Recuperación de Información y Recomendaciones en la Web (2016)

Grupo 13

Integrantes

Lucas Lusardo	4.575.276-9
Leonel Rosano	5.039.791-0
Santiago Ramirez	4.786.391-4
Santiago Nieves	4.744.970-4
Santiago Noguera	4.937.742-6

Introducción

En el día a día a la hora de visitar nuestros sitios de preferencia para mantenernos informados estamos invirtiendo una cantidad de tiempo considerable que podría utilizarse para otras actividades, por otro lado, muchas veces resulta poco práctico acceder a sitios de contenidos similares para complementar la información que deseamos consumir.
Por estas razones nos volcamos a diseñar un sistema mediante el cual sea posible centralizar en un sitio web información de alguna índole de interés, dado el alcance y duración de esta edición del curso se optó por abordar una única temática, se recolecta y presenta información al usuario final acerca de noticias asociadas a sus cuadros de futbol de interés.

Enfoque de solución

La aplicación pone el foco en que los usuarios se informen puntualmente sobre eventos y noticias asociadas a cuadros de fútbol uruguayos.

La obtención de la información se realiza mediante Web Scrapping sobre sitios de la web, tales como "ovaciondigial.com.uy" , "auf.org.uy" y “tenfield.com.uy”.
Desde la página de tenfield se obtienen las noticias relativas a cada cuadro, desde la de la auf las fechas y por último desde ovación se obtiene el fixture de la próxima fecha, el cual los días viernes es enviado por email a los usuarios.

Cabe destacar que si bien se está centralizando la información a partir de tres fuentes distintas como son auf, tenfield y ovación, se optó por obtener información parcialmente centralizada, es decir, en vez de ir a buscar noticias a la página de un club específico se irá a la de tenfield, esto con el fin de mejorar la performance y disminuir el costo de la aplicación, si bien es notorio que los cambios realizados en dicha página impactaran directamente en nuestro sistema, lo cual puede incidir en problemas de disponibilidad.

Diseño e implementación

El diseño del sistema corresponde a un proyecto Ruby on Rails, se usó rails 5.0 y ruby 2.3.1, para el frontend de la aplicación se usó angular js.

Para hacer el scrapping de la webs se utilizó la gema nokogiri la cual de manera sencilla resolvía la conexión a la web y la solicitud de la información en el sitio, brindando luego cierta interfaz a partir de la cual se pueden buscar los datos deseados.

```
2  doc = Nokogiri::HTML(open("http://www.threescompany.com/"))
```

```
1  @doc.css("dramas name").first # => "<name>The A-Team</name>"
```

El framework Rails brinda un ORM (Active Record) el cual usamos con una base de datos mysql 5.6, en la misma alojamos información correspondiente a tres entidades: Usuarios, Cuadros, Estadios.

De los usuarios se sabe su email y cuadros seleccionados además de si éste desea recibir un email semanal con información sobre la próxima fecha.

Sobre los cuadros se cuenta con los nombres de los mismos, algunos alias necesarios para poder hacer scraping en diferentes páginas y las páginas de la cuales se busca información sobre el cuadro.

De los Estadios se aloja la información correspondiente a su ubicación y su nombre.

La información de las fechas de cada cuadro es extraída mediante web scraping desde la página de la auf, los tweets son embebidos mediante un iframe y las noticias se obtienen mediante web scraping sobre la página de tenfield correspondiente al cuadro en cuestión.

