

Recuperación de Información y recomendaciones en la Web

Integrantes:

Juan Medina

4.921.887-8

Germán Costabel

4.778.044-9

Índice

Índice	2
Introducción	3
Problema	3
Enfoque de la solución	3
Diseño	4
Python Scrapper	4
Elasticsearch Module	5
Front-End React	5
Funcionalidades y uso	5
Acceso a corpus de 1000 artículos	5
Búsqueda por texto en título y cuerpo	5
Preprocesamiento	5
Filtrado por rango de fechas	6
Búsqueda por exclusión	6
Orden del resultado	8
Conclusiones	8
Trabajo Futuro	9
Scrapping automático del corpus	9
Categorización	9
Estadísticas	10

Introducción

En este proyecto buscamos generar un buscador de mayor utilidad para la página de noticias de presidencia (<https://www.gub.uy/presidencia/comunicacion/noticias>). El buscador presente en la página no da opciones para realizar búsquedas detalladas, en el ámbito periodístico (y también ciudadano) la información de presidencia es un recurso importante, sin embargo, encontrar información en ella puede presentar dificultades. En este proyecto intentamos generar un prototipo de buscador que pueda solventar estos problemas

Problema

El centro del problema es la gran cantidad de artículos presentes en la web de presidencia sin una forma razonable de poder filtrarlos o buscar algo particular entre ellos. Si no se cuenta con el título exacto de un artículo dado es casi imposible poder llegar a él a través de la página o utilizando herramientas externas como google.

El buscador proporcionado por la web no cuenta con herramientas básicas que ayuden a buscar algo tan simple como conseguir todos los artículos que mencionan covid durante un periodo de tiempo, esto es casi imposible con las herramientas actuales.

Actualmente según la página de presidencia se cuentan con 30.844 artículos, y cada día se agregan más. Contar con una herramienta que permita buscar en un corpus de este tamaño permite un mayor acceso real a la información provista por el estado. Esto beneficia fuertemente tanto a periodistas que busquen información que ha provisto presidencia previamente como al ciudadano común que busque acceder a esta información.

Enfoque de la solución

La solución se enfoca en intentar implementar la mayor cantidad de casos de uso y herramientas a un buscador comprometiendo lo menos posible la escalabilidad del proyecto a futuro. Se busca generar un buscador que pueda lidiar con las necesidades de buscar dentro de los artículos de presidencia. Desde poder filtrar por texto a poder seleccionar en qué rango de fechas se encuentran los artículos, o buscar artículos por exclusión para poder refinar más la búsqueda.

La solución se enfocará en los problemas que le interesa solucionar a un periodista que utilice el buscador ya que inicialmente son los problemas más complejos y una vez que se tiene eso solucionado poder reconvertir la herramienta en algo que funcione para el uso del público general conlleva relativamente poco trabajo.

Diseño

Para el diseño de la aplicación se intentó minimizar la cantidad de overhead de los componentes y minimizar la complejidad para poder dedicar la mayor parte del esfuerzo al desarrollo de las funcionalidades.

