Recuperación de Información y Recomendaciones en la Web

Curso 2018

Grupo 15

Zonas de delitos

Autores:

Federico Balarini

Nicolas Ferraro

Mateo Suburu

Emiliano Videla

Supervisor:

Libertad Tansini

November 29, 2018

Contenido

1 Introducción **2**

2 Problema **3**

3 Enfoque de la solución **4**

4 Diseño y Arquitectura **6**

5 Implementación **8**

6 Funcionalidades y uso **11**

7 Evaluación y resultados **12**

8 Conclusiones **14**

9 Trabajo Futuro **15**

1 Introducción

Hoy en día la delincuencia es un tema que se encuentra en boca de todos. Dado su alto grado de presencia en la sociedad y su continuo crecimiento, es común encontrarse a diario con noticias policiales que denotan dicho fenómeno. Sin embargo, esta información con el tiempo se pierde o pasa desapercibida para la gente.

Por ello, resultó de interés resumir y centralizar la información policial disponible en páginas de noticias y proveerla al usuario de una manera más ilustrativa y resumida de modo que le permita conocer mejor las tendencias delictivas en la ciudad de Montevideo, pudiendo tomar las precauciones pertinentes.

2 Problema

El principal problema es la descentralización de la información relativa a incidentes y delitos en Montevideo. Generalmente, este tipo de noticias son las más recurrentes en radio, televesión y diarios, causando que las noticias particulares rapidamente pierdan trascendencia por nuevos hechos posteriores. Actualmente no se cuenta con una herramienta adecuada que cumpla la función de centralizar toda esta información y hacerla disponible facilmente a las personas, ofreciendo una visión más global de la problematica.

El objetivo principal es recuperar las noticias policiales disponibles en la web y extraer la información más relevante (lugar, fecha, tipo) asociada a las mismas, y en base a esta información y haciendo uso de herramientas de geolocalización ir ubicando los delitos en un mapa de forma que sea intuitivo para cualquier usuario poder tener una idea general de las tendencias delictivas.

3 Enfoque de la solución

Como se mencionó anteriormente, se busca poder centralizar información de incidentes y delitos, y poder hacerla disponible a usuarios a través de una interfaz web.

Para ello, se analizaron diversas fuentes de noticias, evaluando algunos aspectos importantes a la hora de utilizarlos como proveedores de datos para la aplicación en cuestión. Algunos de estos aspectos fueron:

- Confiabilidad de la fuente.
- Categorización de noticias, específicamente se valoraba positivamente la existencia de una categoría "Policiales" y la posibilidad de filtrar las noticias por ella.
- Cantidad de noticias dentro de dicha categoría.
- Dificultad al extraer las noticias y sus meta-atributos.

En base a los puntos anteriores, se terminaron utilizando 3 fuentes: Teledoce, El Pais, y Montevideo Comm, dónde la primera es la principal dada la cantidad de noticias que tiene disponible desde un punto de vista histórico, y la última la que toma menos relevancia debido a la incapacidad de obtener noticias únicamente "Policiales" al utilizarla.

Luego de definidas las fuentes de los artículos, el siguiente paso para lograr una solución apropiada fue analizar las noticias obtenidas. Los datos que se busca extraer de cada una de las noticias son:

- Título
- Desarrollo completo de la noticia
- Fecha

- Barrio
- Calles
- Categoría (Hurto, Rapiña, Asesinato, etc.)

Luego de obtener la mayor cantidad posible de información, esta es mostrada al usuario en un mapa de Montevideo en el cual se pueden observar la cantidad de delitos en cada barrio, así como también la información detallada de cada uno de los delitos y su noticia asociada.

4 Diseño y Arquitectura

El sistema cuenta con 3 componentes principales:

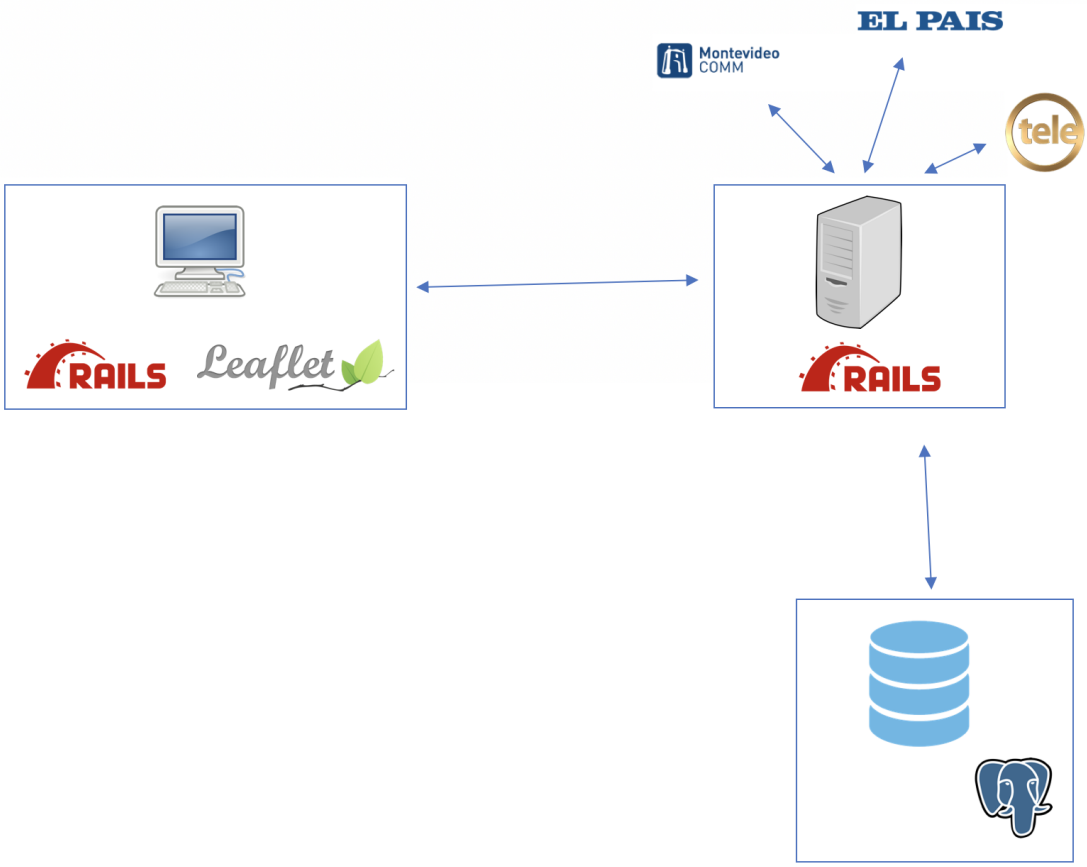
- Base de datos
- Servidor backend
- Interfaz gráfica web

La aplicación se desarrolló utilizando **Ruby on Rails**, un framework muy conocido para desarrollo de aplicaciones web que utiliza Ruby como lenguaje de programación. Este framework tomó mucha importancia en los ultimos años y una de sus principales características es la poca configuración necesaria para crear una aplicación web al utilizarlo, lo que desemboca en relativa rapidez y facilidad en el desarrollo siempre que se sigan ciertas convenciones y buenas prácticas.

Se utilizó la base de datos relacional **PostgreSQL** para persistir los datos extraídos relacionados a los delitos. Rails puede ser usado con una gran variedad de bases de datos pero en este caso fue PostgreSQL la elegida.

El mapa que se visualiza en la interfaz grafica fue implementado con **Leaflet**, una libreria Javascript potente para el trabajo con datos geográficos en la Web. Una de sus principales ventajas es su facilidad para ser integrada con una gran variedad de lenguajes además de contar con una amplia documentación y apoyo de la comunidad.

La arquitectura utilizada es cliente-servidor. El servidor es el encargado de obtener las datos de las fuentes ya mencionadas y almacenar lo necesario en la base de datos. Luego, una vez seleccionado un barrio en la interfaz grafica, el cliente realiza la conuslta al servidor para obtener los delitos de ese barrio.


EL PAIS
Montevideo
COMM
tele
RAILS
Leaflet
RAILS

5 Implementación

Para la etapa de implementación se decidió paralelizar el trabajo, para lo cual se separó el desarrollo de la capa de presentación y la capa de recuperacion y extraccion de informacion.

Para el backend se crearon distintos procesos (rake tasks) ejecutables desde la terminal, para iniciar tanto la obtención de artículos de las distintas fuentes, asi como el análisis de los mismos para extraer la mayor cantidad de información de los mismos. Una vez ejecutados dichos procesos durante cierto tiempo, se cuenta con una base de datos poblada con "Crímenes" y todos los datos relativos a cada uno de ellos.

