Aplicación de minería de texto para el análisis métrico sobre ahorro energético en proyectos del Séptimo Programa Marco

Application of text mining for the metric analysis of energy saving of the Seventh Framework Programme

Carlos García-Zorita
Sergio Marugán-Lázaro
Daniela De Filippo

RESUMEN:
Objetivos. Este estudio analiza la actividad científica y tecnológica sobre ahorro energético a través de los proyectos del Séptimo Programa Marco. Se utiliza la base de datos europea CORDIS y se emplean herramientas de minería de texto para el análisis de contenido.
Diseño/Metodología/Enfoque. El trabajo se desarrolla en dos fases. La primera utiliza un abordaje cienciométrico para obtener información sobre: países que participan y lideran los proyectos, distribución de la financiación y relación entre número de participantes y fondos obtenidos. La segunda, analiza la frecuencia de términos y tópicos principales de los proyectos.
Resultados/Discusión. Se han concedido 256 proyectos siendo los mayores participantes: Alemania, Francia, Reino Unido, España e Italia. La financiación recibida es muy variable y no hay relación entre el número de países e instituciones por proyecto y el monto conseguido. El análisis de contenido muestra que se destacan los proyectos relacionados con Gestión de la Energía Eléctrica; Obtención de energías alternativas; Vehículo Eléctrico y Eficiencia Energética.
Conclusiones. Los países más activos son los grandes productores. No hay relación entre el tamaño de los equipos de investigación y la financiación captada ya que hay proyectos con pocos países e instituciones participantes y una alta financiación. Los temas de los proyectos son muy variados, algunos son muy específicos (Vehículo Eléctrico) y otros generales relacionados con la gestión energética y la eficiencia urbana.
Originalidad/Valor. La aplicación de metodologías mixtas para el estudio de un campo científico se presenta como un abordaje prometedor para describir el área y analizar su dinámica.
PALABRAS CLAVE: Ahorro Energético; Séptimo Programa Marco; Estudios métricos de la información; Minería de textos.

ABSTRACT:
Objectives. *This study analyses the scientific and technological activity on energy saving through the projects of the Seventh Framework Programme. The European CORDIS database is used and text mining tools are used for content analysis.*
Design/Methodology/Focus. *The work is developed in two phases. The first uses a scientometric approach to obtain information on: countries participating and leading the projects, distribution of funding and the*

Carlos García-Zorita: Departamento de Biblioteconomía y Documentación, Universidad Carlos III de Madrid, España.
czorita@bib.uc3m.es;
0000-0002-6860-8069.

Sergio Marugán-Lázaro: Instituto INAECU (UC3M-UAM), Universidad Carlos III de Madrid, España.
smarugan@pa.uc3m.es
0000-0002-7084-8221.

Daniela De Filippo:* Instituto INAECU (UC3M-UAM), Universidad Carlos III de Madrid, España.
dfilippo@bib.uc3m.es;
0000-0001-9297-9970.

Cómo citar: García-Zorita, C.; Marugán-Lázaro, S. & De Filippo, D. (2018). Aplicación de minería de texto para el análisis métrico sobre ahorro energético en proyectos del Séptimo Programa Marco. *Bibliotecas. Anales de Investigación; 14*(2), 149-163.

Recibido: 14 de febrero de 2018
Revisado: 14 de marzo de 2018
Aceptado: 25 de marzo de 2018

* Autora correspondiente.

relationship between the number of participants and the funds obtained. The second phase focuses on the analysis of the frequency of terms to identify the main topics of the projects.

Results/Discussion. *256 projects have been awarded with the largest participants: Germany, France, United Kingdom, Spain and Italy. The funding received is very variable and there is no relationship between the number of countries and institutions per project and the amount achieved. The content analysis shows that among the main topics are those related to Electric Power Management; Obtaining energy from alternative sources (wind and solar); Electric Vehicles and Energy Efficiency.*

Conclusions. *The most active countries in this field are the large producers. There is no relationship between the size of the research teams and the funding raised as there are projects with few participating countries and institutions and high funding. The topics analyzed by the projects are very varied and there are some specific ones such as the Electric Vehicle and others related to energy management and urban efficiency.*

Originality/Value. *The application of mixed methodologies for the study of a scientific field is presented as a promising approach to describe the area and analyze its dynamics.*

KEYWORDS: *Energy saving; Seventh Framework Programme; Information metrics studies; Text mining.*

"…la cienciometría y la bibliometría se han convertido en uno de los modelos centrales para analizar, medir y evaluar diferentes aspectos de la actividad científica…"

Introducción

El análisis de la actividad científica en un área determinada puede ser abordado desde la perspectiva de su producción científica y tecnológica. En este sentido, la cienciometría y la bibliometría se han convertido en uno de los modelos centrales para analizar, medir y evaluar diferentes aspectos de la actividad científica (Callon *et al.*, 1995). Entre sus usos más frecuentes están el análisis de la productividad científica (de investigadores, grupos, instituciones, disciplinas o países), así como del estudio de su impacto en la propia comunidad científica. Sin embargo, ha tenido otras aplicaciones más innovadoras como la detección de nuevos frentes de investigación y campos emergentes, el estudio de la conformación de redes de colaboración y la identificación de nichos de investigación. Además de la utilización de publicaciones y patentes como fuentes de información, el análisis de los proyectos de investigación también resulta una herramienta clave ya que, en el caso de los proyectos realizados al amparo de programas de I+D competitivos, permite valorar el balance entre la investigación básica y la aplicada, la atención prestada a las líneas o a fuentes de investigación de carácter emergente, a la investigación interdisciplinar o a la investigación con componentes transfronterizos, como es el caso de ciertos problemas medioambientales o socioeconómicos (Plaza, 2001).

