MANUAL BÁSICO DE USO DE LA INSTANCIA PÚBLICA DE CORPUS VIEWER

Plan de impulso de las Tecnologías del Lenguaje

07/2019

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y EMPRESA
SECRETARÍA DE ESTADO PARA EL AVANCE DIGITAL

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

ÍNDICE

1. PROPÓSITO DEL DOCUMENTO

Este documento proporciona una guía básica de usuario de la plataforma Corpus Viewer (Visor de corpus) para visualización de colecciones documentales, desarrollada dentro del Plan de Tecnologías del Lenguaje. Permite, mediante el uso de las tecnologías de lenguaje natural y otras técnicas de inteligencia artificial, analizar grandes volúmenes de información textual no estructurada (colecciones de documentos españoles y extranjeros) e inferir relaciones entre estos textos.

Esta aplicación sirve de apoyo a los responsables de las políticas públicas, tanto para el diseño y seguimiento de políticas, como para la gestión de convocatorias de ayudas a partir del tratamiento de las grandes colecciones de datos no estructurados de las que se dispone.

Corpus Viewer es una herramienta que se encuentra en producción en distintas entidades del Sector Público (SEAD, SEUIDI, FECYT), y habitualmente los usuarios reciben una formación de varias horas previas a su acceso a la herramienta. Para los accesos a la instancia (demostrador online) pública no es práctico plantear una formación de dicho tipo, y la herramienta en sí no está diseñada para ser autoexplicativa en toda su funcionalidad, lo que aconseja que los usuarios dispongan de un mínimo de documentación para poder interpretar mejor la información suministrada por la herramienta. Esta guía se ha escrito con dicho propósito.

2. CORPUS DOCUMENTALES DISPONIBLES

Entendemos por corpus, una colección de documentos cuyo contenido está expresado en lenguaje natural.

A fecha de 3 de julio de 2019 los siguientes corpus documentales están disponibles en la instancia pública de Corpus Viewer:

- ACL: Es un corpus de publicaciones científicas en el ámbito de la lingüística computacional (Association of Computational Linguistics).
- CORDIS720: Proyectos de Investigación financiados por la Unión Europea dentro del Séptimo Programa Marco y de Horizonte 2020.
- CORDIS720_AI: Contiene una selección de proyectos del corpus anterior en los que la Inteligencia Artificial se encuentra presente, bien porque el proyecto desarrolla técnicas de

Inteligencia Artificial, o bien porque se utilizan en algún ámbito de aplicación. La selección de los proyectos incluidos en el subcorpus se ha realizado de forma automática empleando técnicas de aprendizaje automático. El uso de estas técnicas permite abordar el etiquetado de un número elevado de proyectos evitando el gran coste en tiempo que supondría un etiquetado manual, pero implica inevitablemente la introducción de un determinado margen de error en cuanto a los proyectos seleccionados.

Próximamente se irán publicando en la plataforma los siguientes corpus documentales:

- Ayudas de la National Science Foundation (NSF)
- Ayudas americanas en el ámbito de ciencias de la salud (NiH)
- Un corpus de publicaciones científicas de mayor tamaño (basado en Semantic Scholar)

La publicación de estos y otros corpus se notificará a los usuarios activos, salvo que hayan expresado su deseo de no recibir ninguna comunicación.

3. ACCESO A CORPUS VIEWER

El acceso al demostrador online se debe solicitar remitiendo un correo a plantecnologiaslenguaje@mineco.es con asunto “Acceso Corpus Viewer”.

Una vez se haya tramitado su solicitud, recibirá un correo electrónico con su usuario y contraseña, permitiéndole el acceso al demostrador a través de la siguiente dirección web:

https://cvdemo.plantl.gob.es/CorpusViewer/#/login

Tras identificarse en el sistema es conveniente que cambie la contraseña de acceso proporcionada inicialmente, para lo que debe acceder a la opción “Editar Perfil” situada en el menú desplegable de la parte superior derecha de la ventana.

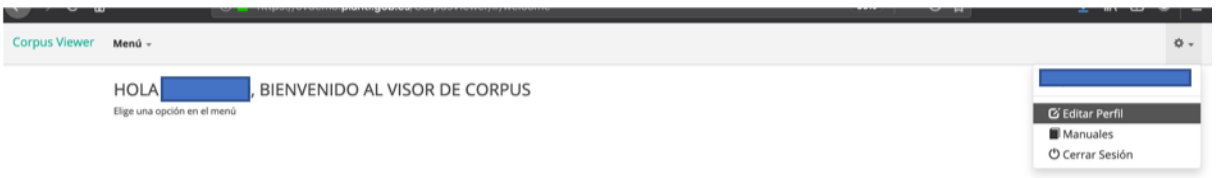

Figura 1: Edición de Perfil del Usuario.

