UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

SISTEMA PARA RECOLECTAR, ARCHIVAR Y VISUALIZAR LA INFORMACIÓN CONTENIDA EN REDES SOCIALES EN CHILE

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL EN COMPUTACIÓN

JAZMINE ALEJANDRA MALDONADO FLORES

PROFESOR GUÍA:
BARBARA POBLETE LABRA

MIEMBROS DE LA COMISIÓN:
JOHAN FABRY
GONZALO NAVARRO BADINO

Este trabajo ha sido parcialmente financiado por Núcleo Milenio Centro de Investigación de la Web Semántica.

SANTIAGO DE CHILE
OCTUBRE 2015

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERA CIVIL EN COMPUTACIÓN
POR: JAZMINE ALEJANDRA MALDONADO FLORES
FECHA: OCTUBRE 2015
PROF. GUÍA: SR. BARBARA POBLETE LABRA

SISTEMA PARA RECOLECTAR, ARCHIVAR Y VISUALIZAR LA INFORMACIÓN CONTENIDA EN REDES SOCIALES EN CHILE

Las redes sociales virtuales son muy utilizadas hoy en día por usuarios en casi todo el mundo. En ellas se puede interactuar con otros usuarios, se puede compartir experiencias y opiniones. En particular, la plataforma social *Twitter*, es muy utilizada como fuente de noticias, ya que su estructura facilita la difusión de la información. Muchos medios de comunicación tradicionales tienen cuenta en *Twitter*, y además, los usuarios informan a través de esta red los sucesos de los cuales están siendo testigos. Por otro lado, profesionales de diversas áreas están interesados en la información que ésta y otras redes sociales contienen, es por ello que, en distintas partes del mundo se están comenzando a buscar opciones para mantener un registro histórico de lo que se publica en las redes sociales.

El presente trabajo consiste en la implementación de un sistema para mantener un registro periodístico e histórico de la presencia chilena en la red social *Twitter*. El sistema desarrollado recolecta, de forma automática y periódica, contenido Web de carácter noticioso relacionado con Chile desde la red social *Twitter*. Además filtra, ordena y georeferencia la información para finalmente presentarla a través de mapas y de forma interactiva a los usuarios.

Las metodologías utilizadas para recolectar y geolocalizar los datos están basadas en investigaciones previas realizadas por alumnos del Departamento de Ciencias de la Computación de la Universidad de Chile. Estas metodologías fueron integradas, modificadas y extendidas para adaptarlas a lo requerido y lograr su uso efectivo sobre datos en español y para la geografía de Chile.

La puesta en marcha de los procesos de recolección y geolocalización de datos resultó exitosa. Hasta el 17 de Agosto del 2015 se logró recolectar 78.413.724 *tweets* relacionados a 17.850 eventos noticiosos. Del total de eventos 9.895 fueron geolocalizados.

Se desarrolló una primera versión de la aplicación Web que permite al usuario visualizar los datos recolectados, brindándole la posibilidad de interactuar aplicando filtros y explorar la información relacionada a diferentes eventos de interés.

También se realiza una caracterización de los datos recolectados durante el periodo comprendido entre el 1 de Noviembre del 2014 y el 30 de Abril del 2015.

A pesar de que fue necesario modificar las fechas dispuestas en la planificación original, los objetivos planteados se cumplieron.

En Chile no existe actualmente una herramienta similar. Por lo que la herramienta que aquí se presenta se desarrolla con la finalidad de convertirse en un servicio complementario a los que provee la Biblioteca Nacional de Chile sobre archivo de contenido Web.

A mi hijo, Alex, que siempre ha sido y seguirá siendo,
el fin último de todos mis proyectos. ♡

Agradecimientos

En primer lugar quiero agradecer a mis padres ya que me apoyaron en todo momento y siempre confiaron en mi, incluso en los momentos en que yo misma dudé. Me alentaron y no dejaron que me rindiera bajo ninguna circunstancia.

A mi hijo, que toma gran parte de mi tiempo pero que desde chiquitito ha demostrado una capacidad sorprendente para entender cuándo necesito concentrarme y pareciera que en esos momentos se esforzara en respetar lo más posible ese espacio.

