Facultad de Ingeniería, Universidad de la República Oriental del Uruguay

Popularidad en los portales web en Uruguay

Proyecto de fin de curso 2014 de la materia

Recuperación de Información y Recomendaciones en la Web

Web del proyecto
webir.mcarbajal.com

Docente
Libertad Tansini

Alumnos
Mauricio Carbajal 4255190-6
Agustín Young 4467691-2

Introducción

En el contexto de la materia, nos proponemos como proyecto extraer información de internet, procesarla, crear nuestra propia base de datos para luego brindar al usuario la posibilidad de hacer consultas sobre dicha información.

Motivación

Queremos conocer la popularidad de los distintos actores de nuestra sociedad. Poder medir y comparar la ponderación que ellos tienen en los diferentes portales locales. Identificar y observar la evolución de las tendencias, y la estabilidad de las mismas a lo largo del tiempo.

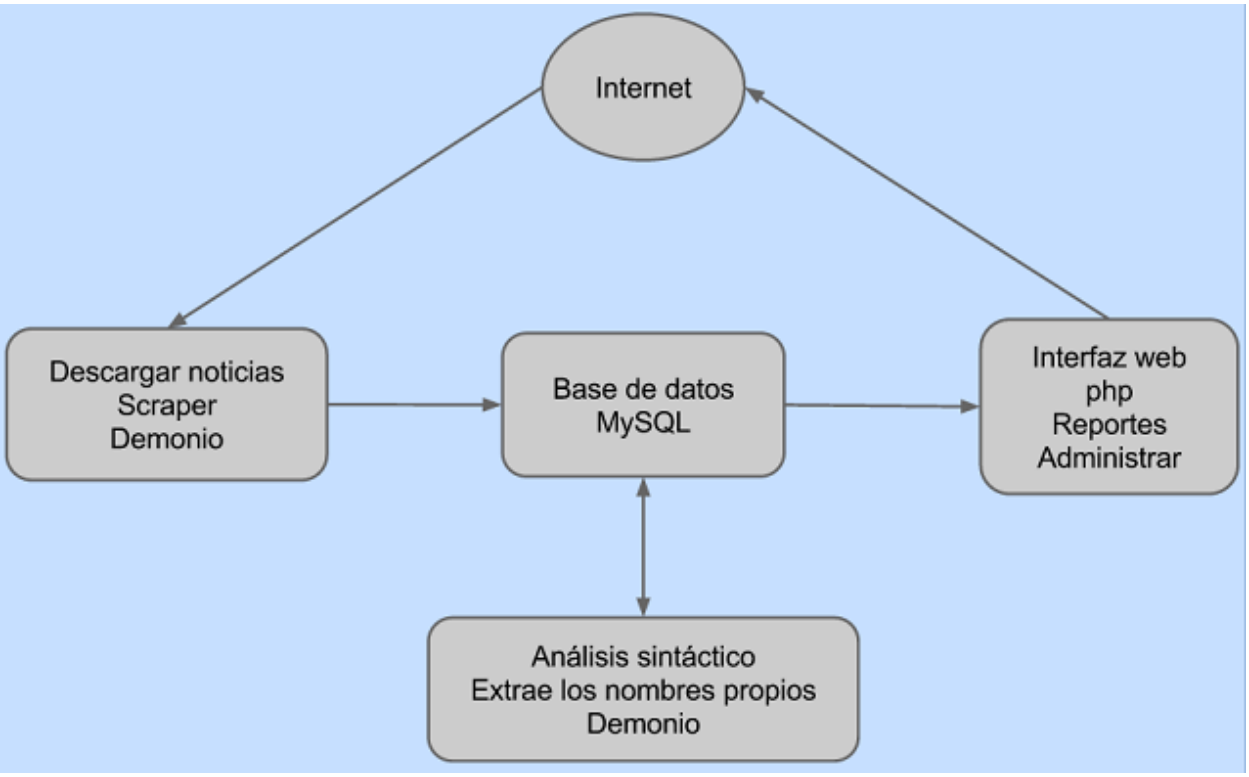
Sobre la Herramienta

Hemos diseñado una herramienta que se ocupa de obtener de forma semi-automática todas las noticias de las portadas de los principales diarios de Uruguay y guardarlas en una base de datos relacional.

Como segunda etapa, un analizador sintáctico recorre las noticias disponibles en la base de datos, y a través de analizar su título y su portada, obtiene los grupos nominales (nombres propios, etc), y se representan como una mención asociada al portal y a la fecha de aparición.

