BASIC MANUAL OF USE OF THE PUBLIC INSTANCE OF CORPUS VIEWER

PLAN FOR THE ADVANCEMENT OF LANGUAGE TECHNOLOGY

July/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

INDEX

1. DOCUMENT PURPOSE

This document provides a basic user guide of the Corpus Viewer platform for analyzing documentary collections, developed within the Language Technology Plan. It allows, through the use of natural language technologies and other artificial intelligence techniques, to analyze large volumes of unstructured textual information and infer relationships between these texts.

This application serves as support for those responsible for public policies, both for the design and monitoring of policies, as well as for the management of projects calls exploiting the large collections of unstructured data available.

Corpus Viewer is a tool that is in production in different entities of the Public Sector in Spain (SEAD, SEUIDI, FECYT), and users usually receive training several hours prior to their access to the tool. For access to the instance (online demonstrator) It is not practical to propose such training, and the tool itself is not designed to be self-explanatory in all its functionality, which suggests that users have a minimum of documentation to better interpret the information provided by the tool. This guide has been written for that purpose.

2. AVAILABLE DOCUMENTARY CORPUS

We understand by corpus, a collection of documents whose content is expressed in natural language.

As of January 18, 2020 the following documentary corpus are available in the public instance of Corpus Viewer:

- ACL: It is a corpus of scientific publications in the field of computational linguistics (Association of Computational Linguistics).
- CORDIS720: Research Projects funded by the European Union within the Seventh Framework and Horizon 2020 Program.
- CORDIS720_AI: Contains a selection of previous corpus projects in which Artificial Intelligence is present, either because the project develops Artificial Intelligence techniques, or because they are used in some scope of application. The selection of the projects included in the subcorpus has been carried out automatically using machine learning techniques. The

use of these techniques makes it possible to address the labeling of a large number of projects, avoiding the high cost in time that manual labeling would entail, but inevitably implies the introduction of a certain margin of error regarding the selected projects.

The following documentary corpus will soon be published on the platform:

- Aid from the National Science Foundation (NSF)
- American aid in the field of health sciences (NiH)
- A corpus of larger scientific publications (based on Semantic Scholar)

The publication of these and other corpus will be notified to active users, unless they have expressed their desire not to receive any communication.

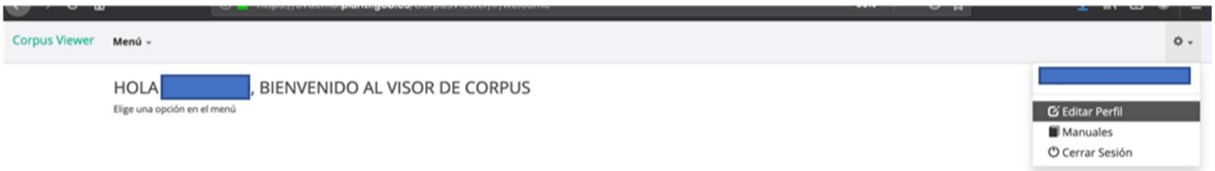
3. ACCESS TO CORPUS VIEWER

Access to the online demonstrator must be requested by sending an email to plantecnologiaslenguaje@mineco.es with subject “Corpus Viewer Access”.

Once your application has been processed, you will receive an email with your username and password, allowing access to the demonstrator through the following web address:

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

After identifying yourself In the system it is convenient that you change the access password initially provided, for which you must access the “Editar Perfil” option located in the drop-down menu in the upper right part of the window.



HOLA [redacted], BIENVENIDO AL VISOR DE CORPUS

Elige una opción en el menú

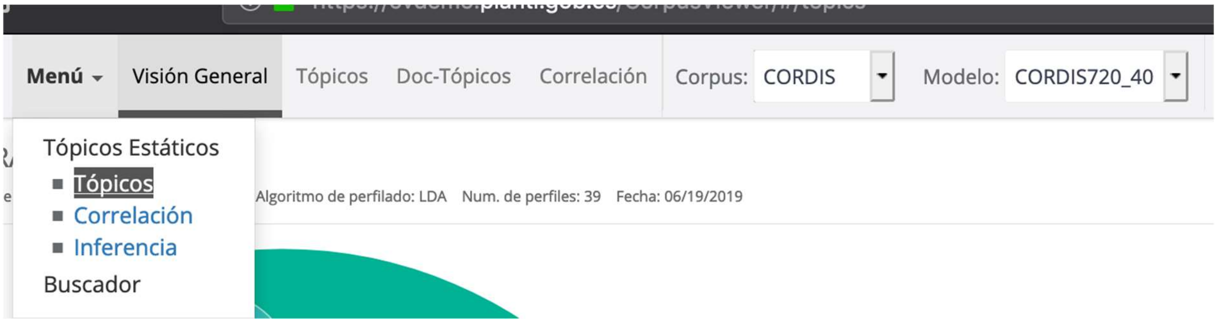
Figure 1: User Profile Edition.

To log out of Corpus Viewer, access this menu again, to the option "Cerrar Sesión".

4. NAVIGATION BY CORPUS VIEWER TOOLS

To use the tool itself, you must access the "Menu" option in the upper tab. Once you have selected any of the available options, the following information appears on that top tab:

- A list of available tabs, each of which provides a different view of the selected documentary corpus.
- A drop-down menu in which you can select the corpus with which you want to work.
- A drop-down menu that allows you to select a model from those associated with the selected corpus.



Corpus: CORDIS

Modelo: CORDIS720_40

Algoritmo de perfilado: LDA Num. de perfiles: 39 Fecha: 06/19/2019

Figure 2: Navigation through the General Menu. Display selection based on topics.

5. A BRIEF INTRODUCTION TO THE MODELING OF TOPICS

The construction of topic models is based on a machine learning technique called *Latent Dirichlet Allocation (LDA)*. There are multiple sources on the Internet that provide information about this technique, some merely intuitive, and others addressing in greater mathematical detail the generation of topics and documents. This Quora entry contains several explanations with different levels of complexity. For reasons of academic recognition we also want to include the original paper by David Blei in which the original algorithm is proposed.

For the purposes at hand, it is possibly enough to explain the following two basic concepts in a **very simplistic** way:

- In LDA a **topic** can be characterized as a **set of words that usually appear together** in many documents. For example: the words *gene, cellular, membrane* usually co-occur frequently. LDA is able to locate these co-occurrences on the complete collection of documents, and define the topics from them. You could say that each set of words represents a possible **thematic area** that is what we call a **topic**.
- In LDA a document can be characterized by a single topic, although often it is really a mixture of topics. Again, LDA provides a vector for each document that indicates the extent to which the document belongs to each of the identified topics.