Para cerrar su sesión en Corpus Viewer, acceda nuevamente a dicho menú, a la opción de "Cerrar Sesión".

4. NAVEGACIÓN POR LAS HERRAMIENTAS DE CORPUS VIEWER

Para hacer uso de la propia herramienta debe acceder a la opción "Menú" que figura en la pestaña superior. Una vez que ha seleccionado cualquiera de las opciones disponibles, aparece en dicha pestaña superior la siguiente información:

- Un listado de pestañas disponibles, cada una de las cuales proporciona una vista diferente del corpus documental seleccionado.
- Un menú desplegable en el que puede seleccionar el corpus con el que desea trabajar.
- Un menú desplegable que permite seleccionar un modelo de entre los asociados al corpus seleccionado.

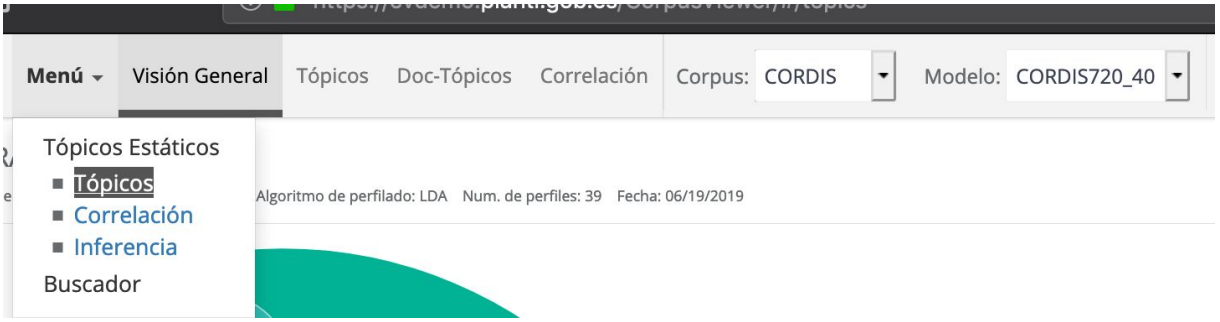

Figura 2: Navegación por el Menú General. Selección de la visualización basada en tópicos.

5. UNA BREVE INTRODUCCIÓN AL MODELADO DE TÓPICOS

La construcción de modelos de tópicos se basa en una técnica de aprendizaje automático denominada *Latent Dirichlet Allocation (LDA)*. Existen múltiples fuentes en Internet que proporcionan información acerca de esta técnica, algunas meramente intuitivas, y otras abordando con mayor detalle matemático la generación de tópicos y documentos. Esta entrada de Quora contiene varias explicaciones con distinto nivel de complejidad. Por razones de reconocimiento académico queremos incluir también el paper original de David Blei en el que se propone el algoritmo original.

A los efectos que nos ocupan, posiblemente resulta suficiente explicar los siguientes dos conceptos básicos de manera **muy simplista**:

- En LDA un **tópico** se puede caracterizar como un **conjunto de palabras que suelen aparecer juntas** en muchos documentos. Por ejemplo: las palabras *gene, cellular, membrane* suelen co-ocurrir frecuentemente. LDA es capaz de localizar estas coocurrencias sobre la colección completa de documentos, y definir los tópicos a partir de ellas. Se podría decir que cada conjunto de palabras representa una posible **área temática** que es a lo que llamamos **tópico**.
- En LDA un documento puede estar caracterizado por un único tópico, aunque frecuentemente es realmente una mezcla de tópicos. Nuevamente, LDA proporciona un vector para cada documento que indica en qué medida el documento pertenece a cada uno de los tópicos identificados.

Las herramientas que se emplean en Corpus Viewer están basadas en *Latent Dirichlet Allocation*, pero incluyen algunas modificaciones realizadas dentro de los distintos contratos ejecutados en el Plan de Tecnologías del Lenguaje. El lector interesado puede remitirse a la página web del plan para consultar más información sobre algunos de estos desarrollos:

https://www.plantl.gob.es/inteligencia-competitiva/resultados/desarrollos-SW/Paginas/desarrollos.apx

6. VISUALIZACIÓN DE LOS TÓPICOS ASOCIADOS A UN CORPUS DOCUMENTAL

Seleccionando "**Menú -> Tópicos estáticos: Tópicos**", tenemos acceso a las siguientes pestañas:

- Visión General: Permite estudiar las temáticas principales del corpus.
- Tópicos: Permite estudiar las temáticas principales del corpus.
- Doc-Tópicos: Permite analizar las temáticas de documentos concretos.
- Correlación: Permite estudiar las relaciones entre temáticas.