La tercera etapa es brindar una interfaz web **webir.mcarbajal.com** que permite al usuario visualizar las menciones más frecuentes, permitiendo diversas consultas (considerando los distintos diarios, y las distintas fechas de aparición)

Con el fin de construir una solución escalable se plantea un esquema de arquitectura distribuida en cuatro módulos:

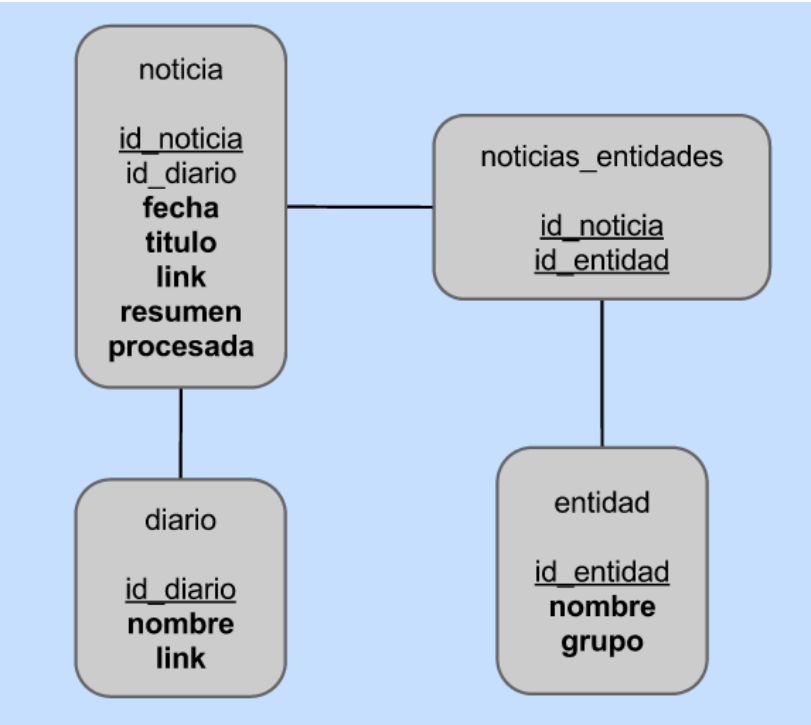


Descargar noticias

El módulo encargado de descargar las noticias lo realiza por medio de técnicas de scraping. Este fue programado con el framework Scrapy para el lenguaje Python.
A nivel de framework se definieron cuatro spiders para capturar las noticias de las portadas de los diarios El País, El Observador, La Diaria y La República. Luego se definió un único pipeline que envía la información a la base de datos MySql controlando previamente la existencia de la misma y en caso de contar con mayor información la actualiza. Esto último se implementó de esta forma debido a que en varios casos (el más evidente el portal de La Diaria) la noticia era mencionada en la parte superior, y luego en la misma portada existe la misma referencia pero con información descriptiva.
Para ponerlo a funcionar se recomienda dejarlo ejecutando de modo desatendido (por ejemplo dentro de un screen o en el crontab de linux), actualmente alcanza con ejecutar el script scrapyTodos.sh y este hace la actualización cada una hora. Esta ingesta permite redundancia con varios nodos funcionando concurrentemente (de hecho, durante este mes estuvo corriendo en varias computadoras personales simultáneamente).

Base de datos

Se utilizó el motor MySQL alojado en el hostingHostGator. A continuación una esquema de las tablas:



La tabla diario esta cargada previamente con los diarios que se esta trabajando.

La tabla noticia es cargada por el modulo DescargarNoticias y se dejan con la bandera procesado apagada.

El analizador sintáctico recorre las noticias que están marcadas como no procesadas, identifica entidades en la misma, si no existen en la tabla entidad les da el alta, y finalmente agrega la relación entre noticia y entidad en la tabla noticias_entidades.

Análisis sintáctico

El proyecto tiene un componente de procesamiento del lenguaje natural. A partir de un texto, necesitamos distinguir cuáles son las entidades allí mencionadas.

Esto no es una tarea menor, ya que en primer lugar no todas las palabras se relacionan a lo que la noticia refiere.

