0 |
| 2005 | 4296.0 | 1531.0 |
| 2006 | 5013.0 | 1696.0 |
| 2007 | 5682.0 | 1998.0 |
| 2008 | 4906.0 | 1942.0 |
| 2009 | 5267.0 | 1855.0 |
| 2010 | 4452.0 | 1924.0 |
| 2011 | 4824.0 | 2559.0 |
| 2012 | 3553.0 | 1637.0 |
| 2013 | 3525.0 | 1365.0 |
| 2014 | 4201.0 | 1381.0 |
| 2015 | 4150.0 | 1541.0 |

**Tabla / Noticias de Ciencia y noticias de Tecnología (categorías excluyentes) por diario y año**

| YEAR | ABC CIENCIA | ABC TECNOLOGÍA | MUNDO CIENCIA | MUNDO TECNOLOGÍA | PAIS CIENCIA | PAIS TECNOLOGÍA |
|---|---|---|---|---|---|---|
| 2002 | 956.0 | 683.0 | 275.0 | 166.0 | 1459.0 | 922.0 |
| 2003 | 1228.0 | 739.0 | 636.0 | 316.0 | 1598.0 | 972.0 |
| 2004 | 1159.0 | 655.0 | 591.0 | 413.0 | 1608.0 | 1033.0 |
| 2005 | 1049.0 | 775.0 | 549.0 | 469.0 | 1678.0 | 1138.0 |
| 2006 | 1152.0 | 712.0 | 1376.0 | 761.0 | 1568.0 | 1041.0 |
| 2007 | 1206.0 | 814.0 | 2138.0 | 918.0 | 1377.0 | 1082.0 |
| 2008 | 1023.0 | 859.0 | 1991.0 | 1014.0 | 925.0 | 857.0 |
| 2009 | 869.0 | 672.0 | 2488.0 | 1166.0 | 1031.0 | 709.0 |
| 2010 | 710.0 | 664.0 | 2233.0 | 1235.0 | 693.0 | 608.0 |
| 2011 | 619.0 | 764.0 | 2738.0 | 1641.0 | 612.0 | 765.0 |
| 2012 | 677.0 | 587.0 | 1599.0 | 1230.0 | 506.0 | 358.0 |
| 2013 | 736.0 | 568.0 | 1374.0 | 867.0 | 683.0 | 471.0 |
| 2014 | 911.0 | 510.0 | 1675.0 | 907.0 | 981.0 | 448.0 |
| 2015 | 811.0 | 537.0 | 1535.0 | 1134.0 | 1029.0 | 487.0 |

