Proyecto Webir

21 de Noviembre de 2016
Facultad de Ingeniería UdelaR

Análisis de la información recabada en la web de inundaciones en el Uruguay

Grupo 7
Camila Irachet
Leticia Vaz

Índice general

Introducción

En el siguiente documento se describe la resolución del proyecto final de la asignatura Recuperación de Información y Recomendaciones en la Web. Se basa en la propuesta y solución por parte de nuestro grupo, al planteamiento de analizar la recuperacion de informacion en la Web. Describiremos qué componentes forman parte de la solución, mencionando que es lo que realizó, dando ejemplos de los resultados obtenidos y por último enumerando algunos de los trabajos a futuros que se podrían realizar.

Propuesta

La motivación inicial del proyecto fue obtener información sobre las inundaciones que se registraron en los últimos 10 años en Uruguay, para luego poder hacer un análisis estadístico en base a los datos recabados. La restricción impuesta por la inexistencia de datos, acotó el alcance del proyecto a un período de 6 meses (6 meses hacia atrás hasta el día de hoy).

Como fuente de información, se tomaron las noticias referentes a inundaciones publicadas en el portal del Observador y tanto las páginas extraídas de internet, como los datos extraídos de los mismos fueron almacenados en una base de datos no relacional.
Para esta extracción de los datos, se tomaron en cuenta una serie de palabras y sinónimos de las mismas y con el resultado de las consultas se pudo hacer una análisis estadístico que se mostró tanto en formato numérico como con gráficas.

El interés por trabajar con datos referidos a inundaciones residió en la importancia del tema en cuanto a las múltiples implicancias que cada inundación trae aparejadas consigo: evacuaciones, pérdidas materiales, problemas sanitarios, movilización de recursos, enfermedades, cortes de luz, riesgos eléctricos, cortes de agua, problemas psicológicos de las familias afectadas, defunciones, etc.

Extracción de Datos

Para la extracción de los datos, se utilizó como fuente de información la página del Observador. Como el contenido en dicha página es muy amplio y variado, se filtraron las noticias que estuvieran relacionadas con Inundaciones, lo que redirige a una página con las inundaciones en los últimos 6 meses (http://www.elobservador.com.uy/inundaciones-s).

Luego, haciendo uso de la herramienta JSoup, se identificaron las noticias y se buscó dentro de los elementos de cada una, el atributo href correspondiente a la url que contiene la misma.

Ya que dentro de la página del Observador, hay muchas noticias que no están relacionadas con inundaciones a nivel nacional, fue necesario examinarlas todas para comprobar si contenían la palabra NACIONAL dentro del cuerpo.


NACIONAL INUNDACIONES

Más de 100 desplazados y nueve rutas cortadas en todo el país

En caso de que la noticia fuera de interés, se almacenaba dentro de una base de datos no relacional (MongoDB). En caso contrario, se descarta para no extraer datos incorrectos que influyan en el análisis final.

Una vez identificadas las noticias relacionadas al estudio de este proyecto, se buscaron los siguientes datos utilizando palabras clave:

- Cantidad de desplazados
- Cantidad de evacuados
- Cantidad de autoevacuadas
- Cantidad de muertos
- Cantidad de rutas cortadas
- Departamentos afectados

Una de las mayores dificultades que se presentaron, fue que, a pesar de que todas las noticias pertenecían al mismo diario y eran de un período corto de tiempo, la forma en que fueron redactadas difería de gran manera entre una y otra.

A pesar de que la extracción automática es más rápida y eficiente, hubo algunos datos que no fue sencillo de extraer con palabras clave, por lo que fue necesario extraerlos manualmente.

Herramientas utilizadas

- **JSoup:** es un parser Java para HTM.Proporciona una API muy conveniente para la extracción y manipulación de datos, utilizando lo mejor de DOM, CSS y jQuery. Se eligió esta herramienta dado que es open source(gratuita), en lenguaje Java y fácil de usar.