6.1 TÓPICOS: PESTAÑA VISIÓN GENERAL

La primera de las visualizaciones disponibles nos lleva a una ventana en la que se nos muestra información general sobre el corpus documental seleccionado, y sobre cada uno de los tópicos identificados para dicho corpus. Se incluye así mismo una visualización gráfica interactiva del modelo. Según se va pasando el ratón por los conjuntos, se muestra una etiqueta con las palabras que

caracterizan cada tópico. Si se hace clic en uno se accede al detalle de dicho tópico. Volviendo a hacer clic se vuelve a la vista general.

En la lista Tópicos del modelo, se ofrece la siguiente información para cada uno de los tópicos:

- Tamaño relativo del perfil (estimado por el modelo LDA; está relacionado con la importancia del tópico en el corpus, pero no puede inferirse una relación directa con el número de documentos asociados al tópico, ya que hemos visto que los documentos pueden pertenecer a varios tópicos en distinta medida).
- Un título propuesto por un anotador experto de la SEAD (texto en negrita)
- La lista de palabras identificadas como más relevantes para el tópico en cuestión (debajo del título de cada tópico).

La lista de tópicos es de tipo deslizante, por lo que deberemos desplazarnos con el cursor por ella para visualizar todos los tópicos.

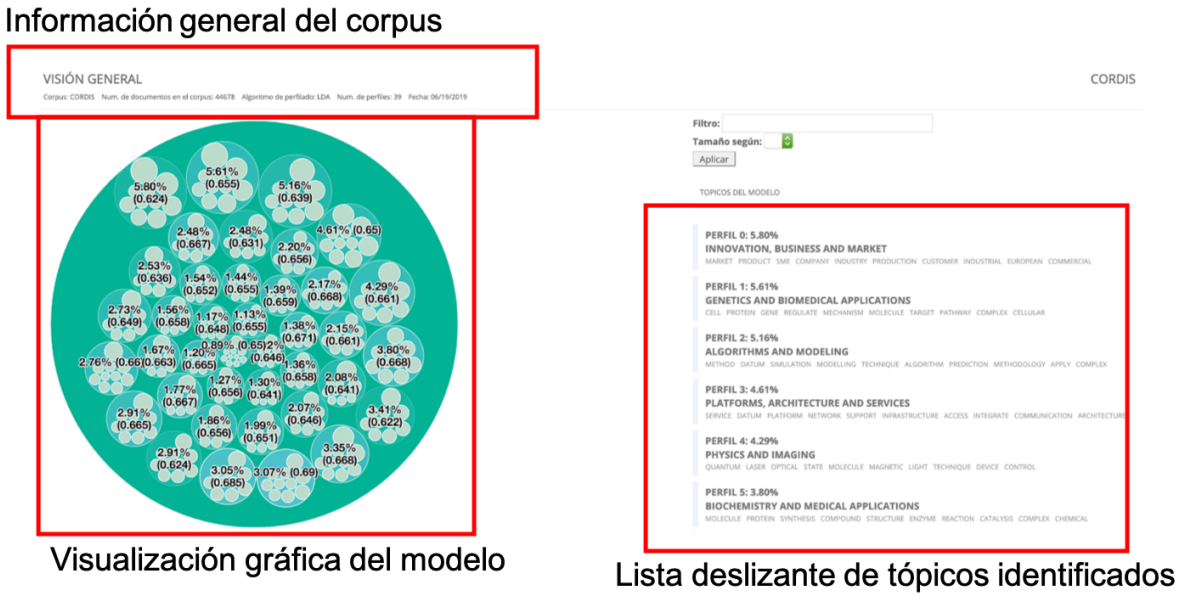

Figura 3: Vista General de Tópicos de Corpus Viewer.

Si hacemos clic sobre cualquiera de los tópicos (tanto en la visualización gráfica como en la lista de tópicos), la vista cambia para enfatizar el tópico seleccionado y se muestra, además:

- Una visualización gráfica de las palabras más relevantes del tópico (tanto sobre el gráfico interactivo de bolas, como en la versión histograma)

- Un listado de los documentos que mejor representan el perfil seleccionado. Haciendo clic sobre el enlace disponible, podemos acceder al texto asociado al documento.

Haciendo clic nuevamente sobre el gráfico de bolas podemos movernos a otro perfil, o regresar a la visualización del modelo general.

