accenture

AI.nnovation Space

POLITÉCNICA

BUSCADOR KEYQ

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

BUSCADOR KEYQ

Noviembre 2020

CONTENIDO

1. RESUMEN EJECUTIVO

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**.

Nuestra empresa es experta 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.

Nuestro proyecto **KeyQ** permite una búsqueda intuitiva usando 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.

Este proyecto propone una búsqueda basada en **términos compuestos**. Siguiendo el ejemplo anterior, serían términos compuestos "circuit breaker" y "landing gear". El usuario teclea algunas letras de un término (no necesariamente las iniciales), y el sistema le muestra los términos simples y compuestos que hay en el corpus, ayudándole a crear una consulta más precisa.

La Figura 1 muestra un ejemplo de búsqueda por términos. En este ejemplo, el usuario ya ha indicado el término simple "repair" y ha tecleado "circ". El sistema identifica los términos (simples y compuestos) que contienen "circ", en este caso uno simple ("circuit") y uno compuesto "circuit breaker" (mucho más específico). Si el usuario selecciona el término compuesto, la búsqueda se realizará por los términos "repair" y "circuit breaker".

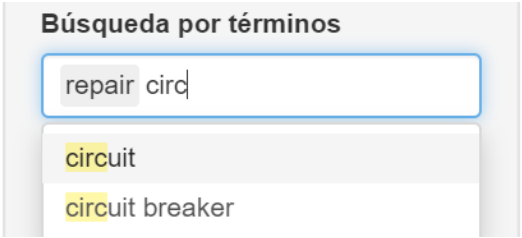

Figura 1. Búsqueda por términos. El sistema permite buscar por términos compuestos.

2. CASO DE USO: DOCUMENTACIÓN AERONÁUTICA

Para este caso de uso, se creó un corpus a partir de 225 documentos (pdf y Word) proporcionados por Accenture. Seis documentos no pudieron procesarse, por lo que se usaron **219 documentos** (9.0 millones de palabras) para el caso de uso. Estos documentos son manuales técnicos en inglés de reparación de aeronaves y normativas de todo tipo (normas de conectores eléctricos, cable, tornillos, remaches, etc.). Algunos de los documentos son pdf compuestos de imágenes que, aunque tenemos la tecnología software para extraer sus contenidos textuales, no han sido utilizados para este caso de uso.

Tampoco se ha realizado una limpieza exhaustiva de los documentos obsoletos (documentos con una “marca de agua” en diagonal). Al no limpiar estos documentos, el texto de la “marca de agua” se incorpora como conjuntos de letras al texto de la página donde se encuentra la “marca de agua”, produciendo frases ilegibles.

La Figura 2 muestra los términos compuestos más relevantes (más frecuentes) del corpus. Debe observarse que solo se muestran los 40 más relevantes identificados por nuestro software, de un total de 121 elementos compuestos que aparecen más de 1000 veces en el corpus.

El tiempo de computación requerido para calcular los términos compuestos está en el orden de las horas, por lo que nos vemos capacitados a procesar volúmenes de documentos 10 veces mayores con solvencia.