También agradezco a mi pololo, Boris Romero, que estuvo junto a mi en todo momento. Él me ayudó en la redacción de este documento, me escuchó en mis momentos de crisis de estrés y siempre que fue necesario tuvo un abrazo contenedor para mi.

A mi profesora guía, Bárbara Poblete, y a Vanessa Peña, que fue como una segunda profesora guía. Ellas me acompañaron a lo largo de este proceso y me brindaron su mentoría y consejos para llevar a cabo exitosamente esta tarea.

Finalmente me gustaría al Núcleo Milenio Centro de Investigación de la Web Semántica que auspiciaron el trabajo realizado.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

1.1. Contexto y Motivación

Desde hace muchos años han existido entidades encargadas de mantener registro histórico de lo que sucede en el mundo. Antiguamente el material almacenado se conformaba principalmente por libros, cartas y periódicos; años después, a lo anterior se le sumó contenido audiovisual y hoy, parte importante del material histórico de interés se conforma por *e-mails* y publicaciones en la Web.

Actualmente existen herramientas para recolectar contenido Web estático. Las organizaciones que recolectan contenido Web las utilizan para guardar las diferentes versiones de los sitios Web de interés histórico. Dos de las más conocidas son *Heritrix*[1] y *Web Curator*[2]. Sin embargo, la Web está cambiando. Muchos sitios son generados dinámicamente y dificultan la tarea de recolección utilizando las herramientas antes mencionadas.

Las redes sociales llegaron para quedarse. Son casos particulares de sitios Web generados dinámicamente y algunas de ellas como *Facebook*[3] y *Twitter*[4] se posicionan como uno de los sitios más visitados por los usuarios de Internet[2].

Twitter en particular, es utilizada por muchos usuarios como fuente de noticias[8]. Es frecuente que los usuarios comunes publiquen en *Twitter* sobre sucesos de los cuales están siendo testigos antes de que la información sea publicada por los medios de comunicación y rápidamente la información se difunde por la red gracias a los *retweets*. En algunos países, como Chile y Estados Unidos, casi todos los medios de comunicación tradicionales tienen una cuenta oficial en *Twitter* a través de la cual publican reportes de las últimas noticias en tiempo real.

Además en las redes sociales se reúnen estudiantes y profesores, trabajadores y empresa-

[1] `http://webarchive.jira.com/wiki/display/Heritrix/Heritrix`
[2] `http://webcurator.sourceforge.net/`
[3] `http://www.facebook.com`
[4] `http://www.Twitter.com`

rios, famosos y fanáticos, acortando distancias y rompiendo jerarquías, lo que las transforma en una parte importante de lo que son las relaciones humanas en la sociedad de hoy en día.

La información contenida en redes sociales podría considerarse como el borrador de un libro de historia, en el que podemos encontrar tanto información objetiva de los sucesos que ocurren día a día, como también opiniones personales sobre esos sucesos. Ambos tipos de información son valiosos ya que permiten saber lo que está sucediendo pero también diferenciar posturas, medir el interés de las personas por un tema en particular, etc.

Aunque el principal uso que se le da a las redes sociales virtuales es el de interactuar con otros usuarios, la información que en ellas se comparte también está siendo utilizada para otros fines. Profesionales de diversas áreas ven en las redes sociales virtuales una fuente importante de información y una oportunidad para aprender más sobre el comportamiento de las personas, es por esto que surge el desafío de generar nuevas opciones para preservar ese contenido.

El trabajo realizado en esta memoria de título consiste en la creación de una aplicación para mantener un registro periodístico e histórico de la presencia Chilena en la red social *Twitter*. El sistema desarrollado recolecta, de forma automática y periódica, contenido Web de carácter noticioso relacionado con Chile desde la red social *Twitter*. Además filtra, ordena y georeferencia la información para finalmente presentarla a través de mapas y de forma interactiva a los usuarios.