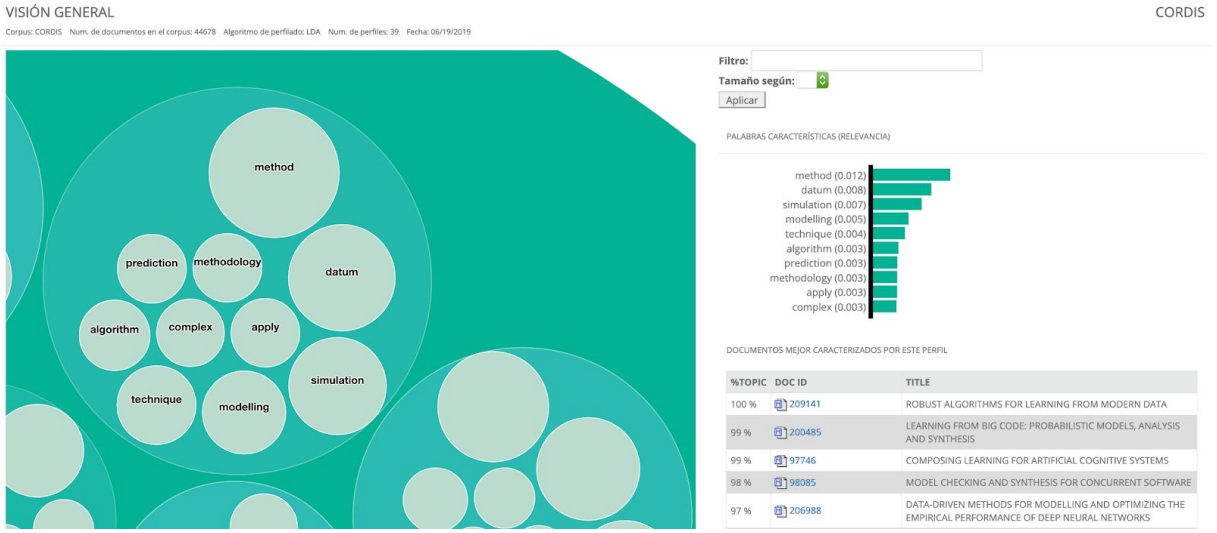

Figura 4: Visualización detallada de tópico incluyendo su descripción basada en palabras, y los documentos más característicos del tópico seleccionado.

6.2 TÓPICOS: PESTAÑA TÓPICOS

Esta segunda pestaña permite una visualización del modelo similar a la descrita en el caso anterior, si bien la selección de tópicos se realiza mediante un menú desplegable en el que se muestra el título de los tópicos y su importancia relativa en el corpus.

Figura 5: Visualización de tópicos en la pestaña "tópicos".

Nuevamente, para el tópico seleccionado se muestran los documentos más representativos del mismo, y el listado de palabras más relevantes, en formato histograma y bolsa de palabras.

Esta ventana ofrece además la posibilidad de enfatizar las palabras más discriminativas del tópico (palabras clave) seleccionando la opción "Con penalización por TF/IDF".

El uso de TF-IDF es habitual en la representación de documentos empleando bolsas de palabras. En este caso, utilizamos una extensión de este concepto para representar el valor de las palabras en cada tópico. Siendo:

- TF: Term Frequency: Mide la probabilidad de una palabra en un tópico dado.
- IDF: Inverse Document Frequency: En este contexto, es un factor inverso a la importancia del término en el conjunto de tópicos del modelo.

De esta manera, si activamos la opción "Con penalización por TF-IDF", el sistema re-ponderará el peso asignado a cada palabra dentro del tópico, y se restará peso a aquellas palabras que son comunes a un mayor número de tópicos (palabras vacías). Dicho de otro modo, enfatizaremos las palabras más discriminativas, en el sentido de que suben de importancia las palabras que únicamente están presentes en el tópico seleccionado.

Por último, cabe mencionar que la pestaña ofrece información sobre la entropía normalizada del tópico, que da idea de la transversalidad del tópico a lo largo de la colección de documentos. Sin embargo, el cálculo de entropías normalizadas actualmente implementado ofrece un bajo rango dinámico, y desde el equipo técnico de la SEAD se está trabajando en mejorar la calidad de este indicador.

6.3 TÓPICOS: PESTAÑA DOC-TÓPICOS

La pestaña "Doc-Tópicos" permite realizar una búsqueda de documentos por palabras clave. Dicho buscador tiene la capacidad de "autocompletar", por lo que al introducir unas palabras se proporcionarán sugerencias de documentos que las contienen.

Una vez seleccionado el documento que se desea analizar, se ofrece una visualización gráfica del contenido temático del mismo. Recordemos que en Latent Dirichlet Allocation cada documento queda caracterizado por su nivel de pertenencia a los tópicos del modelo.