En este trabajo en concreto nos proponemos analizar la actividad científica relacionada con "Ahorro Energético" a través del estudio de los proyectos europeos correspondientes a las diferentes convocatorias del Séptimo Programa Marco. Se ha decidido estudiar esta área porque es clave en el desarrollo de la economía ya que los ahorros energéticos conllevan ahorro económico. Asimismo, se producen impactos socioeconómicos positivos relevantes al generarse nuevas actividades económicas. En este sentido, el desarrollo del mercado

de productos y servicios vinculados con la eficiencia energética y el propio ahorro energético generado se ven reflejados en un incremento del PIB y del empleo. Además, existen otros beneficios de tipo ambiental como el ahorro en el uso de recursos naturales o la reducción de emisiones de dióxido de carbono. Si a esto se suma la reducción de la dependencia energética exterior, se advierte que estamos ante un sector de vital importancia económica y estratégica para todos los países.

Uno de los aspectos que generan mayor interés para el estudio del ahorro y eficiencia energética, es su relación con la sostenibilidad energética y el uso de recursos renovables. Este punto ha sido crucial para el desarrollo de políticas de I+D+i nacionales e internacionales, de hecho, en la Unión Europea existe un gran interés en la promoción del uso de fuentes de energía renovables. Este interés se especifica, por ejemplo, en la legislación a la Directiva 2009/28/CE del Parlamento Europeo sobre el fomento del uso de energía procedente de fuentes de renovables y que se esboza en un conjunto de metas a alcanzar para el año 2020. Por otra parte, la OCDE plantea una toma de conciencia en relación con la importancia de los desarrollos sostenibles a través de diversos programas que permitan mejorar su estudio y promoción, especialmente en el campo del crecimiento sostenible y del llamado "crecimiento verde" (OCDE, 2011; OCDE, 2012). Las decisiones políticas en materia de energía renovable en la UE son cada vez más importantes y un reflejo de ello es el aumento de la generación y el consumo de energía a partir de fuentes alternativas, con un crecimiento continuo en los últimos años, aunque todavía están lejos de alcanzar la meta establecida para el año 2020 (Sanz-Casado *et al.*, 2014). El interés por ahondar en este tema en el entorno europeo, también se ha hecho evidente en la financiación de diferentes programas y convocatorias competitivas de proyectos de I+D+i siendo una de las principales los Programa Marco.

Es por ello que este estudio se realizará analizando los proyectos sobre Ahorro Energético concedidos en las diferentes convocatorias del Séptimo Programa Marco. Si bien el uso de esta fuente focaliza el estudio en el entorno europeo, es importante mencionar que cada vez son más los países externos a la región que participan como socios en estas convocatorias por lo que la cobertura resulta aún mayor.

Teniendo en cuenta estas consideraciones, el objetivo principal de este estudio es:

1) Analizar la actividad científica y tecnológica sobre ahorro energético a través del estudio de los proyectos del Séptimo Programa Marco.

Para ello se pretende alcanzar los siguientes objetivos específicos:

1) Recoger información proveniente de los proyectos europeos para identificar cuáles son los principales países involucrados en la actividad científica sobre Ahorro Energético, cuáles ejercen el liderazgo en este campo temático y cómo se distribuye la financiación obtenida.
2) Detectar los principales tópicos abordados en los proyectos sobre ahorro energético identificando los términos más frecuentes.

3) Probar la validez de técnicas de minería de texto para el estudio del contenido de los proyectos.

Metodología

Para el análisis propuesto se seguirá una metodología basada en los estudios métricos de la información. El trabajo ha sido desarrollado en dos fases que se detallan a continuación.

Primera fase

Para recoger información sobre los proyectos se utilizó como fuente de información la base de datos europea CORDIS (http://cordis.europa.eu/projects/home_es.html), que ofrece información por tipo de convocatoria, tema, país y tipo de resultado. En este caso la búsqueda de documentos se ha limitado al tipo "proyectos", de la convocatoria "Séptimo Programa Marco" y del tema "ahorro energético". Se ha recuperado toda la información disponible para cada proyecto y, en la tabla 1 se muestran los datos obtenidos.

Tabla 1. Información obtenida sobre los proyectos en Ahorro Energético.

Indicador	Descripción
Referencia	Número único que identifica a cada proyecto
Título	Título del proyecto
Convocatoria	Programa marco al que corresponde la convocatoria
Programa	Programa específico en el que se encuadra el proyecto
Fecha inicio	Día, mes y año de inicio
Fecha fin	Día, mes y año de finalización del proyecto
Coordinador	País que lidera el proyecto
Participantes	Nombre de los países que participan
Nº países participantes	Total de países por proyecto
Nº instituciones	Total de instituciones por proyecto
Instituciones de ES	Nombre de las instituciones españolas que participan
Temática	Áreas temáticas a las que pertenece el proyecto
Resumen	Resumen del proyecto
Total €	Presupuesto total en € para la realización del proyecto

Con la información obtenida se construyó una base de datos relacional y se obtuvieron los principales indicadores cienciométricos: número de proyectos por año; por convocatoria y programa; por país participante y país coordinador; sectores institucionales y organismos más activos en proyectos; distribución de la financiación recibida; relación entre el número de países e instituciones firmantes con la financiación. Si bien esta información puede resultar valiosa para comprender la dinámica de la actividad científica sobre Ahorro Energético, no deja de ser puramente descriptiva. Por tanto, para dar un paso más se ha decidido profundizar en el contenido de los proyectos.

