Caso de uso - VIGILANCIA SECTORIAL

Infoday sobre Actuaciones del Plan de Impulso de las Tecnologías del Lenguaje: oportunidades de la compra pública de innovación – 25/04/2017

Timeline proyecto vigilancia sectorial

11/2013 - 06/2014

Pilotos de implementación de sistemas Big Data para el apoyo a la gestión de ayudas

- Piloto tecnologías BigData: Hadoop, Hive, Pig, Hue, etc
- Piloto aplicación de grafos
- Piloto correlación oferta y demanda de empleo en el sector TIC

01/2015 - 12/2015

Sistema de soporte a la evaluación de ayudas de la Secretaría de Estado de Telecomunicaciones y para la Sociedad de la Información

20/10/2015

Presentación oficial del Plan de Impulso de las Tecnologías del Lenguaje

05/2016 - 12/2016

Nuevas funcionalidades e integración de técnicas relacionadas con la generación de modelos de semántica latente, semántica basada de ontologías y aprendizaje automático aplicados a la vigilancia sectorial

Participación y colaboración en el proyecto

satec

GOBIERNO DE ESPAÑA
MINISTERIO DE ENERGÍA, TURISMO Y AGENDA DIGITAL
SECRETARÍA DE ESTADO PARA LA SOCIEDAD DE LA INFORMACIÓN Y LA AGENDA DIGITAL

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA, INDUSTRIA Y COMPETITIVIDAD
SECRETARÍA DE ESTADO DE INVESTIGACIÓN, DESARROLLO E INNOVACIÓN

sgfsi
S.G. DE FOMENTO DE LA SOCIEDAD DE LA INFORMACIÓN

Oficina Española de Patentes y Marcas

CDTI Centro para el Desarrollo Tecnológico Industrial

FECYT FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA

uc3m

ixa

UPC TALP

iic instituto de ingeniería del conocimiento

Ontology Engineering Group

satec IT's easy

GOBIERNO DE ESPAÑA
MINISTERIO DE ENERGÍA, TURISMO Y AGENDA DIGITAL
SECRETARÍA DE ESTADO PARA LA SOCIEDAD DE LA INFORMACIÓN Y LA AGENDA DIGITAL

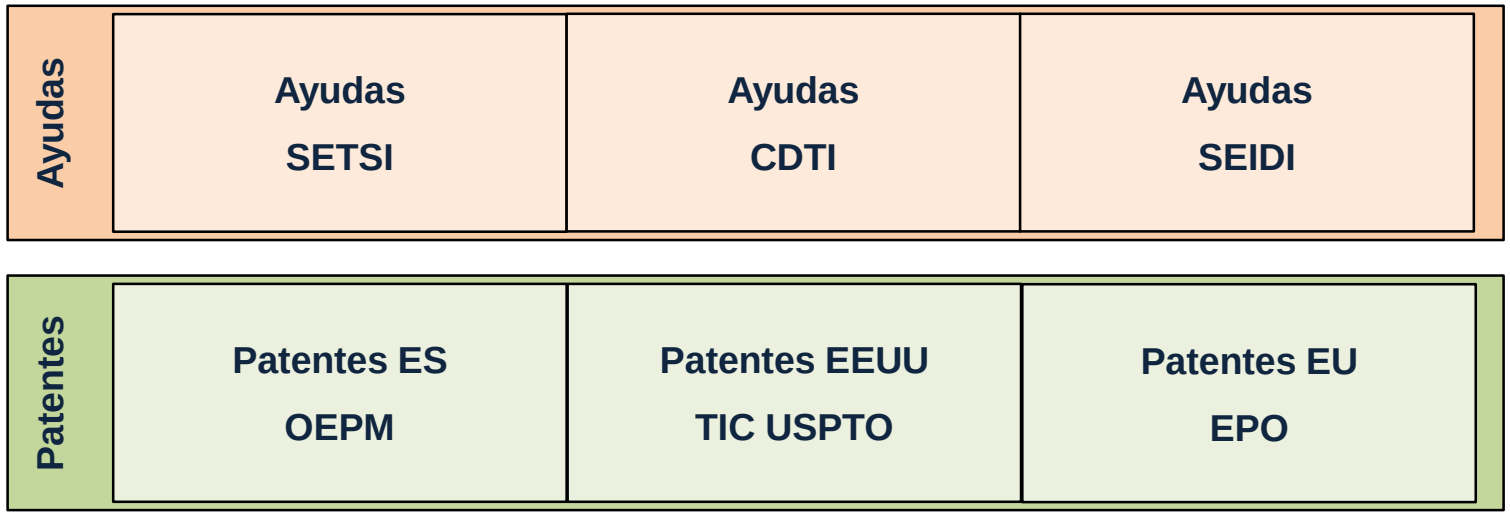
red.es

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Fuentes de información

Corpus documental

- Entendido como un conjunto de documentos de una misma tipología, que se agrupan en base a una serie de criterios explícitos.



En nuestro caso concreto, información textual estructurada y no estructurada para: solicitudes de ayuda, descripciones técnicas de patentes, publicaciones científicas, etc

Aproximación elegida

ETL

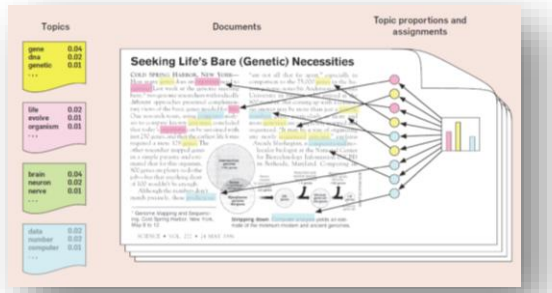
Extracción de datos estructurados y no estructurados

Técnicas básicas PLN

Tokenizado • Detección de categorías gramaticales (PoS) • Lematizado • Wikificado

Modelado de tópicos

Modelo generativo: estático (LDA, CTM), dinámico (DTM), jerárquico (rLDA) • Vectorización • Huella única del documento



Técnicas alternativas

Semántica latente avanzada • Semántica basada en ontologías • Machine learning • Word-embedings • Grafos

Objetivo: facilitar el trabajo de los evaluadores

Mejorar el conocimiento general sobre los documentos contenidos en un corpus

Descubrimiento de los temas tratados por cada documento • Vista global de las temáticas tratadas • Vista jerárquica de temas tratados • Obtención de documentos por temática abordada • Identificación de expertos en áreas concretas • Temas que suelen darse de manera conjunta • Evolución en el tiempo de las temáticas