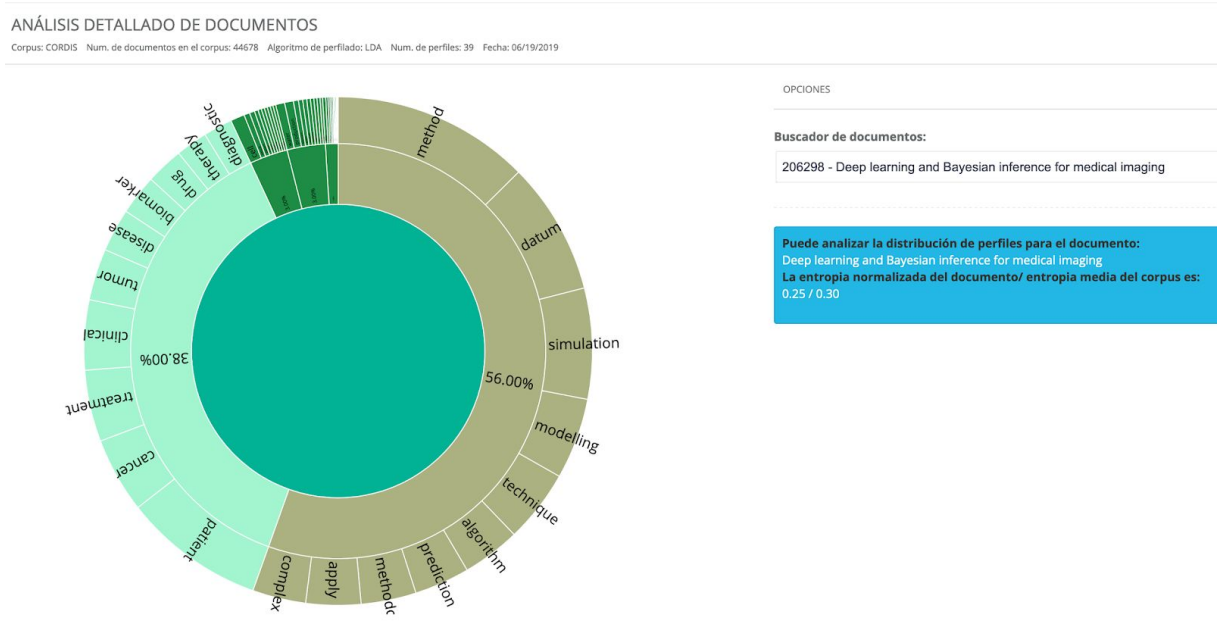

Figura 6: Análisis detallado de documentos basado en las temáticas más relevantes que lo caracterizan.

A modo de ejemplo, la figura incluida muestra que el documento:

"206298 - Deep learning and Bayesian inference for medical imaging"

Pertenece en un 56% al tópico caracterizado por las palabras “method, datum, simulation, ...” (Algorithms and Modeling), en un 38% al tópico caracterizado por las palabras “patient, cancer, treatment, ...” (Cancer and Biomedical Applications), y en menor medida a otros perfiles.

El gráfico es interactivo, lo que permite ampliar para visualizar los tópicos de menor importancia para el documento haciendo clic sobre ellos. Para volver a la vista anterior más general, basta con hacer clic en el centro de la corona circular.

6.4 TÓPICOS: PESTAÑA CORRELACIÓN

Por último, la herramienta permite medir el nivel de correlación entre tópicos. Para ello, se estima que la relación entre dos tópicos es mayor cuando dichos tópicos tienden a ocurrir de forma conjunta en los mismos documentos.

Navegando sobre el gráfico de la izquierda podremos seleccionar cada uno de los tópicos del modelo, y los enlaces con otros tópicos muestra su nivel de concurrencia con otros tópicos del modelo. Dado que en la figura no se dispone de espacio suficiente para mostrar el título completo de los perfiles, se incluye dicha información en formato textual en la parte derecha de la pestaña. Al posicionarse en la figura sobre el nombre de una temática, el título completo se mostrará en la información textual del lado derecho de la página. Al seleccionar una temática sobre la figura se muestran sólo las relaciones con la misma, ocultándose el resto de curvas.