The tools used in Corpus Viewer are based on *Latent Dirichlet Allocation*, but include some modifications made within the various contracts executed in the Language Technology Plan. The interested reader can refer to the plan's website for more information on some of these developments (currently information is only published in Spanish):

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

6. VISUALIZING THE TOPICS THAT CHARACTERIZE A DOCUMENTARY CORPUS

Selecting "**Menú -> Tópicos estáticos: Tópicos**", we have access to the following tabs:

- Visión General: Allows you to study the main themes of the corpus.
- Tópicos: It allows studying the main themes of the corpus.
- Doc-Tópicos: It allows analyzing the themes of specific documents.
- Correlación: It allows studying the relationships between themes.

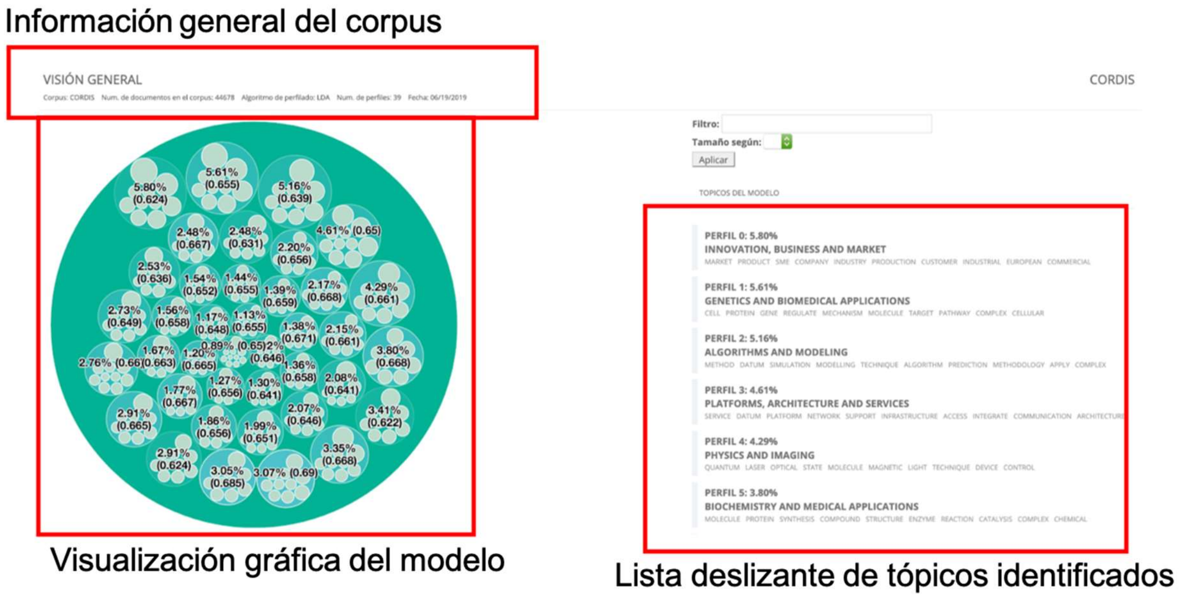
6.1 TÓPICOS: GENERAL VISION TAB

The first of the available visualizations takes us to a window in which we are shown general information about the selected documentary corpus, and about each of the topics identified for said corpus. It also includes an interactive graphic display of the model. As the cursor passes through the sets, a label is shown with the words that characterize each topic. If you click on one you will access the detail of that topic. Clicking again returns to the overview.

In the list “Tópicos del modelo”, the following information is offered for each of the topics:

- Relative profile size (estimated by the LDA model; it is related to the importance of the topic in the corpus, but a direct relationship cannot be inferred with the number of documents associated with the topic, since we have seen that the documents can belong to several topics to a different extent).
- A title proposed by an expert annotator of the SEAD (bold text)
- The list of words identified as most relevant to each topic (below the title of each topic).

The list of topics is of the sliding type, so we must move with the cursor over it to visualize all the topics.



Información general del corpus

VISIÓN GENERAL

Corpus: CORDIS Num. de documentos en el corpus: 44678 Algoritmo de perfilado: LDA Num. de perfiles: 39 Fecha: 06/19/2019

CORDIS

Filtro:

Tamaño según:

Aplicar

TOPICOS DEL MODELO

PERFIL 0: 5.80%
INNOVATION, BUSINESS AND MARKET
MARKET PRODUCT SME COMPANY INDUSTRY PRODUCTION CUSTOMER INDUSTRIAL EUROPEAN COMMERCIAL

PERFIL 1: 5.61%
GENETICS AND BIOMEDICAL APPLICATIONS
CELL PROTEIN GENE REGULATE MECHANISM MOLECULE TARGET PATHWAY COMPLEX CELLULAR

PERFIL 2: 5.16%
ALGORITHMS AND MODELING
METHOD DATUM SIMULATION MODELLING TECHNIQUE ALGORITHM PREDICTION METHODOLOGY APPLY COMPLEX

PERFIL 3: 4.61%
PLATFORMS, ARCHITECTURE AND SERVICES
SERVICE DATUM PLATFORM NETWORK SUPPORT INFRASTRUCTURE ACCESS INTEGRATE COMMUNICATION ARCHITECTURE

PERFIL 4: 4.29%
PHYSICS AND IMAGING
QUANTUM LASER OPTICAL STATE MOLECULE MAGNETIC LIGHT TECHNIQUE DEVICE CONTROL

PERFIL 5: 3.80%
BIOCHEMISTRY AND MEDICAL APPLICATIONS
MOLECULE PROTEIN SYNTHESIS COMPOUND STRUCTURE ENZYME REACTION CATALYSIS COMPLEX CHEMICAL