Bib.An.Invest. Vol. 14 No. 2: 149-163
ISSN: 1683-8947

Segunda fase

Para obtener información sobre los temas tratados en los proyectos fue necesario utilizar herramientas externas ya que la mayoría de los proyectos no ofrecen palabras claves ni clasificaciones temáticas que permitan clasificar los tópicos en que se focalizan. Por el contrario, los proyectos cuentan con un resumen que ha sido usado para identificar tales tópicos.

Actualmente está muy extendido el uso de técnicas de minería de texto *(text mining)* y del aprendizaje automático *(machine learning)* para analizar semánticamente el contenido de documentos. Entre las muchas técnicas existentes en este campo, en este trabajo nos hemos decantado por dos.

Natural Language Toolkit

El análisis de los títulos y los resúmenes de proyectos se ha realizado utilizando técnicas de procesamiento de lenguaje natural. Por "lenguaje natural" nos referimos a un idioma que se utiliza para la comunicación diaria de los seres humanos, en contraposición con lenguajes artificiales como lenguajes de programación y notaciones matemáticas. Los estudios basados en análisis de lenguaje natural pueden ser tan simples como contar frecuencias de palabras para comparar diferentes estilos de escritura, o llegar a "entender" expresiones humanas completas, al menos hasta el punto de poder darles respuestas útiles.

"Las tecnologías basadas en procesamiento de lenguaje natural se están generalizando cada vez más. Por ejemplo, los teléfonos y las computadoras de mano son compatibles con el reconocimiento predictivo de texto y escritura a mano…"

Las tecnologías basadas en procesamiento de lenguaje natural se están generalizando cada vez más. Por ejemplo, los teléfonos y las computadoras de mano son compatibles con el reconocimiento predictivo de texto y escritura a mano; los motores de búsqueda web dan acceso a información encerrada en texto no estructurado; la traducción automática nos permite recuperar textos escritos en chino y leerlos en español; el análisis de texto nos permite detectar sentimientos en *tweets* y *blogs*. Al proporcionar interfaces hombre-máquina más naturales y un acceso más sofisticado a la información almacenada, el procesamiento del lenguaje ha pasado a desempeñar un papel central en la sociedad de la información multilingüe (Steven *et al.*, 2009).

Para tratar los datos provenientes de los proyectos europeos, se ha realizado la programación de un *script* en Python que utiliza la herramienta NLTK *(Natural Language Toolkit de Python)*. NLTK es un proyecto *open source* que constituye un conjunto de librerías de procesamiento de texto: clasificación, tokenización (separación en palabras), lematización, análisis semántico, etc. (Bird *et al.*, 2011). Esta herramienta se ha aplicado al conjunto de resúmenes de los proyectos desde un fichero Excel que contenía tres campos: identificador del documento, título y resumen.

En el análisis de los textos se han seguido los pasos que se detallan a continuación:

- Depuración del texto a tratar: de los resúmenes de los proyectos se han eliminados espacios extra y se ha transformado todo el texto a minúsculas.

- División de cada resumen en palabras *(tokens)* y análisis morfosintáctico para diferenciar los tipos gramaticales de las palabras (NLTK asigna una etiqueta a cada palabra). Las que se han considerado más relevantes fueron las que se presentan en la tabla 2.
- Eliminación de *stop-word* (palabras vacías).
- Generación de *n-grams:* cadenas de palabras formadas por dos o más sustantivos adjetivos consecutivos y asignación de la etiqueta "NGRAM" a este nuevo tipo.
- Análisis de los textos considerando únicamente los tipos de palabras seleccionados: sustantivos, adjetivos y verbos.
- Análisis de la frecuencia de aparición de términos y *n-grams*.
- Visualización a través de una nube de palabras con la librería *wordcloud* de Python (disponible en https://github.com/amueller/word_cloud).

Tabla 2. Tipos de palabras utilizados.

Categoría	Descripción
NN	Nombre común, singular.
NNP	Nombre propio, singular.
NNS	Nombre común, plural.
NNPS	Nombre propio, plural.
JJ	Adjetivos, numerales, ordinales.
VB	Verbo, infinitivo
VBD	Verbo, pasado.
VBG	Verbo, participio de presente o gerundio.
VBN	Verbo, participio pasado
VBP	Verbo, presente, no tercera persona del singular
VBZ	Verbo, presente, tercera persona del singular.
CD	Numerales, cardinales.