**Gráfico / Noticias solo Ciencia recolectadas**

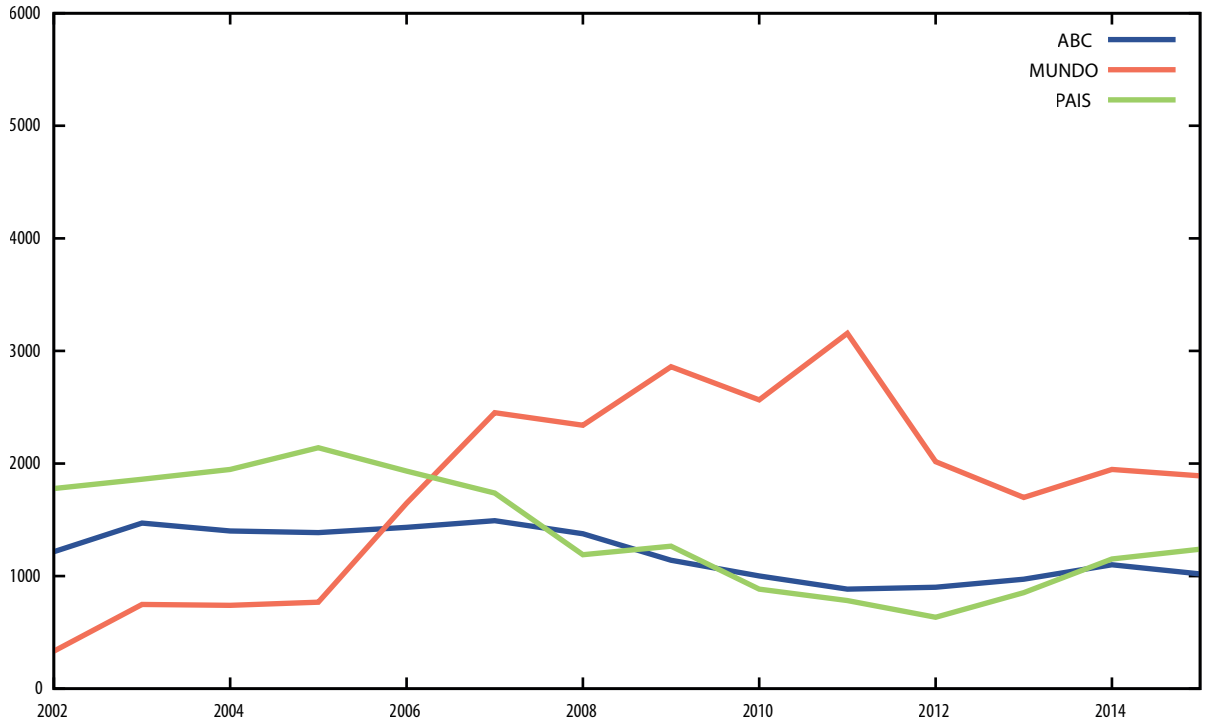



**Gráfico / Noticias solo Tecnología recolectadas**

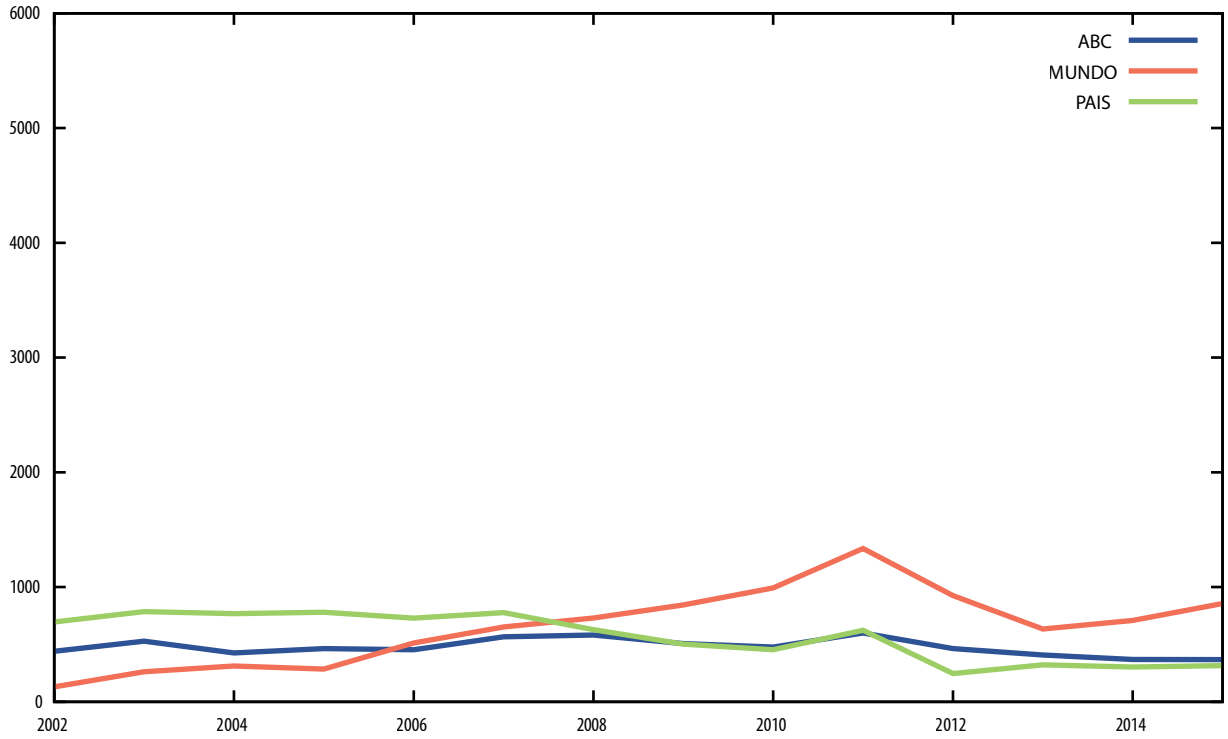


**Gráfico / Noticias solo Ciencia recolectadas (%)**

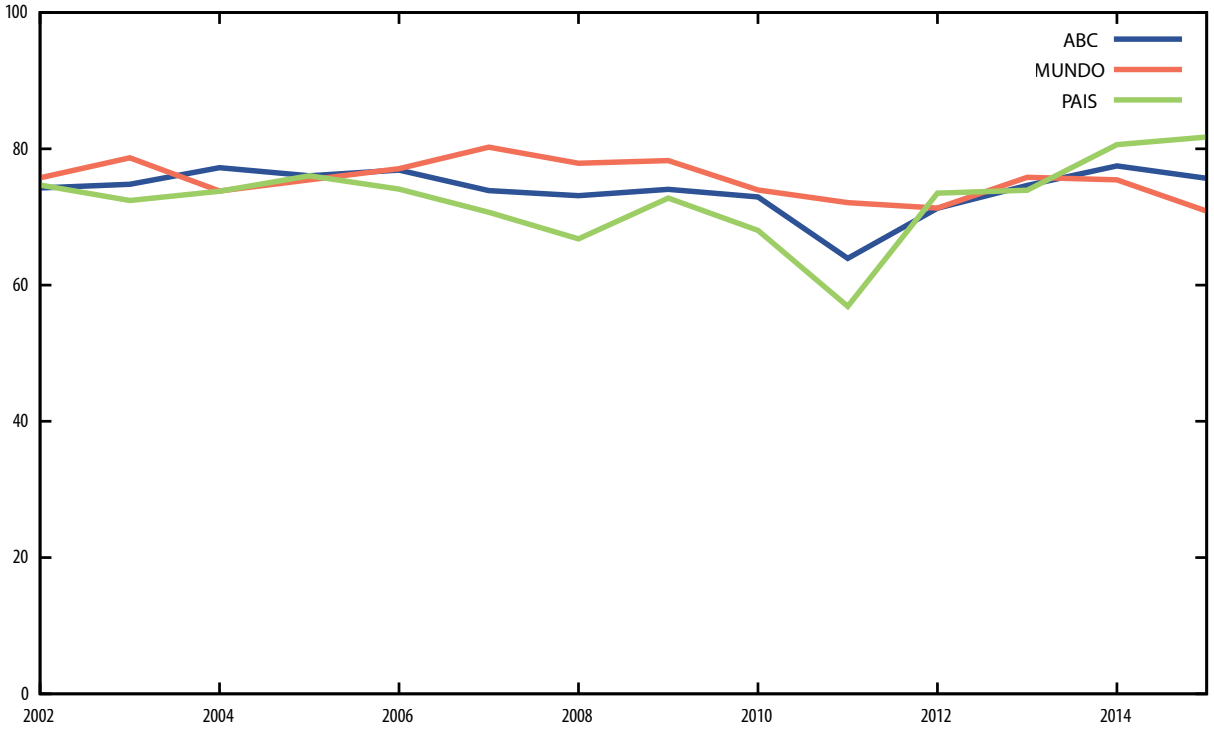



**Gráfico / Noticias solo Tecnología recolectadas (%)**

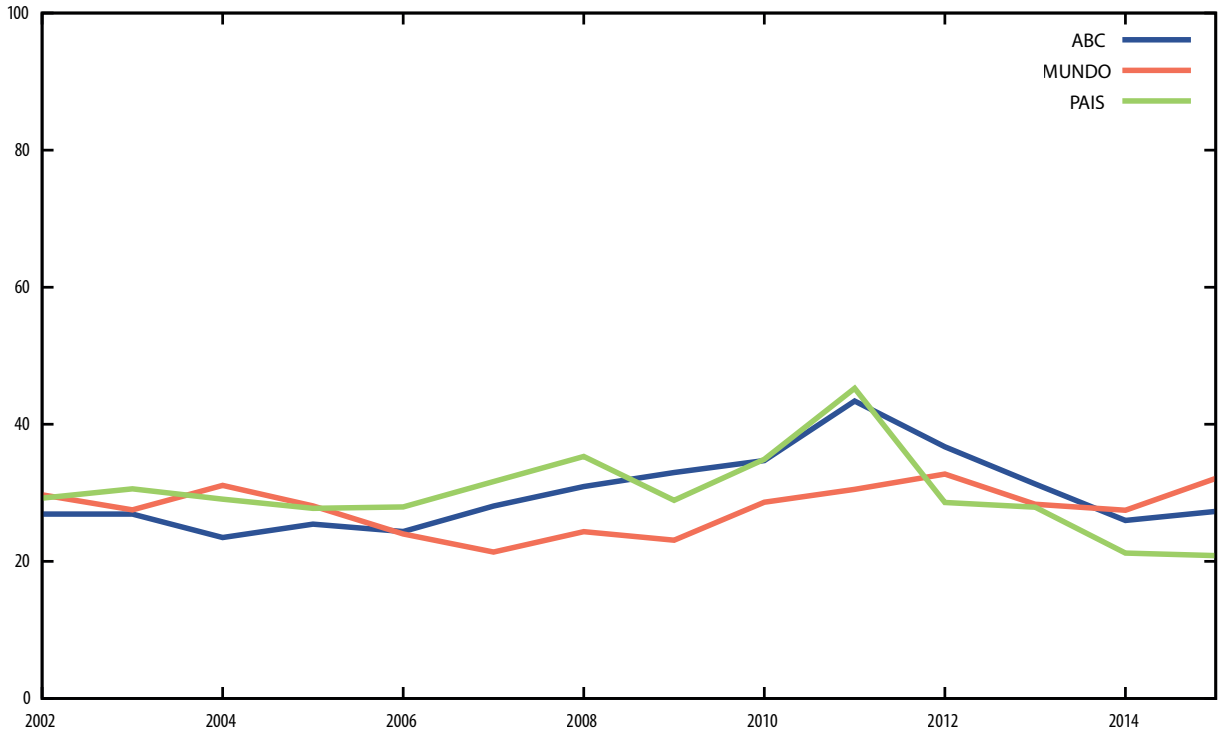


**Gráfico / Noticias Ciencia y Tecnología recolectadas**

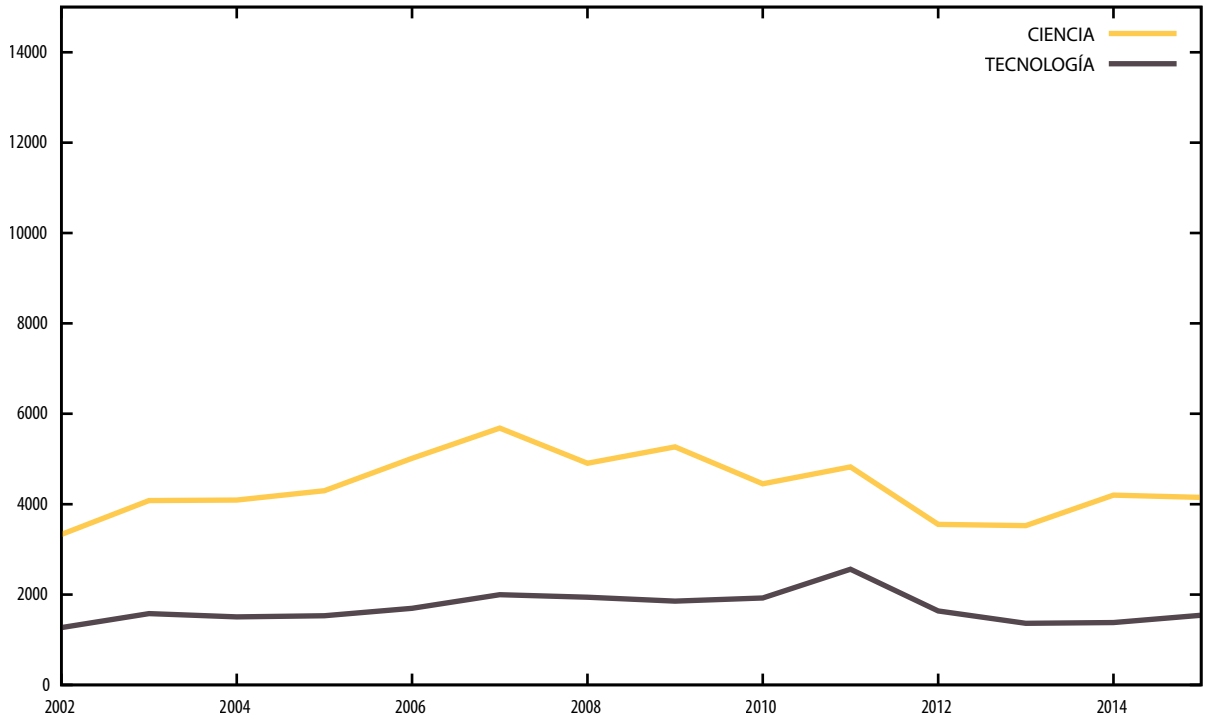



**Gráfico / Noticias Ciencia vs. Tecnología**

# CATEGORÍAS DEL MODELO TEÓRICO DE CULTURA CIENTÍFICA

El Modelo Teórico de Cultuta Científica distingue entre Ciencia Intrínseca y Extrínseca (Quintanilla, 2005; Quintanilla, 2012). Teniendo en cuenta que previamente hemos diferenciado entre noticias sobre Ciencia y noticias sobre Tecnología, nos encontramos con cuatro categorías en las cuales hemos de clasificar las noticias analizadas. Estas categorías, al igual que sucede con Ciencia y Tecnología, no son excluyentes; de manera que una misma noticia puede categorizarse en varias de ellas simultáneamente.

No obstante, la definición de los modelos en términos de categorización automática es menos diáfana que la categorización entre Ciencia/No Ciencia y semás aplicadas hasta ahora. Debido a ello, el categorizador ha sido incapaz de clasificar de forma clara una notable cantidad de noticias: 25.853 (31.25 %). Dado que el categorizador arroja un índice de confianza en la clasificación hecha, siempre es posible establecer un umbral más tolerante para adscribir a uno y/o otro modelo cada noticia. Sin embargo hemos preferido mantener las decisiones originales del categorizador por considerarlas más fiables; el precio, obviamente, es dejar sin adscribir una cantidad importante de noticias.