Visualización gráfica del modelo

Lista deslizante de tópicos identificados

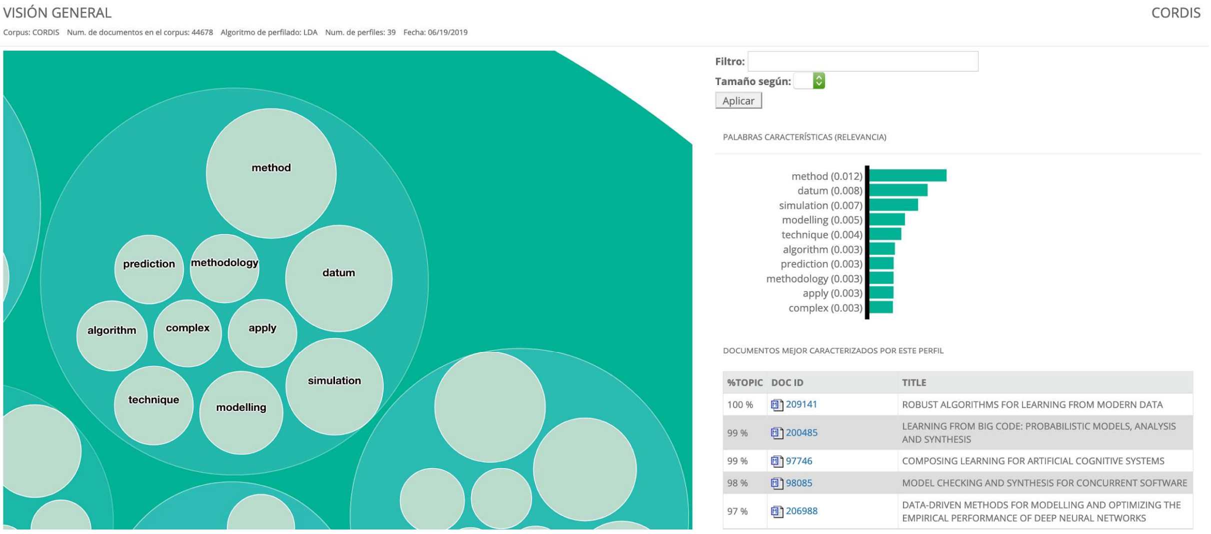
Figure 3: General View of Corpus Viewer Topics.

If we click on any of the topics (both in the graphic display and in the list of topics), the view changes to emphasize the selected topic and also shows:

- A graphic display of the most relevant words of the topic (both on the interactive ball chart, as in the histogram version)

- A list of the documents that best represent the selected profile. By clicking on the available link, we can access the text associated with the document.

By clicking on the ball chart again we can move to another profile, or return to the general model display.



VISIÓN GENERAL

CORDIS

Corpus: CORDIS Num. de documentos en el corpus: 44678 Algoritmo de perfilado: LDA Num. de perfiles: 39 Fecha: 06/19/2019

Filtro:

Tamaño según:

Aplicar

PALABRAS CARACTERÍSTICAS (RELEVANCIA)

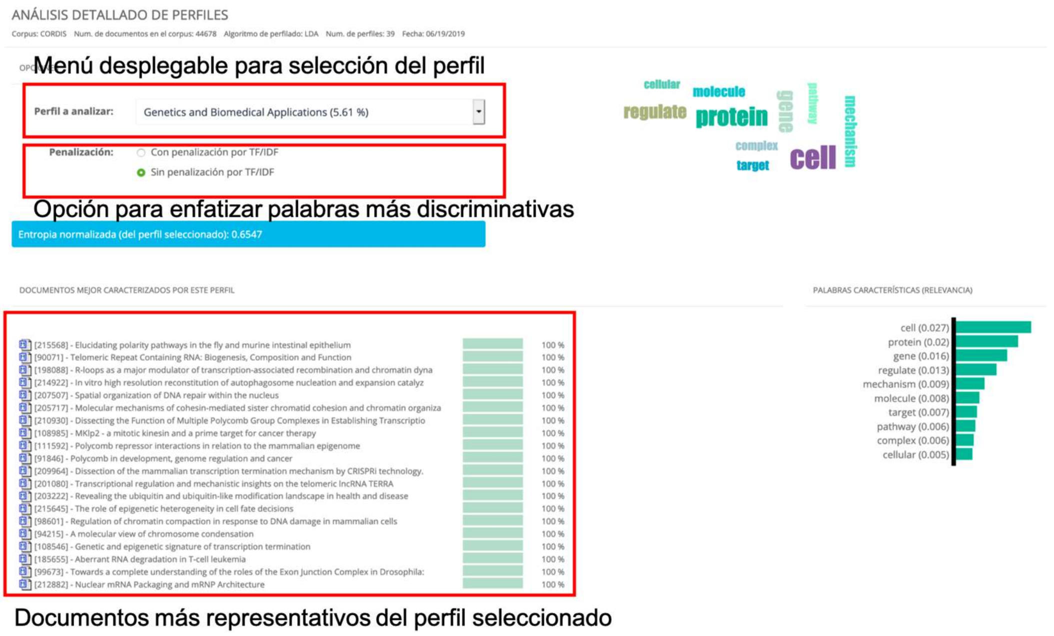
DOCUMENTOS MEJOR CARACTERIZADOS POR ESTE PERFIL

%TOPIC	DOC ID	TITLE
100 %	209141	ROBUST ALGORITHMS FOR LEARNING FROM MODERN DATA
99 %	200485	LEARNING FROM BIG CODE: PROBABILISTIC MODELS, ANALYSIS AND SYNTHESIS
99 %	97746	COMPOSING LEARNING FOR ARTIFICIAL COGNITIVE SYSTEMS
98 %	98085	MODEL CHECKING AND SYNTHESIS FOR CONCURRENT SOFTWARE
97 %	206988	DATA-DRIVEN METHODS FOR MODELLING AND OPTIMIZING THE EMPIRICAL PERFORMANCE OF DEEP NEURAL NETWORKS

Figure 4: Detailed visualization of topic including its description based on words, and the most characteristic documents of the selected topic.

6.2 TÓPICOS: TOPICS TAB

This second tab allows a visualization of the model similar to that described in the previous case, although the selection of topics is done through a drop-down menu in which the title of the topics and their relative importance in the corpus are shown.