Veámoslo con un ejemplo. Texto a procesar: *"Abaratamiento del petróleo permite a ANCAP recuperar sus finanzas y dejar atrás el fuerte déficit"*

Muchas de ellas son elementos gramaticales (preposiciones, determinantes, etc) que tienen una función dentro de la oración, pero no nos dan ninguna pista sobre la noticia.
En la oración: *del, a, sus, y, el*

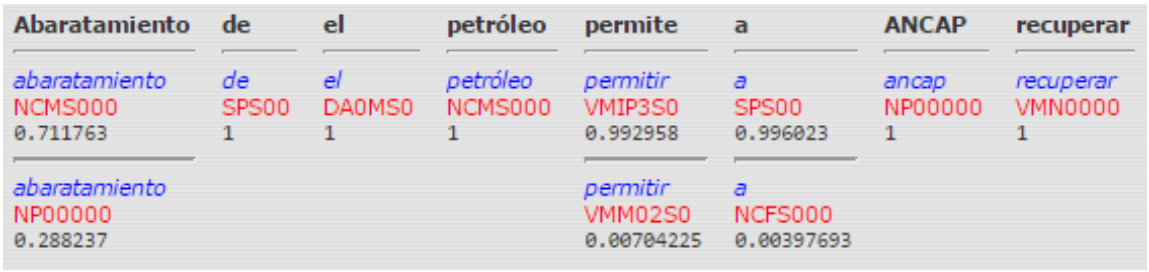
Por razones de alcance del proyecto, nuestra estrategia es bastante simple, y consiste identificar los grupos nominales de las oraciones, y tomarlos como unidad para representar un actor o un tema del cual se está hablando.

*"**Abaratamiento** del **petróleo** permite a **ANCAP** recuperar sus **finanzas** y dejar atrás el fuerte **déficit**"*

En principio, esta oración aportaría una mención al abaratamiento (en general), una al petróleo, una a ANCAP, una a las finanzas en general y otra al déficit.

Para mejorar esto, nosotros ponderamos los nombres propios (en este caso ANCAP) cinco veces más que los otros sujetos que puedan aparecer. De esta forma, controlamos que realmente tengamos estadísticas de los actores, pero también permitimos encontrar temas que si bien no son actores, pueden estar siendo centro de atención en un cierto momento (ej. las finanzas).

La forma de identificar esto de forma automática, es utilizando el analizador sintáctico “freeling”. A continuación podemos ver de qué forma realiza la identificación, con la oración de ejemplo:



Abaratamiento	de	el	petróleo	permite	a	ANCAP	recuperar
abaratamiento NCMS000 0.711763	*de* SPS00 1	*el* DA0MS0 1	*petróleo* NCMS000 1	*permitir* VMIP3S0 0.992958	*a* SPS00 0.996023	*ancap* NP00000 1	*recuperar* VMN0000 1
abaratamiento NP00000 0.288237				*permitir* VMM02S0 0.00704225	*a* NCFS000 0.00397693		

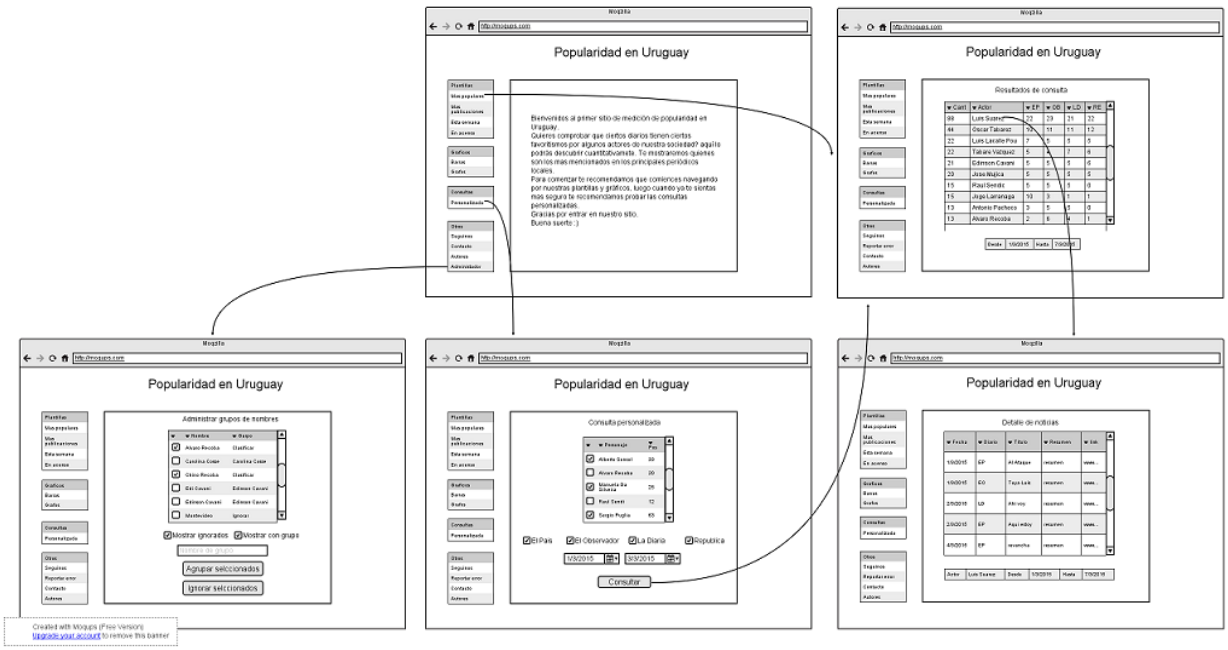
sus	finanzas	y	dejar	atrás	el	fuerte	déficit
su DP3CP0 0.999692	*finanzas* NCFP000 1	*y* CC 0.999962	*dejar* VMN0000 1	*atrás* RG 0.993421	*el* DA0MS0 1	*fuerte* AQ0CS0 0.916667	*déficit* NCMS000 1
sus I 0.000308452		*y* NCFS000 3.76761e-05		*atrás* I 0.00657895		*fuerte* NCMS000 0.0833333	

