INVESTIGADORES ESPAÑOLES REVISAN LA VANGUARDIA DE LAS TECNOLOGÍAS EN MINERÍA DE TEXTOS PARA QUÍMICA

Buena parte de los datos relevantes para el cáncer sólo está disponible de forma no estructurada

En un artículo publicado en *Chemical Reviews*, la Unidad de Minería de Textos en Biología del Centro Nacional de Investigaciones Oncológicas (CNIO), en coordinación con investigadores del Centro de Investigación Médica Aplicada (CIMA) de la Universidad de Navarra y el Barcelona Supercomputing Centre (BSC-CNS), ha publicado la primera revisión exhaustiva sobre las metodologías de vanguardia que impulsan los motores de búsqueda de compuestos químicos, denominados sistemas de reconocimiento de entidades y minería de textos. **Alfonso Valencia**, ex investigador del CNIO y actualmente en el BSC como Director del Departamento de Ciencias de la Vida, fue el investigador principal que supervisó el trabajo. El BSC está particularmente interesado en explorar y desarrollar la aplicación de técnicas de HPC y de *Machine Learning* para la extracción de información en áreas de la química y la biomedicina.

El creciente campo de las aplicaciones de Big Data en la investigación biomédica, junto con el uso del aprendizaje automático y las tecnologías de inteligencia artificial para la minería de textos, ha dado lugar a numerosas herramientas prometedoras. "Esta revisión –señalan los autores– pretende ser una guía práctica para que los investigadores se adentren en el mundo de los datos científicos y también para ayudarles a prever los próximos pasos en este emergente campo".

"A través del lanzamiento de los Gold Standard datasets y de la organización de varios eventos de desafío comunitario, la Unidad de Minería de Textos en Biología ha desempeñado un papel crítico en el desarrollo y evaluación de los sistemas actuales de minería de textos en química", explica Martin Krallinger, jefe de la Unidad y primer autor de la revisión.

UNA GRAN CANTIDAD DE DATOS NO ESTRUCTURADOS

Buena parte de los datos relevantes para el cáncer sólo está disponible de forma no estructurada. Este tipo de datos incluye la literatura científica, las patentes de compuestos de uso médico, registros electrónicos sanitarios o documentos de ensayos clínicos. De hecho, cada año, más de 20.000 nuevos compuestos aparecen en las revistas científicas.

Transformar esta información no estructurada en bases de datos que puedan ser procesadas de forma más eficiente por los ordenadores o consultadas por la gente es crucial para cosas como la identificación de nuevas dianas farmacológicas y de efectos secundarios o encontrar nuevos usos para fármacos ya aprobados. Los compuestos químicos y los fármacos son elementos centrales para la investigación biomédica.

Esta revisión proporciona una descripción completa y detallada de los conceptos fundamentales, aplicaciones técnicas y tecnologías actuales para satisfacer estas demandas de información.

TRABAJO DE REFERENCIA:

Krallinger M, Rabal O, Lourenço A, Oyarzabal, Valencia A (2017). Information Retrieval and Text Mining Technologies for Chemistry. Chemical Reviews