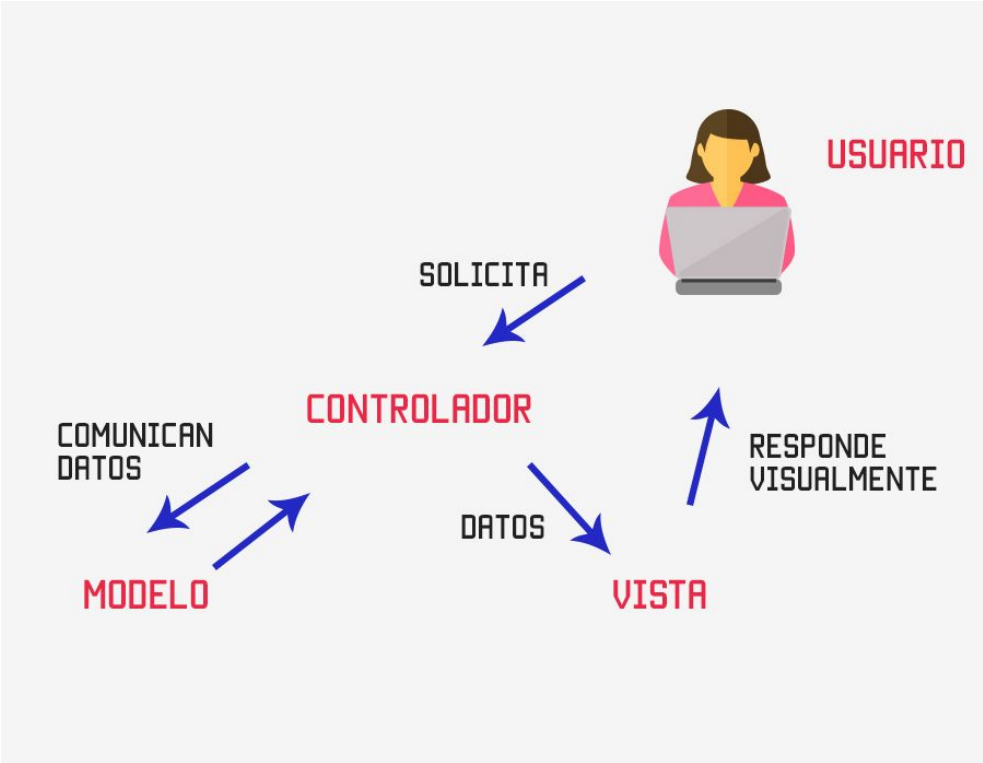
Por último para armar los emails que se envían semanalmente se saca la información ovación digital, portal deportivo uruguayo.

Para el envío de los emails se configuró una crontask que de forma automática manda el email con los datos de la próxima fecha a los usuarios suscritos.

1. Arquitectura del sistema

La arquitectura es cliente servidor, el cliente basado en javascript y html apoyándose en angularjs. El servidor como ya fue mencionado es Ruby, y es donde se hace el scraping de las webs, por último la base de datos es mysql 5.6.

Se utilizó el MVC o Modelo-Vista-Controlador el cual es un patrón de arquitectura de software (sugerido por el framework usado), basado en 3 componentes (Vistas, Modelos y Controladores) claves, dicho patrón separa la lógica de la aplicación de la lógica de la vista en una aplicación. Es una arquitectura importante puesto que se utiliza tanto en componentes gráficos básicos hasta en sistemas empresariales.



2. Tecnologías utilizadas

- Ruby
- Rails
- HTML
- Javascript
- Angularjs
- Nokogiri
- MySQL

Dificultades

Las principales dificultades encontradas mediante el desarrollo del proyecto fueron:

- La página de cada cuadro tiene formatos distintos
- Se requiere de un gran trabajo de procesamiento de lenguaje natural para el desarrollo de la aplicación.
- Es necesario definir de una manera correcta en cuales momentos se saldrá a buscar información por otras páginas y en que momento se utilizará una cache, para no degradar el rendimiento del sistema, con motivo final de obtener un feel del sistema positivo.

Conclusiones

Luego de realizar las pruebas y medir los tiempos con los que se obtienen las respuestas, podemos concluir que Scrap es una herramienta muy poderosa cuando se trata de navegar sobre estructuras HTML de sitios web y obtener información de ellas.

Uno de los problemas que ocurre en este tipo de aplicación es que, en algunos casos, para hacer referencia a un mismo club existen varias maneras (abreviaciones, sobrenombres, con o sin tilde, etc.). Por ejemplo, si en la noticia aparece el nombre “C.N.deF.”, se debe considerar que se está haciendo alusión al Club Nacional de Football. Es decir, existe un gran trabajo de curado de la información para que la aplicación sea útil.