Donde lo etiquetado en rojo es lo que define las categorías gramaticales de la palabra. Son definidas en función de la palabra, y también del contexto en el que está usada.

Son de nuestro interés, las palabras etiquetadas con “NP*****” (5 pts) y “N******” (1 pts)

Interfaz Web

Se realizó un diseño de producto mínimo viable por medio de una herramienta de Mockups con el objetivo de convenir el funcionamiento general y flujos principales de sitio a construir.



Se puede acceder aqui a estos mokups ingresando en esta url https://moqups.com/#!/edit/aguyoung@gmail.com/5OxrwGD0

La idea fue definir, antes de comenzar a programar, un conjunto de plantillas que el administrador va cambiando a lo largo del tiempo según los temas que resultan más interesantes. Estas mostraran un ranking con el recuento de noticias en cada portal para un conjunto de entidades en un periodo de tiempo. Al momento de la entrega estan las plantillas ‘Ranking Semanal’, ‘Más populares del mes’, ‘Semana de Luis Suárez’ y ‘Candidatos al balotaje’.

Por otra parte se podrá hacer una consulta personalizada donde el usuario puede seleccionar un conjunto de entidades y un periodo de tiempo y obtendra un ranking similar al explicado anteriormente.

Una vez en el ranking al que se puede llegar tanto por plantillas como por la consulta personalizadas, el usuario puede seleccionar una entidad y obtendrá una nueva tabla con el detalle de las noticias en que aparece dicho actor con toda la información, como tituto link, etc.

Se creó una sección de administración con **password “webir2014”**. Aquí se agrupan las entidades que representan el mismo individuo. Por ejemplo “Luis Suárez” y “Lucho Suárez” en realidad pertenecen al mismo grupo Luis Suárez. Los resultados siempre muestran grupos, esto quiere decir que en la noticia aparecerá alguna de las formas que se usa para referirse a ese individuo.