ANÁLISIS DETALLADO DE PERFILES

Corpus: CORDIS   Num. de documentos en el corpus: 44678   Algoritmo de perfilado: LDA   Num. de perfiles: 39   Fecha: 06/19/2019

Menú desplegable para selección del perfil

Perfil a analizar: Genetics and Biomedical Applications (5.61 %)

Penalización: ○ Con penalización por TF/IDF
● Sin penalización por TF/IDF

Opción para enfatizar palabras más discriminativas

Entropia normalizada (del perfil seleccionado): 0.6547

DOCUMENTOS MEJOR CARACTERIZADOS POR ESTE PERFIL

[215568] - Elucidating polarity pathways in the fly and murine intestinal epithelium	100 %
[90071] - Telomeric Repeat Containing RNA: Biogenesis, Composition and Function	100 %
[198088] - R-loops as a major modulator of transcription-associated recombination and chromatin dyna	100 %
[214922] - In vitro high resolution reconstitution of autophagosome nucleation and expansion catalyz	100 %
[207507] - Spatial organization of DNA repair within the nucleus	100 %
[205717] - Molecular mechanisms of cohesin-mediated sister chromatid cohesion and chromatin organiza	100 %
[210930] - Dissecting the Function of Multiple Polycomb Group Complexes in Establishing Transcriptio	100 %
[108985] - MKlp2 - a mitotic kinesin and a prime target for cancer therapy	100 %
[111592] - Polycomb repressor interactions in relation to the mammalian epigenome	100 %
[91846] - Polycomb in development, genome regulation and cancer	100 %
[209964] - Dissection of the mammalian transcription termination mechanism by CRISPRi technology.	100 %
[201080] - Transcriptional regulation and mechanistic insights on the telomeric lncRNA TERRA	100 %
[203222] - Revealing the ubiquitin and ubiquitin-like modification landscape in health and disease	100 %
[215645] - The role of epigenetic heterogeneity in cell fate decisions	100 %
[98601] - Regulation of chromatin compaction in response to DNA damage in mammalian cells	100 %
[94215] - A molecular view of chromosome condensation	100 %
[108546] - Genetic and epigenetic signature of transcription termination	100 %
[185655] - Aberrant RNA degradation in T-cell leukemia	100 %
[99673] - Towards a complete understanding of the roles of the Exon Junction Complex in Drosophila:	100 %
[212882] - Nuclear mRNA Packaging and mRNP Architecture	100 %

Documentos más representativos del perfil seleccionado

PALABRAS CARACTERÍSTICAS (RELEVANCIA)

cell (0.027)
protein (0.02)
gene (0.016)
regulate (0.013)
mechanism (0.009)
molecule (0.008)
target (0.007)
pathway (0.006)
complex (0.006)
cellular (0.005)

Figure 5: Display of topics in the "tópicos" tab.

Again, for the selected topic, the most representative documents are shown, and the list of the most relevant words, both in histogram and word bag format.

This window also offers the possibility of emphasizing the most discriminative words of the topic (keywords) by selecting the option “Con penalización por TF/IDF”.

The use of TF-IDF is common in the representation of documents using bags of words. In this case, we use an extension of this concept to represent the value of the words in each topic. Being:

- TF: Term Frequency: Measures the probability of a word in a given topic.
- IDF: Inverse Document Frequency: In this context, it is an inverse factor to the importance of the term in the set of topics of the model.

In this way, if we activate the option “Con penalización por TF/IDF”, the system will reweigh the weight assigned to each word within the topic, and weight will be subtracted from those words that are common to a larger number of topics (common words with little semantic relevance). In other words, we will emphasize the most discriminative words, in the sense that words that are mostly present just in the selected topic are emphasized.

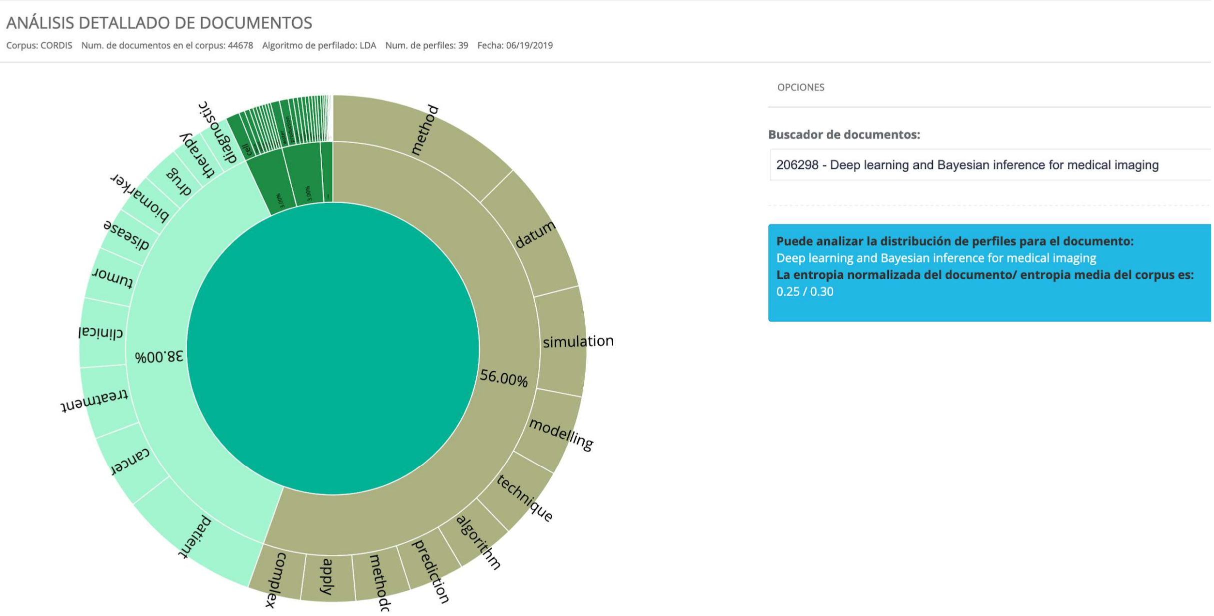
Finally, it is worth mentioning that the tab offers information on the standardized entropy of the topic, which gives an idea of the mainstreaming of the topic throughout the collection of documents.

However, the calculation of standardized entropies currently implemented offers a low dynamic range, and the SEAD technical team is developing new indicators to better characterize horizontal and vertical topics.

6.3 TÓPICOS: DOC-TÓPICOS TAB

The "Doc-Tópicos" tab allows you to search for documents by keywords. This search engine has the ability to "autocomplete", so that by entering some words, suggestions of documents containing them will be provided.

Once the document to be analyzed has been selected, a graphic visualization of its thematic content is offered. Remember that in Latent Dirichlet Allocation each document is characterized by its level of belonging to the topics of the model.



ANÁLISIS DETALLADO DE DOCUMENTOS

Corpus: CORDIS Num. de documentos en el corpus: 44678 Algoritmo de perfilado: LDA Num. de perfiles: 39 Fecha: 06/19/2019

OPCIONES

Buscador de documentos: 206298 - Deep learning and Bayesian inference for medical imaging

Puede analizar la distribución de perfiles para el documento:
Deep learning and Bayesian inference for medical imaging
La entropia normalizada del documento/ entropia media del corpus es:
0.25 / 0.30

Figure 6: Detailed analysis of documents based on the most relevant topics that characterize it.