Otro problema es que cada página tiene su propio código HTML, diferente de los demás, por lo que hay que generar código diferente por cada una de ellas, lo que conlleva un costo asociado. Además existe la posibilidad de que las webs de dónde se obtiene la información modifiquen la forma en que muestran la información, y esto hace que sea necesario modificar la forma en la página estaba siendo parseada.

Dado que se obtuvo un prototipo satisfactorio, cumpliendo con nuestros objetivos, podemos concluir que se puede desarrollar una aplicación de gran porte y genérica, tomando como base dicho prototipo.

Trabajo a futuro

Sería interesante y con proyección en el mercado continuar el desarrollo de la aplicación de modo más general, apuntando a centralizar información de interés para el usuario, la cual es consultada periodicamente, como podrían ser resultados de encuentros deportivos, recitales, cotizaciones de las monedas, horoscopo y algunas otras, en donde al momento de actualizar la información en la web marcada como deseada por el usuario ésta también sea enviada a través de un único correo electrónico periódico (típicamente diario), de manera de enviar en una única instancia el máximo de información actualizada posible.
Teniendo entonces una amplia variedad de categorías de información seleccionables por el usuario, en donde el conjunto de información elegido por un usuario pueda ser modificable en el tiempo, lo cual da la flexibilidad necesaria para satisfacer a distintas clases de usuarios y sin duda alguna aporta a optimizar el tiempo de éstos.

Manual de usuario

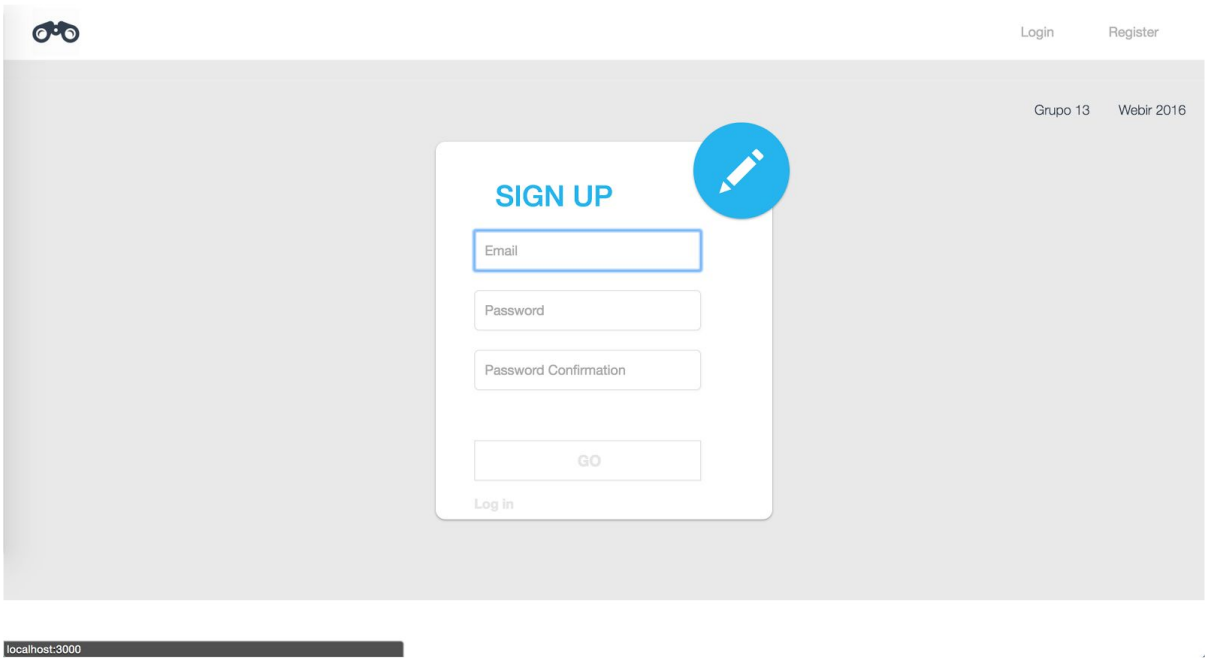
El flujo básico consta de un registro en la aplicación, con posteriores ingresos, mediante los cuales es posible seleccionar los cuadros sobre los que es de interés estar informado.