**Tabla / Noticias Ciencia Intrínseca**

| YEAR | ABC | MUNDO | PAIS |
|---|---|---|---|
| 2002 | 625.0 | 245.0 | 1114.0 |
| 2003 | 836.0 | 562.0 | 1229.0 |
| 2004 | 775.0 | 510.0 | 1187.0 |
| 2005 | 702.0 | 407.0 | 1250.0 |
| 2006 | 787.0 | 1112.0 | 1230.0 |
| 2007 | 851.0 | 1719.0 | 1050.0 |
| 2008 | 716.0 | 1609.0 | 681.0 |
| 2009 | 468.0 | 1705.0 | 626.0 |
| 2010 | 426.0 | 1703.0 | 452.0 |
| 2011 | 395.0 | 2292.0 | 453.0 |
| 2012 | 406.0 | 1202.0 | 391.0 |
| 2013 | 463.0 | 1083.0 | 502.0 |
| 2014 | 610.0 | 1323.0 | 772.0 |
| 2015 | 526.0 | 1190.0 | 784.0 |

**Tabla / Noticias Ciencia Extrínseca**

| YEAR | ABC | MUNDO | PAIS |
|---|---|---|---|
| 2002 | 332.0 | 87.0 | 622.0 |
| 2003 | 339.0 | 153.0 | 649.0 |
| 2004 | 326.0 | 168.0 | 635.0 |
| 2005 | 267.0 | 108.0 | 620.0 |
| 2006 | 282.0 | 278.0 | 466.0 |
| 2007 | 232.0 | 357.0 | 402.0 |
| 2008 | 231.0 | 334.0 | 264.0 |
| 2009 | 172.0 | 306.0 | 236.0 |
| 2010 | 168.0 | 364.0 | 197.0 |
| 2011 | 156.0 | 475.0 | 173.0 |
| 2012 | 148.0 | 247.0 | 142.0 |
| 2013 | 164.0 | 253.0 | 174.0 |
| 2014 | 138.0 | 277.0 | 141.0 |
| 2015 | 154.0 | 266.0 | 126.0 |

**Tabla / Noticias Tecnología Intrínseca**

| YEAR | ABC | MUNDO | PAIS |
|---|---|---|---|
| 2002 | 53.0 | 23.0 | 140.0 |
| 2003 | 53.0 | 39.0 | 103.0 |
| 2004 | 47.0 | 38.0 | 102.0 |
| 2005 | 59.0 | 27.0 | 116.0 |
| 2006 | 68.0 | 83.0 | 146.0 |
| 2007 | 80.0 | 88.0 | 126.0 |
| 2008 | 62.0 | 121.0 | 96.0 |
| 2009 | 39.0 | 156.0 | 124.0 |
| 2010 | 58.0 | 177.0 | 128.0 |
| 2011 | 58.0 | 204.0 | 118.0 |
| 2012 | 30.0 | 114.0 | 44.0 |
| 2013 | 45.0 | 49.0 | 42.0 |
| 2014 | 30.0 | 73.0 | 26.0 |
| 2015 | 19.0 | 78.0 | 38.0 |