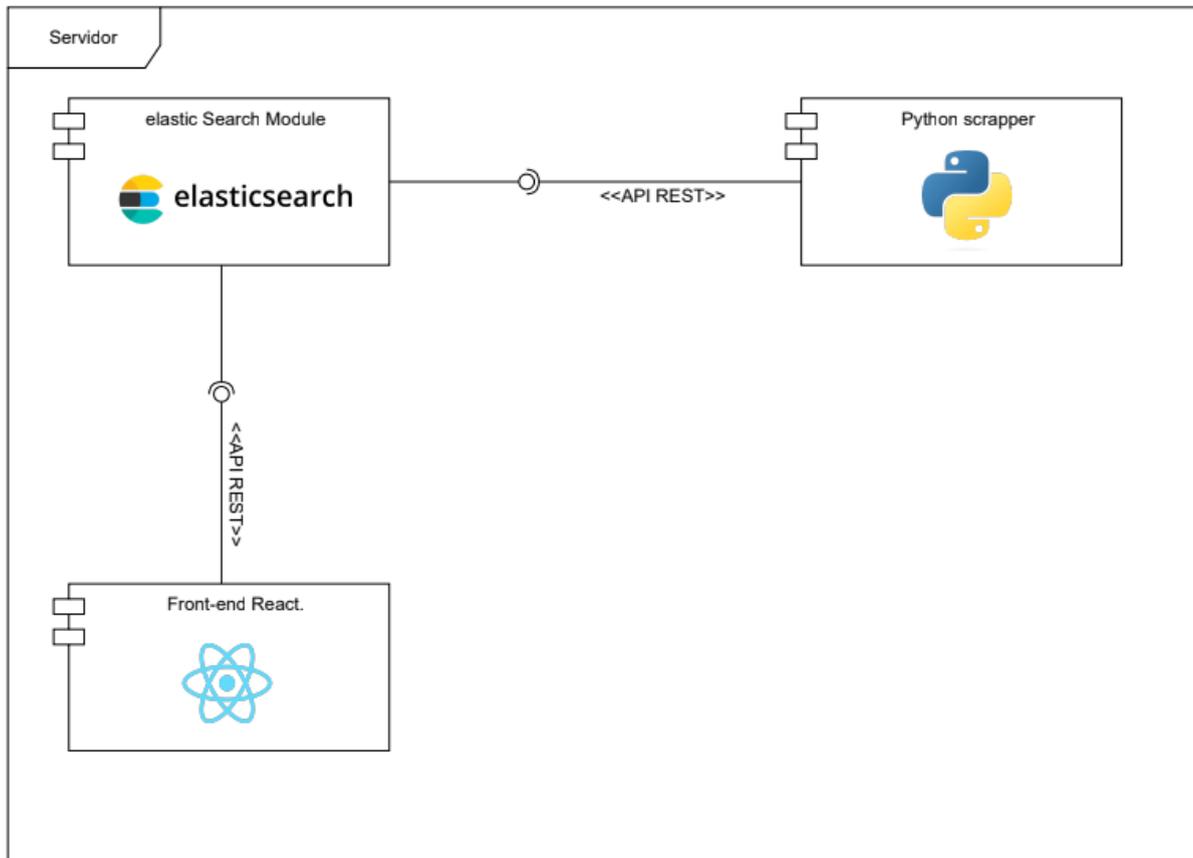


Diagrama de componentes

Python Scrapper

Un scrapper hecho en python con BeautifulSoup que recorre la web de presidencia y almacena en formato CSV los artículos, separado por las columnas de URL (la cual se utiliza como una suerte de clave), fecha de publicación del artículo, título del artículo, y cuerpo del artículo. Se eligió descartar el contenido multimedia de los artículos, usualmente imágenes o videos, para minimizar el coste de almacenamiento, por que no hay interés en utilizarlos para búsqueda y por que son fácilmente recuperables teniendo la URL al artículo si fuese necesario.

Se utilizó CSV para almacenarlos inicialmente porque la librería por defecto para manejo de CSVs de python permite escribir línea a línea lo que permite más fácilmente retomar el scrapeo en caso de fallo de conexión o algún otro problema inesperado.

Una vez se tienen los artículos en formato CSV se convierten al formato json que utiliza Elasticsearch y se cargan a través de la API de Elasticsearch.

Elasticsearch Module

Un contenedor de docker con una instancia de Elasticsearch corriendo que permite a la aplicación cliente realizar las búsquedas. Se alimenta del scraper y maneja tanto el almacenamiento e indizado del corpus como las búsquedas en el mismo. Se eligió por ser el método que requería menor overhead por parte del equipo, el servicio de Elasticsearch provee de todo lo necesario mientras que otras herramientas requieren mucho mayor trabajo de implementación y/o implementar un servidor o wrapper para poder comunicarse con ellas.