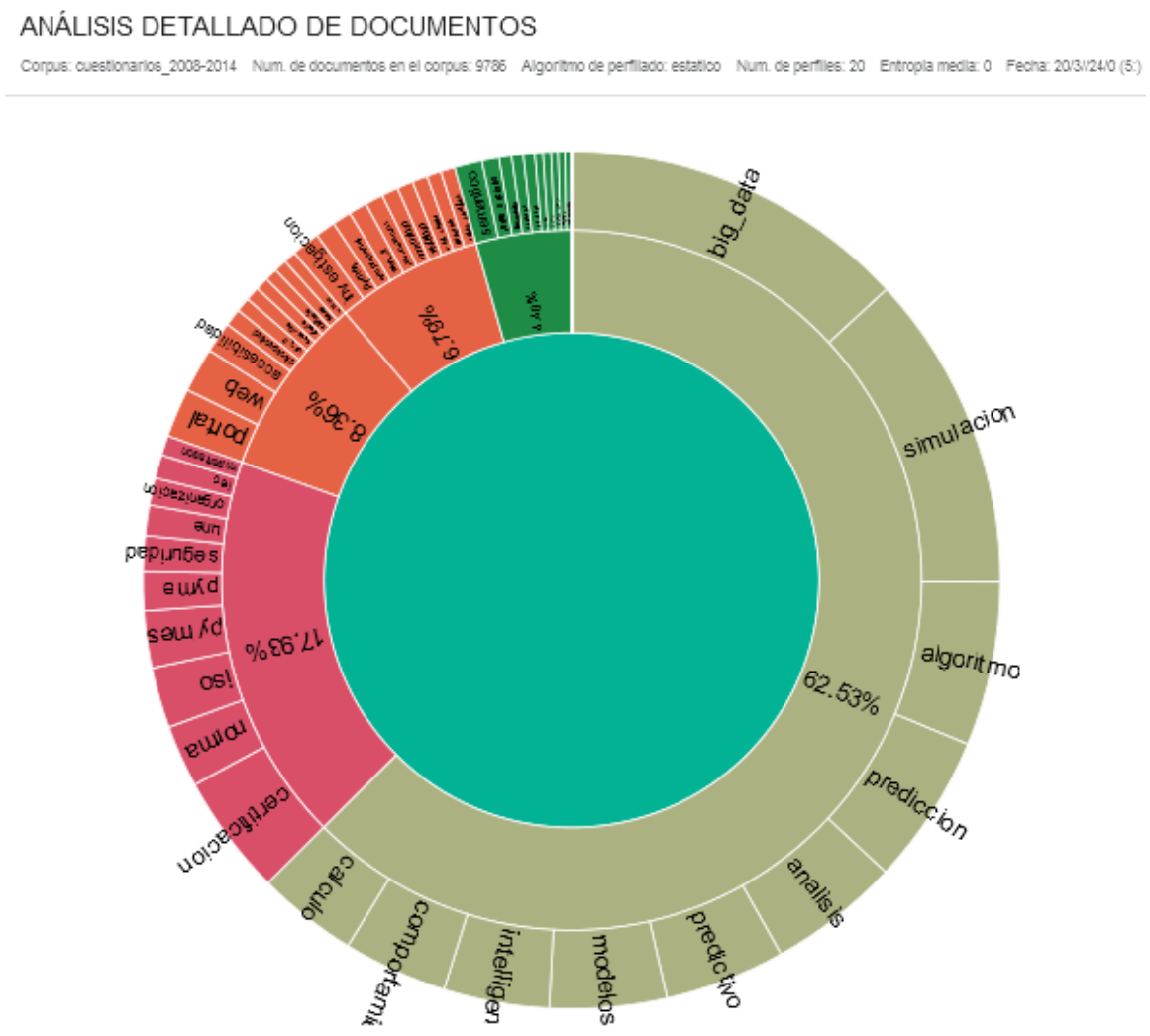
Herramientas que permiten hacer búsquedas eficientes sobre el corpus documental

Búsquedas textuales • Búsquedas de documentos parecidos a otro dado • Localización de parejas de documentos temáticamente similares

Comparación entre distintos corpus documentales

Posibilidad de comparar documentos cuyo contenido se desconoce • Compartiendo modelo de tópicos entre organizaciones, pero sin compartir las solicitudes • Detección de solapamientos • Prevención del fraude • Medición de los resultados de la investigación