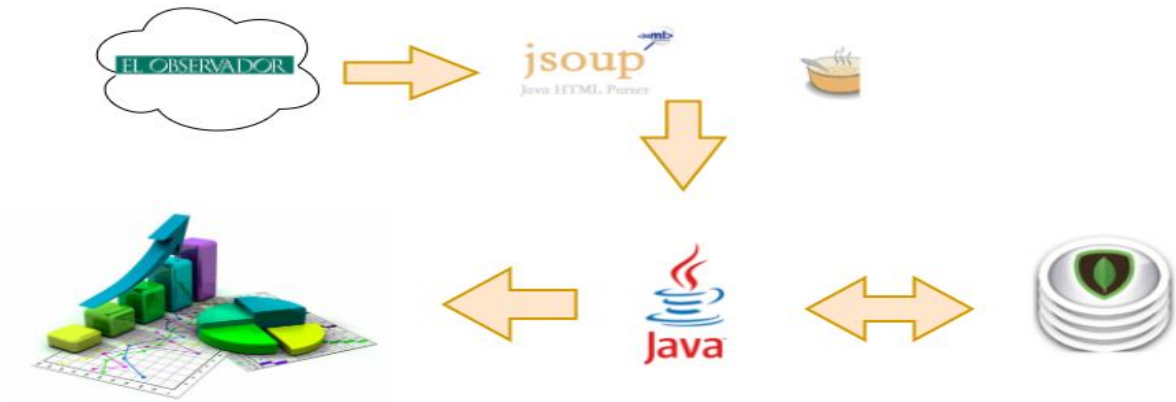
  Web Scraping: es una técnica que es utilizada para extraer información de sitios web. Los programas que utilizan ésta técnica, generalmente simulan la navegación de un humano utilizando el protocolo HTTP. Utilizando la técnica de Web Scraping se implementa un módulo que realiza la recuperación de los avisos de inundaciones desde el sitio web del Observador almacenados localmente.

  El web scraping se realiza sobre la página de El Observador mediante esta librería. Esta técnica se realizó sobre el código HTML de la página .Para esto, se analizó el código fuente de la página y se extrajo la información deseada utilizando las etiquetas HTML del mismo. Web scraping se realiza una vez y luego se interactúa con la información almacenada en la base.

- **MongoDB:** es un sistema de base de datos no relacional(NoSQL). Decidimos persistir los datos mencionados en una base de datos NoSQL, ya que persistimos en formato JSON.

- **Java:** es un lenguaje de programación orientado a objetos. Utilizamos este lenguaje para la implementación de la solución ya que se contaba con experiencia previa con este lenguaje y además de que que existen una amplia variedad de herramientas disponibles que facilitan el trabajo de publicación de servicios y parseo de páginas.

  La siguiente imagen muestra una descripción general de la arquitectura de la solución, identificando los módulos claves necesarios .



Análisis Estadístico

Como se mencionaba en la introducción, la falta de datos condiciona el análisis. El set de preguntas planteado al inicio del proyecto tuvo que ser acotado y la inferencia que se pudo realizar fue nula.

De todas las noticias que aparecían cuatro de ellas corresponden a la inundación de Julio y tres de ellas a la de Abril.

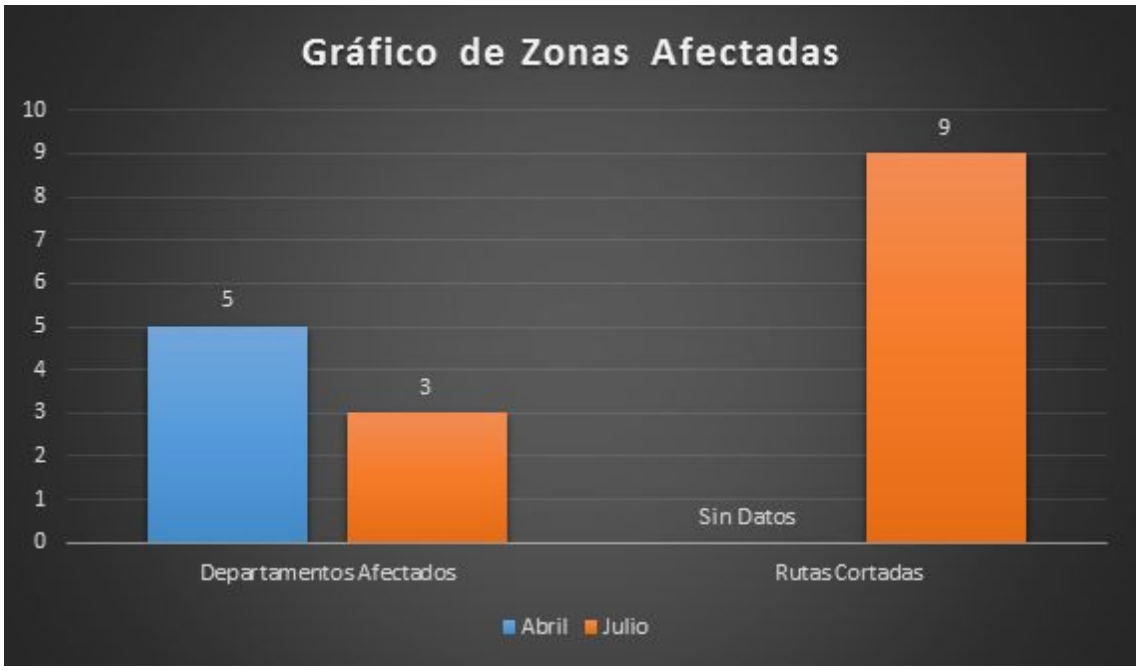
En la primera gráfica se puede apreciar que la cantidad de desplazados en el mes de Abril fue de varios órdenes mayor que en la de Julio. Además si hacemos una análisis de las personas evacuadas y autoevacuadas con respecto a la cantidad de desplazados, se puede notar que en la segunda inundación hay más evacuados. Esto puede deberse a que dado los datos anteriores, se tomaron mayores medidas por parte del gobierno.