Front-End React

Una aplicación cliente construida con react.js que consume directamente las funcionalidades del componente con Elasticsearch a través de API REST. Se eligió react por su facilidad en cuanto a su velocidad de desarrollo además de proveer una forma simple de comunicarse con la API de Elasticsearch.

Funcionalidades y uso

Acceso a corpus de 1000 artículos

Se scrapeo e importó a Elasticsearch mil artículos con los cuales se realizaron todas las pruebas y no se presentó ningún problema en cuanto funcionamiento por tamaño del corpus.

Búsqueda por texto en título y cuerpo

Se utiliza la query "query_string" de Elasticsearch, que consiste en especificar los campos en los que se quiere buscar un determinado texto, además, permite realizar búsquedas booleanas en las que simplemente se escribe una operación booleana y se agregan los parámetros que se desean buscar.

Preprocesamiento

Previo a realizar las consultas a Elasticsearch, se realiza un preprocesamiento de la entrada. Se cuenta con una bolsa de stop words obtenida del sitio countswordsfree (<https://countswordsfree.com/stopwords/spanish>).

Cada palabra ingresada se anida como un OR a la palabra anterior. Por ejemplo si se ingresa "obras en hospitales" la búsqueda será "obras OR hospitales". Se le asignará mejor puntuación al resultado de la búsqueda si cuenta con ambos términos.

Filtrado por rango de fechas

Una de las características de Elasticsearch es la manipulación de fechas, estas se filtran mediante operadores como lt (menor que) o gt (mayor que) permitiendo un filtrado sencillo de implementar. Se toman de dos campos del sitio las fechas para conformar el rango en el que se desea buscar las noticias.

obra en hospital

March 1 2022 March 17 2022

ASSE inauguró sede norte del Hospital de Ojos en Artigas

El norte de ASSE, Ivonne Bruno Vaz iourem; el director del Hospital especializado de Ojos, Felipe Berta, y el del nosocomio de Artigas, Aníbal Fontoura. En su oratoria, Cipriani precisó que la población del departamento de Artigas es de 73.378 habitantes, de los cuales 56.820 son usuarios del prestador público, por lo que es una obligación invertir para brindar una asistencia integral y de calidad. Señaló que la nueva sede del Hospital de Ojos en el norte del país se enmarca en el plan Uruguay Ve, proyecto de descentralización que prevé el desarrollo de una oftalmología integral y de calidad en el interior del país. En ese sentido, aseguró que se brindará asistencia en policlínica, intervenciones quirúrgicas y el seguimiento posoperatorio, lo que permitirá una mayor cercanía con los usuarios y minimizar de forma considerable los tiempos de atención. Por otra parte, el titular de ASSE resaltó la importancia del nuevo tomógrafo de 16 líneas que se instaló en el nosocomio, que posibilitará agilizar los diagnósticos y posteriores tratamientos, al igual que el laboratorio de anatomía patológica inaugurado en la fecha. Asimismo, informó que el hospital recibió una nueva ambulancia de traslado, que se incorpora a la flota ya existente, y se entregaron otras cero kilómetros en Yacaré, Bernabé Rivera, y La Bolsa, respectivamente; así como un micro en la policlínica de Topador, para el traslado de pacientes pediátricos de zonas rurales. Además, con el objetivo de fortalecer la atención en salud bucal, se entregaron sillones odontológicos en el centro departamental y en las policlínicas Sequeira, Baltasar Brum, Gómez Gotuzzo y Cerro Ejido. De esta manera, ASSE continúa