En Chile no existe actualmente una herramienta similar. La Biblioteca Nacional de Chile realiza colecciones digitales de algunos sitios Web para algunos eventos específicos, pero hasta ahora no se ha implementado ningún mecanismo para recolectar información de las redes sociales ni tampoco para mantener procesos de recolección automáticos y continuos en el tiempo. La herramienta que aquí se presenta se desarrolla con la finalidad de convertirse en un servicio complementario a los ya que provee la Biblioteca Nacional de Chile.

1.2. Experiencias Internacionales

HOCKX-YU, Helen en su artículo *Archiving Social Media in the Context of Non-print Legal Deposit*[7] presenta la problemática sobre cómo almacenar la información generada en redes sociales virtuales. A continuación se enumeran las iniciativas que menciona en el artículo, las cuales proponen soluciones para mantener registro parcial de la información generada en redes sociales virtuales y sus principales características:

1. *ArchiveSocial* en Estados Unidos[5] : Servicio que permite a empresas y entidades mantener registros completos de su presencia en variadas redes sociales virtuales. Éste garantiza preservar el contenido a pesar de que éste sea eliminado de la red social, ofrece una firma digital que permite utilizar esos registros con fines legales y además ofrece herramientas de filtro, búsqueda y visualización de contenidos.

[5] `http://archivesocial.com/`

2. Archivo Gubernamental de redes sociales en el Reino Unido[6]: Desde 2014 se almacenan sistemáticamente *tweets* y videos de *YouTube*[7] publicados por los departamentos centrales del gobierno del Reino Unido desde su cuenta oficial de *Twitter* y de *Youtube*.
3. *Archive-it*[8] : Servicio que almacena información obtenida desde la Web agrupada por eventos o por organizaciones.
4. Acuerdo entre la Biblioteca del Congreso en Estados Unidos y *Twitter*: El 2010 *Twitter* liberó el acceso de todos los *tweets* públicos generados desde el 2006 a la Biblioteca del Congreso.[4]

ArchiveSocial y el Archivo Gubernamental del Reino Unido tienen un enfoque de selección de información basado en entidades, es decir, se obtiene la información relacionada con una empresa o institución específica. Además recolectan la información mediante el uso de APIs. El servicio que ofrece *ArchiveSocial* es privado y está orientado principalmente para entidades particulares interesadas en mantener un seguimiento de su presencia en redes sociales, mientras que el servicio que provee el Archivo Gubernamental de redes sociales en el Reino Unido es un servicio público el cual mantiene registro de ciertas cuentas seleccionadas por ser de interés del gobierno del país.

La tercera iniciativa *Archive-it*, aunque recolecta información para entidades específicas, tiene un enfoque basado en eventos, es decir, se obtiene información relacionada a un tema en particular que se desarrolla durante un periodo de tiempo definido. A diferencia de las otras iniciativas antes mencionadas, este servicio recolecta información utilizando Heritrix[9], una herramienta para recolectar contenido web estático, junto a una herramienta desarrollada por ellos llamada Umbra[10].

El método utilizado por la Biblioteca del Congreso de Estados Unidos es probablemente el óptimo, ya que, con un acuerdo entre el publicador y las entidades encargadas de resguardar la información se tiene acceso a todos los datos y resulta ser una forma mucho más completa para mantener un registro histórico. Los datos del archivo aún no han sido liberados y la última actualización respecto a esta iniciativa fue publicada el año 2013, en la que informan que están en proceso de buscar la mejor forma para dar acceso a los datos al público general[1].

1.3. Objetivos

1.3.1. Objetivo general

Desarrollar un sistema que permita recolectar, almacenar y visualizar contenido Web de carácter noticioso generado en la red social *Twitter* en Chile.

