Centro de Información de COVID (CIC): Charlas científicas de relámpago

Transcripción de una presentación de Carlos Badenes-Olmedo (Universidad Politécnica de Madrid), 26 de Julio de 2023.

Título: Drugs4Covid: Gráfico de Conocimiento sobre Fármacos utilizados en el Control Clínico del Coronavirus

Publicación NIH: Lessons learned to enable question answering on knowledge graphs extracted from scientific publications: A case study on the coronavirus literature

Grabación de YouTube con hojas dispositivas

Verano 2023 CIC Información de webinario

Transcripción Editada: Karem Coca

Transcripción

Hoja 1

Carlos Badenes-Olmedo:

Ok, gracias, Lauren. Y gracias a ti, Florence por invitarme. Ok, voy a compartir mi pantalla. Gracias por invitarme a participar en este webinar. Soy Carlos Badenos-Olmedo, investigador del Grupo de Ingeniería Ontológica y también soy profit profesor ayudante en la Universidad Politécnica de Madrid. Voy a presentar el proyecto Drugs4COVID.

Este es un proyecto que proponemos para crear un grafo de conocimiento. Se trata de representar los fármacos que utilizamos durante la pandemia para definir el control clínico del coronavirus.

Hoja 2

La idea es que durante la pandemia de COVID-19, instituciones de todo el mundo intentaron identificar o crear conjuntos de datos con artículos de investigación científica relacionados con el coronavirus. Este tipo de información podría ser útil para las farmacias de los hospitales, para los médicos clínicos. Nuestro grupo de investigación, el Ontology Engineering Group, intentaba promover cómo podemos recuperar información, recuperar conocimiento a partir de este tipo de datos. El primer paso fue identificar cuál es el conjunto de datos más importante que podemos utilizar. En este caso, la Unión Europea proporcionó el portal europeo de datos COVID-19, así como el conjunto de datos COVID [inaudito]. Principalmente, el conjunto de datos COVID-19 con ajustes preestablecidos. Genera los datos más importante que combina información de PubMed BioRxiv, MedRxiv e información de arXiv. También combina información de la

Organización Mundial de la Salud y proporciona más de 400.000 artículos que necesitamos, con la meta que podemos aprovechar para aportar conocimiento. Combinando este tipo de conocimiento con el autoservicio de Madrid, el [inaudito]. Proporcionamos los mecanismos para crear grafos de conocimiento que podemos explotar y proporcionar conocimiento a partir de este tipo de proceso de información.

Hoja 3

Así, en primer lugar, definimos un flujo de trabajo con diferentes pasos que proporcionan no sólo el paso, sino también las recomendaciones los pasos que puede seguir para crear un gráfico de análisis final y también facilitar la explotación. Definimos un proceso de trabajo de seis pasos. El primer paso es la recolección. En este paso la idea es que usted necesita para identificar el conjunto de datos que está relacionado con coronavirus, y también para evaluar si los datos están completamente disponibles o no. El segundo paso es el pre procesamiento, ya que es necesario organizar los datos que se proporcionan, en este caso un conjunto de datos, ya que estos datos no son totales porque estamos trabajando con pruebas. Tenemos que modelizar una determinada forma de [evaluar] los datos. A continuación, pasamos a la tercera etapa, que es la extracción de información. En este paso tenemos que descubrir los elementos principales de este tipo de datos que podemos utilizar para crear el gráfico de análisis. En nuestro caso, los elementos principales son los conceptos biomédicos (por ejemplo, los fármacos, las enfermedades y la información genética). A continuación, tenemos que definir una descripción formal del dominio, que es el cuarto paso, que es la semantificación. En este caso, necesitamos crear una ontología para definir las relaciones entre los conceptos biomédicos y definir todos los conceptos, todos los elementos, que finalmente aparecen en el grafo de conocimiento. El siguiente paso es la generación del grafo de conocimiento. En este paso necesitamos definir las reglas para crear las instancias en el grafo de conocimiento. Y finalmente podemos proporcionar los mecanismos para explotar - para facilitar el uso de la información que contiene el grafo de conocimiento.