As an example, the included figure shows that the document:

“206298 - Deep learning and Bayesian inference for medical imaging”

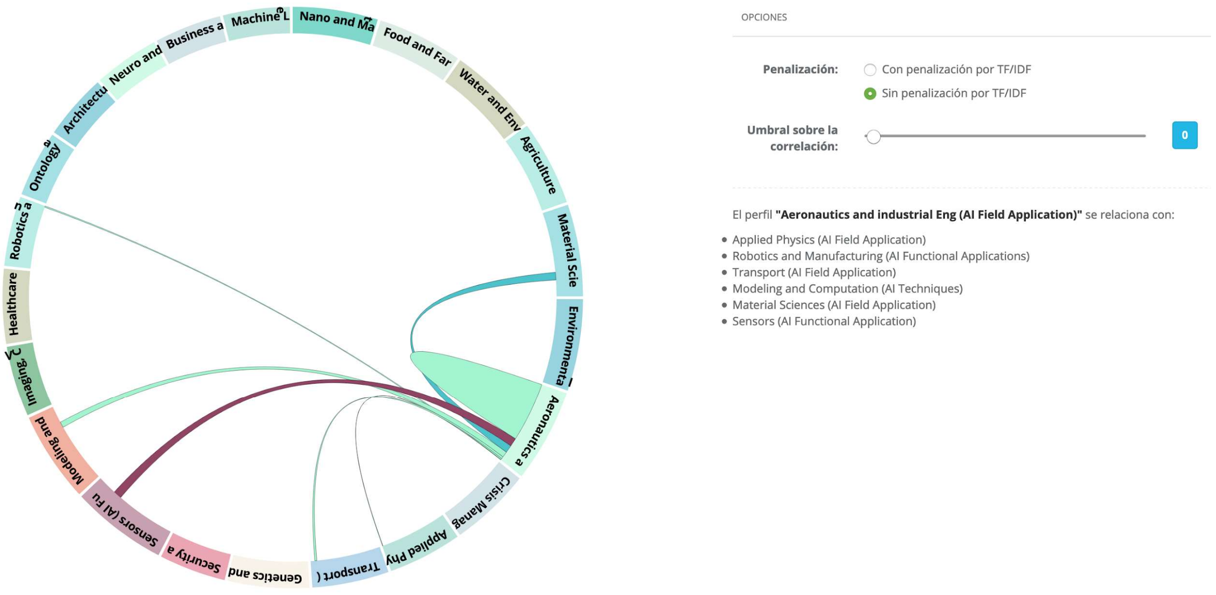
belongs in 56% to the topic characterized by the words “method, datum, simulation,…” (Algorithms and Modeling), in a 38% to the topic characterized by the words "patient, cancer, treatment, ..." (Cancer and Biomedical Applications), and to a lesser extent to other profiles.

The graphic is interactive, which allows to expand to visualize the topics of minor importance for the document by clicking on them. To return to the more general previous view, just click on the center of the circular crown.

6.4 TÓPICOS: CORRELACIÓN TAB

Lastly, the tool allows you to measure the level of correlation between topics. For this, it is estimated that the relationship between two topics is greater when these topics tend to occur together in the same documents.

Navigating on the graph on the left we can select each of the topics of the model, and the links with other topics show their level of concurrence with other topics of the model. Since the figure does not have enough space to show the full title of the profiles, this information is included in textual format on the right side of the tab. When positioning in the figure on the name of a subject, the complete title will be shown in the textual information on the right side of the page. Selecting a topic on the figure shows only the relationships with it, hiding the rest of the flows.



OPCIONES

Penalización: ○ Con penalización por TF/IDF
● Sin penalización por TF/IDF

Umbral sobre la correlación: 0

El perfil **"Aeronautics and industrial Eng (AI Field Application)"** se relaciona con:

- Applied Physics (AI Field Application)
- Robotics and Manufacturing (AI Functional Applications)
- Transport (AI Field Application)
- Modeling and Computation (AI Techniques)
- Material Sciences (AI Field Application)
- Sensors (AI Functional Application)

Figure 7: Visualization of the correlation between model topics. For each topic other topics that frequently co-occur are highlighted.

Additionally, you can select the option "Con penalización por TF/IDF" that has already been explained in the previous section, as well as, choose a higher or lower threshold for correlation, so that only those relationships that exceed the threshold will be displayed.

7. STUDY OF RELATIONS BETWEEN DOCUMENTS BASED ON THEIR TOPICS

As already mentioned, the topic modeling algorithm used allows each document to be represented based on its level of belonging to the different topics. This representation allows to measure “semantic distances” between documents. According to this distance, two documents are more similar to each other if their topic vectors are similar as well, that is, if they belong to the same topics to similar extents.

Corpus Viewer incorporates tools that allow to exploit this semantic relationship between documents. Selecting the option "**Menú -> Tópicos estáticos: Correlación**" we access two tabs that exploit this information:

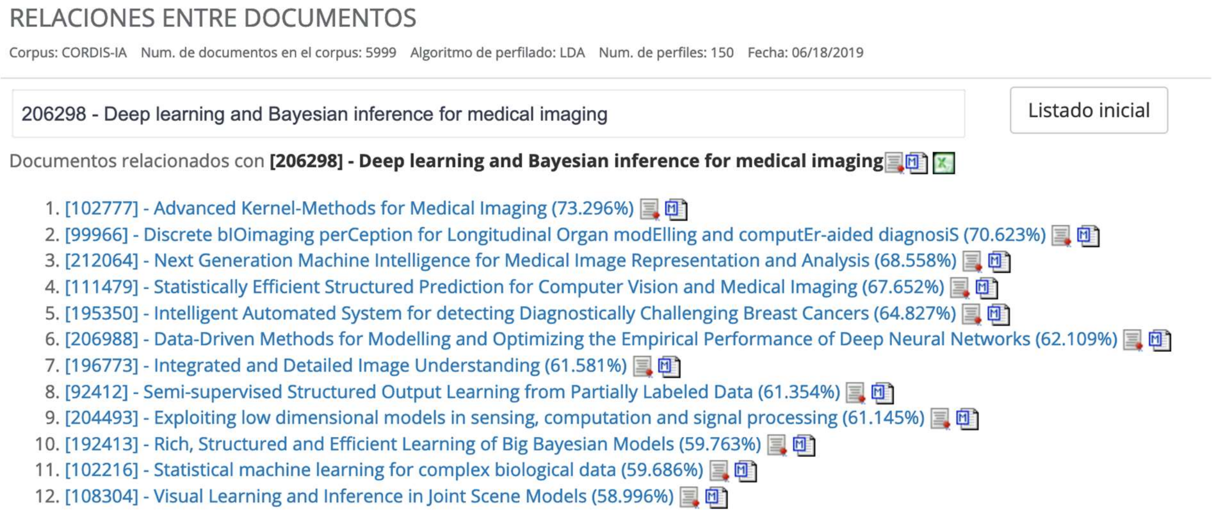
- Documentos: Document search tool by semantic similarity.
- Alarmas: Search tool for pairs of documents with very high semantic similarity.