Para el reconocimiento de barrios, se utilizo el listado de barrios de la capa obtenida en SIG Montevideo para tener una correspondencia 1-1 entre los barrios reconocidos y los que se presentaran ultimamente al usuario. En el caso de las calles, se obtuvo mediante Web Scrapping un listado de la siguiente web oficial:
`http://intgis.montevideo.gub.uy/sit/tmp/v_mdg_nombres_vias_mvd.txt`
Por su parte, para la identificacion de las categorias de los crimenes, se creo manualmente un set de categorias. El set incluye algunas de las distintas formas en que la categoria pudiera aparecer en los articulos, de forma tal que el reconocimiento fuera mas eficaz. Las categorias utilizadas fueron las siguientes: Rehénes, Copamiento, Asalto, Robo, Rapiña, Feminicidio, Asesinato, Violacion, Terrorismo, Fuga.

Para el frontend se tomó como objetivo presentar una interfaz gráfica simple y amigable que permita visualizar los datos de forma gráfica y atractiva, para ello luego de una etapa previa de investigación de herramientas y posibilidades, se terminó decidiendo presentar un mapa interactivo de Montevideo con sus respectivos barrios.

Durante la etapa de investigación se decidió hacer uso de la librería Leaflet, la cual ya fue explicada en detalle en la sección de diseño y arquitectura de la aplicación.

Antes de continuar con el desarrollo de la implementación del mapa es importante aclarar algunos conceptos muy utilizados en el área de información geográfica, mas específicamente, vamos a aclarar los conceptos de:

- Sistema de Coordenadas
- Capas Geográficas
- GeoJson

Sistema de Referencia de Coordenadas : Es quien provee la ayuda necesaria para poder representar un punto de la tierra en un plano según un conjunto de coordenadas.

Capas Geográficas: En el uso de mapas se suelen trabajar con distintas capas de datos, las cuales se pueden ver como distintos planos superpuestos uno encima del otro, cada uno de ellos con distintos tipos de datos representados.

GeoJson: Formato Json con una estructura ajustada para el trabajo de datos geográficos, es muy común encontrarse con capas enteras de datos representados en este formato, y es el que se utilizará para trabajar en este proyecto.

Una vez decidida la herramienta que se utilizaría para el muestreo de los datos, sólo queda obtener los datos y visualizarlos con la ayuda de Leaflet, para ello es importante tener en cuenta que el sistema de referencias en el que se encuentran los datos y en el que está configurado leaflet deben ser iguales.

La capa de datos referentes a los barrios fue obtenida de SIG Montevideo, página con una gran variedad de datos geográficos sobre Montevideo. Luego con la ayuda de QGIS, herramienta de escritorio para el procesamiento de datos geográficos, se convirtió el formato de dicha capa a GeoJson utilizando un sistema de coordenadas igual al configurado en Leaflet (EPG 4326).

Una vez realizados todos estos pasos y ya integrado Leaflet en nuestro proyecto es muy fácil agregar la capa de barrios generada a nuestro mapa, para así finalmente poder asignarle a cada barrio un evento "onClick" con el objetivo de realizar peticiones al servidor, y así obtener los crímenes asociados a dicho barrio (previamente procesados en el backend) y visualizarlos en una tabla.

6 Funcionalidades y uso

La aplicación web posee principalmente dos funcionalidades, una de consulta y otra de validación, en la primera el usuario será capaz de obtener y visualizar crímenes con su respectiva información, mientras que en la segunda se le permitirá al administrador poder releer las noticias, validar la información referente a ellas así como también modificar dicha información en caso de que lo considere pertinente.

Para las consultas, se hace uso de un mapa interactivo de Montevideo con sus barrios delimitados, el usuario tendrá dos funcionalidades básicas de consultas, una en la cual podrá ir visualizando el nombre de cada barrio al posicionar el mouse en cada uno de ellos, y la otra, donde podrá obtener un listado de todos los crímenes realizados en un determinado barrio al hacer clic en dicho barrio, este listado de crímenes se desplegará en una tabla ubicada a la derecha del mapa.

Para la función de validación se realizó otra vista la cual posee una tabla con todos los crímenes del sistema y sus respectivos datos recuperados, esta tabla esta paginada y en ella se podrá modificar cada crimen permitiendo rellenar aquellos campos que no fueron posibles identificar, o modificar aquellos que se identificaron incorrectamente, también será posible marcar un crimen como validado.

7 Evaluación y resultados

Consideramos que los objetivos que nos planteamos inicialmente fueron logrados. Si bien nos hubiera gustado lograr una detección más precisa de las calles de los delitos de forma automática, esto no afecta a la funcionalidad global del sistema ya que los usuarios pueden visualizar en el mapa los barrios con mayor cantidad de delitos según las noticias de la web. Creemos que la solución alternativa implementada para dicho problema, en la cual un usuario "administrador" puede editar y validar manualente los crímenes accediendo al artículo original, fue relativamente buena.