Hoja 4

Así que nos enfocamos en el primer paso. El objetivo es identificar las fuentes de datos relevantes y también evaluar la disponibilidad de los datos. Nuestra propuesta es que hay que realizar una revisión sistemática de la literatura, teniendo en cuenta los principales conceptos del coronavirus, y luego definir el conjunto de datos. Por ejemplo, a partir de repositorios digitales, PubMed, BioRxiv, etc., pero también combinando con otras fuentes, por ejemplo, colecciones clínicas del corpus de coordinación, el conjunto de datos principal COVID y también recursos adicionales. Por ejemplo, patentes, artículos enciclopédicos de Wikipedia. Y todos estos datos se organizan en este primer paso.

Hoja 5

En el siguiente paso, tenemos que transferir estos datos no estructurados, que son texto, a tablas, que son datos estructurados. Entonces, la metodología que proponemos es identificar la información mínima que necesitas. La forma más fácil de transformar un texto en estructurado es definir como datos el texto completo del artículo. En nuestra opinión, esta no es la mejor manera

de hacerlo y nuestra propuesta es definir la información mínima que necesites: el párrafo de los artículos. En esa zona puedes descubrir todas las referencias o la relación entre los conceptos biomédicos.

Hoja 6

A continuación, el siguiente paso es la estructura informativa. En este caso, la idea es crear anotaciones basadas en esos párrafos que descubran fármacos, enfermedades e información genética. En nuestra experiencia, ponemos a punto distintos modelos lingüísticos para cada concepto biomédico. La idea es que hay que definir un modelo lingüístico específico para identificar fármacos y también normalizar los fármacos según los distintos vocabularios, porque en los distintos países utilizamos códigos estándar diferentes.

Hoja 7

Una vez que tenemos las anotaciones con las entidades y también los códigos podemos definir el espacio formal para describir toda esta información. Este es el paso que necesitamos para crear una ontología. En el ámbito biomédico existen muchas ontologías, así que la idea no es crear una ontología desde cero. La idea es reducir las ontologías del sistema y proporcionar la información que falta en la nueva ontología. En nuestro caso, la ontología era EBOCA y la información que faltaba era proporcionar las pruebas que sustentan las relaciones entre fármacos, entre enfermedades y entre información genética. En nuestro caso utilizamos el sistema de lenguaje médico unificado y también la plataforma DISNET. Toda esta información está combinada. También proporcionamos, esta es la zona morada, la información sobre las pruebas. ¿Qué es la evidencia? La evidencia es el párrafo en el que se informa de la relación entre estos elementos en el artículo científico.

Hoja 8

Una vez que hemos definido el dominio formal, la ontología, necesitamos identificar las instancias, las afirmaciones, las declaraciones recuperadas de los artículos científicos para crear instancias en el grafo de conocimiento. Este es el paso de generación del grafo de conocimiento por lo que nuestra metodología se propone crear reglas para identificar utilizando el lenguaje del modelo anterior las entidades y también las relaciones entre ellas. Finalmente, una vez tenemos la ontología, tenemos las instancias, somos capaces de identificar los nodos en el grafo. Por ejemplo, los azules son los elementos, los naranjas son las relaciones entre ellos y los morados son las evidencias que soportan este tipo de relaciones. La evidencia es la unidad mínima de información que es el párrafo y también los artículos.

Hoja 9

Cuando obtenemos el grafo de conocimiento, por último, estamos disponibles para facilitar la explotación de la información. Lo mejor - el primer paso es, por supuesto, podemos utilizar consultas SPARQL - este es un modelo de lenguaje específico que puede crear consultas para explotar el lenguaje, el gráfico de análisis. Esto requiere un experto en este tipo de dominio.

Hoja 10

La metodología que utilizamos en segundo lugar consiste en crear una interfaz pregunta-respuesta que proporcione la información no sólo a partir del grafo de conocimiento, sino también combinando fuentes externas, de otros y

grafos de conocimiento de otros, documentar conexiones y luego soportar preguntas en lenguaje natural para proporcionar respuestas, también en lenguaje natural. Así que nuestra plataforma es para una plataforma COVID.

Hoja 11

Toda esta información, pólizas: los gráficos de conocimiento, los modelos, los conjuntos de datos y los servicios son totalmente gratuitos y públicos. Están disponibles en estas URL. Gracias por su atención y estoy en condiciones de responder a cualquier pregunta que tengan.