Revista de Humanidades Digitales

Grupo TALN, Instituto IRIF, Fundación EGPV. TecSemHu: Tecnologías Semánticas para las Humanidades Digitales.

http://edusearch.cat

Reseña realizada por:
Anna VILLAR COLELL
Universidad Nacional de Educación a Distancia
annavillarc@gmail.com

La cantidad de información sobre cualquier materia a la que los estudiantes de hoy tienen acceso es ingente. Si bien se puede celebrar el fácil acceso a todas las fuentes y a tipos distintos de informaciones, también es para ellos una ardua tarea saber identificar dónde está la aguja en ese inmenso pajar que es Internet. Las materias humanísticas que cualquier alumno de secundaria, bachillerato o universidad esté cursando están representadas en la red en multiplicidad de formatos, idiomas y estructuras. Tal cantidad de información termina por abrumar, entorpeciendo así la tarea y el objetivo. No existe guerra, obra de arte o poeta del que no se pueda indagar en webs o redes sociales para hacer un trabajo académico, la dificultad está en el manejo de esta infinidad casi intimidatoria de datos.

Este problema no pasó inadvertido al investigador Toni Badia del grupo de investigación Tratamiento Automático de Lenguaje Natural (TALN)[1] de la Universidad Pompeu Fabra de Barcelona, investigador principal de TecSemHu. El programa nace con la voluntad de facilitar el estudio y la investigación en el campo de las humanidades mediante la aplicación de las tecnologías de análisis semántico a distintos documentos textuales y audiovisuales. La idea madre, sin embargo, la tenemos que ir a buscar en un proyecto europeo previo, coordinado por el grupo de la UPF, el Event Understanding Through Multimodal Social Stream Interpretation (EUMSSI)[2], en el que se desarrollaron las técnicas que se usan en TecSemHu. Estas se basan en la aplicación de tecnologías de análisis textual, de transcripción automática del habla y de reconocimiento facial y de voz en audio y vídeo a contenidos multimedia a gran escala para integrarlo todo en una misma plataforma. Los metadatos generados se retroalimentan y hacen posible la obtención de información valiosa y relacionada de distintas fuentes desde este único soporte. El público al que se dirigía era el colectivo de periodistas y editores profesionales de

RHD 7 (2022)
ISSN
2531-1786

Recibido 25/04/2021 – Aceptado 04/08/2022

distintos medios.

El grupo de la UPF vio las potencialidades de este trabajo y sus aplicaciones más allá del ámbito periodístico. Las tecnologías semánticas pueden ser aplicadas también al campo educativo en el cual destacan distintos ámbitos humanísticos como las ciencias sociales. De aquí nace TecSemHu. Al grupo investigador se le unieron dos colaboradores más, el Instituto de Recursos i Investigación para la Formación SL (IRIF)[3], una empresa privada centrada en la innovación para la formación, y la Fundación de escuelas Garbí Pere Vergés (EGPV)[4], propietaria de dos escuelas punteras en innovación pedagógica. El equipo presenta su propuesta el año 2018 al programa de becas RecerCaixa 2018[5] y de ahí surge su financiación. Inicialmente se plantea como un proyecto de dos años, de 2018 a 2020. Sin embargo las circunstancias derivadas de la pandemia del COVID-19 hacen que se alargue hasta marzo de 2021.

El proyecto crea el portal edusearch.cat, un entorno destinado a hacer trabajos e investigaciones en ciencias sociales en grupos de Secundaria (el piloto se llevó a cabo en alumnos de 3º de la ESO). La interfaz[6] consta, por un lado, de una ventana con un editor de texto que incluye una herramienta de procesamiento del texto para generar contenido relacionado y, por otro, una segunda ventana en la que se nos muestra este contenido en su multiplicidad de formatos. ¿En qué ayuda esta herramienta a los alumnos y con qué base se ha implementado? Bien, la respuesta a ambas cuestiones tiene que ver con las tecnologías de análisis semántico. Al pasar el analizador de texto, éste opera sobre conceptos destacados, no sólo palabras, de modo que el estudiante recibe datos valiosos que pueden dar lugar a relaciones inesperadas y potenciar, así, el conocimiento sobre el tema en cuestión.

