

KeyQ

KeyQ es un sistema que permite obtener los términos (simples y compuestos, esto es, la terminología del dominio de aplicación) de un conjunto de documentos. También permite la búsqueda por dichos términos, devolviendo los párrafos donde se encuentra dicho término compuesto, ordenados por relevancia.



Información de contacto

Dirección: ETSI Informáticos - UPM
Ontology Engineering Group (OEG)
Departamento de Inteligencia Artificial
Universidad Politécnica de Madrid
Campus de Montegancedo, sn
Boadilla del Monte, 28660 Madrid
Teléfono: 910672915
Página web: demo.inno.oeg-upm.net
Correo electrónico: mariano.rico@upm.es

- Disponible

Tipo de oferta tecnológica

Soluciones tecnológicas

Áreas de investigación e innovación

- Tecnologías digitales, Inteligencia Artificial, ciberseguridad, 5G, robótica

ODS



Disponible desde: 2019

¿Dónde?

Centro de I+D+i en Inteligencia Artificial (AI.nnovation Space) Ontology Engineering Group

Documentación

[Descargar documentación adicional \(pdf\)](#)

Palabras clave: | [búsqueda inteligente](#) | [información](#) | [retrieval](#) | [terminología](#) | [términos](#)

Descripción de la base tecnológica

KeyQ es un sistema que permite obtener los “términos compuestos” (terminología del dominio de aplicación) de un conjunto de textos (corpus). También permite la búsqueda por dichos términos, devolviendo los párrafos donde se encuentra dicho término compuesto, ordenados por relevancia.

La figura 1 muestra los términos compuestos identificados en un corpus de 112 documentos legales en inglés del proyecto europeo Lynx (<http://lynx-project.eu/>), con el que fue probado el sistema.

Figura 1. Ejemplo de “términos compuestos” identificados en un corpus de documentos legales ordenados por número de apariciones (frecuencia), de más frecuente (parte superior) a menos frecuente (parte inferior).

La búsqueda por “termino compuesto” permite identificar las ubicaciones en el corpus mediante **gráficos de dispersión** como el mostrado en la figura 2. Cuantos más términos, compuesto o simples, se añadan a la búsqueda, menos líneas verticales habrá en el gráfico de dispersión y, por tanto, en menos párrafos habrá que buscar la respuesta a la consulta realizada.

Figura 2. Ejemplo de *gráfico de dispersión* para el término “supreme court” sobre el corpus descrito.

La solución se basa en tecnologías de Procesamiento de Lenguaje Natural preparadas para procesar textos en distintos idiomas, así como en tecnologías de Inteligencia Artificial y Aprendizaje profundo (*deep learning*). Se ha trabajado con textos en castellano, catalán, e inglés.

Necesidades de negocio / aplicación

- La búsqueda de información se ha convertido en una actividad habitual de nuestra vida diaria. No solo en el entorno laboral, también en nuestro ocio buscamos dónde ver una película, dónde ir a cenar, o qué colegios tenemos cerca de casa.
- Todos tenemos interiorizado el “estilo Google” de búsqueda de información: las palabras clave. Pero hay situaciones en las que echamos de menos que Google entienda la semántica de las preguntas. Por ejemplo, si preguntamos por “libros que citen a libros de García Márquez”, Google nos devolverá enlaces a páginas web que tengan libros de García Márquez, en lugar de libros que citen a libros del autor de Cien años de Soledad. Google no entiende las preguntas, no entiende su semántica.

Ventajas competitivas

- Multiguismo: Trabaja con textos en español o en inglés. Se han hecho prototipos con otros idiomas, como catalan.
- Búsqueda intuitiva de los términos más usados en un corpus de documentos. Por ejemplo, en un corpus de manuales de reparación de aeronaves, tenemos términos como “circuit”, o “gear”. Una búsqueda por alguno de estos términos mostraría los documentos en los que se encuentran. Sin embargo, cuando el corpus es grande, o cuando el término es frecuente, la búsqueda por palabra clave se hace inviable.
- Propone una búsqueda basada en **términos compuestos**. Siguiendo el ejemplo anterior, serían términos compuestos “circuit breaker” y “landing gear”. El usuario teclearía algunas letras de un término (no necesariamente las iniciales), y el sistema le mostraría los términos simples y compuestos que hay en el corpus, ayudándole a crear una consulta más precisa. La búsqueda por término compuesto obtiene resultados más precisos que la búsqueda tradicional por keyword (estilo Google).
- Escala linealmente (usando producto KeyQ-solr) y se puede llevar a la nube (se ha llevado a Azure).
- Protegido por un registro de software de la CAM.

¿No encuentra lo que busca en un mar de documentos? Nuestra tecnología permite una búsqueda más eficiente mediante la identificación automática de términos compuestos.

Referencias previas de prestación

- Las terminologías generadas por esta herramienta alimentan el sistema KeyQ-solr, un sistema desarrollado bajo el paraguas del AInnovation center de UPM, un centro mixto de UPM y Accenture, entre 2020 y 2021. Se han presentado prototipos al Ministerio de Justicia y a la Generalitat catalana.
- El grupo de investigación es experto en tecnologías de la Web Semántica y Datos Enlazados, estándares bien establecidos avalados por organismos internacionales de estandarización como el W3C. Estas tecnologías nos permiten hacer un tratamiento semántico de la pregunta y proporcionar resultados más precisos.

Protección

- **Registro sw** M-007053, marzo 2019)

Grado de desarrollo



Contacto

Contacto KEYQ

Mariano Rico

ETSI Informáticos UPM, Ontology Engineering Group (OEG)

e: mariano.rico@upm.es