**Tabla / Noticias Tecnología Extrínseca**

| YEAR | ABC | MUNDO | PAIS |
|---|---|---|---|
| 2002 | 304.0 | 88.0 | 528.0 |
| 2003 | 379.0 | 185.0 | 603.0 |
| 2004 | 320.0 | 201.0 | 552.0 |
| 2005 | 307.0 | 195.0 | 565.0 |
| 2006 | 312.0 | 441.0 | 560.0 |
| 2007 | 392.0 | 530.0 | 542.0 |
| 2008 | 346.0 | 565.0 | 478.0 |
| 2009 | 320.0 | 667.0 | 419.0 |
| 2010 | 268.0 | 695.0 | 352.0 |
| 2011 | 360.0 | 943.0 | 500.0 |
| 2012 | 243.0 | 557.0 | 172.0 |
| 2013 | 230.0 | 355.0 | 229.0 |
| 2014 | 226.0 | 438.0 | 211.0 |
| 2015 | 202.0 | 552.0 | 204.0 |

**Tabla / Categorías del Modelo Teórico, datos globales**

| TH_MODEL | NEWS |
|---|---|
| Intr_Sci | 36969 |
| Extr_Sci | 11629 |
| Intr_Tech | 3320 |
| Extr_Tech | 16536 |

**Tabla / Categorías del Modelo Teórico**

| YEAR | INTRINSEC_SCI | EXTRINSEC_SCI | INTRINSEC_TECH | EXTRINSEC_TECH |
|---|---|---|---|---|
| 2002 | 1984.0 | 1041.0 | 216.0 | 920.0 |
| 2003 | 2627.0 | 1141.0 | 195.0 | 1167.0 |
| 2004 | 2472.0 | 1129.0 | 187.0 | 1073.0 |
| 2005 | 2359.0 | 995.0 | 202.0 | 1067.0 |
| 2006 | 3129.0 | 1026.0 | 297.0 | 1313.0 |
| 2007 | 3620.0 | 991.0 | 294.0 | 1464.0 |
| 2008 | 3006.0 | 829.0 | 279.0 | 1389.0 |
| 2009 | 2799.0 | 714.0 | 319.0 | 1406.0 |
| 2010 | 2581.0 | 729.0 | 363.0 | 1315.0 |
| 2011 | 3140.0 | 804.0 | 380.0 | 1803.0 |
| 2012 | 1999.0 | 537.0 | 188.0 | 972.0 |
| 2013 | 2048.0 | 591.0 | 136.0 | 814.0 |
| 2014 | 2705.0 | 556.0 | 129.0 | 875.0 |
| 2015 | 2500.0 | 546.0 | 135.0 | 958.0 |