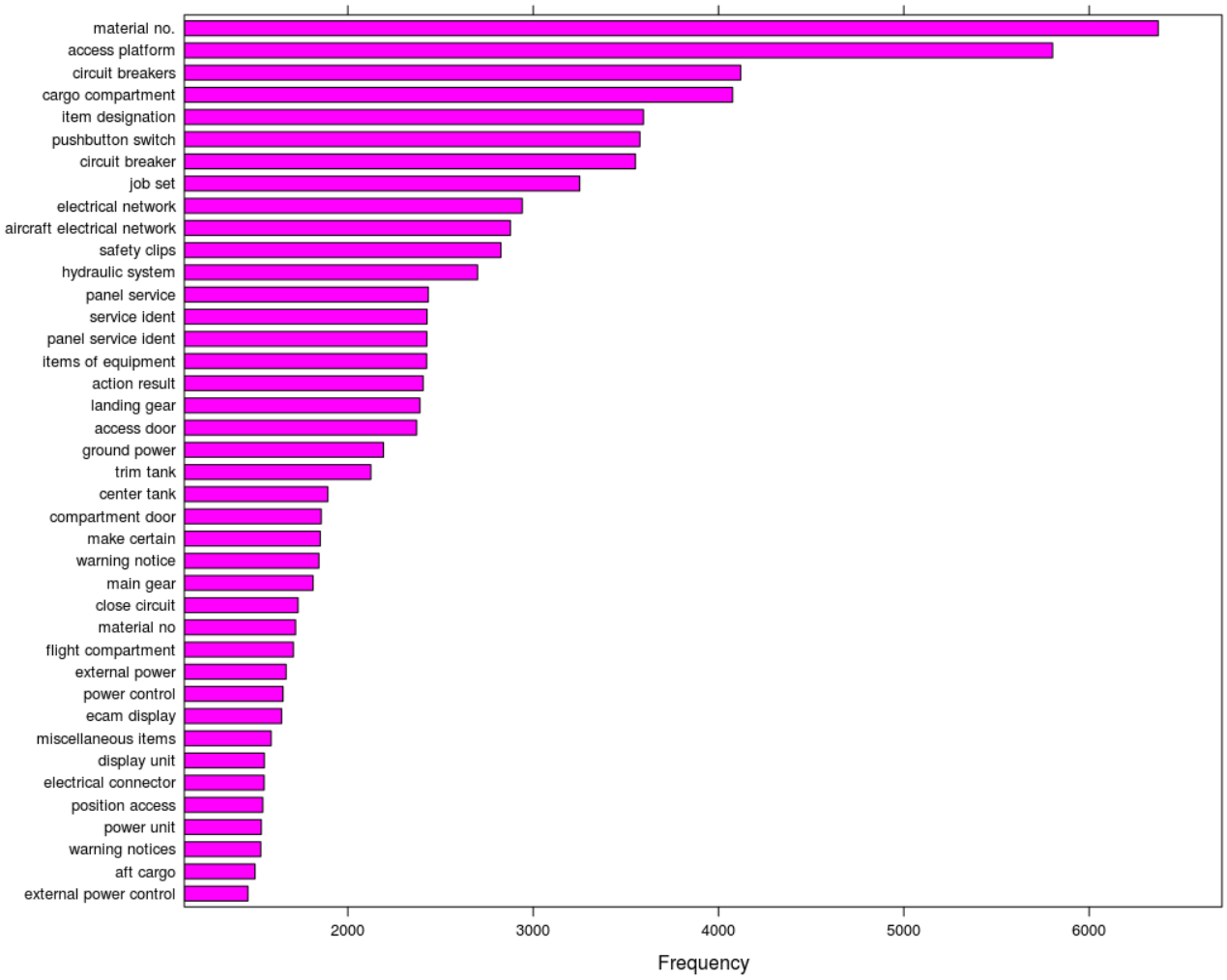

Figura 2. Términos compuestos identificados automáticamente en el conjunto de documentos proporcionado para el caso de uso. Se muestran los 40 **términos compuestos** más utilizados en el corpus, de un total de 121 términos compuestos que aparecen más de 1000 veces.

3. CASO DE USO: DOCUMENTACIÓN LEGAL

Para este caso de uso, se creó un corpus de 112 documentos legales en inglés del proyecto europeo Lynx (http://lynx-project.eu/), con el que fue probado el sistema. La Figura 3 muestra los términos compuestos más relevantes (más frecuentes) del corpus.