Fecha: 04/03/2022
Score: 10.1

ASSE inauguró sede centro del Hospital Especializado de Ojos en Durazno

Asistieron a la inauguración el presidente de ASSE, Leonardo Cipriani; el vicepresidente de esa institución, Marcelo Sosa; el vocal, Julio Micak; el intendente de Durazno, Carmelo Vidalín; el presidente de OSE, Raúl Montero; el director de UTE, Felipe Algorta; el director del Hospital Especializado de Ojos, Felipe Berta, y la directora del Hospital de Durazno Dr. Emilio Penza, Cristina Coirolo. Durante su oratoria, Cipriani explicó que la inauguración responde a uno de los objetivos centrales, que apunta a descentralizar la atención en salud. Se refirió específicamente a la evolución de la atención oftalmológica, a través una mejora en la gestión de los recursos humanos y del equipamiento. Además, resaltó que esta iniciativa permite estar más cerca para asegurar accesibilidad a los habitantes de la región. El jerarca explicó que este hospital cuenta con un block quirúrgico con equipo de facoemulsificación, un microscopio especial para oftalmología y consultorios para control y diagnóstico. En la ocasión, Cipriani adelantó que en el Hospital de Durazno también habrá, próximamente, un nuevo laboratorio y se ampliará la farmacia. Además, informó que, en coordinación con el Banco de Previsión Social (BPS), se prevé la puesta en marcha de un ómnibus para pesquisar a pacientes que padezcan patologías oftalmológicas y que por diferentes motivos no puedan acceder a las consultas en policlínica. El director del Hospital Especializado de Ojos, Felipe Berta, precisó que el centro contará con el apoyo de la dirección del hospital departamental, de la Intendencia de Durazno y de las comises de la región, y trabajará en coordinación con el Hospital de Ojos de

Fecha: 17/03/2022
Score: 9.8

Hospital de Clínicas presentó equipamiento para que pacientes que reciben quimioterapia no pierdan el cabello

Este martes 15 en el Hospital de Clínicas, la vicepresidenta de la República, Beatriz Argimón, senadores y diputados nacionales fueron recibidos por el director, Álvaro Villar, y el titular de la Cátedra de Oncología, Gabriel Krygier, para inaugurar en forma oficial el Dignicap, primer aparato de enfriamiento del cuero cabelludo aprobado internacionalmente que previene la caída de pelo por quimioterapia en casos de tumores sólidos. Argimón dijo que este evento es un mensaje que destaca el abordaje humano y tecnológico para que los tratamientos oncológicos se desarrollen de la mejor manera posible y aporten a la calidad de vida de los pacientes. El dispositivo permite enfrentar la enfermedad, desde el punto de vista anímico, de una mejor manera, señaló la vicepresidenta. Se trata del primer equipo de América Latina de la marca Dignicap. Las beneficiarias en esta instancia serán las usuarias del Hospital de Clínicas y de la Administración de los Servicios de Salud del Estado (ASSE). Las gorras podrán ser utilizadas solamente por pacientes con cáncer de mama u ovarios. El consultorio con el nuevo sistema

Fecha: 15/03/2022
Score: 8.7

Búsqueda por exclusión

Se implementó con la funcionalidad de búsqueda booleana de Elasticsearch, con el parámetro “must_not” que descarta los artículos que cumplan con una condición dada.

Por ejemplo:

obras

January 1 2022

November 9 2022

Transporte inauguró obras desarrolladas mediante el programa Convenios Sociales

La visita a las inauguraciones se realizó este viernes 3, asistieron, además de Falero, el subsecretario Juan José Olaizola y el director nacional de Arquitectura, Santiago Borsari, entre otras autoridades departamentales y locales. "Desde el ministerio, a través del programa Convenios Sociales, queremos apoyar a este tipo de instituciones que tienen un fuerte impacto social", manifestó José Luis Falero y aclaró que, además de dinero, se otorga asesoramiento técnico para la construcción de obras comunitarias en todo el territorio nacional. En la ocasión, informó que en este período de gobierno se duplicó la partida para el programa, lo que significa una inversión aproximada a los 140.000.000 de pesos anuales. Asimismo, aseguró que la prioridad es apoyar a las organizaciones abocadas a la atención de niños y adolescentes, adultos mayores y discapacitados. "Muchas veces, si no es por el apoyo del Estado, estas cosas no se lograrían", agregó. En Onpli el ministerio aportó cerca de 2.500.000 pesos para la reparación de baños, refacción de cocina y aulas. También se realizaron restauraciones en la infraestructura edilicia. La