El frontend se encuentra publicado en **webir.mcarbajal.com**

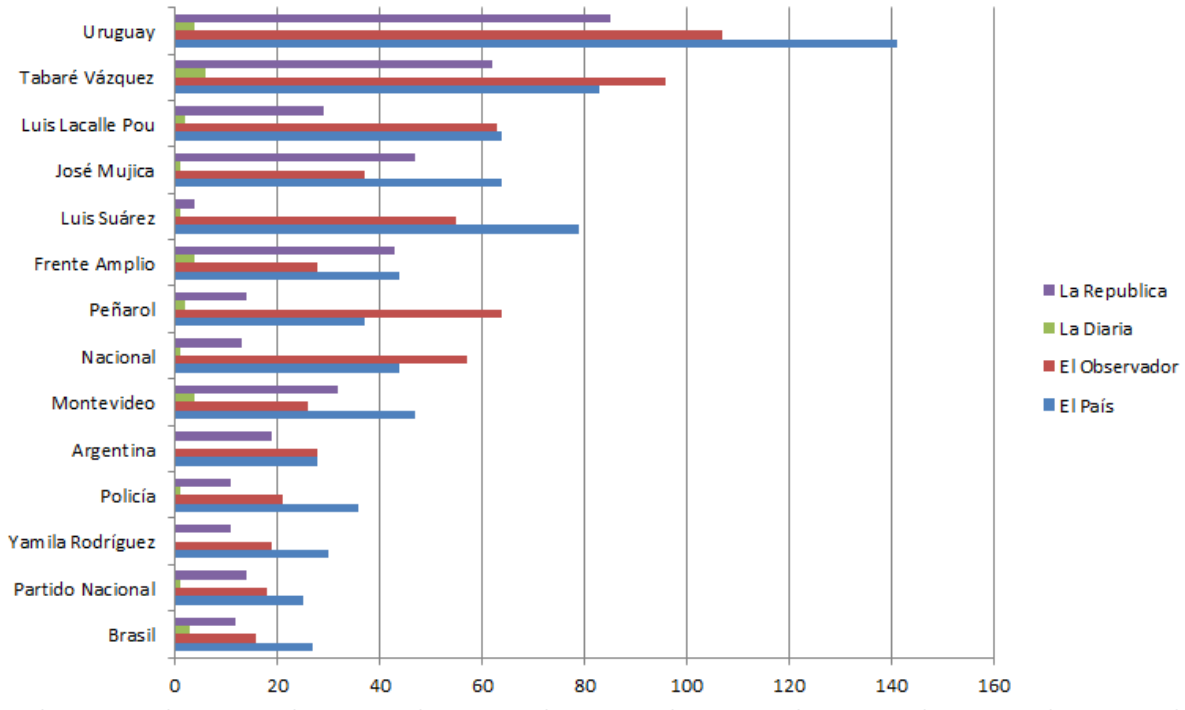
Resultados

Gracias a las técnicas de scraping y de procesamiento de lenguaje natural, en un periodo de un mes, se consiguió construir una base de datos de tamaño considerable (8.385 noticias, ver imagen siguiente)

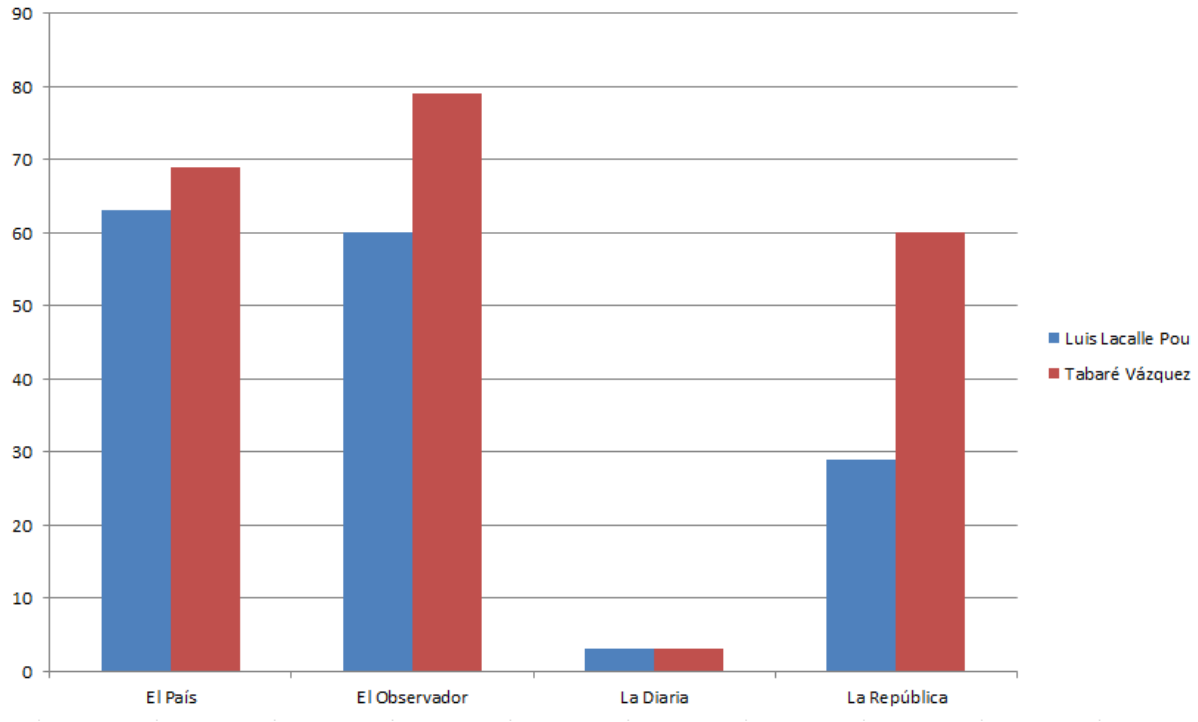
Entidades/Actores	Menciones	Noticias	El País	Observador	La Diaria	La República
5252	13280	8385	4772	2301	196	1116

El objetivo es que nuestra información sea punto de partida para investigaciones. Entendemos que fácilmente se puede obtener información de relevancia. A continuación damos cinco ejemplos visuales que se pueden generar rápidamente a partir de los datos que nuestra herramienta proporciona.