**Gráfico / Noticias Ciencia Intrínseca**

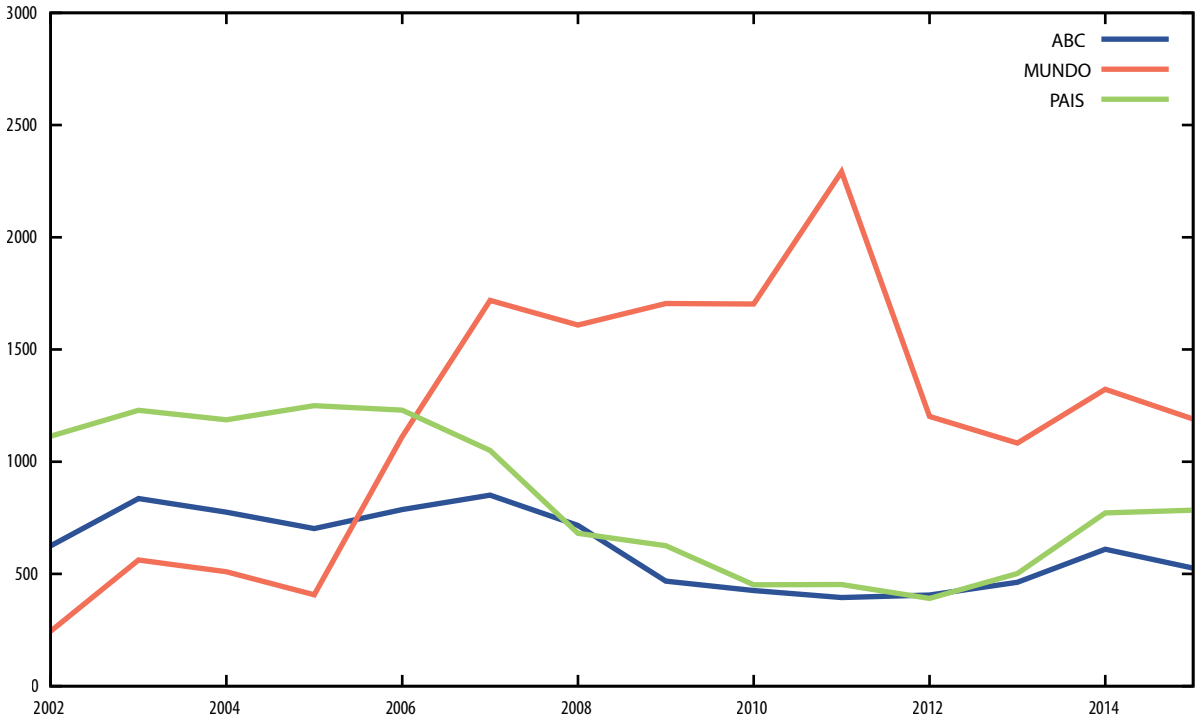


**Gráfico / Noticias Ciencia Extrínseca**

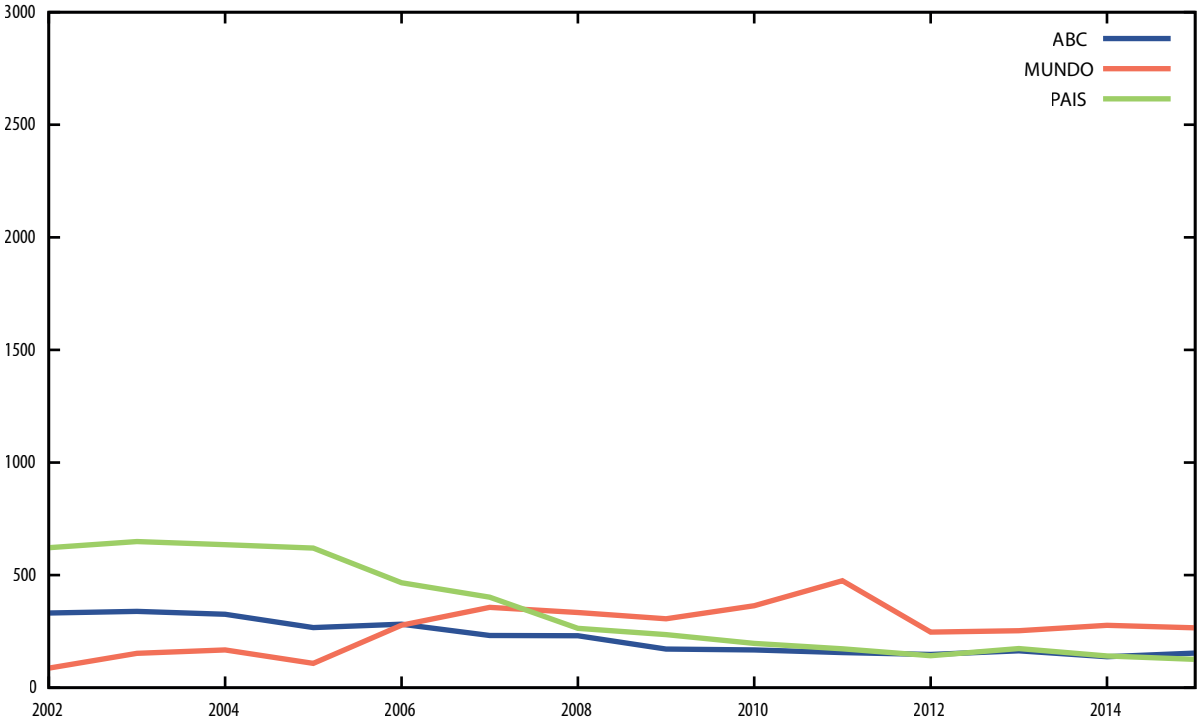



**Gráfico / Noticias Tecnología Intrínseca**

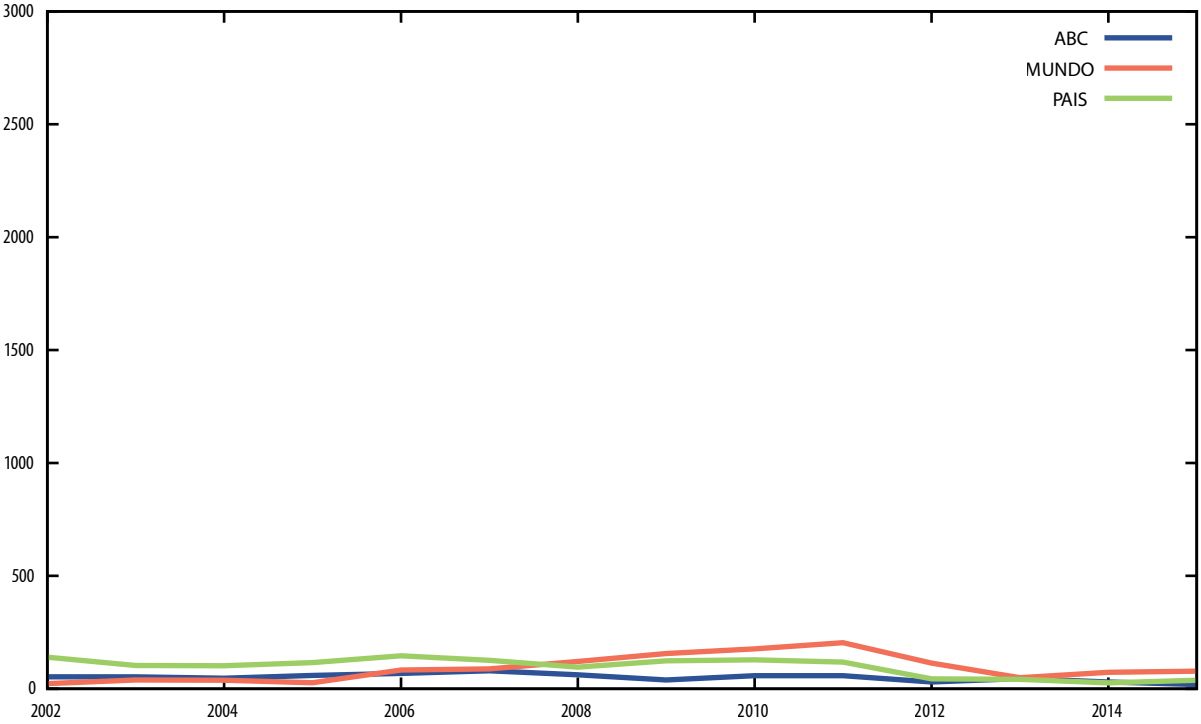


**Gráfico / Noticias Tecnología Extrínseca**

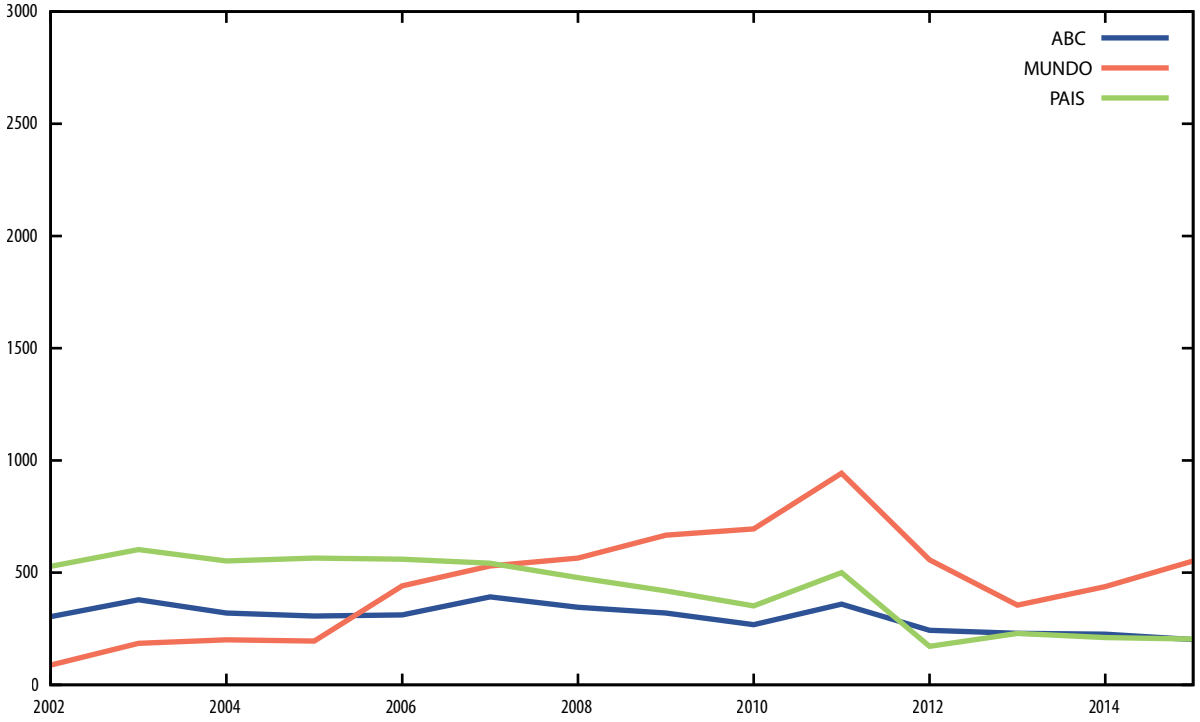



**Gráfico / Theorethic Model**

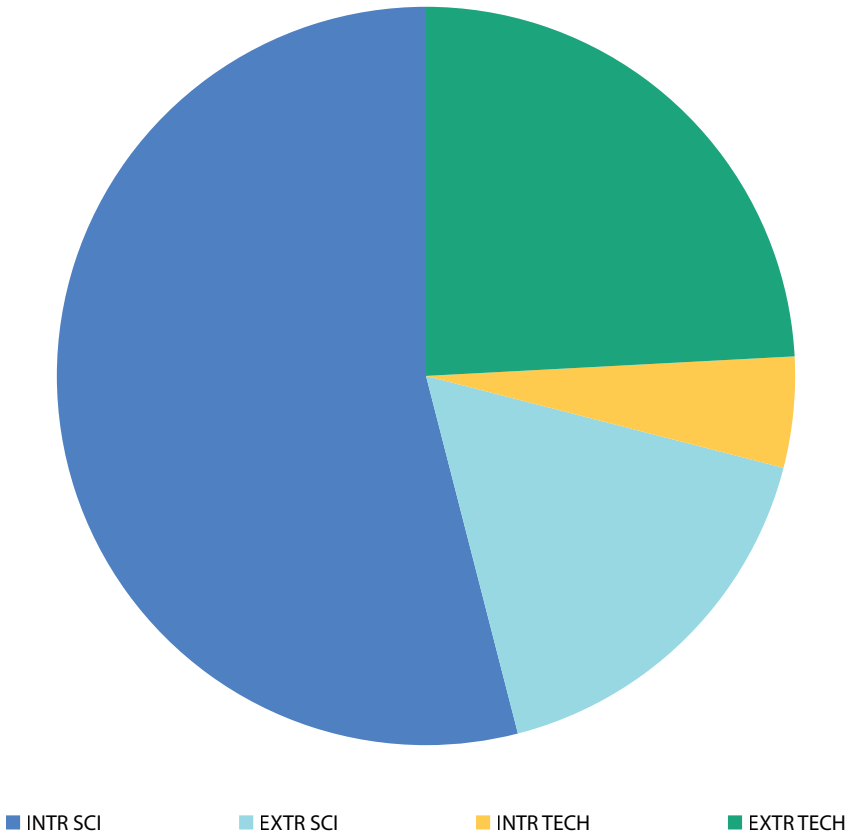
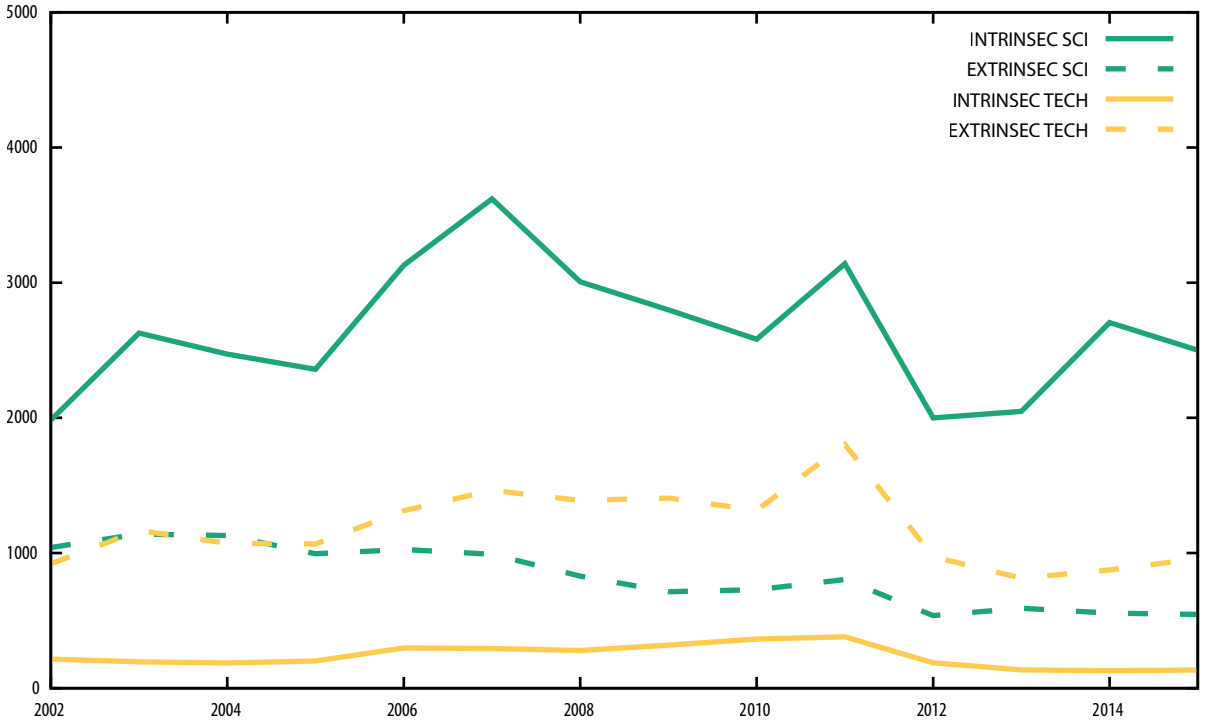


**Gráfico / Ciencia y Tecnología Intrínseca vs. Extrínseca**



# DISTRIBUCIÓN TEMÁTICA

Un aspecto interesante es el de los temas tratados por las noticias; es decir, dentro de lo que podemos considerar Ciencia y/o Tecnología las noticias se ocupan de forma más o menos preferente de determinados temas que no necesariamente coinciden con las disciplinas clásicas en que suele organizarse la Ciencia; de hecho coinciden poco.

Desde un punto de vista metodológico, el problema  es determinar de forma automática los temas tratados ene cnjunto de noticias que conforman el SCSC. Diversas tecnologías están disponibles; una de las más frecuentes es la aplicación de Topic Modeling, a través de alguna de las implementaciones de lo que se conoce como Latent Dirichlet Allocation (LDA) (Blei et al., 2003; Chaney and Blei, 2012). La idea básica es que cada documento trata de diversos temas en una proporción o porcentaje variable; las técnicas como LDA permiten descubrir cuáles son los temas tratados en el conjunto de la colección de noticias, y, además, la proporción de cada uno de esos temas presenta en cada documento particular.

Otra alternativa es la aplicación de técnicas de clustering de documentos, en la confianza de que los documentos (noticias, en nuestro caso) con contenidos temáticos más o menos parecidos serán agrupados juntos; una revisión manual de los clusters conseguidos y su correspondiente etiquetado debería entregarnos un listado de los temas tratados en las noticias del SCSC, y además su intensidad, representada por el tamaño de cada cluster; es decir por el número de noticias contenidas en cada uno de esos clusters.