Fecha: 03/06/2022
Score: 8.4

UTE inauguró obras de servicio de energía eléctrica en Toledo

La inauguración se efectuó este viernes 11 y participaron integrantes del directorio de la empresa, encabezadas por la titular del ente, Silvia Emaldi; la representante del Ministerio de Desarrollo Social (Mides) de Canelones, Maricarmen Suárez, y el alcalde de Toledo, José Luis Gini. En diálogo con Comunicación Presidencial, Emaldi señaló la importancia de esta obra y añadió que es parte del trabajo que realiza el organismo para la inclusión social de la población. Además, informó que con este proyecto se regulariza el servicio de 170 familias de la ciudad de Toledo, un total de 500 personas, entre ellas, unos 150 menores de 12 años, que podrán utilizar la energía eléctrica de manera segura y evitar accidentes. Todos accederán también al bono social que ofrece UTE, con descuentos del 80% en la factura para clientes que registren un consumo de hasta 250 kilovatios. Asimismo, indicó que los beneficiarios de la tarjeta Uruguay Social del Mides o, en su defecto una asignación familiar, obtendrán deducciones del 85 o 90%. La jerarca manifestó que a estas personas se les brinda talleres de eficiencia energética y, por este motivo, durante la inauguración se les entregó a los vecinos un kit con lámparas de consumo eficiente y material educativo. La inversión total fue de unos 13 millones de pesos y las obras fueron efectuadas en el marco del convenio que suscribió UTE con el Instituto Nacional de Cooperativismo. UTE y Correo firman acuerdo para mantener la atención en pequeñas localidades. En otras oratorias, Emaldi se refirió al convenio que el ente desarrollará con el Cerro Troncoso y que comenzará a realizarse este viernes 11 en la ciudad de Progreso, en el departamento

Fecha: 11/02/2022
Score: 8.1

Lacalle Pou inauguró obras de policlínica de ASSE en Villa Constitución

Acompañaron a Lacalle Pou el secretario de la Presidencia, Álvaro Delgado; el presidente de ASSE, Leonardo Cipriani; el intendente de Salto, Andrés Lima; el alcalde de Villa Constitución, Carlos Souto; el presidente de la Comisión Técnica Mixta de Salto Grande (CTM), Carlos Albisu, y representantes de la Embajada de Japón. Un contrato suscripto por ASSE y la sede diplomática del referido país facilitó la remodelación del centro de salud, que permitirá desarrollar telemedicina, con el objetivo de mejorar la accesibilidad a los diagnósticos oportunos y a consultas con especialistas. El presidente de la República, Luis Lacalle Pou, enumeró las realizaciones en la localidad y consideró que le aportan una "importante perspectiva" de desarrollo. A las obras en la policlínica local se le sumará una ambulancia por estrenar y las inversiones en la comisaría local, que incluyen la incorporación de una camioneta y la reciente instalación de un cajero automático. El mandatario recordó en su oratoria que, en una visita anterior, el 9 de setiembre de 2016, recibió el reclamo de mejoras edilicias en el CAIF Pequeños Brillantes de esta

Y luego quitando los que contengan el texto "Lacalle Pou" se observa el resultado:

obras

Lacalle pou

January 1 2022

November 9 2022

Centro Tiburcio Cachón recibe inversión de 20 millones de pesos en reacondicionamiento