Composición de tópicos del documento

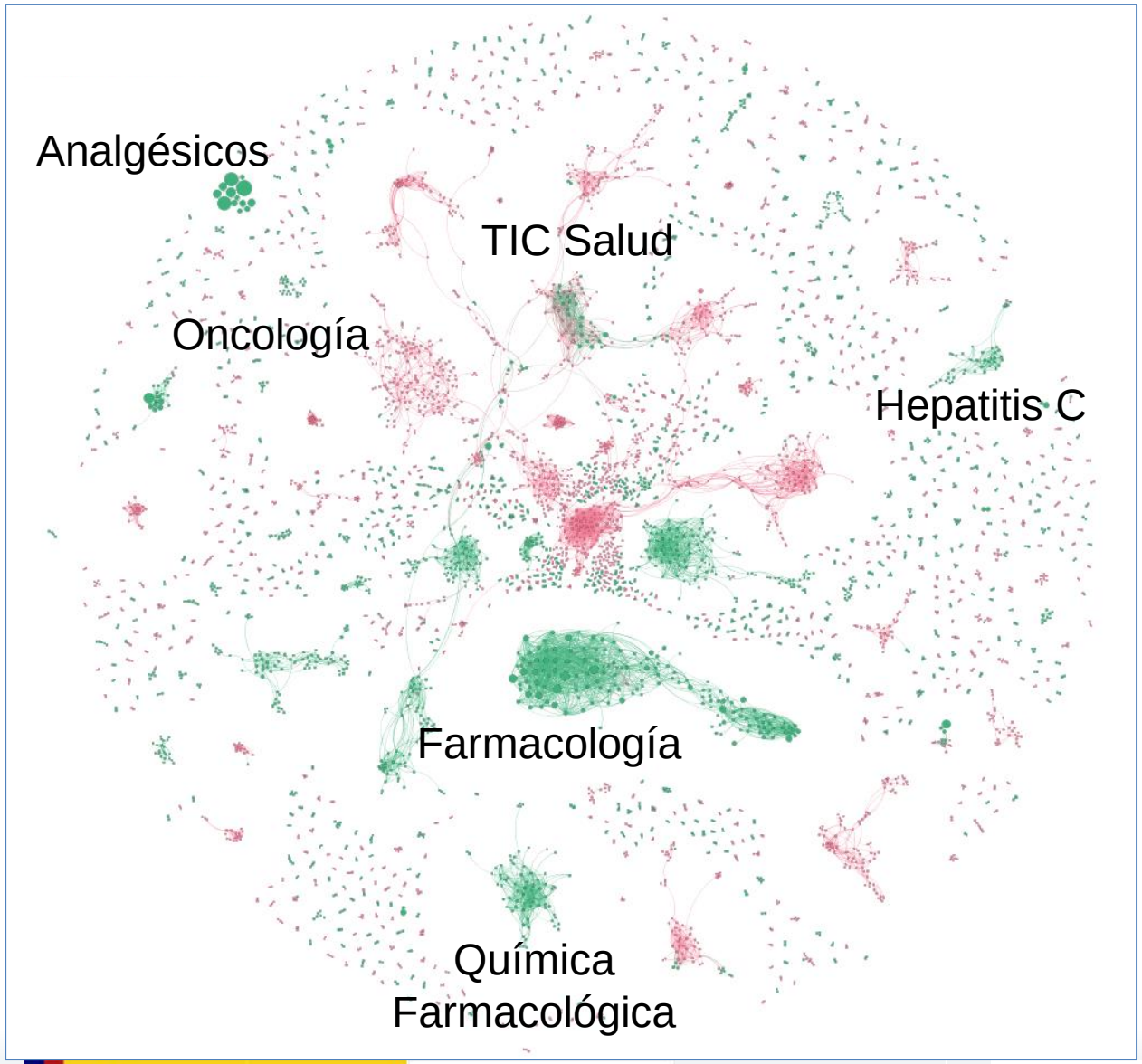


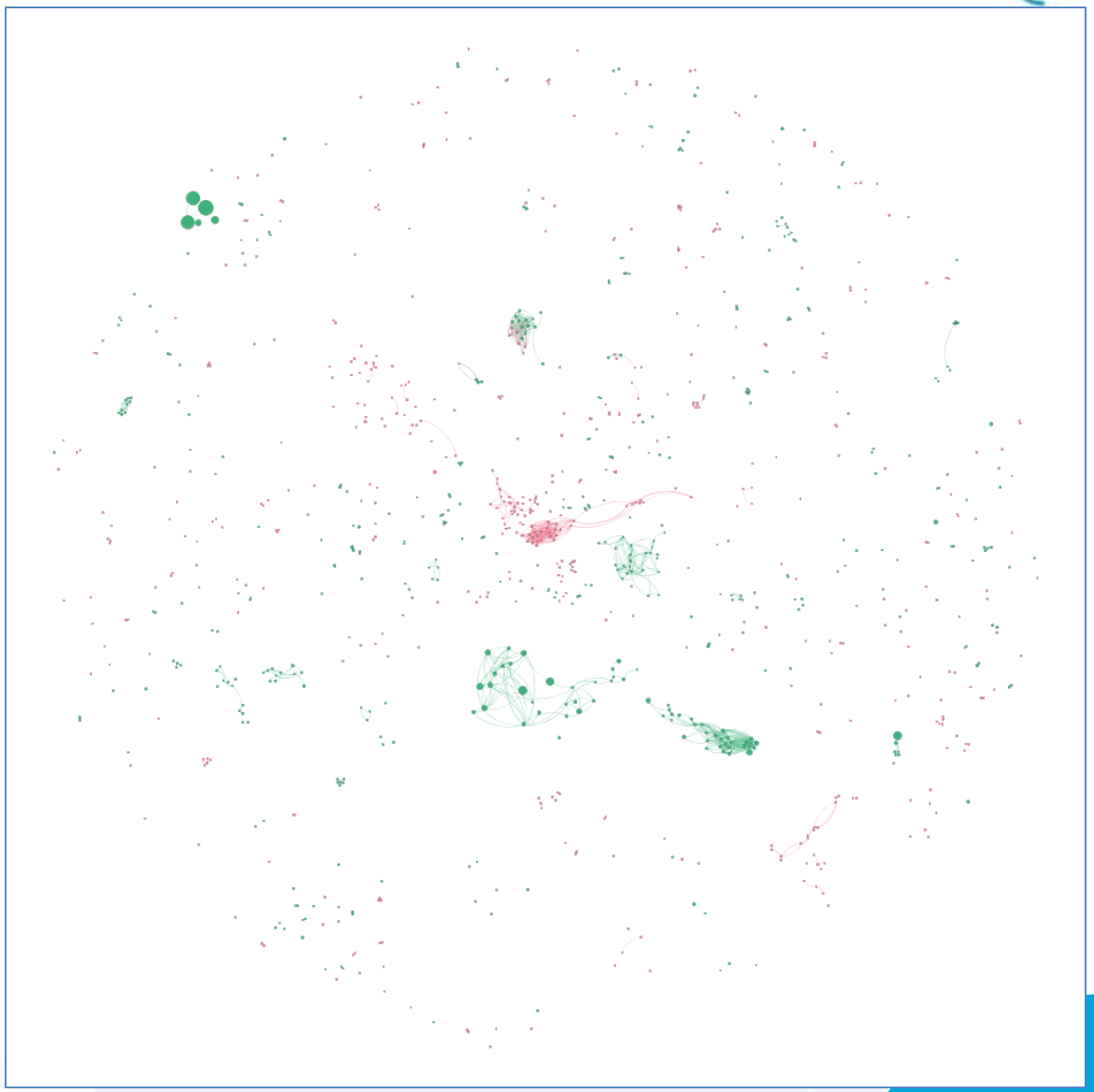
Representación de la huella única del documento para un modelo de tópicos

Distribución de los tópicos en el documento, porcentaje de texto de la sección técnica de la solicitud:

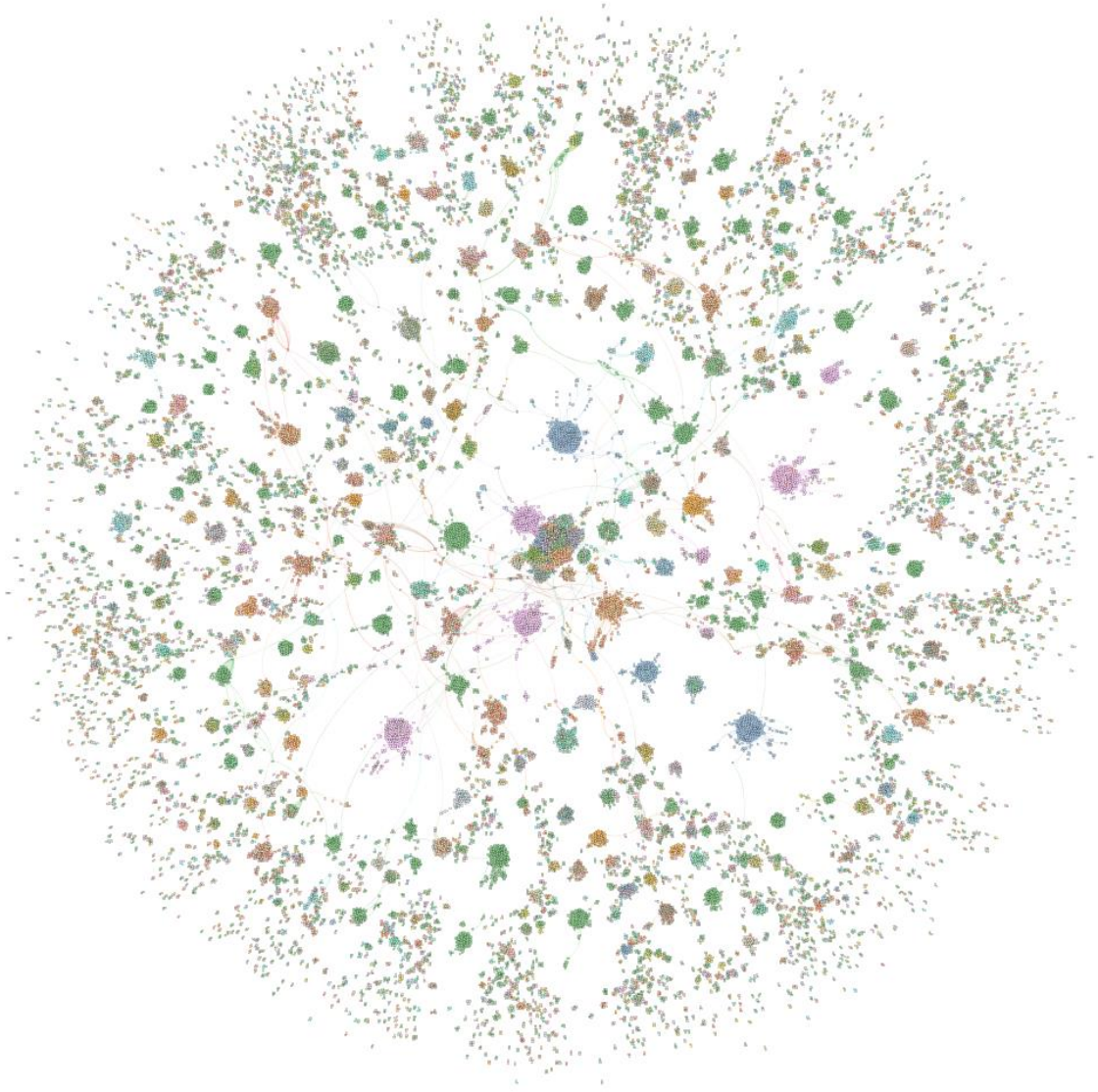
- 62% al tópico “Big data”
- 17% al tópico “Certificaciones”
- 8% al tópico “Portales web”
- 6% al tópico “Investigación”
- 4% al tópico “Semántica”
- …