Cabe destacar que se encontró una única noticia donde hablaba del fallecimiento de una persona. Esto puede deberse a que afortunadamente las inundaciones no fueron tan graves a nivel de vidas humanas o a que no hay la suficiente información disponible sobre un hecho que ocurrió hace meses en el portal del Observador.

En esta segunda gráfica se muestra la cantidad de departamentos que fueron afectados en cada inundación y la cantidad de rutas cortadas que se mencionan en las noticias.

Para la obtención de los departamentos hicimos una consulta buscando el nombre de 19 departamentos. En la primera inundación los departamentos más afectados fueron Soriano, Rocha, Paysandú, Durazno y Treinta y Tres. Lo que notamos en la recolección de datos es que se mencionaba que los departamentos afectados fueron 16, pero solo se nombraba estos 5. En la segunda inundación, se tomó el mismo criterio para obtener los departamentos afectados, siendo estos Florida, Canelones y Rivera.

Para el caso de las rutas, solo se obtuvo información en las noticias referentes a la inundación de Julio.



Trabajo a futuro

Como trabajo futuro podemos comentar que sería interesante volver a evaluar alguna de estas estadísticas calculadas para observar si hay diferencias significativas con los datos recabados al día de hoy. Y poder notar si a medida que pasa el tiempo hay un incremento o no de la información obtenida desde la web.

Por otro lado, se podría realizar un análisis de gran variedad de datos que se pueden obtener de las noticias, como por ejemplo de las discrepancias entre las distintas fuentes de información de una misma noticia, poder hacer una estadísticas descriptivas de los milímetros de lluvia que cayeron, la causa de las inundaciones, ver si hay más información a medida que pasa el tiempo o es la misma información pero replicada, cantidad de "m2" afectados, las zonas o departamentos que son más afectados en el territorio,etc. De esta manera se podrían obtener más datos y así poder contar con un análisis más completo de la inundaciones en el Uruguay.

Además, se podrían agregar más fuentes de noticias, principalmente de otros países, sería una mejora interesante, pues permitiría ver la importancia que se le dio a los eventos de la inundaciones en distintos países.

A su vez también quedó pendiente contar con una interfaz gráfica que integre la parte de extracción automática de datos y las herramientas de análisis estadísticos. Esto se intentó hacer para este trabajo pero no fue posible realizarlo a tiempo.

Conclusiones

En primer lugar, observamos la falta de datos, tanto a nivel de la web como a nivel “oficial”. Como consecuencia, es imposible establecer métricas útiles a la hora de evaluar no sólo datos referidos a las consecuencias de una inundación particular sino acerca del manejo que se hace de la misma. En lo referente a las estimaciones del costo asociado a las inundaciones, si bien entendemos que es algo extremadamente complejo de llevar a cabo dada la cantidad de variables intervinientes, no parece haber un esfuerzo serio por aproximarse a algún número razonable.

Si bien se pudieron recopilar datos de algunas noticias, no son suficientes para hacer un análisis serio ni para poder estudiar una tendencia a lo largo del tiempo. De todas maneras, se logró extraer la información de manera automática y se adquirieron conocimientos de las herramientas y metodología de web scrapping.

Referencias

- https://jsoup.org/
- http://blog.mongodb.org/post/183689081/storing-large-objects-and-files-in-mongodb
- https://www.mongodb.com/document-databases
- http://www.elobservador.com.uy/inundaciones-a2684