«COVID-19 en español: Investigación interdisciplinar sobre terminología, temáticas y comunicación de la ciencia»

Proyecto intramural CSIC 202010E241. COVID19 en español: investigación interdisciplinar sobre terminología, temáticas y comunicación de la ciencia

terminología · ontologías · argumentación · supercomputación · información científica · estructuras editoriales

César González-Pérez, Jose Ignacio Vidal-Liy, Pablo Calleja, Fernando Aguilar, Elea Giménez-Toledo

PTI ES_CIENCIA (ILIA-CSIC, Incipit CSIC, IFCA, USAL, OEG-UPM, ÍnDICEs-CSIC, IH-CCHS)

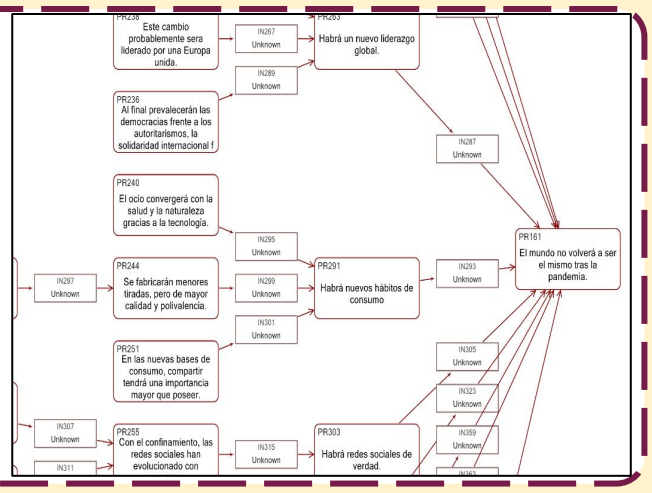
https://pti-esciencia.csic.es/

Análisis terminológicos

- Extracción de términos del corpus mediante las herramientas LogosLink y Sketch Engine y análisis de su valor terminológico
- Análisis cuantitativo y cualitativo de las principales unidades terminológicas: monoléxicas (ej. covid-19, coronavirus) y poliléxicas (ej. estado de alarma, crisis sanitaria)
- Estudio de las relaciones y combinaciones entre las unidades terminológicas y su evolución temporal
- Mapeo de términos en función de las áreas temáticas para la creación de una ontología que establezca las interrelaciones entre las distintas disciplinas científicas, en particular de las Ciencias Sociales y Humanas

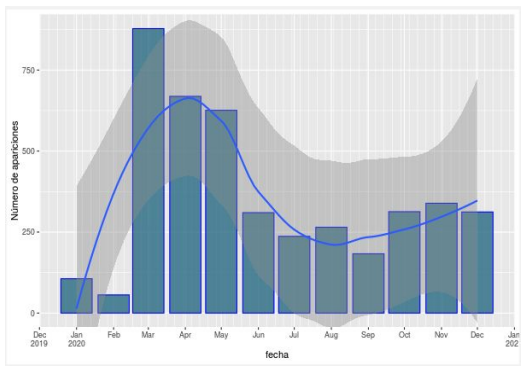
Análisis de argumentación

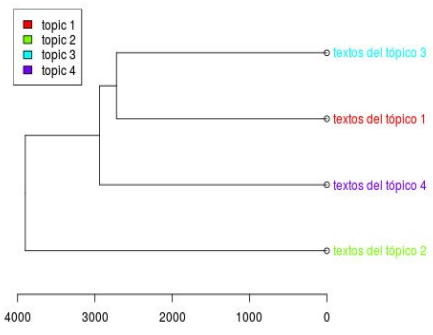
- Ejemplo sobre el artículo “COVID-19: nueve razones por las que el mundo no volverá a ser el mismo”
- El diagrama muestra proposiciones individuales conectados por relaciones de inferencia
- Facilita la comprensión de la argumentación y la evaluación de la fortaleza argumental.
- Metodología IAT/ML (www.iatml.org), desarrollada en el Incipit CSIC



Clasificación de textos por temas emergentes y *clusters*

- Agrupación no supervisada de documentos por palabras y sus contextos
- Herramienta para carga de corpus y visualización de diagramas
- Visualización temporal del uso de términos y diagramas para la visualización del modelado de tópicos





Corpus en español sobre COVID-19 de *The Conversation* (2020)
877 documentos
907 autores/as

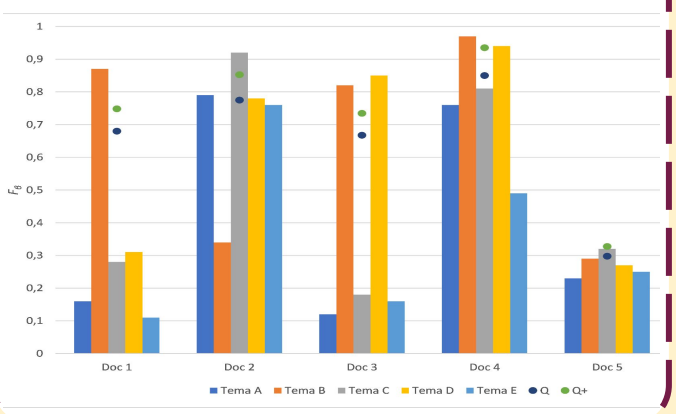
Análisis del modelo The Conversation para divulgar ciencia

- La columna científica: escrita por científicos, cuidada por periodistas y leída por todos los públicos.
- Pluralidad de autores, temas y enfoques
- Lenguaje cercano, atractivo y comprensible
- Ciencia muy leída: más de 5000 lecturas de media y con proyección a muchos ámbitos por las republicaciones.
- Narradores y narratarios. El análisis de las columnas científicas muestra que los destinatarios son lectores cultos e implicados

En desarrollo

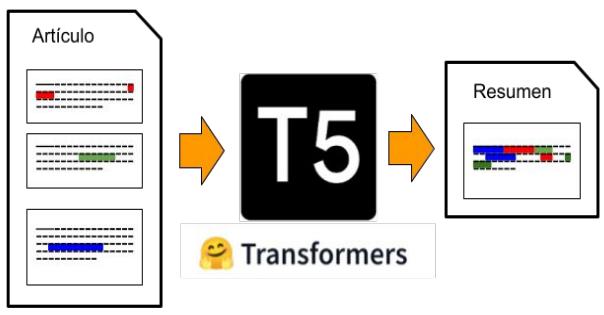
Clasificación temática automática

- Creación de un algoritmo para la clasificación temática automática
- Basado en el análisis terminológico de los documentos con vocabularios específicos de cada tema y frecuencias de formas léxicas en habla común
- Mejora los resultados respecto a modelos de lenguaje en español (MarIA)



Generación de resúmenes multilingües

- Utilizando el modelo de lenguaje multilingüe T5 (Google)
- Generación de un corpus de entrenamiento de pares de artículos en español (documento) y su *abstract* en inglés (resumen).
- Entrenamiento en servidores GPU



Corpus e infraestructuras en abierto

- Base de datos COVID-19 para la identificación de autores, afiliaciones y temas.

EUROPEAN OPEN SCIENCE CLOUD

OPERAs

Triple

Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration

GOBIERNO DE ESPAÑA · MINISTERIO DE CIENCIA E INNOVACIÓN · PTI ES-CIENCIA · CSIC Consejo Superior de Investigaciones Científicas

Agradecimientos
CSIC Interdisciplinary Thematic Platform (PTI). El español como lengua de comunicación científica. ES CIENCIA.

CLARIN CENTRE K · INTELE · DARIAH-EU · Infraestructura de Tecnologías del Lenguaje