Para detallar más, en la segunda ventana aparecen, principalmente, páginas de la Viquipèdia (Wikipedia en catalán) y vídeos educativos, ordenados y con información relevante. También desde el mismo lugar se puede generar una visualización en nubes de palabras que se puede insertar en el editor. Se obtienen, además, mapas, que permiten filtrar por países, y personajes destacados. Por lo que respecta a los videos, encontramos que algunos están segmentados, es decir, empiezan y se dirigen directamente a la parte importante o de interés para el estudiante, ahorrándole así tener que hacer un visionado de todo el contenido. Por último, está también la transcripción del audio correspondiente.

Hay un trabajo concreto detrás de todo esto que no se puede ignorar. El equipo científico del proyecto desarrolla una tarea que hasta ahora estaba sin hacer, la creación de una DBpedia[7] en catalán. Se trata de un proyecto que consiste, a grandes rasgos, en la extracción de datos de Viquipèdia para analizarlos y así poder ser interpretados por una máquina. Es gracias a ello que

[1] Para más información de TALN: https://www.upf.edu/web/taln.
[2] Página web del proyecto: https://www.eumssi.eu/.
[3] Página web de IRIF: https://www.grao.com/es/sobre-grao.
[4] Página web de EGPV: https://www.escolesgarbi.cat/.
[5] Para más información: https://fundacionlacaixa.org/ca/investigacion-salud/investigacion-biomedica/recercaixa/que-es-recercaixa.
[6] Interfaz del proyecto: http://demo.edusearch.cat/demo/.
[7] Proyecto DBpedia: https://www.dbpedia.org/.

se pueden establecer relaciones entre entidades, personas y lugares y, asimismo, se pueden hacer conexiones entre distintos idiomas. El entorno eduserach.cat se nutre de todo esto para facilitar la misión de elaborar una investigación académica en el ámbito de las ciencias sociales. Cada una de las entradas de la Viquipèdia que genera el analizador de texto está enriquecida con esas personas, lugares y entidades relacionadas que mencionábamos.

¿Facilita la herramienta el trabajo al alumno? La respuesta es un sí, pero con matices. El entorno tiene un potencial innegable, pero también algunas carencias. Uno de los principales retos para el futuro sería poder abastecer el repositorio con muchos más datos analizados, ya que se echa en falta más contenido multimedia e información más allá de la Wikipedia. Quizás sea este un primer paso para ir más allá, obtener licencias, reunir más fuentes y engrosar el instrumento. Se hablaba al inicio de la reseña de la dificultad que tienen un alumno o un investigador para lidiar con tanta cantidad de información y fuentes y es precisamente este problema el que el proyecto pretende resolver. Si bien es cierto que disponer de todos los documentos enriquecidos, analizados y ordenados supone una descarga de trabajo y tiempo sumamente valiosa y nos puede incluso dirigir hacia caminos fructíferos desconocidos, es necesario disponer de un criterio mínimo, estar acostumbrado a tareas de discriminación de información, características que cualquier adulto mínimamente formado posee. Quizás con los alumnos de 15 años sea necesario un trabajo previo al uso de la herramienta. Este es, precisamente, uno de los problemas que surgió en la aplicación del piloto. Los profesores de los grupos que trabajaron en él coincidieron en decir que los estudiantes de uno o dos cursos más son más autónomos y pueden aprovechar mejor las posibilidades del instrumento.

Para recapitular, pese a que edusearch.cat podría tener más contenido analizado, permite a investigadores o estudiantes en humanidades unas opciones que hasta ahora no existían en lengua catalana. Pueden acceder a una búsqueda enriquecida para la obtención de información, descubrir conceptos significativos relacionados, aplicar filtros y ver transcripciones. Aun así, aunque los datos proporcionados son relevantes y estructurados y están conectados, sigue siendo una cantidad notable de información. Para usar la herramienta con efectividad hay que poseer una mínima competencia para poder discernir dentro de lo ya destacado y enlazado qué es lo que requiere nuestro propio trabajo o investigación.