[6] http://www.nationalarchives.gov.uk/webarchive/twitter.htm
[7] http://www.youtube.com
[8] http://www.archive-it.org/
[9] http://webarchive.jira.com/wiki/display/Heritrix/Heritrix
[10] http://archiveitblog.wordpress.com/2014/03/13/introducing-archive-it-4-9-and-umbra/

1.3.2. Objetivos específicos

1. Recolectar contenido de interés generado en Chile en la red social *Twitter* de forma continua y periódica automáticamente. Este objetivo incluye filtrar y procesar la información para identificar eventos noticiosos que sean interesantes para el registro Web de la Biblioteca Nacional. Se considerarán eventos noticiosos sobre Chile y su análisis de impacto se realizará sobre *tweets* escritos en español.
2. Almacenar la información de forma apropiada, esto quiere decir, que quede registrada de forma ordenada y accesible, de modo que esté disponible para buscar información específica y realizar estudios sobre ella. Todo esto debe cumplirse sin quebrantar la condiciones de *Twitter* respecto al almacenamiento de *tweets*.
3. Ofrecer una forma de visualización interactiva geotemporal que ubique los eventos en el mapa y muestre el impacto generado por la noticia en la red social. La visualización incluirá otro tipo de información relacionada a los eventos como titulares de noticias y/o conjuntos de palabras que caracterizan el evento. La finalidad de la visualización es que los usuarios sean capaces de entender la información rápidamente, realizar análisis generales y ejecutar búsquedas fácilmente.
4. Ofrecer una solución estructurada y escalable, que permita, adaptar la misma solución para otros países.

1.4. Metodología de trabajo

La metodología de trabajo utilizada para desarrollar el proyecto aquí descrito se divide en tres etapas, las que se detallan a continuación:

1. Desarrollo del sistema de recolección de datos:
   - Estudio de la investigación realizada por Mauricio Quezada sobre la detección de eventos en *Twitter* [11].
   - Adaptación del sistema anterior para el caso Chileno y para el idioma Español.
   - Puesta en marcha del sistema recolector.
2. Desarrollo del sistema de visualización de datos:
   - Estudio de la investigación realizada por Vanessa Peña sobre la visualización geotemporal de eventos [9].
   - Implementación de un sistema, basado en la investigación anterior, para etiquetar geográficamente los usuarios y los eventos recolectados.
   - Puesta en marcha del sistema de geolocalización para que se ejecute diariamente.
   - Adaptar las visualizaciones para el caso de Chile.
3. Desarrollo de la aplicación Web: Esta etapa consiste en la implementación del primer prototipo de la aplicación Web que da acceso a la información recolectada. Las funciones de la aplicación Web son:
   - Permite al usuario identificar visualmente los lugares de Chile que se mencionan en el evento.

- Permite al usuario identificar el impacto que tuvo el evento en la red social *Twitter*, medido en base al número de *tweets* relacionados al evento.
- Permite al usuario identificar visualmente a qué lugar del mundo pertenecen los usuarios que *twittearon* sobre el evento.
- Presenta tanto los titulares asociados al evento, como los *tweets* relacionados.
- Permite al usuario realizar búsquedas por fecha y por región.

1.5. Principales Desafíos

Los principales desafíos que presenta el desarrollo de esta memoria son:

1. Construcción de un listado de cuentas de *Twitter* para ser utilizadas como fuentes de noticias Chilenas en el proceso de detección de eventos de impacto en *Twitter*. El listado debe ser diverso y abarcar todas las regiones del país.
2. Adaptación de los algoritmos utilizados en los procesos de limpieza de datos para que funcionen correctamente con datos en idioma español.
3. Adaptación de los algoritmos de geolocalización para que no sólo geolocalicen datos en Chile, sino que, en lo posible, permitan geolocalizar diferenciando regiones y/o ciudades dentro de Chile.
4. Puesta en marcha de los procesos de recolección, procesamiento y geolocalización de datos para que se ejecuten de forma automática y periódica.
5. Adaptación de la visualización de los *tweets* cuidando de no quebrantar los términos y condiciones de uso de los datos de *Twitter*.
6. Adaptación de las visualizaciones geográficas para el caso de Chile.

1.6. Resultados Obtenidos

Se logra cumplir con todos los objetivos propuestos, pero fue necesario realizar modificaciones en los periodos de tiempo destinados a cada parte respecto a la planificación original.

Se pone en marcha tanto el sistema de recolección de datos como los procesos de tratamiento y geolocalización de datos. Con esto se logra recolectar información de 17.850 eventos de interés Chileno acontecidos durante un periodo de 10 meses, de los cuales 9.895 fueron geolocalizados.