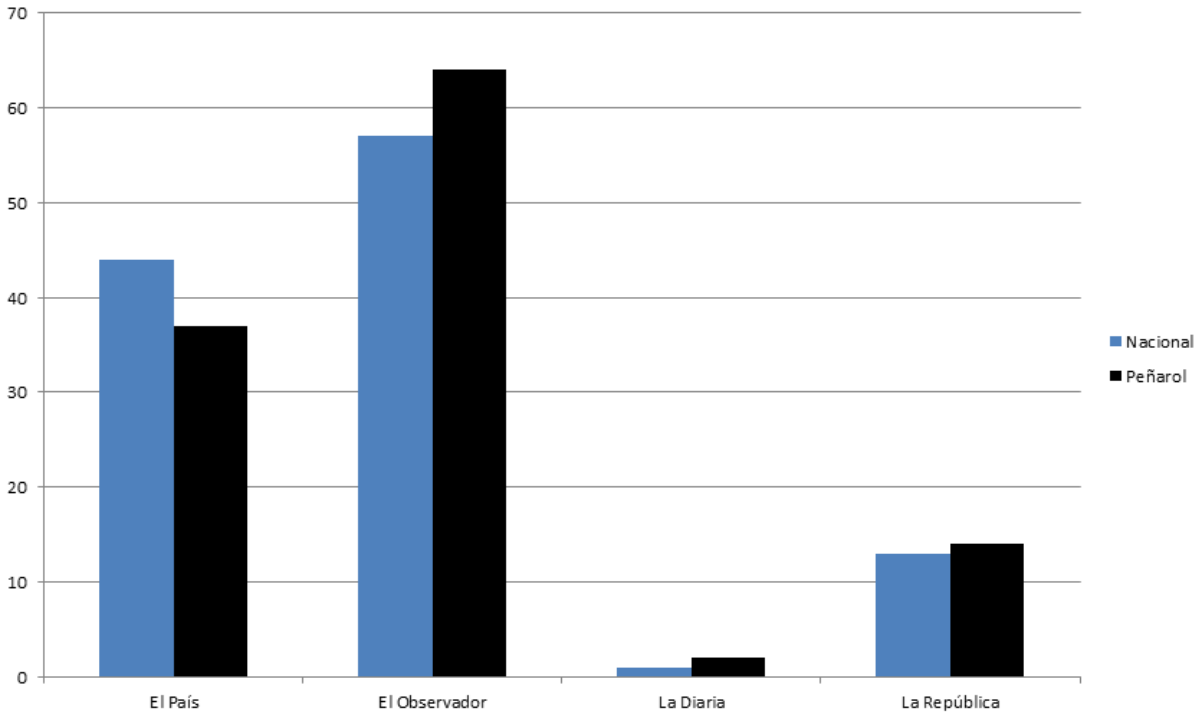
Entidades más mencionados durante un mes discriminado por diario