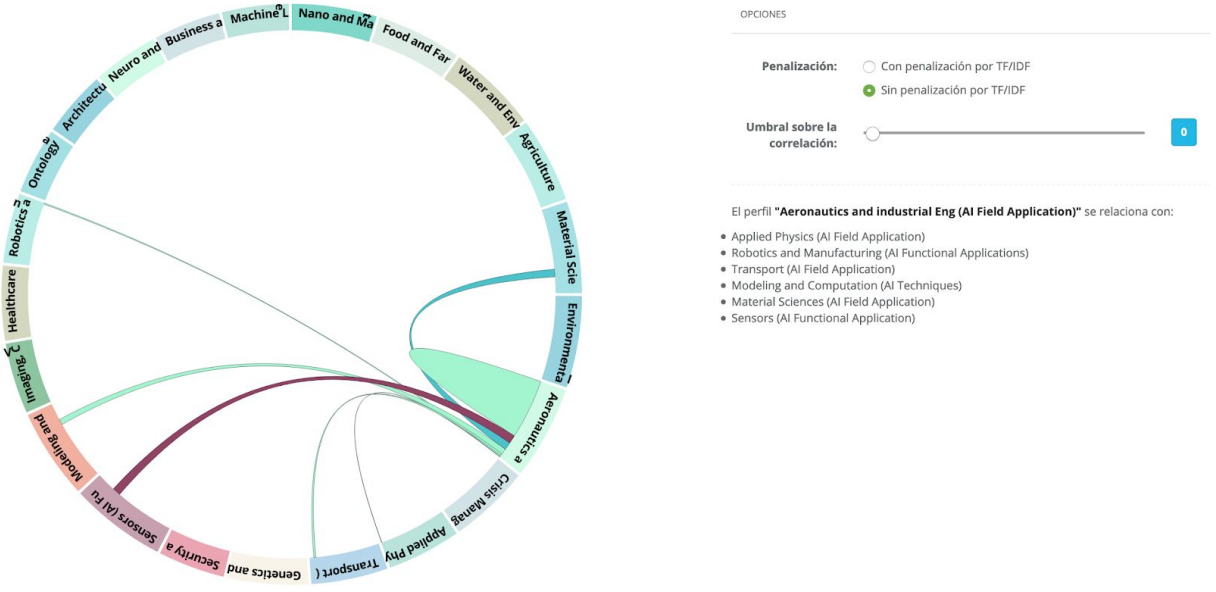

Figura 7: Visualización de la correlación entre tópicos del modelo. Para cada tópico se resaltan otros tópicos que concurren frecuentemente con él.

Adicionalmente, se puede seleccionar la opción “Con penalización por TF/IDF” que ya se ha explicado en el apartado anterior, así como, elegir un mayor o menor umbral para la correlación, de manera que se visualizarán sólo aquellas relaciones que superen el umbral indicado.

7. ESTUDIO DE RELACIONES ENTRE DOCUMENTOS EN BASE A SUS TEMÁTICAS

Como ya se ha comentado, el algoritmo de modelado de tópicos empleado permite representar cada documento a partir de su nivel de pertenencia a los distintos tópicos. Esta representación permite medir “distancias semánticas” entre documentos. Según dicha distancia, dos documentos son más parecidos entre sí cuanto más se parecen sus representaciones basadas en el modelo de tópicos, es decir, cuanto más parecidas son sus temáticas.

Corpus Viewer incorpora herramientas que permiten explotar esta relación semántica entre documentos. Seleccionando la opción “**Menú -> Tópicos estáticos: Correlación**” accedemos a dos pestañas que explotan esta información:

- Documentos: Herramienta de búsqueda de documentos por similitud semántica.
- Alarmas: Herramienta de búsqueda de pares de documentos con muy alta similitud semántica.

7.1 CORRELACIÓN: PESTAÑA DOCUMENTOS

La primera de las pestañas disponibles ofrece un buscador de documentos que permite seleccionar un documento concreto. Una vez seleccionado, se ofrece un listado de hasta 20 documentos que guardan una relación semántica alta con el documento seleccionado.

Para cada uno de los documentos listados, haciendo clic sobre los distintos iconos que aparecen a su derecha, podemos acceder a:

- consultar sus metadatos, incluído el título y el texto completo de cada documento.
- exportar a excel el listado completo de documentos.