Comparativa investigación sanitaria



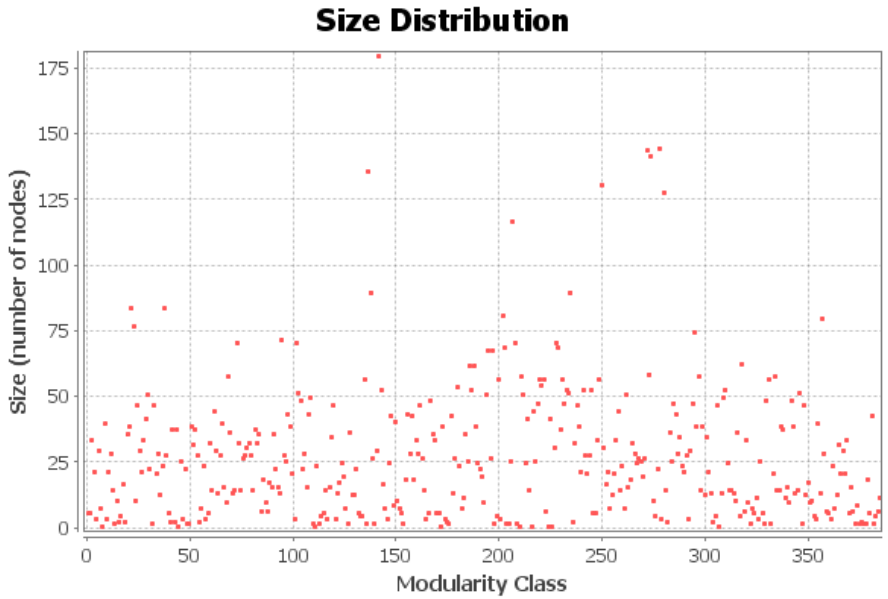


Plan Estatal 2004-2016 proyectos aprobados



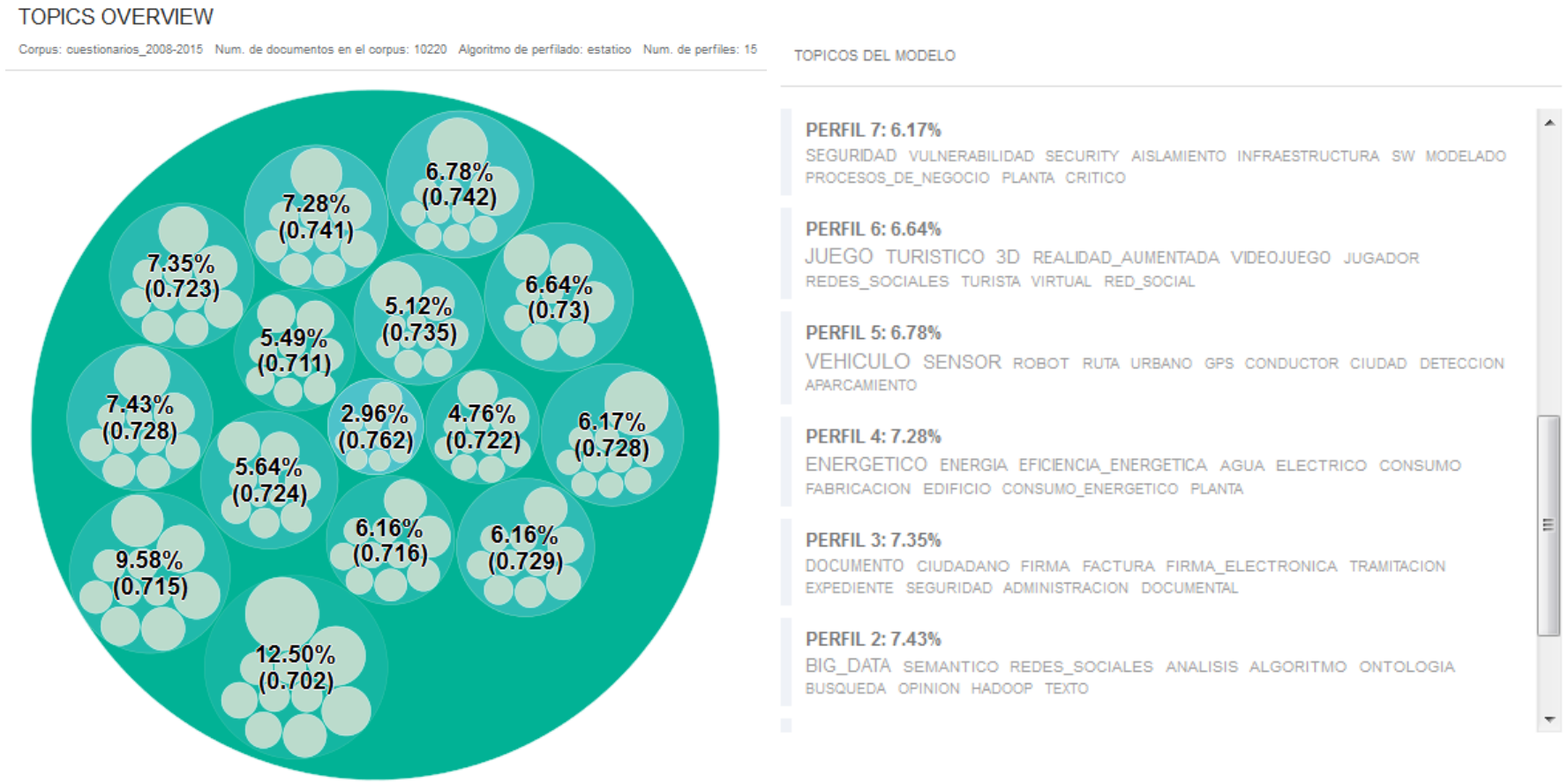
Results:

Modularity: 0,983
Modularity with resolution: 49,983
Number of Communities: 385



33,5Kdocs de diferentes programas con el objetivo de identificar temáticas financiadas por varios programas o cambios de nombre del programa a lo largo del tiempo. 385 comunidades.