Mallet

Además del análisis de contenido simple se ha indagado en los denominados *modelos* (probabilísticos) *de tópicos* o *Probabilistic Topic Models* o simplemente *Topic Models* que fueron introducidos por Hofmann (1999) y Griffiths y Steyvers (2002) y en su versión más popular, usando el algoritmo LDA, por Blei, Ng y Jordan (2003). En esta metodología cada documento es observado como un conjunto de palabras (uni-gramas) que, una vez excluidas aquellas que no tienen significado semántico *(stop-words)*, pueden combinarse en conjuntos o sacos de palabras *(bags of words* o *topics)* formados por aquellas que aparezcan más juntas de manera más frecuente. De este modo, un *corpus* o conjunto de documentos (con textos no estructurados) puede ser a su vez observado como una mixtura de estos tópicos. Cada documento del *corpus* tendrá una cierta probabilidad de poder ser clasificado en cada uno de los tópicos hallados. Eligiendo un cierto umbral de probabilidad, podremos clasificar cada documento en uno o varios tópicos. En este trabajo

Bib.An.Invest. Vol. 14 No. 2: 149-163
ISSN: 1683-8947

se ha utilizado para la generación de tópicos la herramienta MALLET *"Machine Learning for Language Toolkit"*, que incluye el algoritmo LDA, y que es una colección integrada de código Java útil para procesamiento estadístico del lenguaje natural, la clasificación de documentos y otras aplicaciones de aprendizaje automático aplicadas al texto. Esta herramienta fue desarrollada principalmente por Andrew McCallum, de la Universidad de Massachusetts Amherest con la asistencia de estudiantes de posgrado y profesores tanto de la propia universidad como de la Universidad de Pensilvania (McCallum, 2002).

Para trabajar con el conjunto de resúmenes de proyectos es necesario generar un fichero en el que cada fila sea un proyecto y que contenga el título y el resumen. Este fichero se utilizará como "fichero de entrada" sobre el que operará Mallet. A continuación, se establecen los parámetros para el análisis, como el número de tópicos a obtener, las *stop-words* a excluir y el número de iteraciones del algoritmo. En este caso, tras probar diferentes opciones, se han generado para el *corpus* de resúmenes y títulos de los proyectos un grupo de 10 tópicos que de forma no supervisada (es decir, sin la necesidad de un conjunto de entrenamiento) ha permitido clasificar temáticamente dichos proyectos. Se ha evidenciado que con un número de tópicos inferior a diez la clasificación resultaba muy genérica. Por el contrario, con un número mayor, la especificidad de la clasificación era demasiado alta. La ejecución del programa genera ficheros de salida que contienen información sobre: la probabilidad de pertenencia de cada proyecto a cada tópico, así como las palabras y frases asociadas a cada uno. Estos datos permiten otorgar contenido semántico a cada tópico, identificándolos mediante etiquetas.

Para visualizar las relaciones entre tópicos se ha utilizado el programa Gephi. Utilizando el algoritmo *Force Atlas2* se han representado los tópicos y documentos como nodos y las relaciones entre ellos como aristas. El grosor de los nodos está dado por su grado y el tamaño de las aristas es proporcional al porcentaje de pertenencia de ese del documento al tópico.

Resultados

Primera fase

Los resultados preliminares muestran que, al analizar la evolución del número de proyectos en cada una de las convocatorias de los Programa Marco, el Séptimo es el que concentra un mayor número: 25.630. Estos pueden clasificarse en Programas y Temáticas, dentro de la que se encuentra "Ahorro Energético" que cuenta con 256 proyectos. Estos proyectos, a su vez, son transversales y pueden pertenecer a diferentes líneas y programas entre los que destacan ENERGY. Entre los países que sobresalen por su participación en proyectos sobre este tema se encuentran Alemania, Francia, Reino Unido, España e Italia. A pesar del extenso número de países participantes, sólo 24 han sido líderes de proyectos y, entre los de mayor actividad, destacan por liderazgo España y Alemania (Tabla 3).

Entre los sectores institucionales con mayor participación destacan las empresas, centros de investigación y las universidades. A

Tabla 3. Distribución del número de proyectos y del liderazgo por país.

País	Nº Proy. en Ahorro Energético	Nº de proyectos liderados	% Liderazgo
Germany	197	44	22,34
France	142	29	20,42
United Kingdom	137	24	17,52
Spain	135	31	22,96
Italy	128	27	21,09
Netherlands	109	15	13,76
Belgium	84	15	17,86
Switzerland	66	6	9,09
Sweden	61	8	13,11
Austria	59	12	20,34
Denmark	58	7	12,07
Greece	45	5	11,11
Poland	45	2	4,44
Finland	42	5	11,90
Norway	42	5	11,90
Portugal	33	5	15,15
Ireland	25	7	28,00
Israel	14	1	7,14
Slovakia	9	2	22,22
Australia	8	1	12,50
Cyprus	8	1	12,50
Iceland	4	2	50,00
Luxembourg	3	1	33,33
Liechtenstein	2	1	50,00

nivel de organismos el instituto FRAUNHOFER para la Ciencia y la Tecnología, de Alemania es el que muestra mayor participación con una presencia en 53 proyectos de esta temática. Le siguen la Fundación TECNALIA de España (34 proyectos), el *Commissariat a l'energie atomique et aux energies alternatives* de Francia (28 proyectos), la Universidad Técnica de Dinamarca (27 proyectos) y el *VTT Technical Research Centre of Finland Ltd* (27 proyectos).[1]

En cuanto a la financiación se observa que ha sido muy variable con un mínimo de 260.000€ por proyecto y un máximo de 35.500.000€. Se ha analizado también la relación entre la financiación y el número de países participantes, y se ha observado que muchos proyectos con alta financiación tienen un número medio de países participantes por lo que la correlación entre ambas variables es baja. Algo similar ocurre con la relación entre financiación y número de instituciones participantes ya que la gran mayoría de proyectos cuentan con entre 5 y 10 instituciones (entre las que predominan las empresas) y reciben en promedio uno de los mayores valores de financiación (Tabla 4).