7.1 CORRELACIÓN: DOCUMENTS TAB

The first of the available tabs offers a document search engine that allows you to select a specific document. Once selected, a list of up to 20 documents that have a high semantic relationship with the selected document is offered.

For each of the documents listed, by clicking on the different icons that appear on your right, we can:

- Check their metadata, including the title and the full text of each document.
- export the complete list of documents to excel.



RELACIONES ENTRE DOCUMENTOS

Corpus: CORDIS-IA   Num. de documentos en el corpus: 5999   Algoritmo de perfilado: LDA   Num. de perfiles: 150   Fecha: 06/18/2019

206298 - Deep learning and Bayesian inference for medical imaging

Documentos relacionados con **[206298] - Deep learning and Bayesian inference for medical imaging**

1. [102777] - Advanced Kernel-Methods for Medical Imaging (73.296%)
2. [99966] - Discrete bIOimaging perCeption for Longitudinal Organ modElling and computEr-aided diagnosiS (70.623%)
3. [212064] - Next Generation Machine Intelligence for Medical Image Representation and Analysis (68.558%)
4. [111479] - Statistically Efficient Structured Prediction for Computer Vision and Medical Imaging (67.652%)
5. [195350] - Intelligent Automated System for detecting Diagnostically Challenging Breast Cancers (64.827%)
6. [206988] - Data-Driven Methods for Modelling and Optimizing the Empirical Performance of Deep Neural Networks (62.109%)
7. [196773] - Integrated and Detailed Image Understanding (61.581%)
8. [92412] - Semi-supervised Structured Output Learning from Partially Labeled Data (61.354%)
9. [204493] - Exploiting low dimensional models in sensing, computation and signal processing (61.145%)
10. [192413] - Rich, Structured and Efficient Learning of Big Bayesian Models (59.763%)
11. [102216] - Statistical machine learning for complex biological data (59.686%)
12. [108304] - Visual Learning and Inference in Joint Scene Models (58.996%)

Figure 8: List of documents semantically similar to the document selected by the user.

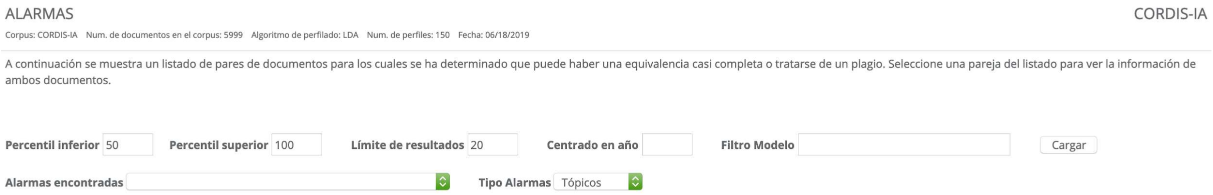
Finally, it is worth mentioning that the list allows iterative document browsing: if we click on the title of the documents in the list of similar documents, we will select that document and the tool will update the list of similar documents with those corresponding to the new document selected.

To return to the complete list, just click on the "Listado inicial" button.

7.2 CORRELACIÓN: ALARMAS TAB

This tool allows you to search for pairs of documents with very high semantic similarity. This similarity can be used to search for duplicates, or documents that have been submitted multiple times for evaluation.

It should be stressed that the tool provided is not based on a search for textual similarity (as turnitin tools, etc.), but semantic similarity. Two documents can be very similar to each other as long as they combine the same topics in similar proportions. For this reason, this search tool is very robust against the presence of synonyms, rewrites of texts, etc., because the representation of the document in the topic model remains relatively stable when the text goes through revision or minor changes.



ALARMAS

CORDIS-IA

Corpus: CORDIS-IA   Num. de documentos en el corpus: 5999   Algoritmo de perfilado: LDA   Num. de perfiles: 150   Fecha: 06/18/2019

A continuación se muestra un listado de pares de documentos para los cuales se ha determinado que puede haber una equivalencia casi completa o tratarse de un plagio. Seleccione una pareja del listado para ver la información de ambos documentos.

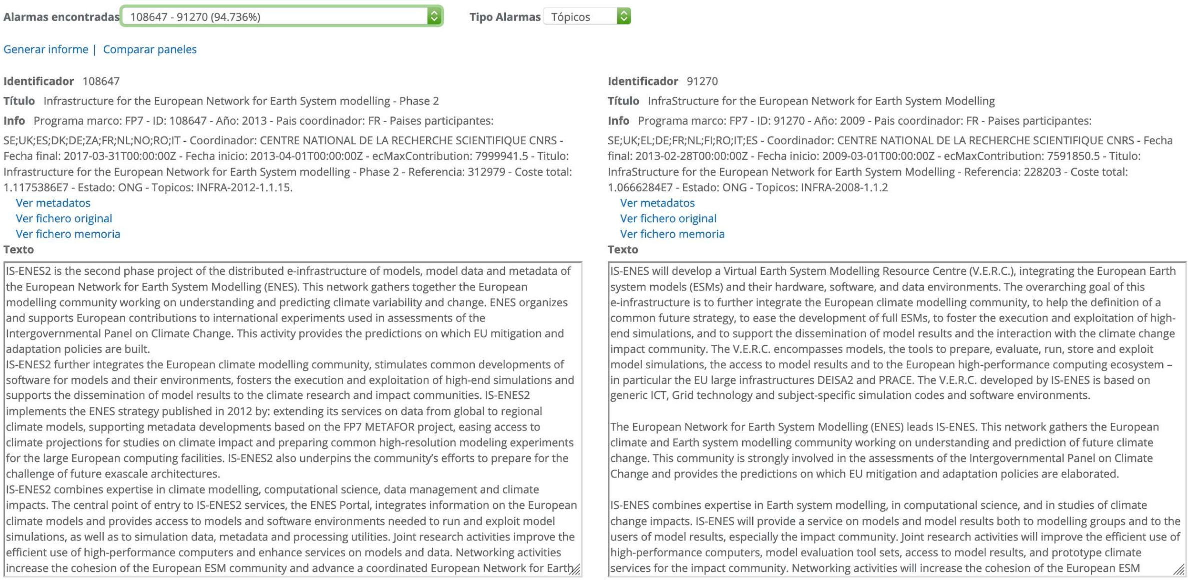
Percentil inferior 50   **Percentil superior** 100   **Límite de resultados** 20   **Centrado en año**   **Filtro Modelo**

Alarmas encontradas   **Tipo Alarmas** Tópicos

Figure 9: Options for searching for "Alarms" based on semantic similarity between documents.