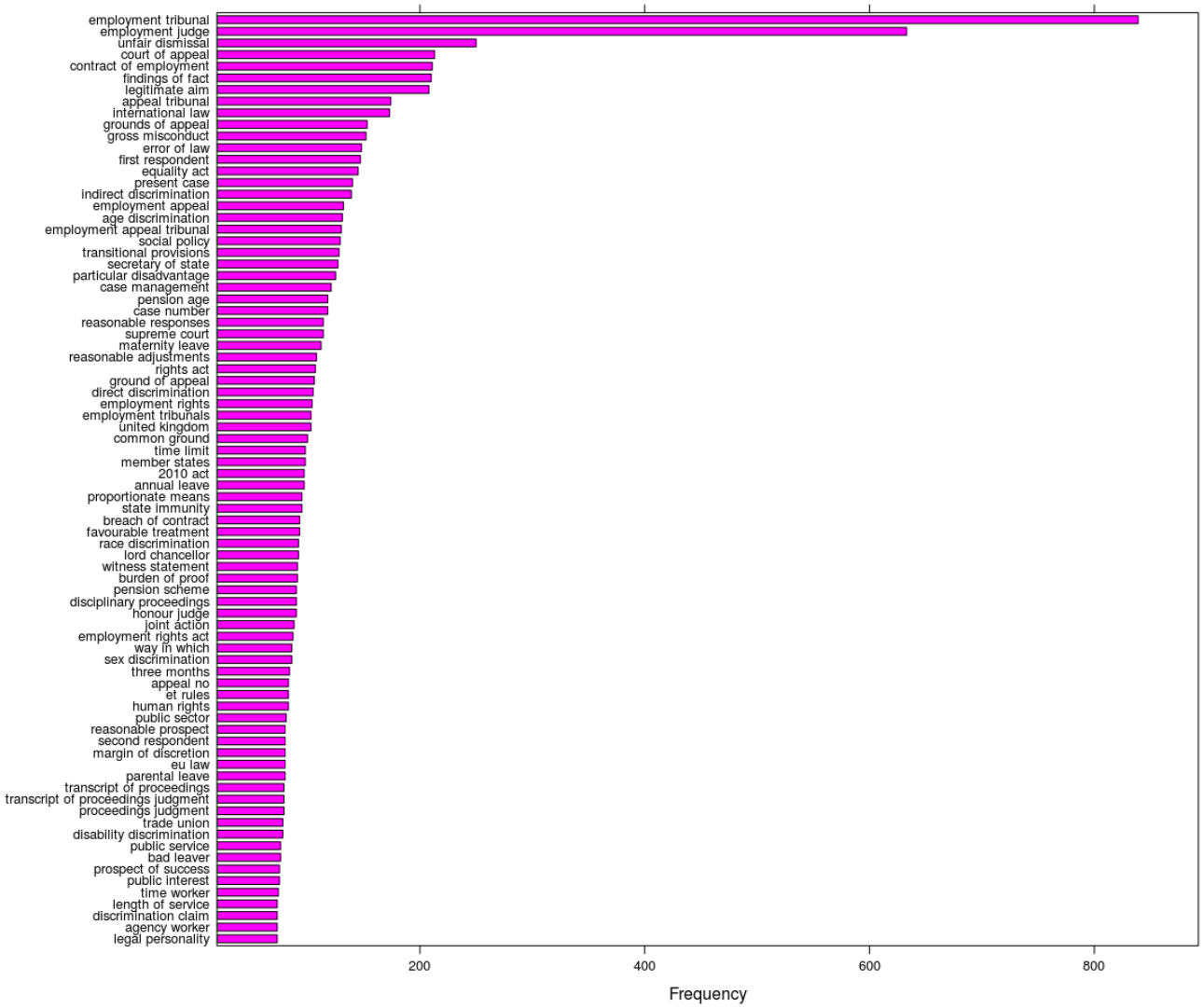

Figura 3. Términos compuestos identificados en un corpus de documentos legales ordenados por número de apariciones (frecuencia), de más frecuente (parte superior) a menos frecuente (parte inferior).

La búsqueda por “termino compuesto” permite identificar las ubicaciones en el corpus mediante gráficos de dispersión como el mostrado en la Figura 4. 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 4. Ejemplo de gráfico de dispersión para el término “supreme court” sobre el corpus descrito en el caso de uso del dominio legal.

4. HERRAMIENTA

Los casos de uso ha mostrado la viabilidad de la herramienta KeyQ para los dominios de los casos de uso. También podemos afirmar que **escala adecuadamente** con el número de documentos y que puede ser aplicada a otros dominios en los que se utilicen grandes volúmenes de documentos.