Para el almacenamiento de los datos se utilizó un modelo de datos relacional, el que permite explorar los datos y realizar análisis sobre ellos. Además se realiza una caracterización de los datos recolectados durante el periodo comprendido entre el 1 de Noviembre del 2014 y el 30 de Abril del 2015, el que se presenta en el Capítulo 6.

El almacenamiento de datos no quebranta las normas de *Twitter*. En el documento del contrato y políticas de uso de la API REST para desarrolladores de *Twitter* se asume que

los datos recolectados a través de ésta serán almacenados[3].

Se construyó una primera versión de la aplicación Web que permite visualizar información geotemporal de los eventos. En ella se presenta información asociada a los eventos como palabras clave, número de *tweets* asociados, lugares relacionados, los titulares de las noticias a partir de los cuales se identifica el evento y un conjunto de los *tweets* relacionados al evento.

En la aplicación se despliegan los *tweets* respetando las normas de visualización de *tweets* de *Twitter* especificados en [6], pero se espera poder mejorar la técnica utilizada en el futuro, ya que la utilizada es lenta y permite al usuario revisar sólo un conjunto limitado de *tweets* asociados a un evento de forma eficiente.

La solución propuesta se estructura utilizando módulos, lo que permite reutilizar partes del software en proyectos futuros similares. También permite modificar las fuentes de noticias y el idioma de los datos (español o inglés) para un correcto funcionamiento en el caso de que se quisiera adaptar la solución a otros países similares a Chile. El código se mantiene en un repositorio *git* y cuenta con una *wiki* en la que se explica cómo preparar el entorno de desarrollo y los pasos para la puesta en marcha de los procesos.

1.7. Organización del documento

En este documento se presenta el trabajo realizado para desarrollar cada una de las partes que conforman el sistema. En el Capítulo 2 se presenta la arquitectura lógica del sistema en completitud y las funciones de los componentes principales, también se describen las tecnologías utilizadas y los servicios externos que proveen información. En el Capítulo 3 se presenta la primera versión de la aplicación Web, donde se explican las visualizaciones utilizadas y los casos de uso de la aplicación Web. En el Capítulo 4 se presenta la metodología utilizada para modelar eventos noticiosos extraídos de *Twitter*. Además de describe el proceso de implementación del sistema de recolección de eventos haciendo énfasis en lo que hubo que modificar y/o extender. En el Capítulo 5 se explican las metodologías del los procesos de geolocalización de eventos y usuarios. Finalmente en el Capítulo 6 se presentan estadísticas generales de los datos recolectados desde la puesta en marcha del sistema de recolección de datos y una caracterización de los eventos noticiosos chilenos realizada a partir de los datos recolectados durante un periodo de seis meses.