Existen diferentes técnicas de clusering aplicables a documentos de texto (Steinbach et al., 2000; Croft et al., 2010) como las noticias del SCSC. En general suelen ser bastante exigentes en lo que a capacidad de proceso necesario se refiere. Algunas de las más frecuentes -las basadas en el algoritmo k-means (Jain, 2010), por ejemplo- exigen prefijar de antemano el número de clusters (temas) deseado, lo cual implica, de una u otra forma, un sesgo final hacia temas más abundantes en noticias ( o viceversa, si se establece un número elevado de clusters).

Por nuestra parte, después de experimentar con varias de estas técnicas, hemos aplicado una metodología basada en las Técnicas de Análisis de Redes (Otte y Rousseau, 2002; Scott, 2013). Muy brevemente, una red es un conjunto de nodos enlazados entre sí por arcos, que representan algún tipo de relación entre tales nodos. Los arcos, además, pueden tener, entre

otros atributos, peso; es decir, un valor numérico que representa la intensidad o fuerza de esa relación.

Si representamos esa red de forma que los nodos enlazados con arcos de más peso (relacionados con mayor intensidad) tiendan a situarse más próximos, y viceversa; nos encontraremos con comunidades de nodos que enlazan con intensidad entre ellos y de forma más débil con los nodos ajenos a esa comunidad.

Podemos utilizar un artefacto de este tipo para modelar el corpus de noticias científicas SCSC, de forma que cada noticia sea un nodo, y los arcos expresen el parecido o similitud semántica entre cada par de nodos o documentos (noticias). Esta similitud semántica puede calcularse mediante técnicas aplicadas y bien conocidas en el campo de la Recuperación de Información (Salton y Buckley, 1988). En una red de estas características, cada comunidad de nodos representa un conjunto de noticias similares semánticamente; es decir, que tratan sobre temas parecidos. Parece razonable interpretar cada comunidad como un área temática; el tamaño (número de nodos o noticias) de cada uno de esas comunidades-áreas temáticasnos estaría diciendo la intensidad conque ese área temática es tratada en el conjunto del SCSC.