1. Para ampliar información sobre las características de los proyectos en cuanto a patrones de colaboración y redes que se establecen, es posible consultar un estudio previo (De Filippo *et al.*, 2016).

Bib.An.Invest. Vol. 14 No. 2: 149-163
ISSN: 1683-8947

Tabla 4. Distribución de los proyectos de Ahorro energético en relación al número de países e instituciones participantes y la financiación obtenida.

Rango de países/instituciones	Nº proyectos	Min €	Max €	Prom. €
Países				
<5	70	261.451,8	29.697.930,0	3.289.474,5
>5 <10	160	499.709,0	27.805.038,0	4.710.445,4
>10 <15	22	869.000,0	35.499.975,6	9.694.443,4
>15 <20	4	3.999.629,0	9.997.207,0	8.420.632,0
Instituciones				
<5	17	261.451,8	2.994.389,0	1.332.714,3
>5 <10	106	499.709,0	29.697.930,0	3.752.293,5
>10 <15	78	869.000,0	12.515.552,0	3.880.949,6
>15 <20	26	1.200.000,0	27.004.954,9	5.735.187,2
>20 <25	16	2.899.857,0	25.189.520,0	10.117.136,4
>25 <30	6	8.898.432,0	27.451.972,0	16.983.685,5
>30 <45	5	3.999.629,0	35.499.975,6	16.683.979,7

Tabla 5. Términos simples y compuestos de mayor frecuencia en los proyectos de Ahorro Energético.

Término simple	Frecuencia	Término compuesto	Frecuencia
energy	809	energy-efficiency	93
project	404	energy-consumption	69
new	279	energy-efficient	50
system	263	renewable-energy	42
systems	224	energy-management	41
power	220	long-term	40
technology	211	real-time	35
research	208	electric-vehicles	34
development	201	cost-effective	33
efficiency	199	low-cost	33
design	186	co2-capture	33
european	186	large-scale	33
based	185	solar-cells	26
technologies	173	solar-thermal	25
management	161	energy-storage	23

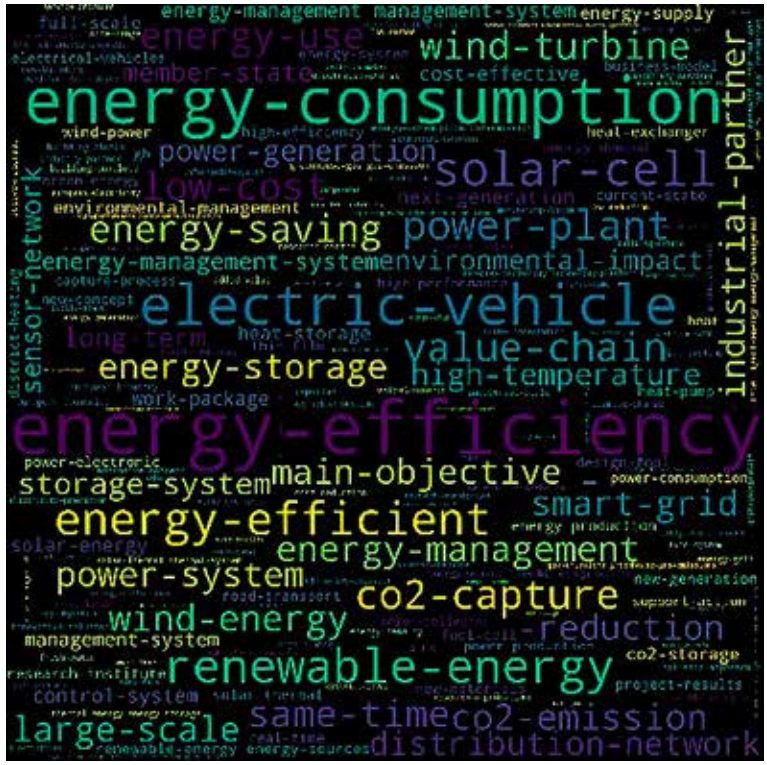
Segunda fase

Tras la aplicación de la herramienta NLTK de Python se ha trabajado con un corpus documental ya depurado en el que solo aparecen los tipos morfológicos seleccionados y los *n-grams* identificados. Así el análisis de frecuencia muestra cuáles son los términos simples y compuestos más frecuentes. En la tabla 5 se presentan ambos grupos y se puede observar que, si bien los términos simples son más

frecuentes, la generalidad de algunos términos como "energy", "project" o "system" hacen que no resulte sencillo identificar las temáticas que abordan los proyectos. Por el contrario, un análisis de *n-grams* permite mayor precisión en la definición de los temas. Es así que, aunque los términos más frecuentes (energy efficiency, energy efficient) son comunes a todos los proyectos, aparecen temas muy concretos como "vehículo eléctrico", "almacenamiento de energía", "energía solar", "captura de carbono", etc que ofrecen una visión más detallada del contenido.

La visualización de los términos más frecuentes se puede apreciar en forma de nube de palabras en la figura 1.

Figura 1. Términos compuesto más frecuentes en los proyectos de Ahorro Energético.



Para seguir profundizando en el análisis de contenido se ha utilizado también Mallet que ha permitido identificar 10 tópicos en los que pueden agruparse los proyectos en función de los temas que tratan. Se muestran a continuación las principales frases y palabras identificadas en cada uno de los tópicos (Tabla 6).