Capítulo 2

Arquitectura del sistema

2.1. Arquitectura lógica

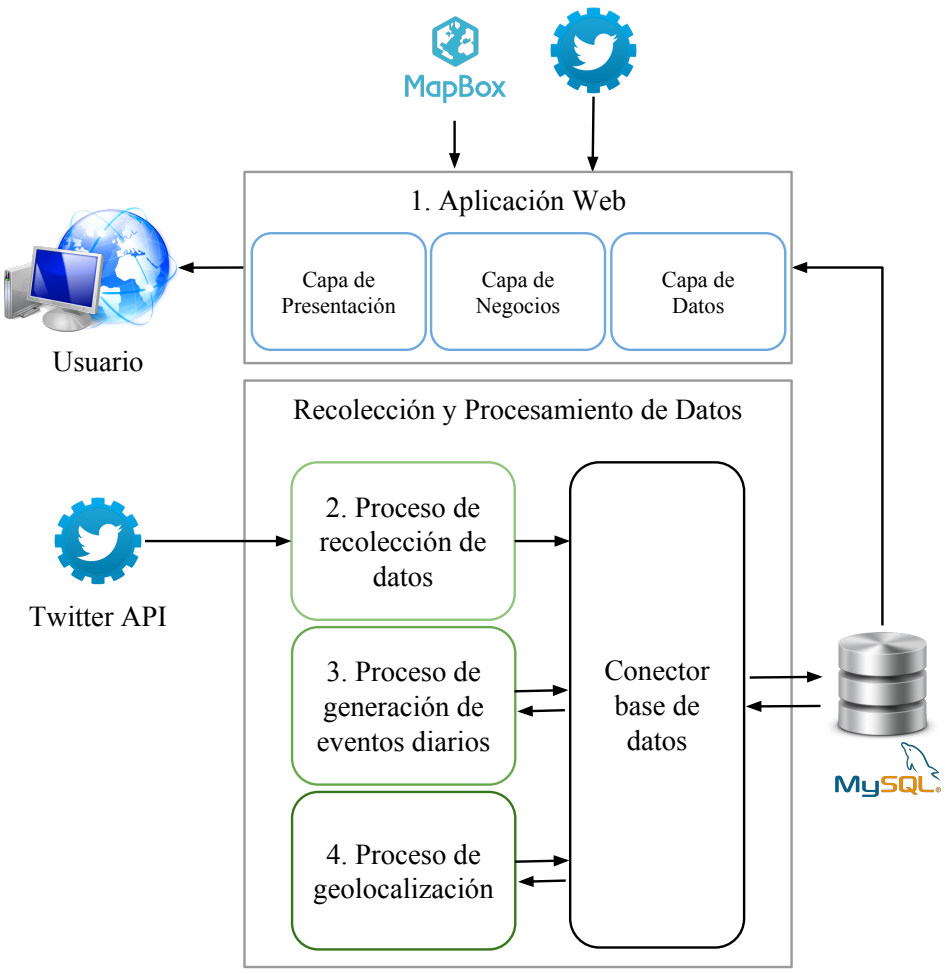


Figura 2.1: Diagrama de componentes del software

El diagrama de la Figura 2.1 muestra la arquitectura y los componentes que conforman el sistema. Éste se divide en dos subsistemas principales, uno correspondiente a la aplicación Web; y otro, compuesto por los procesos que recolectan y procesan los datos.

El subsistema correspondiente a la aplicación Web está diseñado como una estructura de tres capas con responsabilidades definidas, a modo de facilitar la escalabilidad y mantenibilidad de la solución.

El subsistema encargado de recolectar y procesar los datos se construye usando una arquitectura modular. Cada módulo corresponde a un proceso que se ejecuta y se comunica con la base de datos de forma independiente. La comunicación con la base de datos se realiza a través de un conector, el que añade un nivel de indirección para facilitar el trabajo que pueda ser necesario realizar en el futuro a nivel de base de datos. La Figura 2.2 muestra la estructura de la implementación del subsistema de recolección y procesamiento de datos.