Existen diversos algoritmos que permiten descubrir comunidades en una red de estas características; el principal problema es la cantidad de memoria y tiempo de proceso que requieren (Clauset et al., 2004; Leskovec et al., 2010). Uno de los algoritmos más recomendables para descubrir comunidades en redes grandes es InfoMap, y éste es el que hemos aplicado en este trabajo (Bohlin et al., 2014; Edler y Rosvall, 2015).
InfoMap es parametrizable en diferentes aspectos; pero en una configuración estándar produce un listado de comunidades en dos niveles (comunidad y subcomunidad; tema y subtema en nuestro caso), en número variable según las afinidades y grupamientos de documentos efectuados por el mismo algoritmo.

En el caso del SCSC, se obtuvieron siete comunidades de primer nivel, cada una de ellas con más o menos subcomunidades. Las seis comunidades de primer nivel más grandes (en número de documentos), después de una revisión manual, representan seis grandes áreas temáticas. La séptima consiste en una agrupación de noticias difćilmente encuadrables en alguna de las otras seis, de un carácter más bien heterogéneo.

El análisis de las subcomunidades parece, en una primera aproximación, representar también un conjunto de subtemas bastante coherente; pero el elevado número de ellos hace aconsejable postergar su análisis, pues requiere un tiempo considerable.

**Tabla / Grandes temas y número de noticias**

| TOPIC | NEWS |
|---|---|
| Aerospace_and_Astronomy | 8385 |
| Energy_and_Environment | 28569 |
| Health_Sci | 28725 |
| Information_Tech | 7805 |
| Others | 3115 |
| Paleontology_Evolution | 2845 |
| Science_Policy | 3236 |

**Tabla / Distribución longitudinal de noticias por grandes temas**

| YEAR | HEALTH SCI | INFORMATION TECH | ENERGY ENVIRONMENT | AEROSPACE | SCIENCE POLICY | PALEONTOLOGY | OTHERS |
|---|---|---|---|---|---|---|---|
| 2002 | 1702.0 | 1436.0 | 452.0 | 371.0 | 240.0 | 115.0 | 144.0 |
| 2003 | 2327.0 | 1455.0 | 693.0 | 402.0 | 343.0 | 129.0 | 138.0 |
| 2004 | 2038.0 | 1586.0 | 695.0 | 373.0 | 388.0 | 155.0 | 222.0 |
| 2005 | 1773.0 | 2143.0 | 661.0 | 325.0 | 335.0 | 185.0 | 234.0 |
| 2006 | 2341.0 | 2445.0 | 640.0 | 575.0 | 213.0 | 184.0 | 205.0 |
| 2007 | 2640.0 | 2881.0 | 638.0 | 680.0 | 198.0 | 287.0 | 207.0 |
| 2008 | 2135.0 | 2465.0 | 623.0 | 773.0 | 169.0 | 243.0 | 260.0 |
| 2009 | 2786.0 | 2244.0 | 543.0 | 754.0 | 163.0 | 239.0 | 202.0 |
| 2010 | 1925.0 | 2265.0 | 500.0 | 820.0 | 161.0 | 238.0 | 233.0 |
| 2011 | 2409.0 | 2605.0 | 562.0 | 857.0 | 187.0 | 241.0 | 274.0 |
| 2012 | 1238.0 | 1866.0 | 555.0 | 527.0 | 237.0 | 214.0 | 316.0 |
| 2013 | 1372.0 | 1730.0 | 543.0 | 365.0 | 241.0 | 199.0 | 247.0 |
| 2014 | 2193.0 | 1586.0 | 625.0 | 460.0 | 163.0 | 200.0 | 205.0 |
| 2015 | 1846.0 | 1862.0 | 655.0 | 523.0 | 198.0 | 216.0 | 228.0 |

**Gráfico / Main Topics**

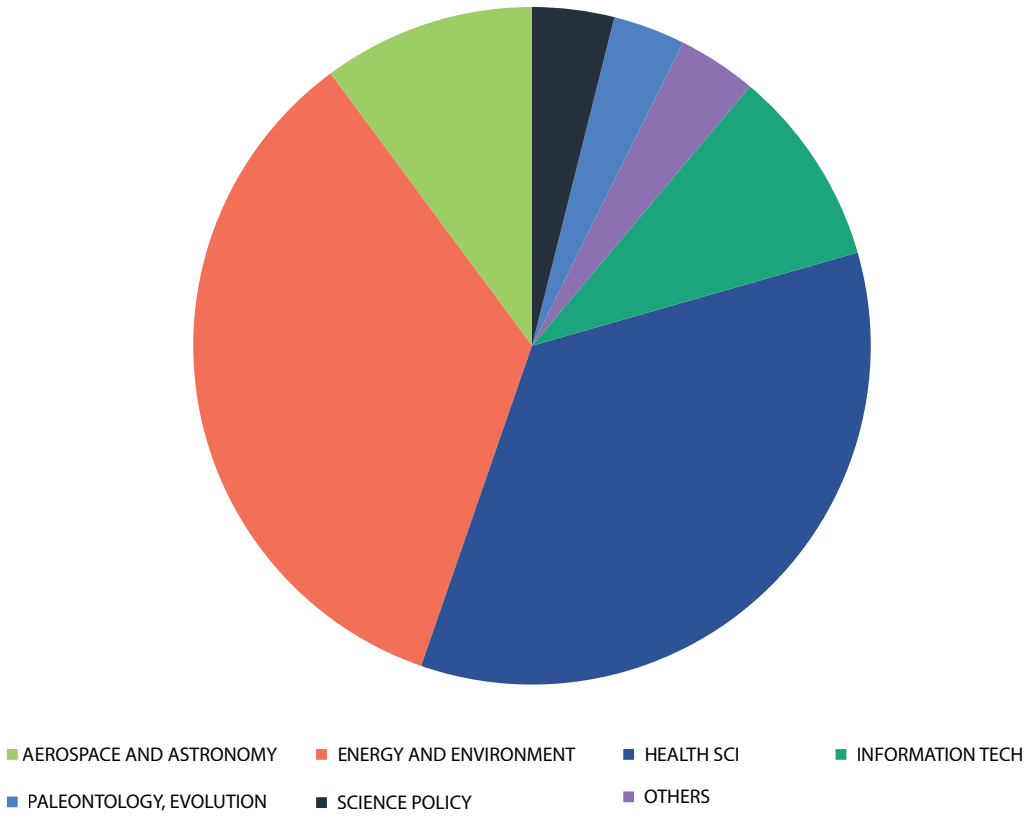
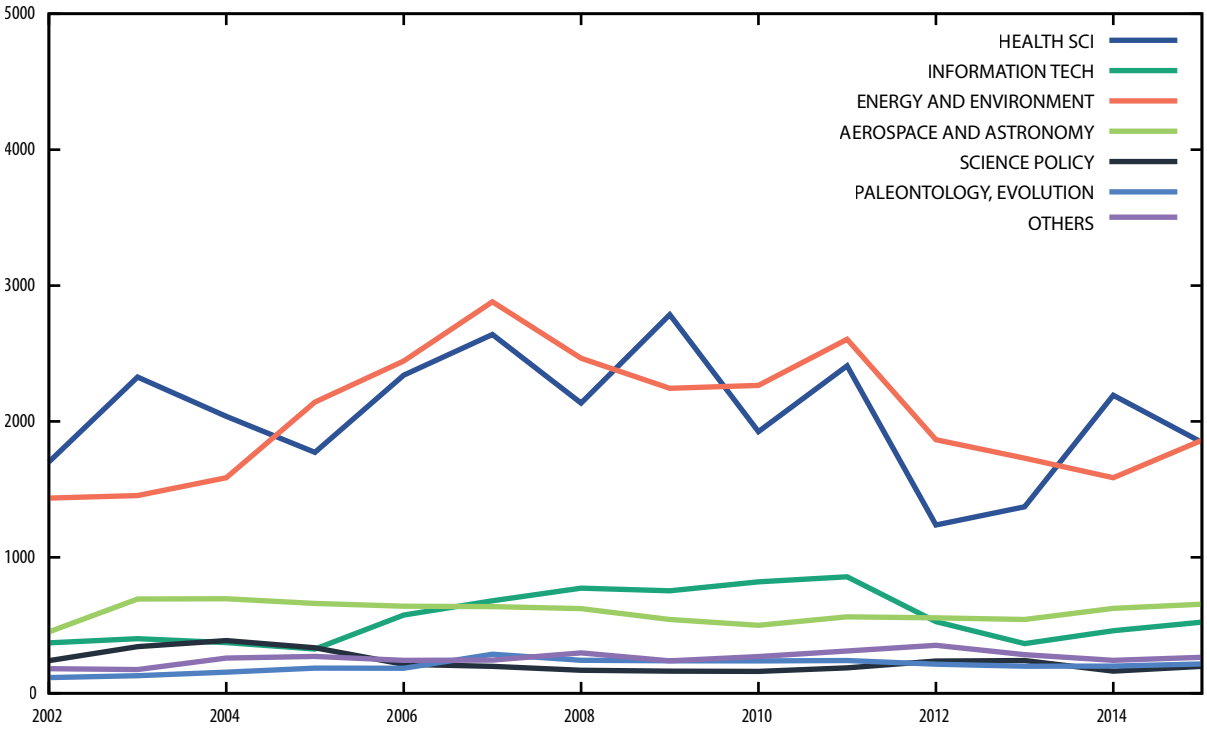