Por otro lado una pantalla en la cual se despliega la información actualizada de los cuadros de interés, como las fechas de los cuadros con su fecha y ubicación en un mapa y los últimos tweets oficiales de los cuadros. Además se muestran las últimas noticias de cada cuadro seguido obtenidas de tenfield.uy .

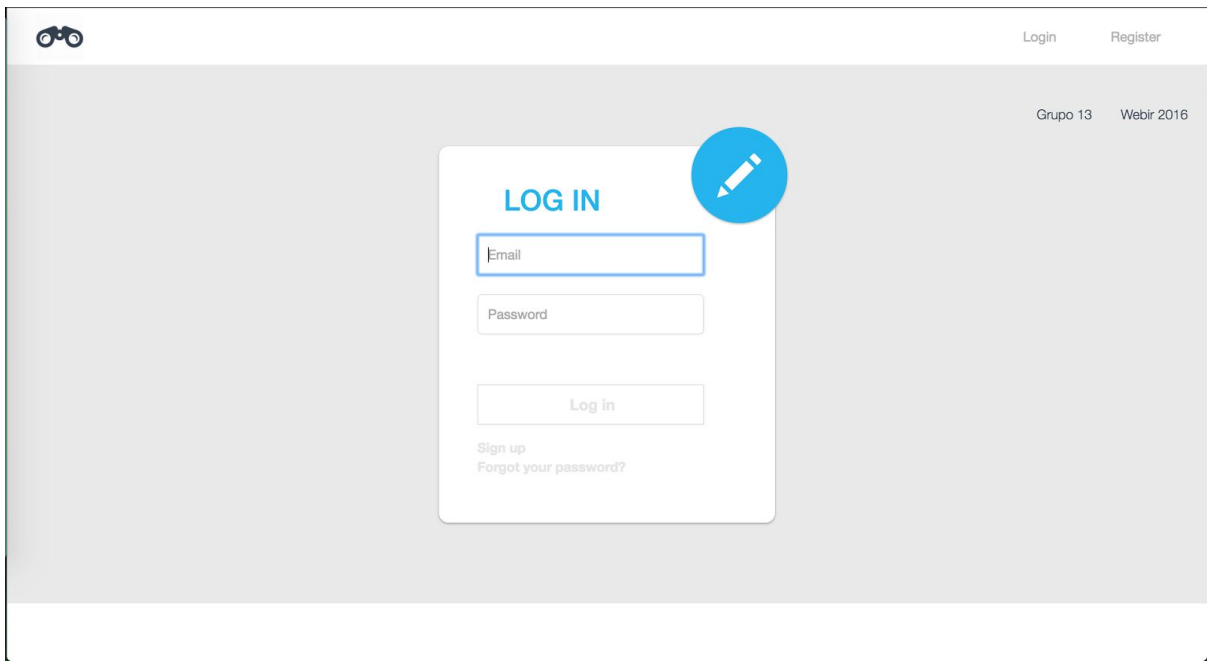
Finalmente los usuarios también reciben notificaciones vía email conteniendo los horarios y ubicación de la fecha correspondiente a la semana en curso.

A continuación se muestra el flujo de la aplicación mediante capturas de pantalla de la misma.

Registro:



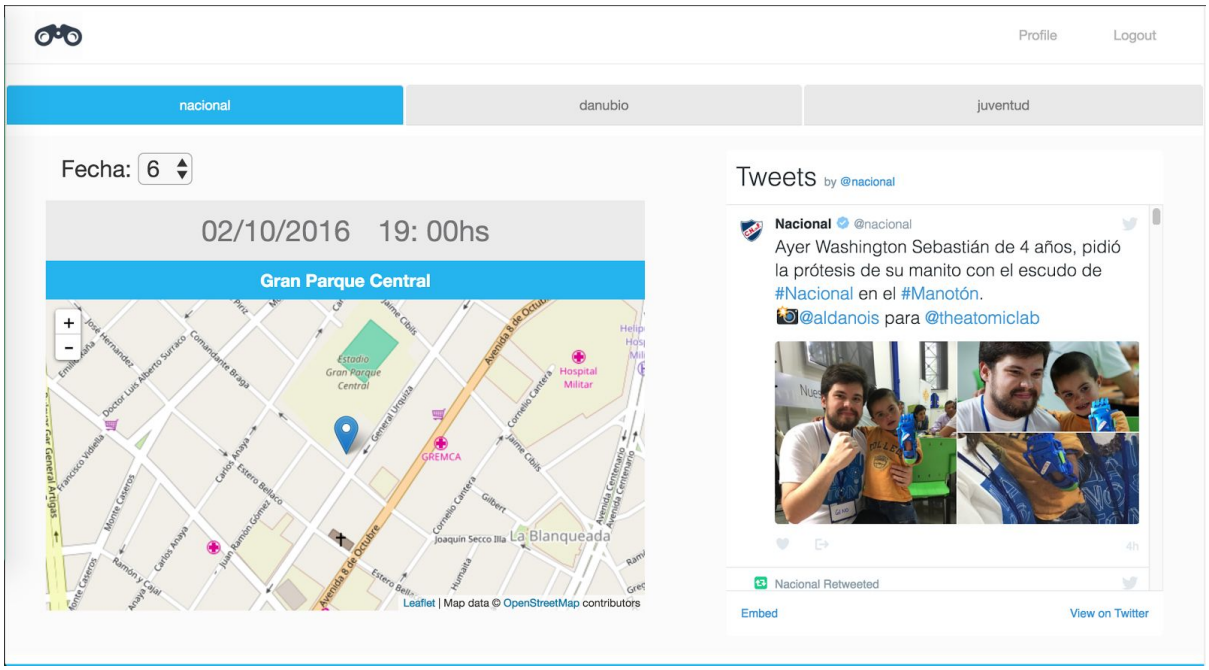
Login:



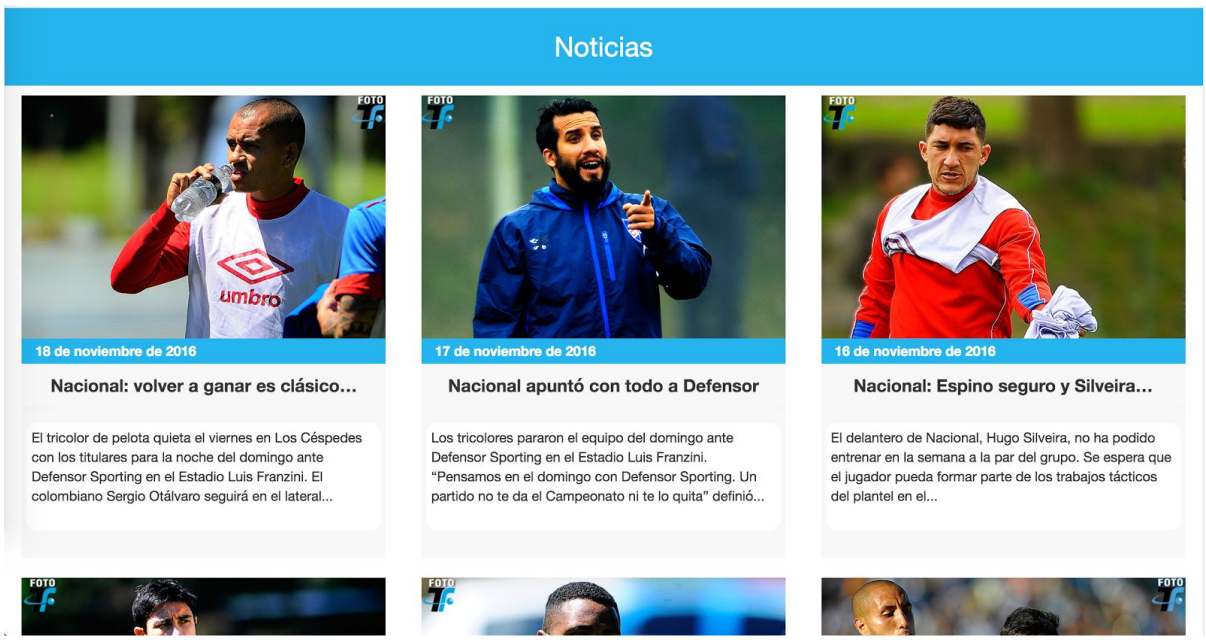
Registro:


Profile
Logout
Seleccione los cuadros que quiere seguir
nacional
penarol
danubio
wanderers
liverpool
cerro
boston_river
juventud
fenix
racing
sud_america
rampla_juniors
defensor
plaza_colonia
river
villa_espanola
Listo
Grupo 13
Webir 2016

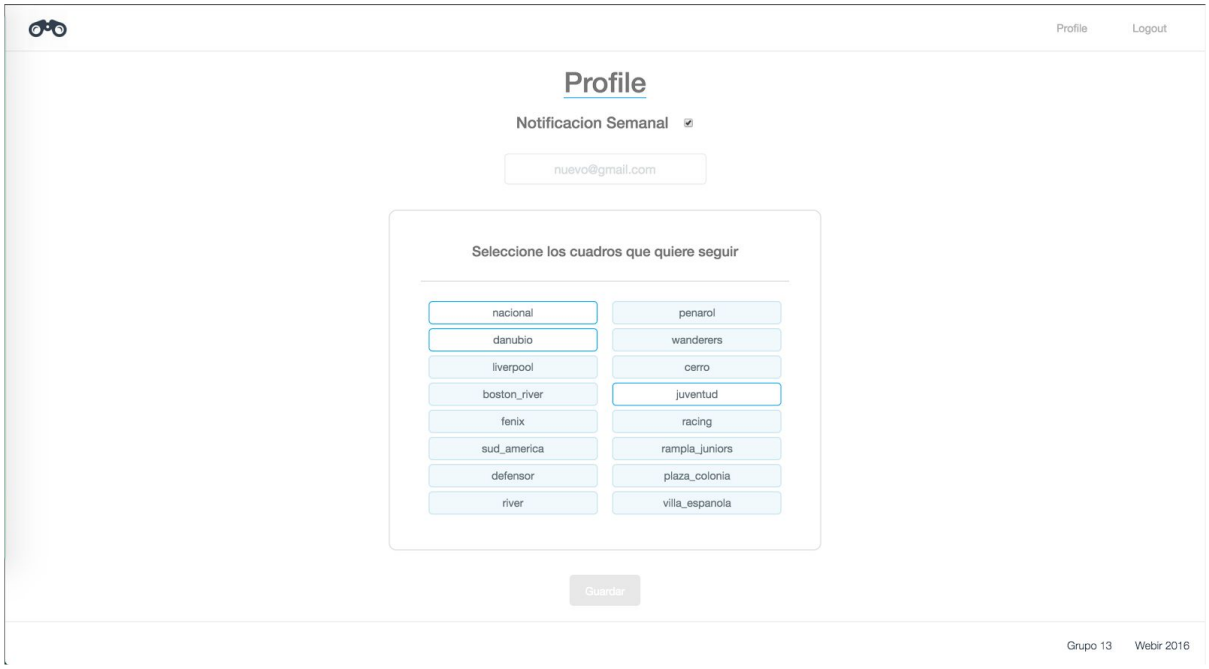
Vista Principal:



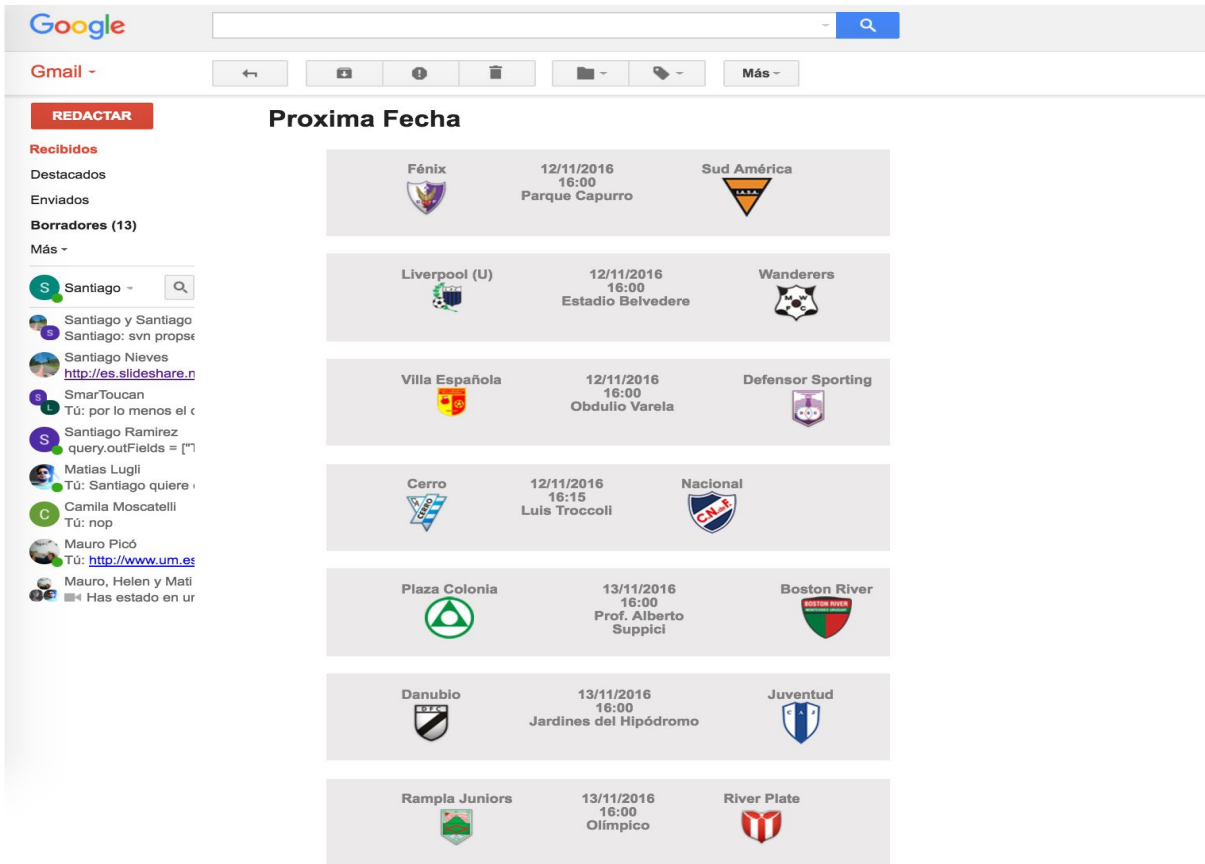
En esta vista el usuario puede elegir entre los cuadros que eligió seguir y ver de ellos, la información de cada fecha (ubicación y horario) un timeline de la cuenta oficial de Twitter de cada cuadro y por último como se muestra en la siguiente imagen, una sección de noticias extraídas en tiempo real de la página de tenfield del cuadro correspondiente.



Perfil de usuario:



En esta vista se pueden actualizar los cuadros que se desean seguir, cambiar el email y seleccionar si se desea recibir una notificación semanal por email. De seleccionar esta última opción el usuario recibirá todos los viernes un email con los datos de la próxima fecha como se muestra a continuación.



References

- Ruby. Retrieved November 20, 2016, from https://www.ruby-lang.org/en/
- Ruby on Rails . Retrieved November 20, 2016, from http://rubyonrails.org/
- Nokogiri. Retrieved November 20, 2016, from http://www.nokogiri.org/
- AUF - Asociación Uruguaya de Fútbol. Retrieved November 20, 2016, from http://auf.org.uy/
- TENFIELD . Retrieved November 20, 2016, from http://www.tenfield.com.uy/
- Ovación Digital. Retrieved November 20, 2016, from http://www.ovaciondigital.com.uy/