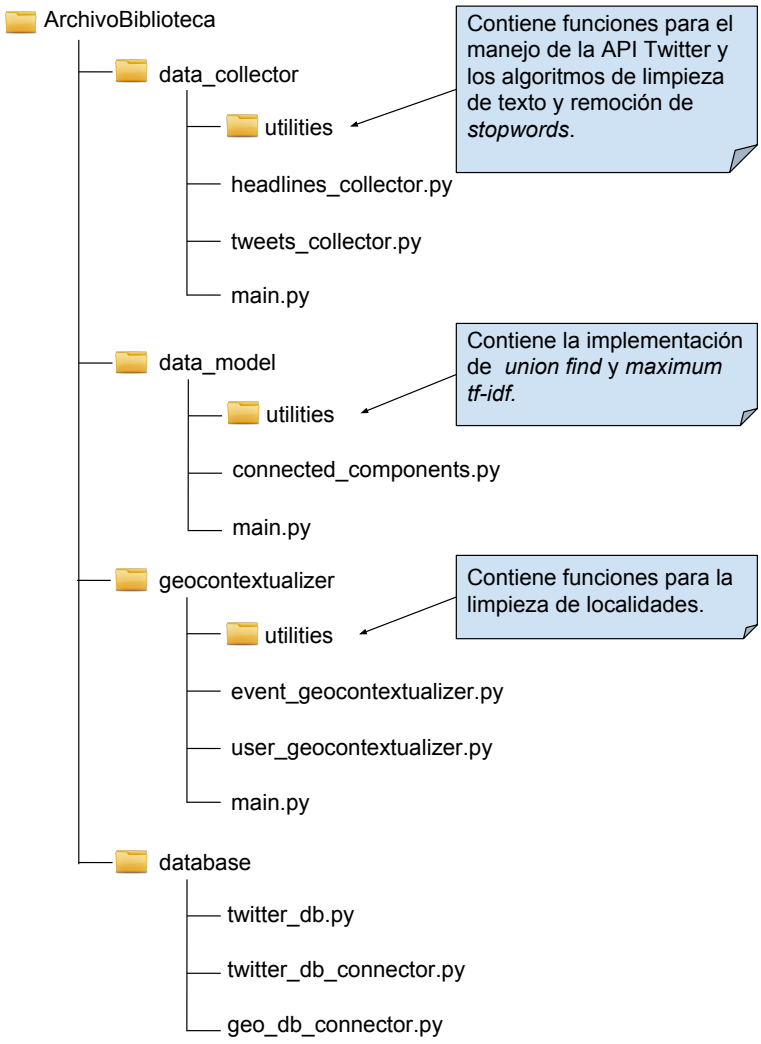


Figura 2.2: Esquema de implementación del subsistema de recolección y procesamiento de datos.

2.2. Componentes del sistema

El sistema se compone por cuatro componentes:

1. Aplicación Web: La finalidad de la aplicación es presentar la información recolectada de forma interactiva a los usuarios finales, permitiéndoles efectuar búsquedas y observar la información agregada de los datos mediante visualizaciones y mapas.
2. Proceso de recolección de datos: Identifica eventos a partir de titulares noticieros y recolecta *tweets* relacionados a cada evento. Este proceso se ejecuta cada hora de cada día y de esa forma recolecta información actualizada sobre las última noticias en todo momento.
3. Proceso de generación de eventos diarios: Procesa los datos recolectados durante un día para agrupar los eventos que traten del mismo tema y se hayan identificado en diferentes horas. Se ejecuta una vez al día luego de terminar la recolección de datos.
4. Proceso que geolocalización de datos: Etiqueta geográficamente los eventos a partir del texto de los *tweets* relacionados y a los usuarios que publicaron *tweets* a partir de la información de sus perfiles en *Twitter*. El proceso se ejecuta una vez al día luego de terminar el proceso de generación de eventos diarios.

2.3. Tecnologías utilizadas

2.3.1. Aplicación Web

La aplicación Web está desarrollada utilizando el *framework* web.py[1] para Python. Se escoge por ser simple pero poderoso y porque da libertad al desarrollador.

La capa de presentación utiliza elementos *javascript/AJAX* para mejorar el aspecto y despliegues de la información. Para las visualizaciones se utilizan las bibliotecas leaflet.js[2] y d3.js[3]. También se utiliza el servicio de *Twitter* para sitios Web que permite embeber *tweets* y presentarlos en un formato acorde a las especificaciones de visualización.

Las capas de negocios y de datos se desarrollan en lenguaje *Python* 2.7.

La capa de datos utiliza *Peewee*[4] como ORM para comunicarse con la base de datos. Se decide usar un ORM porque permite manejar los datos de forma más abstracta y portable.

[1] `http://webpy.org/`
[2] `http://leafletjs.com/`
[3] `http://d3js.org/`
[4] `http://peewee.readthedocs.org/en/latest`

2.3.2. Recolección y Procesamiento de Datos

El lenguaje utilizado para los procesos de recolección de datos, procesamiento de datos y geolocalización de datos es *Python* 2.7, el que cuenta con una gran gama de bibliotecas para realizar el procesamiento de datos de manera muy precisa. Además es un lenguaje con una comunidad amplia y activa que facilita la búsqueda de información y soluciones.