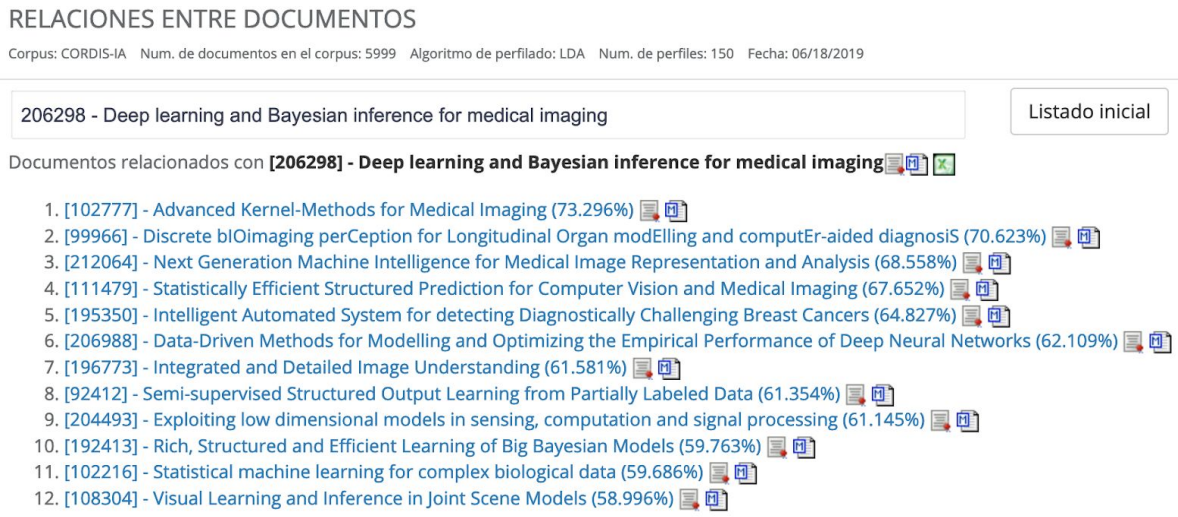

Figura 8: Listado de documentos semánticamente parecidos al documento seleccionado por el usuario.

Finalmente, cabe mencionar que el listado permite la navegación iterativa por documentos: si pinchamos sobre el título de los documentos en la lista de documentos similares, seleccionaremos dicho documento y la herramienta actualizará la lista de documentos similares con los correspondientes al nuevo documento seleccionado.

Para regresar al listado completo, basta con pulsar sobre el botón “Listado inicial”.

7.2 CORRELACIÓN: PESTAÑA ALARMAS

Esta herramienta permite buscar pares de documentos con muy alta similitud semántica. Dicha similitud puede emplearse para buscar duplicados, o documentos que han sido remitidos múltiples veces para evaluación.

Hay que insistir en que la herramienta proporcionada no está basada en búsqueda de similitud textual (como hacen herramientas tipo turnitin, etc), sino similitud semántica. Dos documentos pueden ser muy similares entre sí siempre que combinen las mismas temáticas en proporciones similares. Por este motivo, esta herramienta de búsqueda es muy robusta frente a la presencia de sinónimos, reescrituras de textos, etc., ya que, aunque cambie el texto la representación del documento en el modelo de tópicos permanece relativamente estable en estos casos.

Figura 9: Opciones para la búsqueda de "Alarmas" basada en similitud semántica entre documentos.

La herramienta permite determinar el nivel de similitud exigido para la detección de alarmas (percentil inferior y superior), o exigir que uno de los dos documentos seleccionados pertenezca a un año concreto (campo centrado en año)[1]. Una vez hemos establecido los ajustes deseados, pulsaremos el botón "cargar" y la herramienta cargará los pares de documentos similares en el menú desplegable "Alarmas encontradas".

A modo de ejemplo, si seleccionamos el corpus CORDIS-IA y utilizamos los parámetros por defecto, la primera alarma encontrada (con una similitud del 94%) proporciona la vista de la siguiente figura. Podemos comprobar que se trata de dos proyectos solicitados en los años 2009 y 2013, y que fundamentalmente son continuación el uno del otro.

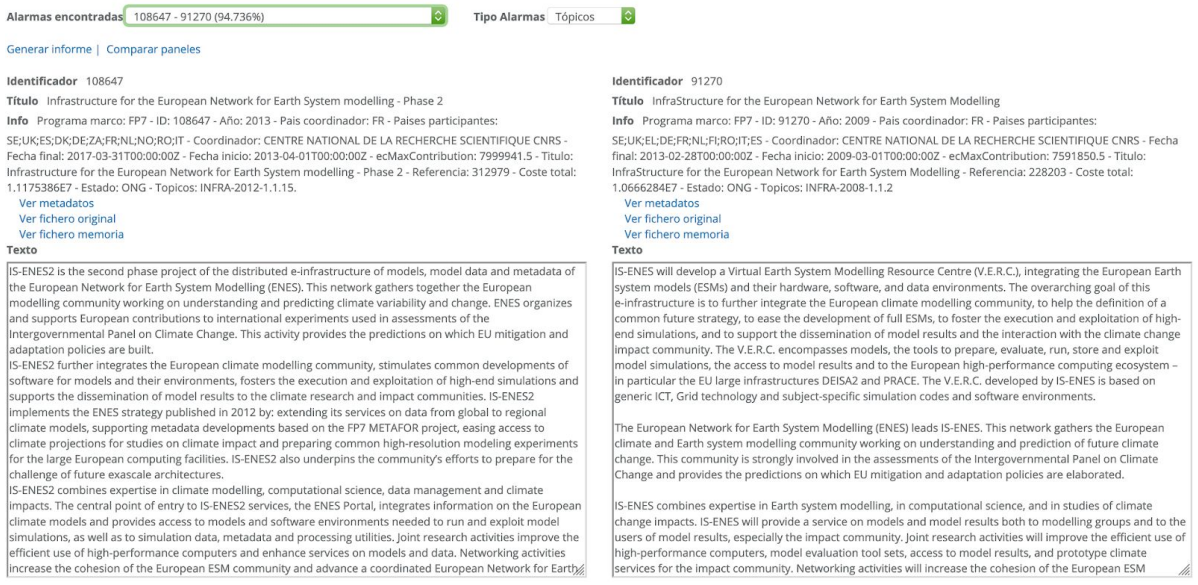

Figura 10: Desplegable con las "Alarmas" encontradas por la aplicación, y vista paralela de dos documentos identificados como muy similares (semánticamente).