El sistema puede funcionar como un servidor que se reentrena cada vez que se aportan nuevos documentos, y un cliente web (móvil, tablet) que permite realizar las consultas. Los resultados pueden ser los **párrafos** (o las **páginas**) más relevantes donde aparecen los términos de la consulta.

La UPM tiene registrado este software en el Registro de la Propiedad Intelectual de Madrid (M-007053, marzo 2019). Sobre este software se ha construido una versión con mayor funcionalidad en el centro de investigación mixto (UPM-Accenture) AInnoSpace. Esta herramienta proporciona la siguiente funcionalidad:

- Gestor de documentos, para seleccionar un corpus (ver Figura 5).
- Análisis estadístico del corpus seleccionado.

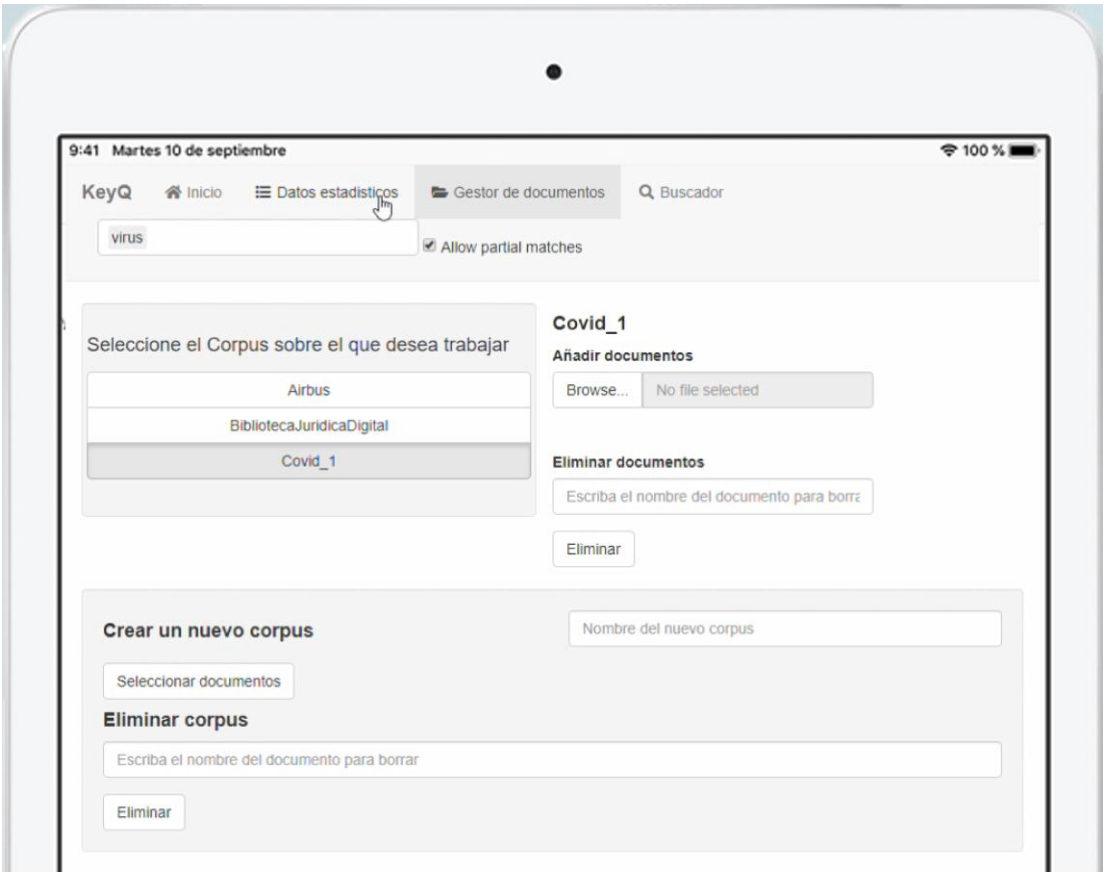

Figura 5. Gestor documental. Funcionalidad desarrollada en AInnovationSpace.