The tool allows to determine the level of similarity required for the detection of alarms (lower and upper percentile), or to require that one of the two selected documents belong to a specific year (field "centered on year")[1]. Once we have established the desired settings, we have to press the "cargar" button and the tool will load the pairs of similar documents in the drop-down menu "Alarmas encontradas".

As an example, if we select the CORDIS-IA corpus and use the default parameters, the first alarm found (with a similarity of 94%) provides the view of the following figure. We can verify that these are two projects requested in years 2009 and 2013, and that they are basically a continuation of each other.



Alarmas encontradas 108647 - 91270 (94.736%)   **Tipo Alarmas** Tópicos

Identificador 108647

Título Infrastructure for the European Network for Earth System modelling - Phase 2

Info Programa marco: FP7 - ID: 108647 - Año: 2013 - Pais coordinador: FR - Paises participantes: SE;UK;ES;DK;DE;ZA;FR;NL;NO;RO;IT - Coordinador: CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE CNRS - Fecha final: 2017-03-31T00:00:00Z - Fecha inicio: 2013-04-01T00:00:00Z - ecMaxContribution: 7999941.5 - Titulo: Infrastructure for the European Network for Earth System modelling - Phase 2 - Referencia: 312979 - Coste total: 1.1175386E7 - Estado: ONG - Topicos: INFRA-2012-1.1.15.

Texto

IS-ENES2 is the second phase project of the distributed e-infrastructure of models, model data and metadata of the European Network for Earth System Modelling (ENES). This network gathers together the European modelling community working on understanding and predicting climate variability and change. ENES organizes and supports European contributions to international experiments used in assessments of the Intergovernmental Panel on Climate Change. This activity provides the predictions on which EU mitigation and adaptation policies are built.
IS-ENES2 further integrates the European climate modelling community, stimulates common developments of software for models and their environments, fosters the execution and exploitation of high-end simulations and supports the dissemination of model results to the climate research and impact communities. IS-ENES2 implements the ENES strategy published in 2012 by: extending its services on data from global to regional climate models, supporting metadata developments based on the FP7 METAFOR project, easing access to climate projections for studies on climate impact and preparing common high-resolution modeling experiments for the large European computing facilities. IS-ENES2 also underpins the community's efforts to prepare for the challenge of future exascale architectures.
IS-ENES2 combines expertise in climate modelling, computational science, data management and climate impacts. The central point of entry to IS-ENES2 services, the ENES Portal, integrates information on the European climate models and provides access to models and software environments needed to run and exploit model simulations, as well as to simulation data, metadata and processing utilities. Joint research activities improve the efficient use of high-performance computers and enhance services on models and data. Networking activities increase the cohesion of the European ESM community and advance a coordinated European Network for Earth

Identificador 91270

Título InfraStructure for the European Network for Earth System Modelling

Info Programa marco: FP7 - ID: 91270 - Año: 2009 - Pais coordinador: FR - Paises participantes: SE;UK;EL;DE;FR;NL;FI;RO;IT;ES - Coordinador: CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE CNRS - Fecha final: 2013-02-28T00:00:00Z - Fecha inicio: 2009-03-01T00:00:00Z - ecMaxContribution: 7591850.5 - Titulo: InfraStructure for the European Network for Earth System Modelling - Referencia: 228203 - Coste total: 1.0666284E7 - Estado: ONG - Topicos: INFRA-2008-1.1.2

Texto

IS-ENES will develop a Virtual Earth System Modelling Resource Centre (V.E.R.C.), integrating the European Earth system models (ESMs) and their hardware, software, and data environments. The overarching goal of this e-infrastructure is to further integrate the European climate modelling community, to help the definition of a common future strategy, to ease the development of full ESMs, to foster the execution and exploitation of high-end simulations, and to support the dissemination of model results and the interaction with the climate change impact community. The V.E.R.C. encompasses models, the tools to prepare, evaluate, run, store and exploit model simulations, the access to model results and to the European high-performance computing ecosystem – in particular the EU large infrastructures DEISA2 and PRACE. The V.E.R.C. developed by IS-ENES is based on generic ICT, Grid technology and subject-specific simulation codes and software environments.

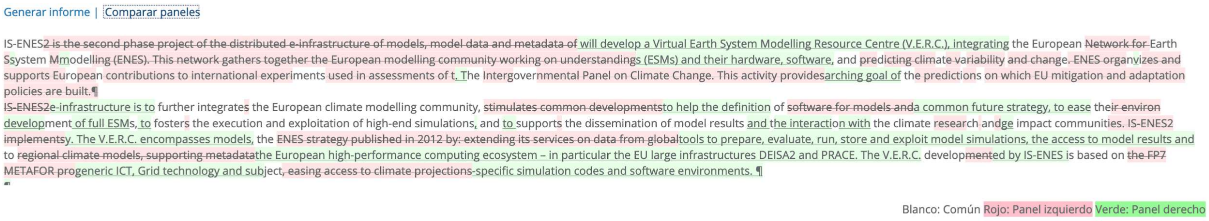
The European Network for Earth System Modelling (ENES) leads IS-ENES. This network gathers the European climate and Earth system modelling community working on understanding and prediction of future climate change. This community is strongly involved in the assessments of the Intergovernmental Panel on Climate Change and provides the predictions on which EU mitigation and adaptation policies are elaborated.

IS-ENES combines expertise in Earth system modelling, in computational science, and in studies of climate change impacts. IS-ENES will provide a service on models and model results both to modelling groups and to the users of model results, especially the impact community. Joint research activities will improve the efficient use of high-performance computers, model evaluation tool sets, access to model results, and prototype climate services for the impact community. Networking activities will increase the cohesion of the European ESM

Figure 10: "Alarms" found by the application, and parallel view of two documents identified as (semantically) very similar.