[1] *En ocasiones resulta interesante disminuir el valor de percentil superior a un valor inferior al 100% o centrar el análisis en un año concreto. Esto puede ser importante especialmente en aquellos casos en los que los documentos han estado sujetos a un proceso de OCR (este es el caso del corpus ACL), ya que en determinados casos puede haber documentos temáticamente idénticos por estar asociados a la presencia de caracteres ruidosos que proceden de un funcionamiento defectuoso del reconocimiento de caracteres.*

Si pinchamos en la opción “comparar paneles”, podemos observar como la similitud **textual** de ambos proyectos es relativamente baja, aunque se haya detectado una similitud semántica alta. En lo referente a similitud textual, las oraciones marcadas en rojo (verde) aparecen únicamente en el texto del documento del panel izquierdo (derecho), mientras que el texto en blanco es el que aparece simultáneamente en ambos documentos. Este ejemplo ilustra claramente la diferencia existente entre esta herramienta basada en similitud semántica frente a otras herramientas basadas en similitud textual.

Figura 11: Panel de comparación textual entre pares de documentos con alta similitud semántica.

8. DOCUMENTOS SIMILARES A UN TEXTO ARBITRARIO

Todas las funcionalidades descritas en la sección anterior permiten explotar similitudes semánticas, pero su uso está restringido a aquellos documentos que pertenecen a las colecciones de documentos cargadas en Corpus Viewer. En ocasiones puede ser interesante buscar similitudes con otros textos nuevos proporcionados por el usuario. Esto es posible seleccionando la opción “**Menú -> Tópicos estáticos: Inferencia**” en el menú principal de Corpus Viewer.

Figura 12: Pestaña para inferencia temática sobre un texto libre proporcionado por el usuario, y búsqueda de documentos con una temática similar indexados en Corpus Viewer.

La herramienta de Inferencia se basa en los siguientes pasos:

1. El texto proporcionado se preprocesa utilizando las mismas herramientas que se emplearon para el preprocesamiento de los documentos del corpus activo.

2. El texto proporcionado se "proyecta" sobre el modelo de tópicos asociado al corpus activo. De esta manera, obtenemos una representación basada en tópicos similar a la disponible para todos los documentos del corpus cargado en Corpus Viewer.

3. Se calcula la similitud semántica entre el texto proporcionado y cada uno de los documentos del corpus seleccionado, y se muestran al usuario los documentos más similares.

Cabe mencionar que esta herramienta requiere la ejecución de ciertos cálculos en los servidores de Corpus Viewer, por lo que el tiempo de respuesta puede ser de algunos segundos (mayor cuanto mayor es el número de documentos en el corpus seleccionado).

También es necesario resaltar que la representación semántica del texto tendrá mayor calidad cuanto mayor sea la longitud del texto proporcionado. Por ello podemos esperar resultados de mayor calidad cuanto más extenso sea el texto empleado para la consulta.

9. BUSCADOR

Seleccionando **"Menú -> Buscador**" se accede a la última de las opciones actualmente activas en Corpus Viewer, que consiste en una herramienta basada en Solr y Banana. Dicha herramienta ofrece la funcionalidad de una herramienta tipo BI, si bien, integra los metadatos disponibles con la representación basada en tópicos de los documentos.

Actualmente el buscador se encuentra en fase de desarrollo, por lo que no se incorpora toda la información que estará disponible en la versión final, y es de esperar cambios en los paneles que finalmente se incorporen en cada corpus.

Aun no estando el desarrollo finalizado, se ha optado por dejar activa dicha pestaña en la instancia abierta de Corpus Viewer, para que los usuarios puedan obtener una primera impresión del tipo de funcionalidad que se proporcionará una vez el desarrollo esté concluido.

Puede consultar la versión de demostración con datos de CORDIS para Inteligencia Artificial desarrollada sobre Javascript.

El funcionamiento del buscador que se incorporará en Corpus Viewer será similar al del demostrador proporcionado, e incluirá toda la potencia de búsqueda y agrupamiento proporcionado por la tecnología de indexado proporcionada por Solr.