A partir de la lectura de cada grupo de palabras y frases, es posible identificar los temas sobre los que trata cada conjunto documental y etiquetarlos. En la tabla 7 se puede observar la temática asignada a cada uno indicando la probabilidad de que un documento sea asignado a ese tópico. Se aprecia que en el tópico central (Tópico 5), con el mayor número de proyectos asignados, la mayoría de los proyectos incluidos tratan sobre temas generales vinculados con el Ahorro Energético y a problemas relacionados con su suministro, distribu-

Bib.An.Invest. Vol. 14 No. 2: 149-163
ISSN: 1683-8947

Tabla 6. Principales tópicos detectados. Palabras y frases más frecuentes en cada tópico.

Tópico	Palabras	Frases
Topic 0	Power, photonic, chip, low, integrated, device, operation, circuit, arrays, mode, wavelength, level, consumption, galactico, optical	power consumption; message passing; circuit operation; paradigm shift; photonic components; optical processing; induced electrical coupling
Topic 1	Energy, building, consumption, data, information, management, efficiency, existing, ict, time, environmental, service, tools, monitoring	intelligent energy; service providers; sensor networks; user behaviour; service centres
Topic 2	Wind, storage, solar, heat, thermal, offshore, plants, scale, power, water, turbines, materials, plant, turbine, temperature	wind turbine; solar termal; renewable energy; altitude wind; offshore wind farms
Topic 3	Solar, materials, cells, high, material,l cost, production, efficiency, thin, cell, based, light, manufacturing, building, film	solar cell; sensitized solar cells; low cost; cell encapsulation; solar cells fabricated
Topic 4	Site, storage, interfaces, methods, soil, tools, security, remediation, assessment, sites, materials, techniques, understanding, mineral, chemical	term behavior; soil remediation; water protection; situ remediation; modelling tool
Topic 5	Energy, technology, based, technologies, cost, high, integration, innovative, industrial, efficiency, approach, demonstration, performance, market, efficient	industrial partners; energy efficiency; cost effective; energy management; energy consumption
Topic 6.	Energy, ict, support, stakeholders, implementation, public, information, sector, national, industry, knowledge, programme, platform, network, infrastructure	general public; public authorities; energy efficiency; relevant stakeholders; implementation plan
Topic 7	Grid, power, energy, network, electricity, demand, smart, distribution, control, renewable, time, management, district, real, generation	distribution network; district heating; smart grid; electricity generation; electricity price
Topic 8	Electric, vehicles, fev, vehicle, battery, range, control, drive, transport, mobility, safety, power, road, electrical	electric vehicles; road transport; electric motor; driving range; battery packs
Topic 9	Capture, gas, fuel, combustion, production, biomass, plant, emissions, materials, separation, scale, engine, reduction, plants, fuels	oxyfuel combustión; flue gas; jet fuel; post combustión; bioethanol production; bioenergy carriers

Tabla 7. Número de proyectos y probabilidad de pertenencia a cada tópico.

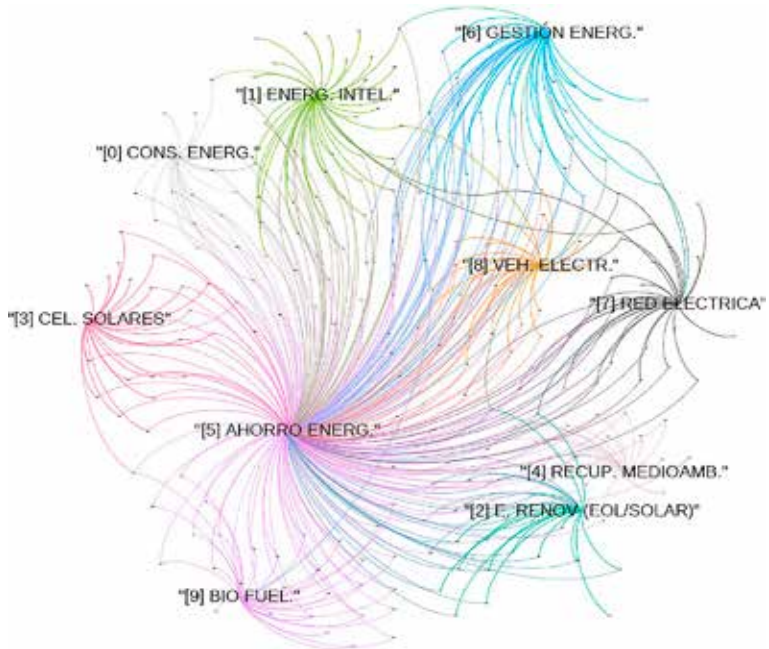
Tópico	Etiqueta	Probabilidad de pertenencia
Topic 0	Consumo Energético	0,03819
Topic 1	Energia Inteligente	0,15725
Topic 2	Energ. Renov. (Eolica Solar)	0,08137
Topic 3	Células Solares	0,05652
Topic 4	Recuperación Medioambiental	0,04361
Topic 5	Ahorro Energético	0,76052
Topic 6	Gestión Energética	0,13149
Topic 7	Red Eléctrica	0,10277
Topic 8	Vehículo Eléctrico	0,06528
Topic 9	Bio-Fuel / Bio-Masa	0,06002

ción energética, etc. Es por esto que se ha etiquetado como “Ahorro Energético” (un proyecto elegido al azar tendría una probabilidad del 76% de incluirse en este tópico).