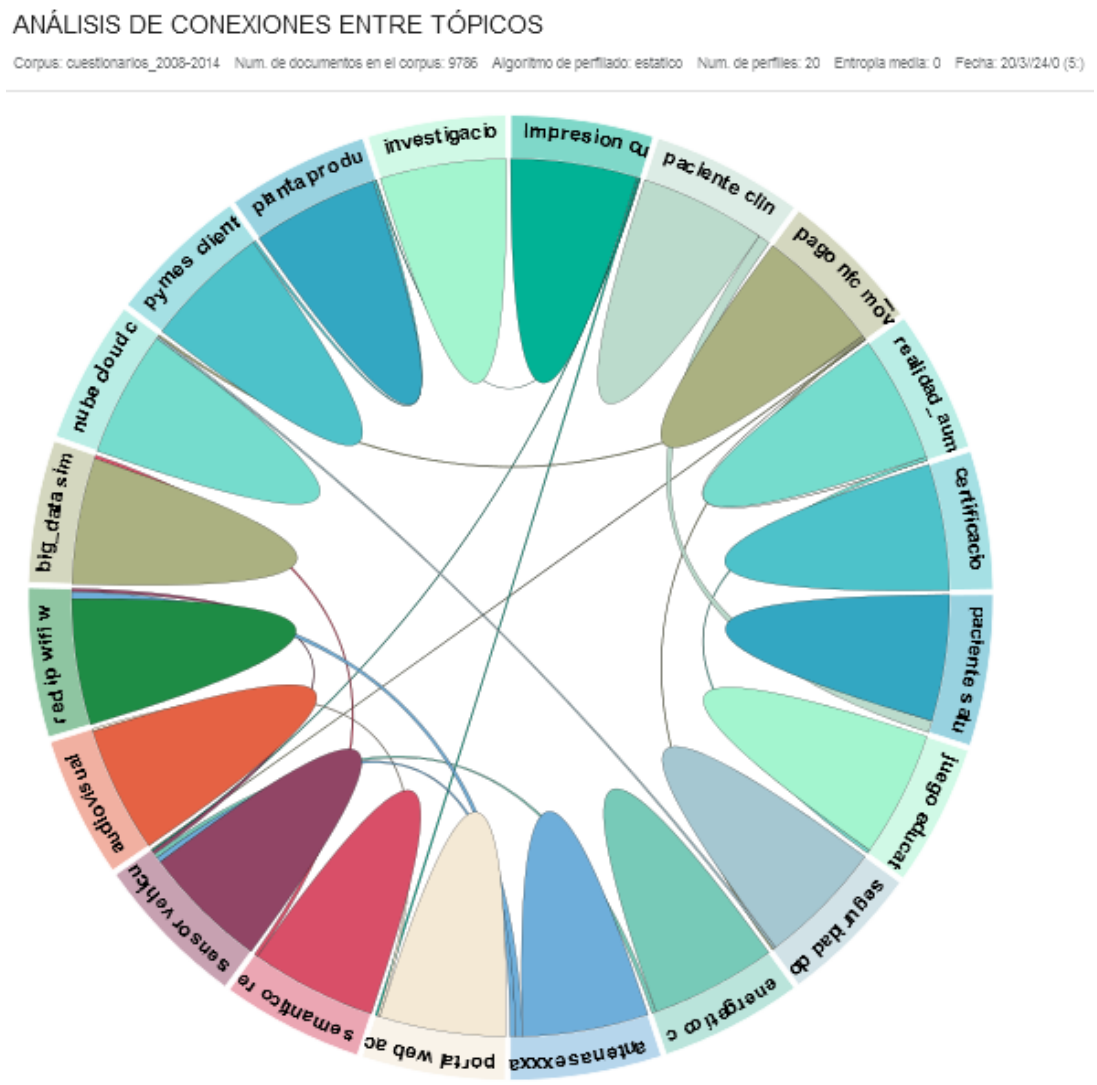
Clasificación de proyectos



Visión global del corpus, identificación de temas tratados en los documentos

Visión global del corpus según los tópicos reconocidos, detalle de palabras más frecuentes del tópico y documentos del corpus que mejor se adaptan al tópico

Reparto de proyectos (correlación)



Reparto de proyectos entre evaluadores

Muestra las relaciones entre tópicos según los documentos analizados.

Permite identificar conocimientos relacionados.

Por ejemplo:

- antena / señal
- red / ip / wifi
- sensor / vehículo

Buscador

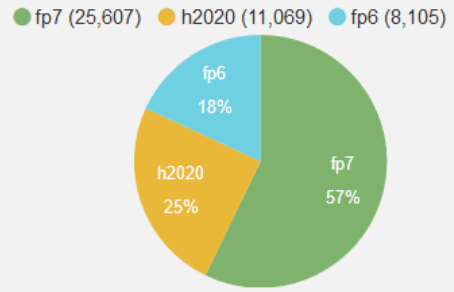
SEARCH | QUERY

:

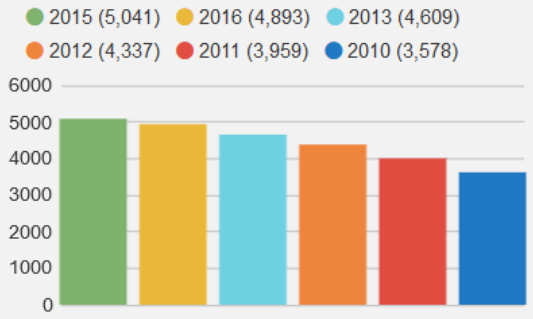
TOTAL HITS | HITS

44,781

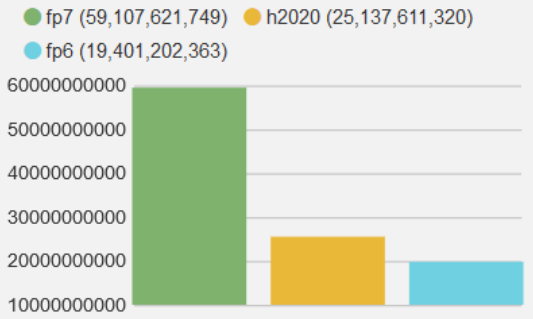
PROYECTOS / PROGRAMA | TERMS



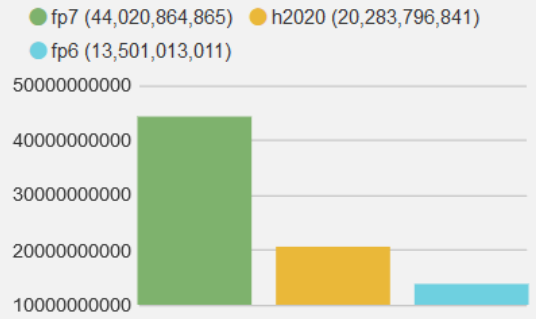
PROYECTOS / AÑO | TERMS



COSTE TOTAL (€) | TERMS



EC MAX CONTRIBUTION (€) | TERMS



FILTERING | FILTERING

No filters available

EVENTS | TABLE

0 to **10** of 5,000 available for paging

id	programa	title	totalCost	year
100107	fp6	Symbolic Computation Infrastructure for Europe	3351831	2006