"Queríamos recuperar este centro, que está destinado a personas con discapacidad visual y que estaba en una situación muy compleja desde el punto de vista edilicio, lo que obstaculizaba una rehabilitación de calidad", enfatizó Lema este jueves 1.º de setiembre, al término de una recorrida por las instalaciones en obras. El jerarca estuvo acompañado por la subsecretaria de la cartera, Andrea Brugman; su par del Ministerio de Transporte y Obras Públicas (MTOP), Juan José Olaizola; el secretario nacional de Cuidados y Discapacidad, Nicolás Scarela, y el director nacional de Arquitectura, Santiago Borsari. La reforma comenzó en marzo, mediante un convenio entre el Mides, que aporta 14 millones de pesos, y el MTOP, que destinó 6 millones para completar los 20. Incluye obras en todo el predio, el edificio y el subsuelo, y abarca un total de 1.237 metros cuadrados. La lista de tareas comprende la instalación de una rampa accesible en la entrada, pavimentación vehicular para ingresar, estacionamiento accesible, planos y maquetas en relieve. Además, se incorporarán dispositivos para la detección de incendios con alarma

Fecha: 01/09/2022
Score: 7.4

Hogar de Melo con capacidad para 65 adultos mayores dispone de nuevas instalaciones tras inversión del MTOP

Falero realizó la recorrida por el departamento de Cerro Largo, este lunes 19, acompañado por el intendente, José Yuramendi, y otras autoridades nacionales y departamentales. Allí, el ministro participó en la inauguración de las obras de la Asociación para la Integración del Adulto. Gracias a la firma de un convenio social, la institución podrá iniciar la tercera etapa de construcción de la sede, que incluye dormitorios y áreas de servicio. En esta oportunidad, el aporte del Gobierno asciende a 2.400.000 pesos. Asimismo, el Ministerio de Transporte y Obras Públicas (MTOP) firmó un convenio social con la Fundación Celeste y con el Club Atlético Porvenir, que permitirá la construcción en Melo de una cancha multipropósito para cada institución, por una inversión total 4.000.000 de pesos por parte de esa cartera. En horas de la tarde, concurrió al evento de la empresa Núñez Transporte, donde se anunció que comenzará a funcionar una nueva línea de ómnibus entre Melo y Punta del Este, con cuatro frecuencias semanales. En materia de infraestructura vial, el ministro informó sobre la próxima firma de un contrato para

Fecha: 19/09/2022
Score: 7.3

Obras en rotonda de rutas 2 y 24 de Río Negro quedarán habilitadas en dos semanas, anunció Falero

Tras finalizar la recorrida por algunas zonas del departamento de Río Negro, Falero, acompañado por el intendente, Omar Lafluf; el director nacional de Arquitectura, Santiago Borsari, y el director nacional de Vialidad, Hernán Ciganda, se expresaron en conferencia de prensa en la sede de la intendencia. El ministro visitó las obras de la rotonda que une las rutas 2 y 24, las que, según explicó, permitirán mejorar la seguridad en ese cruce. Los trabajos, que culminarán en las próximas dos semanas representarán "una solución que hacía mucho tiempo se venía reclamando", apuntó. También anunció que firmará un contrato para mejoras en unos 25 kilómetros de la ruta 25, entre las rutas 3 y 24, en Río Negro, por una inversión de 22 millones de dólares. Los trabajos, que comenzarán en setiembre, se realizarán en hormigón de 22 centímetros de espesor, para una mayor duración. "Esperemos poder seguir haciendo ese tipo de reconstrucción en la red vial para tener la garantía de durabilidad", indicó. Además, se refirió a las obras que se ejecutan en la ruta 20, que unirán las rutas 5 y 24, en 160 kilómetros. Según

Orden del resultado

El orden por defecto que se asigna al set resultado está dado por el score que brinda Elasticsearch para cada artículo encontrado, con la salvedad que se le da el doble de peso a la función de ranking utilizada si uno de los términos buscados está en el título del artículo.

Por defecto el score dado por Elasticsearch está basado en el algoritmo B25, sin entrar en mayor detalle ya que no es un algoritmo particularmente simple lo que se hace con el mismo es darle prioridad a los siguientes puntos:

- Frecuencia de los términos (TF): A más veces aparece un término buscado en el documento mayor importancia tiene el documento.
- Frecuencia inversa de los documentos (IDF): A más documentos contienen el término buscado menos importante es el término.
- Largo del campo: Si un documento contiene el término buscado y su cuerpo es pequeño es más relevante que si tiene el término y su cuerpo es grande.