El resto de tópicos reflejan la alta especificidad de los proyectos, relacionados con diferentes aspectos de la sostenibilidad energética, como las energías renovables (Solar, y Eólica, (Tópico 2), el vehículo eléctrico (Tópico 8), la gestión energética (Tópico 6) y la recuperación medioambiental (Tópico 4). Los valores indicados de probabilidad de pertenencia indican las posibilidades que tiene un documento de pertenecer a cada uno de los Tópicos.

Las relaciones entre tópicos y proyectos se muestran en el grafo de la figura 2, en el que se pueden apreciar 266 nodos (256 representando a cada proyecto, más 10 nodos de los tópicos identificados). Sobre las 2560 aristas iniciales que relacionan todos los nodos, se ha realizado una poda y sólo se muestran las relaciones con una probabilidad >25% de pertenencia de un proyecto a un tópico. En dicho grafo se aprecian 10 subredes identificadas cada una de ellas por un nodo central que es el tema identificado por el análisis de *topic modelling* y que está etiquetado según se muestra en la tabla 7. El tópico con un mayor número de proyectos asociados a él (Tópico 5), toma una posición central en el grafo dado su carácter más general y está compuesto por proyectos que a su vez comparten interés temático con aspectos más específicos. Sin embargo, entre éstos tópicos (1 a 9) las relaciones son menores, y el algoritmo LDA sólo clasifica unos pocos proyectos que pueden ser identificados en dos de ellos, por ejemplo, algunos proyectos enlazan el vehículo eléctrico (Tópico 8) con la gestión energética (Tópico 6).

Figura 2. Visualización de tópicos.



Discusión

El análisis del campo de Ahorro Energético a través de los proyectos del Séptimo Programa Marco resulta interesante para conocer aspectos complementarios a los aportados por los estudios tradicio-

Bib.An.Invest. Vol. 14 No. 2: 149-163
ISSN: 1683-8947

nales sobre producción científica y tecnológica basados en publicaciones y patentes. Asimismo, al ser estos programas altamente competitivos, su análisis permite conocer los temas de interés a los que la comunidad científica da mayor relevancia y en los que participan las instituciones de mayor prestigio internacional.

A nivel metodológico, la base de datos CORDIS ha demostrado ser una herramienta muy válida para la recolección de una amplia información sobre proyectos. Asimismo, al ser pública y de acceso sencillo, es factible de ser utilizada para realizar estudios diversos, tanto a nivel general para analizar las políticas científicas, como para valorar la investigación que se está desarrollando en cada campo del conocimiento. Por otro lado, el uso de técnicas de minería de texto empleando herramientas como NLTK y MALLET es uno de los aspectos más destacados del estudio ya que permite detectar cuáles son los temas de interés en un campo tan concreto como el del Ahorro Energético. La utilización de ambas ha permitido complementar y validar el estudio, identificando términos relevantes por ambas vías.

Los resultados obtenidos ponen de manifiesto que son los grandes países los que lideran la participación y coordinación de proyectos, entre los que España tiene un rol central. La activa participación de empresas hace que este sea el sector institucional predominante, seguido de universidades públicas y centros de investigación. Aunque existen grandes redes institucionales, no se ha evidenciado relación entre el tamaño de estas y la financiación captada. Tampoco se advierte relación entre el monto concedido por proyecto y el número de países implicados dado que existen proyectos con pocos centros participantes pero que han obtenido una financiación elevada.

"Los resultados obtenidos ponen de manifiesto que son los grandes países los que lideran la participación y coordinación de proyectos, entre los que España tiene un rol central."

El análisis de contenido ha permitido detectar que una de las principales temáticas, que presenta un carácter transversal, es la del Ahorro Energético (generalidades), que incluye desde aspectos vinculados con la gestión y la distribución hasta el impacto medioambiental. Dentro de los temas relacionados con fuentes de energía alternativa, se estudian en concreto la solar y la eólica, siendo la primera la más relevante. Tópicos muy específicos como el vehículo eléctrico, aparecen también entre los temas tratados por los proyectos europeos, con claras relaciones con la gestión energética, la energía eléctrica y la eficiencia. Temáticas similares aparecen entre las detectadas en un análisis sobre patentes verdes, lo que evidencia la preocupación constante por el cambio de un modelo de producción tradicional —responsable del agotamiento de los recursos naturales y el incremento de la contaminación— hacia otro en el que la economía verde no es sólo una opción, sino más bien una necesidad (Macedo-Santos *et al*., 2017).

Sin dudas la aplicación de técnicas como las presentadas en este trabajo, puede resultar útil, no sólo para describir el estado de la ciencia y la tecnología en un momento dado, sino para realizar análisis temporales y detectar la dinámica de los campos científicos a lo largo de determinados períodos. Si bien el análisis probabilístico de contenido tiene ciertas limitaciones -como el hecho de generar resultados diferentes tras cada aplicación del algoritmo- permite identificar los temas concretos que se mantienen en todas las iteraciones y, por lo tanto, lograr un nivel mayor de detalle sobre el

contenido de los proyectos. Estas técnicas resultan adecuadas para su aplicación al estudio de contenido de un volumen importante de documentos, dado que no es factible la lectura en profundidad de cada uno.