Detección de duplicados (alarmas)

Percentil inferior 50 **Percentil superior** 99 **Límite de resultados** 20 **Centrado en año** Cargar

Alarmas encontradas (98.957%) **Tipo Alarmas** Tópicos

Generar informe | Comparar paneles

[[SMART_CHARGEREINVEL. ReSistema de carga inteligente de vehículopara automóviles eléectricos Amplia. Solución de la plataforma SmartCity para incluir la gestión y /explotación TIC municipal de infraestructuras de recarga integrada en los sistemas de pago y control de accesos de los aparcamientokings públicos actuales, como lugar idóneao para la recarga y estancia de los vehículos eléctricosV.E. Vinculación de Smart Charging Infraestrucuture y Smart Parking Management en una plataforma de Smart City.Para desarrollar los objetivos de este proyecto es conveniente resolver las siguientes problemáticas actuales detectadas en los parkings públicos: a) Las soluciones de recarga actuales no están integradas en infraestructuras de pago de vía púia publica o de parking municipal ni utilizan un protocolo estandarizado y reconocido a nivel europeo de gestión de infraestructuras de recarga (OCPP) en toda la arquitectura. b) La plataformasolución debe contemplar el uUso de distintas formas de identificación de usuario y pago para aproximarse a la realidad de los parkings púublico medianos y grandes actuales c) No existen plataformas municipales que gestionan la utilización de empresas de car-sharing de forma eficiente. d) Utilizaciónso de un amplio abanico de sistemas deSmart Phones como terminal para realizar la identificación, pago y gestión de usuario.la plaza y de la infraestructura de recarga. La oportunidad de mercado se encuentra en los parkings urbanos medianos (hasta 250 plazas) y grandes (más de+ 250 plazas) y las infraestructuras de car-sharing localizadas en vía púia publica y parkings de interior.a) Uso del protocolo estandarizado europeo OCPP de gestión de infraestructuras de recarga que permitan garantizar la interoperabilidad, universalidad y crecimiento de la solución en toda su arquitectura (Soluciones de recarga, terminal de identificación, maquinas de pago de parking y Aplicativo de Gestión) Este hecho posibilitara a la plataforma, la gestión futura de redes heterogéneas de puntos de recarga y evitar la

Blanco: Común Rojo: Panel izquierdo Verde: Panel derecho

Identificador

Título REINVEL. Recarga inteligente de vehículos eléctricos

Info Empresa

Ver metadatos
Ver fichero original
Ver fichero memoria

Texto

[[REINVEL. Recarga inteligente de vehículos eléctricos Ampliación de la plataforma SmartCity para incluir la gestión y explotación de infraestructuras de recarga integrada en los sistemas de pago y control de accesos de los aparcamientos públicos actuales, como lugar idónea para la recarga y estancia de los vehículos eléctricos.Para desarrollar los objetivos de este proyecto es conveniente resolver las siguientes problemáticas actuales detectadas en los parkings públicos:a) Las soluciones de recarga actuales no están integradas en infraestructuras de pago de vía pública o de parking municipal ni utilizan un protocolo estandarizado y reconocido a nivel europeo de gestión de infraestructuras de recarga (OCPP) en toda la arquitectura.b) La plataforma debe contemplar el uso de distintas formas de identificación de usuario y pago para aproximarse a la realidad de los parkings público medianos y grandes actualesc) No existen plataformas municipales que gestionan la utilización de empresas de car-sharing de forma eficiente.d) Utilización de un amplio abanico de sistemas de identificación de usuario.La oportunidad de mercado se encuentra en los parkings urbanos medianos (hasta 250 plazas) y grandes (más de 250 plazas) y las infraestructuras de car-sharing localizadas en vía pública y parkings de interior.a) Uso del protocolo estandarizado europeo OCPP de gestión de infraestructuras de recarga que permitan garantizar la interoperabilidad, universalidad y crecimiento de la

Identificador

Título SMART_CHARGE. Sistema de carga inteligente para automóviles electricos.

Info Empresa:

Ver metadatos
Ver fichero original
Ver fichero memoria

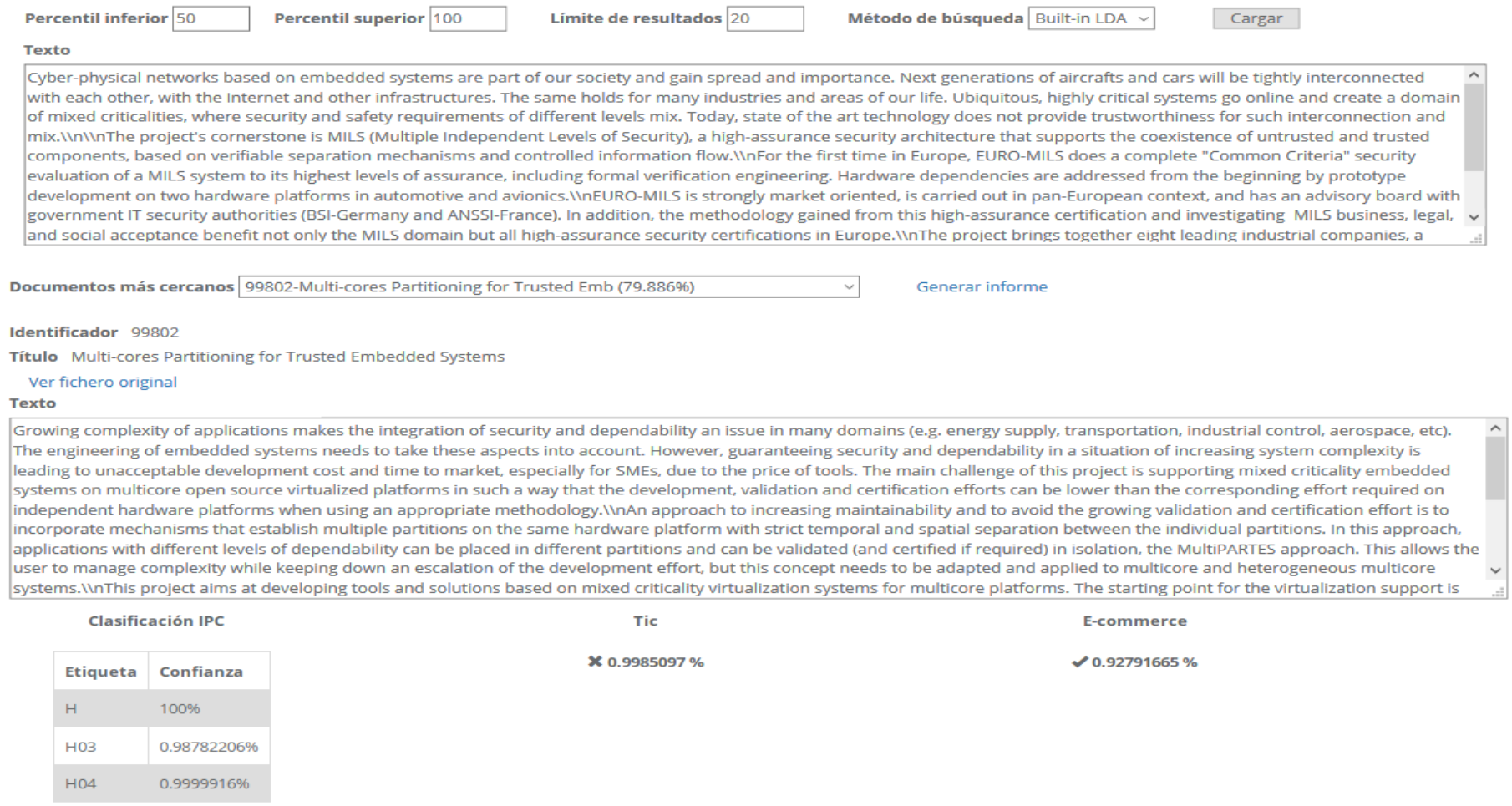
Texto

[[SMART_CHARGE. Sistema de carga inteligente para automóviles electricos. Solución de gestión/explotación TIC municipal de infraestructuras de recarga integrada en los sistemas de pago y control de accesos de los parkings públicos actuales, como lugar idóneo para la recarga y estancia de V.E. Vinculación de Smart Charging Infraestrucuture y Smart Parking Management en una plataforma de Smart City.Para desarrollar los objetivos de este proyecto es conveniente resolver las siguientes problemáticas actuales detectadas en los parkings públicos a) Las soluciones de recarga actuales no están integradas en infraestructuras de pago de via publica o de parking municipal ni utilizan un protocolo standarizado y reconocido a nivel europeo de gestión de infraestructuras de recarga (OCPP) en toda la arquitectura. b) La solución debe contemplar el Uso de distintas formas de identificación de usuario y pago para aproximarse a la realidad de los parking publico medianos y grandes actuales c) No existen plataforma municipales que gestionan la utilización de empresas de car-sharing de forma eficiente d) Uso de Smart Phones como terminal para realizar la identificación, pago y gestión de la plaza y de la infraestructura de recarga. La oportunidad de mercado se encuentra en los parkings urbanos medianos (hasta 250 plazas) y grandes (+ 250 plazas) y las infraestructuras de car-sharing localizadas en via publica y parkings de interior.a) Uso del protocolo estandarizado europeo OCPP de gestión de infraestructuras

Basado en la distancia semántica entre documentos

GOBIERNO DE ESPAÑA MINISTERIO DE ENERGÍA, TURISMO Y AGENDA DIGITAL SECRETARÍA DE ESTADO PARA LA SOCIEDAD DE LA INFORMACIÓN Y LA AGENDA DIGITAL red.es Plan TL Plan de Impulso de las Tecnologías del Lenguaje

Búsqueda por inferencia temática

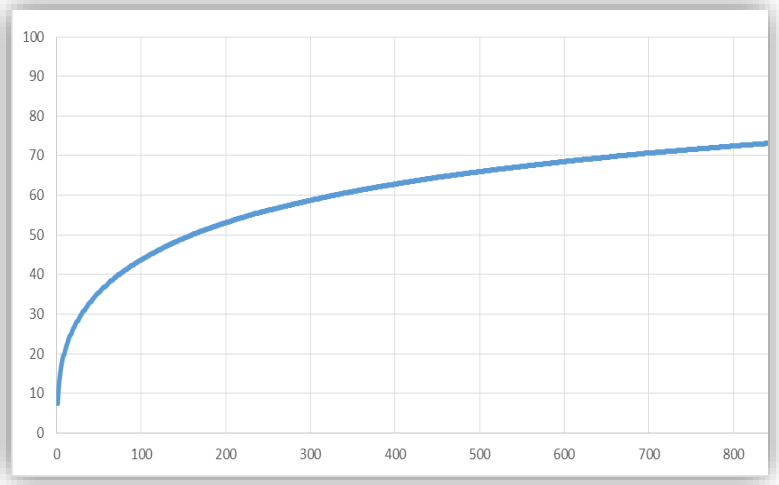


Permite la búsqueda de documentos del corpus cuyo contenido es más parecidos a un texto dado

Se emplea una búsqueda no textual basada en la distancia de sus proyecciones sobre el modelo de tópicos