[1] *Sometimes it is interesting to decrease the upper percentile to a value less than 100% or to focus the analysis on a specific year. This can be important especially in those cases in which the documents have been subject to an OCR process (this is the case of the ACL corpus), since in certain cases there may be thematically identical documents because they are associated with the presence of noisy characters that come from a malfunction of character recognition.*

If we click on the “comparar paneles” option, we can see how the **textual similarity** of both projects is relatively low, although a high semantic similarity has been detected. Regarding textual similarity, the sentences marked in red (green) appear only in the text of the document in the left (right) panel, while the white text is the one that appears simultaneously in both documents. This example clearly illustrates the difference between this tool based on semantic similarity versus other tools based on textual similarity.



IS-ENES~~2 is the second phase project of the distributed e-infrastructure of models, model data and metadata of~~ will develop a Virtual Earth System Modelling Resource Centre (V.E.R.C.), integrating the European ~~Network for~~ Earth ~~S~~system ~~M~~modelling ~~(ENES). This network gathers together the European modelling community working on understanding~~s (ESMs) and their hardware, software, and ~~predicting climate variability and change. ENES organ~~v~~izes and supports European~~ ~~contributions to international experi~~ments ~~used in assessments of t~~. The ~~Interg~~over~~nmental Panel on Climate Change. This activity provides~~arching goal of the ~~predictions on which EU mitigation and adaptation policies are built.¶~~

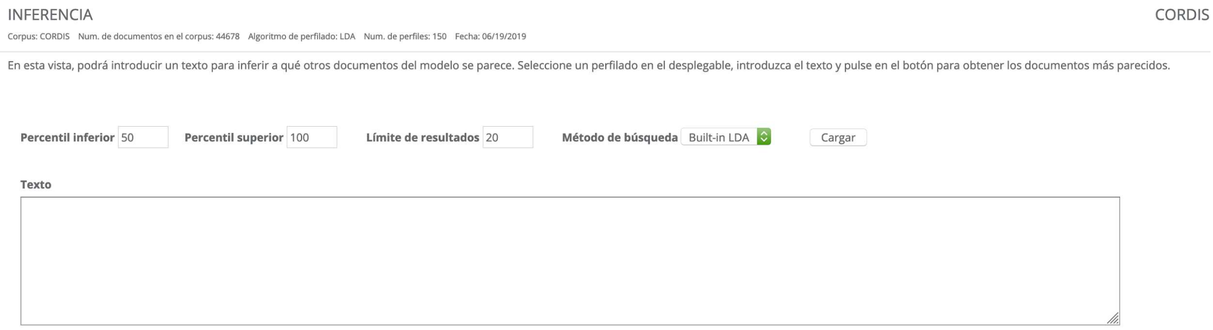
~~IS-ENES2~~e-infrastructure is to further integrate~~s~~ the European climate modelling community, ~~stimulates common developments~~to help the definition of ~~software for models and~~a common future strategy, to ease thei~~r environ~~ development of full ESMs, to foster~~s~~ the execution and exploitation of high-end simulations, and to support~~s~~ the dissemination of model results and the interaction with the climate ~~resear~~ch ~~an~~dge impact communiti~~es. IS-ENES2 implements~~y. The V.E.R.C. encompasses models, the ~~ENES strategy published in 2012 by: extending its services on data from global~~tools to prepare, evaluate, run, store and exploit model simulations, the access to model results and to ~~regional climate models, supporting metadata~~the European high-performance computing ecosystem – in particular the EU large infrastructures DEISA2 and PRACE. The V.E.R.C. develop~~ment~~ed by IS-ENES is based on ~~the FP7 METAFOR pro~~generic ICT, Grid technology and subject~~, easing access to climate projections~~-specific simulation codes and software environments. ¶

Blanco: Común 🟥 Rojo: Panel izquierdo 🟩 Verde: Panel derecho

Figure 11: Textual comparison panel between pairs of documents with large semantic similarity.

8. DOCUMENTS SIMILAR TO AN ARBITRARY TEXT

All the functionalities described in the previous section allow to exploit semantic similarities, but their use is restricted to those documents that belong to the collections of documents already loaded in Corpus Viewer. Sometimes it can be interesting to look for similarities with other new texts provided by the user. This is possible by selecting the option "**Menú -> Tópicos estáticos: Inferencia**" in the main menu of Corpus Viewer.



INFERENCIA

CORDIS

Corpus: CORDIS Num. de documentos en el corpus: 44678 Algoritmo de perfilado: LDA Num. de perfiles: 150 Fecha: 06/19/2019

En esta vista, podrá introducir un texto para inferir a qué otros documentos del modelo se parece. Seleccione un perfilado en el desplegable, introduzca el texto y pulse en el botón para obtener los documentos más parecidos.

Percentil inferior: 50

Percentil superior: 100

Límite de resultados: 20

Método de búsqueda: Built-in LDA

Figure 12: Tab for thematic inference about free text provided by the user, and search for documents with a similar theme indexed in Corpus Viewer.

The Inference tool is based on the following steps:

1. The text provided is preprocessed using the same tools that were used for preprocessing the documents of the active corpus.
2. The text provided is “projected” on the topic model associated with the active corpus. In this way, we obtain a representation based on topics similar to that available for all the corpus documents loaded in Corpus Viewer.
3. The semantic similarity between the text provided and each of the documents of the selected corpus is calculated, and the most similar documents are shown to the user.

It is worth mentioning that this tool requires the execution of certain calculations on Corpus Viewer servers, so the response time may be a few seconds (larger when the number of documents in the selected corpus is also very large).

It is also necessary to highlight that the semantic representation of the text will have better quality the longer the length of the text provided. Therefore we can expect higher quality results the longer the query text.

9. SEARCH TOOL

selecting **"Menú -> Buscador**" you can access the last of the options currently active in Corpus Viewer, which consists of a tool based on Solr and Banana. This tool offers the functionality of a BI type tool, although it integrates the available metadata with the document-based representation of topics.

Currently, the search engine is in the development phase, so all the information that will be available in the final version is not incorporated, and changes in the panels that are finally incorporated in each corpus are expected.

Although the development is not finished, it has been decided to leave this tab active in the open instance of Corpus Viewer, so that users can get a first impression of the type of functionality that will be provided once the development is completed.

You can check the demo version with data from CORDIS for Artificial Intelligence developed on Javascript (takes a while to load).

The operation of the search engine that will be incorporated into Corpus Viewer will be similar to that of the demonstrator provided, and will include all the search and grouping power provided by Solr indexing technology.