Por otro lado, todo el sistema de recolección y reconocimiento de crímenes y sus datos tiene el gran potencial de que si la aplicación web fuera a dejarse hosteada, se podría dejar todos los procesos automatizados para que se ejecuten, por ejemplo, una vez al día, y así mantener la base de datos actualizada y poder captar las nuevas tendencias delictivas.

Un dato interesante es que se logra reconocer barrios en aproximadamente el 45% de las noticias, mientras que en prácticamente el 100% se reconoce al menos una calle. Analizando este resultado, podemos decir que en el caso de las calles los reconocimientos suelen ser mucho menos exactos, dada la amplitud del "diccionario" de calles utilizado y que los nombres de las calles muchas veces puedan corresponderse a palabras que se usan en un artículo sin hacer referencia a la calle. Por otro lado, notamos que en muchos artículos no se menciona el barrio en el cual se efectuó el crímen, y dentro del porcentaje en que sí se menciona, el reconocimiento fue bastante bueno.

En cuanto a las categorias, las mismas se reconocieron en un 56% de los crimenes, lo que resulta bastante bueno considerando que no siempre se explicita el tipo de

delito y que el set de categorias de delito que usamos es reducido.

Quizás como una evaluación pendiente queda valorar la precisión en el reconocimiento de barrios, calles y categorías de los crímenes con datos más precisos, pero esto probablemente requeriría un estudio basado en aprendizaje automático que podría ser un proyecto en sí mismo.

8 Conclusiones

Se cumplieron los objetivos propuestos de presentar la información de interés al usuario de una forma interactiva y gráfica, además fue posible centralizar la información (de varias fuentes) y persistirla en una base de datos. Permitiendo generar así una colección de datos de fácil acceso, que dispone de varios métodos de filtrado no existentes en las fuentes originales de la información (cómo por ejemplo lo son, los barrios, y las categorías definidas por nosotros).

Si bien se cumple con los objetivos, se tiene una solución básica que tiene mucho lugar a mejora. Teniendo en cuenta que ya se dispone de la información, consideramos importante iterar en el manejo y categorización de esta en trabajo futuro.

9 Trabajo Futuro

Diferentes mejoras y funcionalidades han sido pensadas para este producto como posibles trabajos futuros. No solo han sido consideradas por hacer mas eficiente y performante a la aplicación sino que también mejoran la calidad del producto y la experiencia de usuario.

Una de ellas es agregar Elasticsearch. Elasticsearch es generalemente usado como motor de búsqueda y filtros por ser una herramienta de alta performance en la cual se realizan búsquedas complejas de texto completo. Haciendo uso de ella se agregarian varios filtros en la intefaz grafica de forma de mejorar la experiencia de usuario a la hora del listado de delitos por barrios. Los posbiles filtros a agregar son:

- Tipo de delito (robo, hurto, rapiña, asesinato,...)
- Fecha (ultimo, mes, año,...)
- Calles (filtrar por determinada calle ingresada por el usuario)

Otra de las mejoras propuestas es agregar más fuentes de forma de poder tener mas datos, no sólo para comparar los datos ya obtenidos de otra fuente sino que también se pueden obtener más delitos o ampliar la información de uno ya extraído en otra fuente. Para ello, se pensó en agregar una fuente mas *informal* por así decirlo, como puede ser Twitter o Facebook.

Otra mejora aparte que fue discutida es la idea de *crowdsourcear* el reporte de delitos a usuarios de la aplicación. Utilizando mecanismos parecidos a los que proveen otras aplicaciones, cómo por ejemplo lo hace Waze . Esto permitiría tener datos en tiempo real que el usuario puede visualizar; teniendo la opción de elegir ver los datos obtenidos de fuentes más oficiales, aquellos provistos por los usuarios, o ambos.

Por último, de la forma que fue implementada la solución, hay una gran cantidad de usuarios que no fueron tenidos en cuenta como pueden ser usuarios de otros departamentos del país. Con esto, la mejora al producto sería poder ampliar la solción para considerar otros departamentos ademas de Montevideo y poder obtener delitos de otras zonas del país.

References

[1] "Ruby on Rails - https://rubyonrails.org/"

[2] "PostgreSQL - https://www.postgresql.org/"

[3] "Leaflet - https://leafletjs.com/"

[4] "Elasticsearch - https://www.elastic.co/"

[5] "QGIS - https://www.qgis.org"

[6] "SIG Montevideo - http://sig.montevideo.gub.uy"