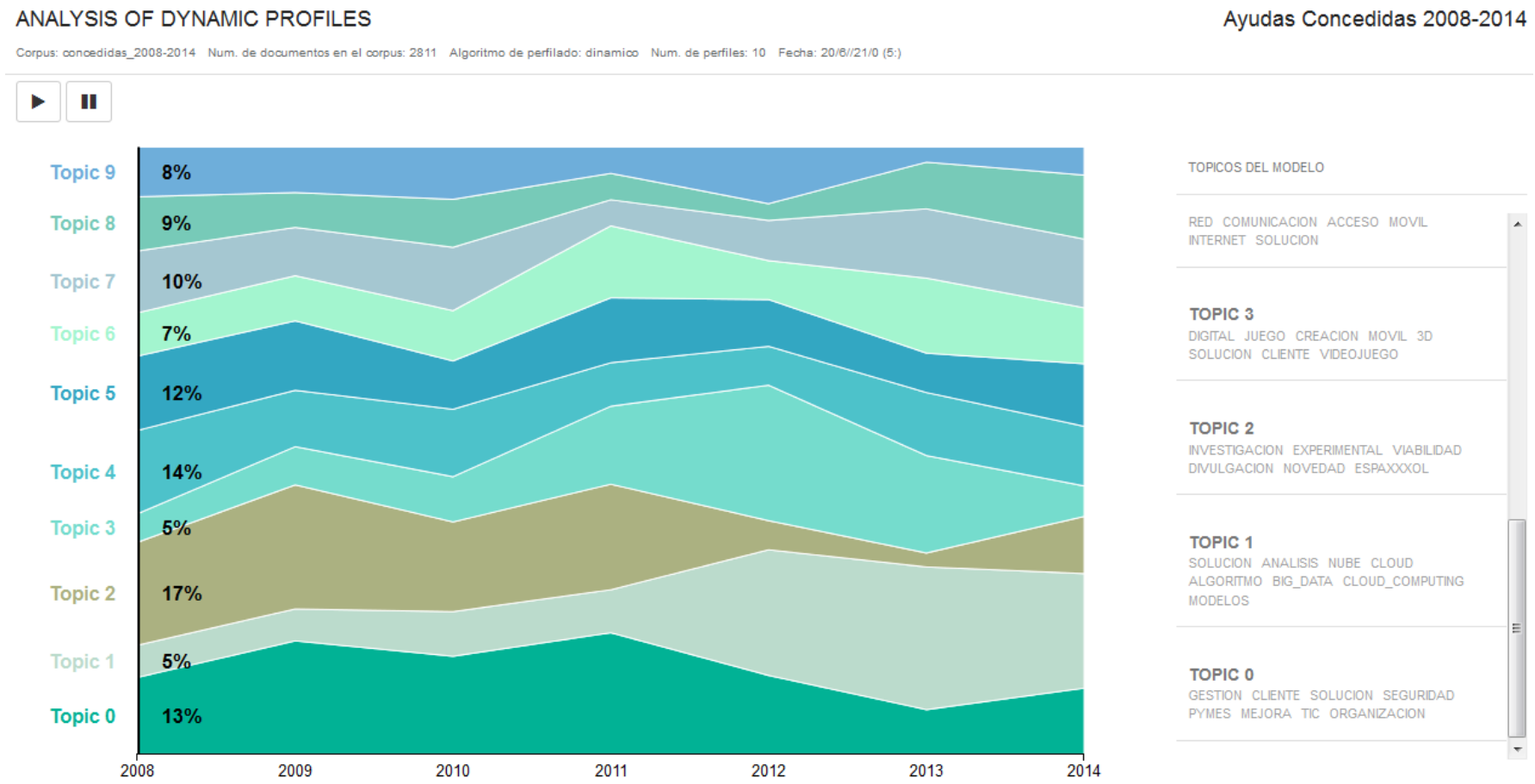
Recuperación de información OEPM

- **Un 7.5% de las citas buscadas aparecieron como primer resultado de la búsqueda automática.**
- Resultados detallados de la búsqueda de citas:
  - El 20.8% aparecieron entre los 10 primeros resultados
  - El 35.5% entre los 50 primeros
  - El 43.7% entre los 100 primeros
  - **El 50% aparecieron entre los 161 primeros resultados**



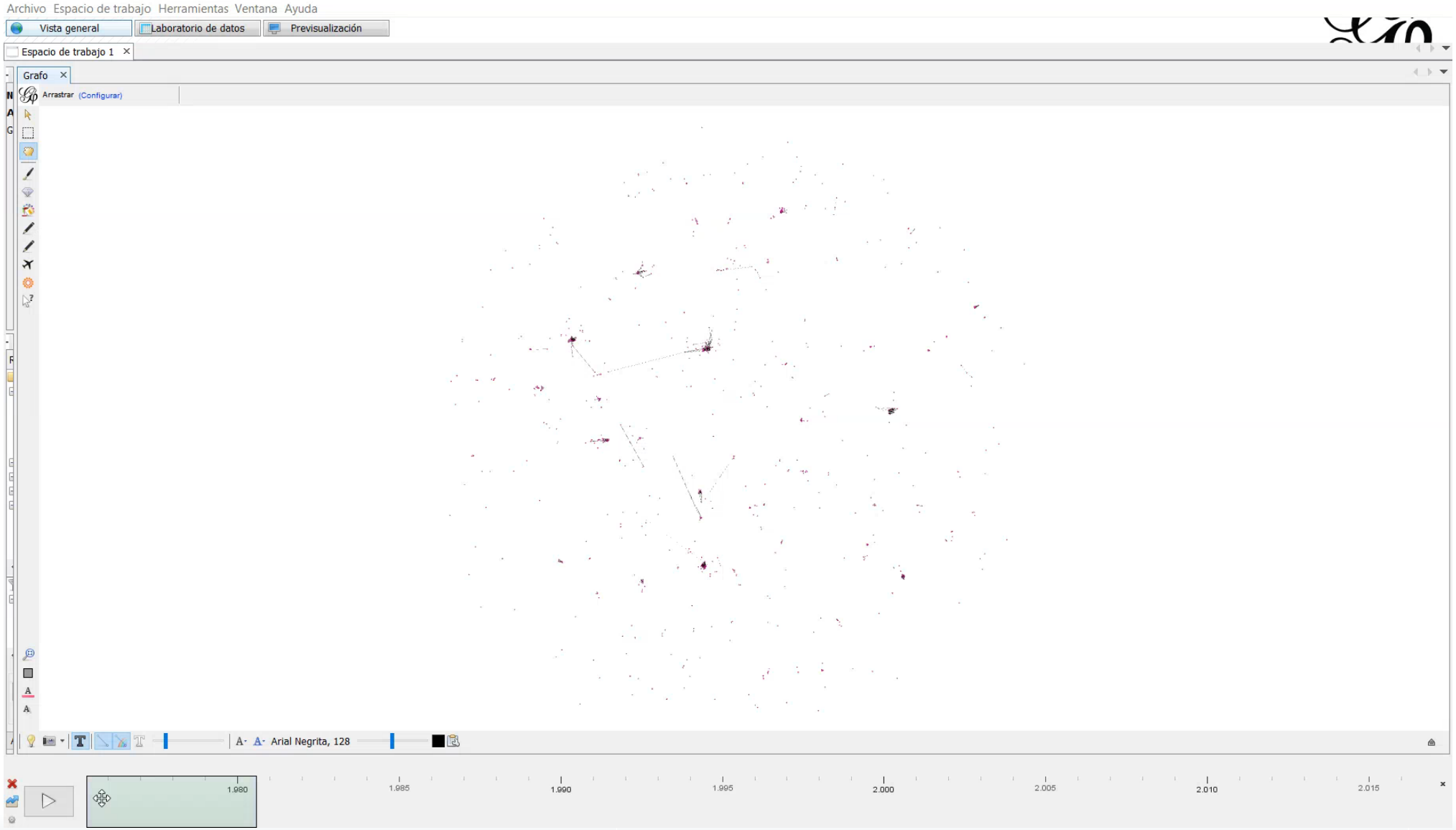
- El 87.7% de todas las citas buscadas aparecieron entre las primeras 2.700 patentes seleccionadas por el sistema. **Esto es, para una amplia mayoría de los casos, se redujo la búsqueda al 1% del total de patentes analizadas.**
- Estos resultados se obtuvieron sin aplicar ninguna restricción a la búsqueda. **Aplicando filtros adicionales por metadatos (IPC u otros) se podría reducir más aún el ámbito de búsqueda.**

Evolución temporal



Evolución de la temática contenida en los documentos del corpus

Evolución temporal National Science Foundation



Herramientas opensource

satec

Java

python

ANGULARJS FRAMEWORK

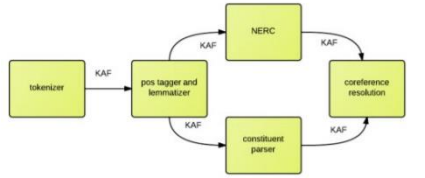
D3

{JSON}

talend

DBpedia Spotlight

cassandra



IXA pipes: ready to use NLP tools

MALLET

machine learning for language toolkit

APACHE Spark

Apache Solr

docker

kubernetes

Gephi

neo4j

elastic

Líneas de actuación futuras

- Uso de técnicas **semánticas** empleando un conocimiento base (ontologías específicas dominio)
- Profundizar en técnicas para **independencia idioma**
- Análisis de **redes semánticas**
- **Análisis temporal** sobre modelos de tópicos
- **Reducción del peso** de los modelos
- Aplicación de nuevas técnicas en cálculo de distancia
- Incorporación de **nuevos corpus documentales** y compartición de modelos
- Publicación REST APIs - PaaS / Interoperabilidad (IaaS)
- Avanzar hacia un posible **análisis de impacto** (utopía o realidad??)

Muchas gracias!!

david.nieto@satec.es

GOBIERNO DE ESPAÑA
MINISTERIO DE ENERGÍA, TURISMO Y AGENDA DIGITAL
SECRETARÍA DE ESTADO PARA LA SOCIEDAD DE LA INFORMACIÓN Y LA AGENDA DIGITAL
red.es
Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

www.satecgroup.com