**Gráfico / Main Topics**



# BIBLIOGRAFÍA

Bauer MW (2009) The evolution of public understanding of science – Discourse and comparative evidence. Science, Technology and Society 14(2): 221–240

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022. http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Bohlin, L., Edler, D., Lancichinetti, A., & Rosvall, M. (2014). Community detection and visualization of networks with the map equation framework. In Measuring Scholarly Impact (pp. 3-34). Springer International Publishing.

Castillo, C. (2004). Effective Web Crawling. (Ph.D. thesis). University of Chile. http://chato.cl/research/crawling_thesis

Chaney, A. J. B., & Blei, D. M. (2012, March). Visualizing Topic Models. In ICWSM. http://www.cs.columbia.edu/~blei/papers/ChaneyBlei2012.pdf

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. Physical review E, 70(6), 066111. http://arxiv.org/pdf/cond-mat/0408187

Croft, W. B., Metzler, D., & Strohmann, T. (2010). Search engines. Pearson Education. https://pdfs.semanticscholar.org/4c9f/afa3b1bed97bb00b8bc68db39a9ad48490f1.pdf
Edler, D., & Rosvall, M. (2015). The infomap software package. http://www.mapequation.org/code.html

Eyheramendy, S., Lewis, D. D., & Madigan, D. (2003). On the naive bayes model for text categorization. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.20.4949

Figuerola, C.G.; Groves, T. and Quintanilla, M.A. (2014): Science and Technology in digital newspapers, II International Seminar on Indicators of Scientific and Technologycal Culture, Salamanca, November 2014. http://diarium.usal.es/figue/files/2015/01/figuerola2014science.pdf

Hsu, Chih-Wei; Chang, Chih-Chung & Lin, Chih-Jen (2003). A Practical Guide to Support Vector Classification (PDF) (Technical report). Department of Computer Science and Information Engineering, National Taiwan University. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666. http://www.ppgia.pucpr.br/~fabricio/ftp/Roges/JainClustering_PRL10.pdf

Joachims, T. (1998, April). Text categorization with support vector machines: Learning with many relevant features. In European conference on machine learning (pp. 137-142). Springer Berlin Heidelberg. https://eldorado.tu-dortmund.de/bitstream/2003/2595/1/report23_ps.pdf

Kim, S. B., Han, K. S., Rim, H. C., & Myaeng, S. H. (2006). Some effective techniques for naive bayes text classification. IEEE transactions on knowledge and data engineering, 18(11), 1457-1466.http://ir.kaist.ac.kr/papers/2006/some%20effective%20techniques%20for%20naive%20bayes%20text%20classification.pdf

Leskovec, J., Lang, K. J., & Mahoney, M. (2010, April). Empirical comparison of algorithms for network community detection. In Proceedings of the 19th international conference on World wide web (pp. 631-640). ACM. http://arxiv.org/pdf/1004.3539

McCallum, Andrew; Nigam, Kamal (1998). A comparison of event models for Naive Bayes text classification (PDF). AAAI-98 workshop on learning for text categorization. 752. http://www.kamalnigam.com/papers/multinomial-aaaiws98.pdf

Olston, C. and Najork, M. Web Crawling. Invited survey article. Journal of Foundations and Trends in Information Retrieval, 4(3):175-246,2010.http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.163.9242&rep=rep1&type=pdf

Otte, E., & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. Journal of information Science, 28(6), 441-453. http://www.academia.edu/download/42254790/Social_Network_Analysis_A_Powerful_Strat20160206-25456-1pc1lcl.pdf

Quintanilla MA (2005) Tecnología: Un enfoque filosófico y otros ensayos de filosofía de la tecnología [Technology: A philosophical approach and other essays on the philosophy of technology]. México, D.F., México: Fondo de Cultura Económica

Quintanilla MA (2012) Cultura, Tecnología e innovación [Culture, Technology and Innovation]. In: Aibar E, Quintanilla MA (eds) Ciencia, tecnología y sociedad. Enciclopedia Iberoamericana de Filosofía [Science, Technology and Innovation. Iberoamerican Encyclopaedia of Philosophy]. Madrid: Trotta, pp. 103–136.

Quintanilla, M.A.; Figuerola, C.G. and Groves, T. (2014): Ten years of Science News, 13th International Public Communication of Science and Techmology Conference (PCST 2014), Salvador, Brazil

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), 513-523. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.101.9086

Scott, J. (2013). Social network analysis. Thousand Oaks, CA, US: Sage Publications, Inc

Steinbach, M., Karypis, G., & Kumar, V. (2000, August). A comparison of document clustering techniques. In KDD workshop on text mining (Vol. 400, No. 1, pp. 525-526). https://pdfs.semanticscholar.org/c110/0f525044b2b926f7bd7f407ce7b0157bcfd8.pdf

Steinwart, Ingo; and Christmann, Andreas; Support Vector Machines, Springer-Verlag, New York, 2008. ISBN 978-0-387-77241-7

Vapnik, Vladimir N.; The Nature of Statistical Learning Theory, Springer-Verlag, 1995. ISBN 0-387-98780-0