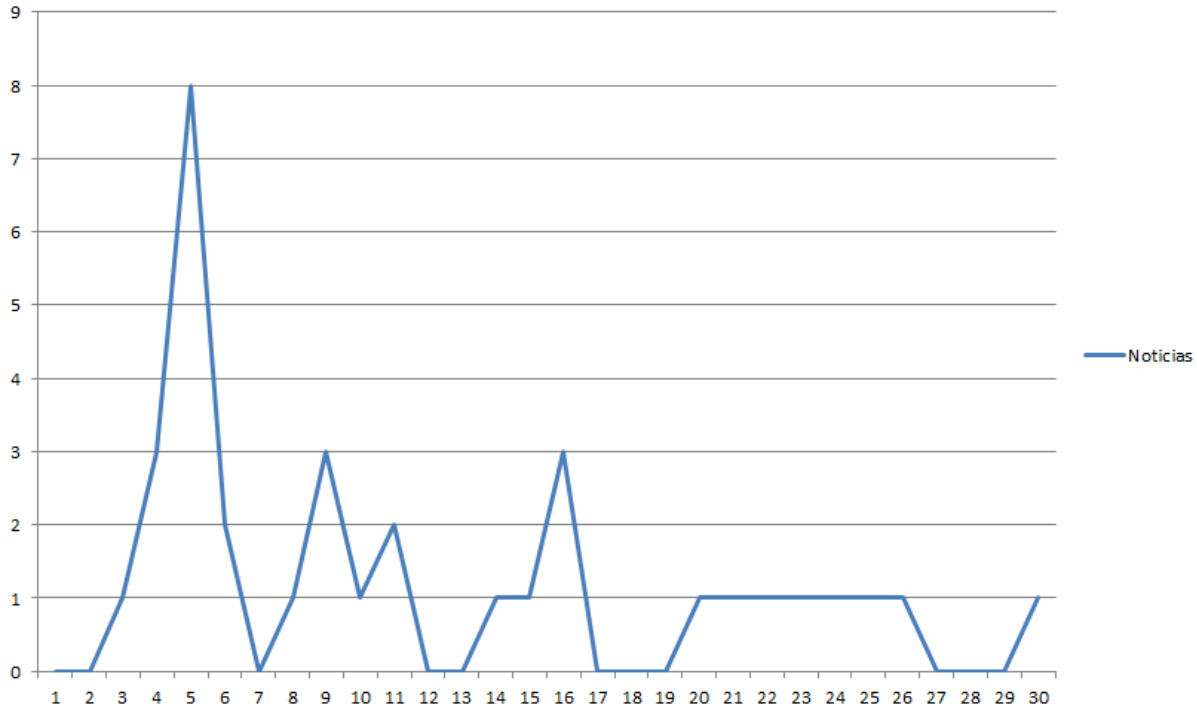
Menciones de los candidatos a presidente durante el periodo de balotaje



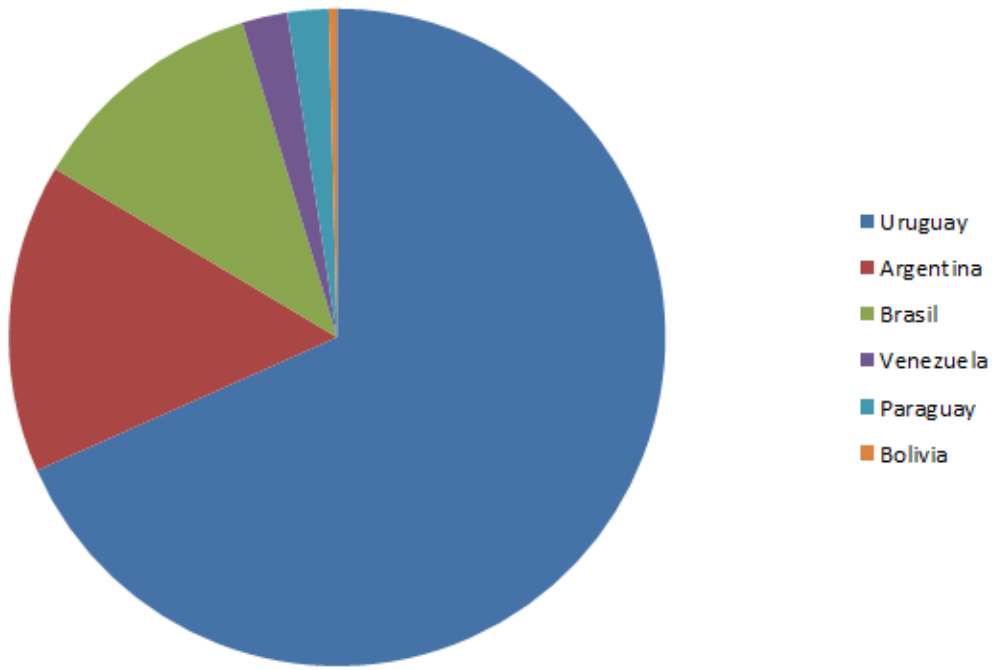
Menciones de Nacional y Peñarol durante un mes



Menciones de Barack Obama durante Noviembre



Cantidad de noticias de cada país del Mercosur en un mes



Conclusiones

Como resultado del proyecto, no solamente se logró explorar cuáles eran los actores más mencionados en cada portal (primer objetivo), sino que pudimos implementar una interfaz sencilla, disponible online, y presentar la información de una forma más amigable, así como permitir consultas personalizadas y administrar los actores para resolver ambigüedades. Para lograrlo, aprendimos a utilizar varias herramientas y lenguajes (detallados más abajo), por lo cual fue un trabajo bastante formativo para nosotros. Comprobamos experimentalmente la importancia y potencialidad que tienen las técnicas de scraping.