En este sentido un estudio en el que se analizan memorias institucionales de actividad sobre sostenibilidad, muestra la utilidad del uso de herramientas de procesamiento del lenguaje natural como estrategia de investigación para detectar los principales temas de interés en la sostenibilidad ambiental, económica y social (Szekely & Brocke, 2017). Asimismo, estudios previos han demostrado que en áreas temáticas interdisciplinares y con alto desarrollo tecnológico e innovador, la utilización de técnicas de *Topic model,* resulta adecuada para poder tener un conocimiento que no es factible de obtener por otras vías, como la evolución temática dentro de un campo (Zhai *et al.*, 2017). Otros trabajos similares en los que se aplican estas técnicas a campos como el de Información y Documentación, muestran su utilidad para identificar subdisciplinas y temas emergentes a lo largo del tiempo (Figuerola *et al.*, 2017).

En cuanto al aporte de este estudio, se ha evidenciado que el sector del ahorro energético es transversal a varios ámbitos de la economía (industria, transporte, edificación), por lo que su definición no resulta sencilla. Asimismo, tal como se menciona en estudios previos, la medición de la dimensión científica del ahorro energético es compleja y no se dispone de información precisa en relación a su alcance y potencial de crecimiento (IDAE, 2011). Por otra parte, se trata de un sector dinámico y en constante desarrollo debido a la fuerte innovación tecnológica en la que se ve inmerso. Esto dificulta la aplicación de técnicas tradicionales de bibliometría ya que no es posible recuperar publicaciones con gran precisión. Sin embargo, el uso de minería de textos posibilita la identificación de contenidos, por lo que otras fuentes de información, como los proyectos de investigación, pasan a ser susceptibles de análisis.

Esto resulta relevante ya que, en general, los estudios multidimensionales sobre energía corresponden a informes elaborados por distintos departamentos ministeriales y consejerías. Dichos informes inciden con frecuencia en la elaboración y análisis del contexto económico, social, administrativo y sobre las políticas energéticas, los sistemas de información ambiental y los estudios de vigilancia tecnológica. Sin embargo, no es sencillo encontrar estudios de relevancia sobre el contexto científico y tecnológico en el que se desarrollan las actividades. Por ello, conocer los proyectos europeos que están abordando problemas de ahorro y eficiencia energética, permite tener una noción más clara sobre los desarrollos reales que se están produciendo en la actualidad.

Agradecimientos

Este trabajo ha sido realizado en el marco del Proyecto "Detección de nuevos frentes de investigación e innovación en Eficiencia Energética. Análisis de los flujos de conocimiento entre el ámbito científico, la industria y la sociedad" (REF: CSO2014-51916-C2-1-R) financiado por el Ministerio de Economía y Competitividad de España (MINECO). ■

Bibliografía

Bird S., Edward, L., & Ewan K. (2009), *Natural Language Processing with Python.* O'Reilly Media Inc.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research; 3,* 993-1022.

Callon, M., Courtial, J. P., & Penan, H. (1995). *Cienciometría: la medición de la actividad científica: de la bibliometría a la vigilancia tecnológica.* Gijón: Trea.

De Filippo, D., Marugán, S., & García-Zorita, C. (2016). *Actividad española en Ahorro Energético. Participación, captación de recursos y liderazgo en los proyectos europeos del Séptimo Programa Marco.* Madrid, España: CONAMA.

Figuerola, C, García-Marco, F., & Pinto, M. (2017). Mapping the evolution of library and information science (1978-2014) using topic modeling on LISA. *Scientometrics,* 112:1507-1535. doi: 10.1007/s11192-017-2432-9.

Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. En *Proceedings of the 24th Annual Conference of the Cognitive Science Society.*

Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. En *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence.*

IDAE. (2011). *Informe de sostenibilidad ambiental del plan de energías renovables 2011-2020.*

McCallum, A. K. (2002). MALLET: *A Machine Learning for Language Toolkit.* Recuperado de http://mallet.cs.umass.edu.

Nonato Macedo dos Santos, R., Pandiella-Dominique, A., Lascurain-Sánchez, M. L., & Sanz-Casado, E. (2017). Tecnologias verdes para um mundo autossustentável: um olhar sobre Brasil e Espanha. *Em Questão, 23*(2), 277-294.

OECD. (2011). *Towards Green Growth,* OECD, Paris.

OECD. (2012). *Greening Development: Enhancing Capacity for Environmental Management and Governance,* OECD, Paris.

Plaza, L. (2001). Obtención de indicadores de actividad científica mediante el análisis de proyectos de investigación. En Albornoz, M. (compilador) *Indicadores Bibliométricos en Iberoamérica* (pp. 63-70). Buenos Aires, Argentina: RICYT.

Sanz-Casado, E., Serrano-López, A., De Filippo, D., & Lascurain-Sánchez, M. L. (2014). The SpainChina scientific cooperation in renewable energy (2003-2012). *Science Focus, 9*(2), 43-52

Székely, N., & vom Brocke, J. (2017). What can we learn from corporate sustainability reporting? Deriving propositions for research and practice from over 9,500 corporate sustainability reports published between 1999 and 2015 using topic modelling technique. *PLoS ONE, 12*(4), e0174807. doi: 10.1371/journal.pone.0174807.

Zhai, Y., Ding, Y., & Wang, F. (2017). Measuring the diffusion of an innovation: a citation analysis. *Journal of the Association for Information Science and Technology, 69*(3), 368-379.