Conclusiones

Se considera que el resultado del proyecto fue más que satisfactorio. Solo uno de los casos de uso planteados inicialmente no se implementó, el agregado automático de artículos al corpus, mayormente por que se determinó que no agrega demasiado al prototipo y sería algo que agregaría valor una vez que el proyecto esté funcionando, pero no en su estado inicial.

Se encontró en Elasticsearch una herramienta poderosa para implementar búsquedas que además combinándolo con React resulta en una infraestructura sólida y simple para llevar a cabo este proyecto y que puede ser útil para otros proyectos similares.

Trabajo Futuro

Se logró el desarrollo de muchas de las características deseadas pero quedan funcionalidades que serían interesantes de agregar, tanto algunas que se encontraban previamente como otras que surgieron durante el desarrollo.

Scrapping automático del corpus

Hay dos funcionalidades en cuanto a la adquisición del corpus que serían útiles de agregar. En este momento el corpus es estático y sin intervención manual no crece de ninguna manera, mientras que los artículos de presidencia si crecen, además que los artículos ya importados están lejos de ser todos los presentes en la web.

Primero, la importación automática de artículos nuevos al corpus. Desarrollar un proceso que chequee los nuevos artículos presentes en la web de presidencia y los agregue automáticamente al corpus, por el volumen de artículos esto sería suficiente correrlo una vez al día, o incluso una vez a la semana. La forma más simple sería agregar otro componente a la arquitectura que corra este proceso en los tiempos indicados, existe gran variedad de servicios que permiten realizar algo así por un costo bajo por lo que probablemente sería la forma más eficiente de solucionarlo.

Segundo, incorporar el resto de los artículos ya presentes en la página de presidencia al corpus. Desde el punto de vista de desarrolla es un costo similar al punto anterior con la salvedad de que desde el punto de vista de hardware o contratación de servicios esta parte si puede volverse costosa ya que aumentan rápidamente el tamaño del corpus demandando más recursos de Elasticsearch además que por el volumen de artículos también tiene un costo no ignorable el procesamiento inicial.

Categorización

Generar una categorización de los artículos en el corpus agrega valor al buscador. Poder filtrar entre artículos de promulgación de leyes, informes, artículos sobre un departamento específico o por cualquier otra de las categorías que puedan ser útiles sería una función útil en el buscador.

En cuanto a la parte técnica de esto lo primero sería determinar las categorías en las que interesa categorizar los artículos. Aunque hay algunas obvias puede ser de interés entrevistar usuarios (o potenciales usuarios) para tener un idea más concreta de qué categorías serían las útiles. Una vez se cuente con las categorías se puede determinar el mejor método para realizar el filtrado, algunas como las que traten sobre promulgación de leyes probablemente se puedan determinar de forma facil a traves de aplicar un regex mientras que otras pueden tomar mayor trabajo necesitando recurrir a algún tipo de aprendizaje automático para clasificarlas, este último método puede beneficiarse también de

tener el buscador ya funcionando con usuarios para contar con mayor información que pueda alimentar un algoritmo de ese tipo.

Estadísticas

Una vez se tiene todo el corpus incorporado en Elasticsearch se generan estadísticas del mismo que pueden ser útiles o de interés. Teniendo acceso a las búsquedas que realizan los usuarios se puede conseguir saber qué artículos son de más interés para el público, o que temas están llamando más la atención. ¿Qué tipos de leyes generan más interés en el público? ¿Qué tipo de noticias generan más tráfico? ¿Hay más interés en las noticias sobre la apertura de nuevos hospitales? ¿O sobre cambios en la reglamentación impositiva?

Que parte de las estadísticas serían más importantes o útiles no es obvio inicialmente, pero tener acceso a más información puede ser un valor agregado interesante del proyecto.