El sistema completo se encuentra versionado en un repositorio git sobre bitbucket, y los módulos pueden ejecutarse en cualquier máquina que cuente con las librerías adecuadas. Nuestro proyecto requiere ser monitoreado, los formatos de los sitios y las entidades cambian en el tiempo.

Creemos que el volumen de datos procesados fue suficiente para que la información que brinda el sistema esté de acuerdo con la sensación que el usuario tiene sobre "de qué se está hablando". Los resultados que obtuvimos respecto a los actores más mencionados durante noviembre, a decir verdad, no nos sorprendieron demasiado. Esto fue una forma de verificar que cada módulo estaba haciendo su trabajo de forma correcta, ya que si no aparecían Vázquez, Lacalle Pou, Suárez, Peñarol y Nacional en los primeros lugares, definitivamente estábamos cometiendo errores.

En términos generales, consideramos que logramos implementar la idea que habíamos conversado en los primeros monitoreos, y algunos detalles extra que teníamos como metas secundarias.

Más adelante se detallan algunas posibles mejoras a futuro.

Herramientas Utilizadas

En la siguiente tabla se resumen las herramientas y tecnologías utilizadas en este proyecto:

MySQL	Motor de base de datos.
php	Lenguaje de programación para desarrollo web.
Bitbucket	Servidor de repositorio de control de versiones git.
python	Lenguaje de programación de alto nivel.
Scrapy	Framework de Python para extraer. información de páginas web.
HostGator we eat up the competition	Hosting con php y mysql.
Google Drive	Servicio para compartir archivos.
moqups	Página web para hacer mockups online.
FreeLing	Analizador sintáctico.

Posibles mejoras

- Configurar el módulo de scrapy de envio de mails (scrapy.contrib.statsmailer.StatsMailer) para recibir estadísticas de la ingesta.
- Incorporar mas portales, por ejemplo los de los principales noticieros televisados en Uruguay, www.subrayado.com.uy www.teledoce.com.uy www.telenocheonline.com
- Extraer el cuerpo entero de cada noticia y encontrar más relaciones ponderadas como menos relevantes por no estar en el resumen.
- Generación de gráficas y grafos dinámicos.
- En un futuro, si se siguen incorporando fuentes, seria convenente implementar un modulo administrdor de las spiders de scrapy para administrarlas desde un frontend por medio de webservices con scrapyd (http://scrapyd.readthedocs.org/en/latest/overview.html)
- Respecto a la parte de procesamiento del lenguaje natural, sería bueno resolver las correferencias (cuando una palabra en realidad hace referencia a un sujeto mencionado anteriormente, debemos contabilizar de alguna manera especial esa segunda referencia)

Otros trabajos

- Piso Crawler, proyecto open source publicado en github, unifica los portales de inmobiliarias en Uruguay, tales como apartamentos.com.uy, buscandocasa.com.uy, etc.
- Sitio Manuela Lucas, arte visual sobre el comportamiento social de las personas, básicamente generan grafos de relaciones, por ejemplo popularidad en Twitter o Facebook.

Referencias

Definición Scraping
http://es.wikipedia.org/wiki/Web_scraping

Herramienta de Scraping
http://scrapy.org/

Ambiente virtual python
http://virtualenv.readthedocs.org/en/latest/virtualenv.html

Lenguaje que direcciona y parsea los archivos xml,XPath
http://www.w3.org/TR/xpath/

Otra herramienta de Scraping (no utilizada) con interface visual (versión paga y version gratis)
http://scrapinghub.com/ y https://github.com/scrapinghub/portia

Proyecto relacionado, generación de grafos visuales a partir de twitter
http://www.manuelalucas.com/

Proyecto de scraping que unifica portales de inmobiliarias en Uruguay
https://github.com/gdelfresno/pisocrawler

Repositorio de código gratuito privado, bitbucket con git
https://bitbucket.org/

Sitio oficial de lenguaje de programación web php
http://php.net/

Hostgator, hosting contratado
http://marketing.hostgator.com/

Mockups online
https://moqups.com/

Motor de base de datos MySQL
www.mysql.com

Página oficial de Python
https://www.python.org/