Las herramientas utilizadas más importantes son:

- NLTK[5] (Natural Language Toolkit): Herramienta disponible para *Python* utilizada para el trabajo de procesamiento de texto.
- Biblioteca requests: Biblioteca escrita en *Python* utilizada para realizar las solicitudes a la API para desarrolladores de *Twitter*.
- Biblioteca simplejson: Biblioteca escrita en *Python* utilizada para manejar las respuestas de *Twitter*, las que son entregadas en formato JSON[6].
- Biblioteca geopoint: Biblioteca escrita en *Python* utilizada para representar puntos geográficos a partir de coordenadas.
- CLAVIN [7] (Cartographic Location And Vicinity INdexer): Un paquete de software de código abierto escrito en *Java*, usado para extraer localidades a partir de texto no estructurado y para resolverlas utilizando un diccionario geográfico.
- Biblioteca logging: Biblioteca escrita en *Python* que facilita la administración del sistema de registro.

2.3.3. Base de datos

Tanto los datos recolectados como información geográfica se almacena en una base de datos MySql[8]. El diagrama del modelo de datos se encuentra disponible en los anexos (Figura 6.9).

El proceso de geolocalización de datos utiliza otra base de datos local MySql, la que contiene un índice geográfico de localidades chilenas y de países. Este índice es usado para etiquetar lugares dentro de Chile y para complementar la información geográfica cuando sólo se tienen las coordenadas geográficas.

2.4. Servicios externos

API para desarrolladores de *Twitter*

Para recolectar la información de la red social *Twitter* se utiliza la API REST para desa-

[5] http://www.nltk.org/
[6] http://json.org/
[7] http://clavin.bericotechnologies.com/
[8] http://www.mysql.com/

rrolladores[9] Ésta provee acceso programático para leer datos desde *Twitter* bajo ciertas restricciones. La API REST identifica aplicaciones de *Twitter* y usuarios utilizando OAuth[10].

La aplicación Web utiliza los servicios de *Twitter* para sitios Web[11]. Estos permiten integrar *widgets*, botones y herramientas de *scripting* del lado del cliente para integrar *Twitter* y mostrar los *tweets* en un sitio Web, incluyendo el botón de *tweet*, el botón de Seguimiento, *Tweets* incrustados, y líneas de tiempo incrustadas.

MapBox

MapBox[12] es una plataforma de mapeo para desarrolladores. Es utilizada para obtener los mapas utilizados como base en las visualizaciones geográficas de la aplicación Web.

GitHub

Como herramienta de versionamiento se utiliza git y se utiliza el servicio para estudiantes de *GitHub*[13] que permite tener repositorios privados. En la misma plataforma se mantiene una *Wiki* donde se documenta la instalación de las dependencias requeridas, las fuentes de noticias utilizadas durante la recolección de datos y los procedimientos para la puesta en marcha de los sistemas de recolección, procesamiento y geolocalización de datos.

[9] `http://dev.twitter.com/rest/public`.
[10] `http://oauth.net/`
[11] `http://dev.twitter.com/web/overview`
[12] `http://www.mapbox.com/`
[13] `http://www.github.com`

Capítulo 3

Aplicación Web

3.1. Descripción general

La aplicación Web tiene como finalidad presentar la información recolectada desde *Twitter* sobre eventos noticiosos chilenos.

La información que describe un evento noticioso está compuesta por: titulares relacionados, *tweets* relacionados, palabras clave, información geográfica e información temporal.

Además de la información que describe a un evento se muestra información agregada, como la cantidad de *tweets* relacionados al evento, la cantidad de *tweets* por país (considerando la información geográfica del usuario que lo publicó) y la cantidad de veces que se menciona un lugar en los *tweets* relacionados.

Para presentar la información descrita de forma ordenada se utiliza la visualización propuesta por Vanessa Peña en un estudio realizado en paralelo al periodo de realización de la memoria y que se detalla en el artículo *Galean: Visualization of Geolocated News Events from Social Media* [9]. Esta forma de visualización permite presentar al usuario información geográfica y temporal de los eventos a través de mapas y marcadores interactivos. La visualización ofrece una vista general de los diferentes eventos para un día específico y cuando el usuario hace *click* en el marcador de un evento se despliega la información asociada al mismo.

Las ventajas que entrega esta forma de visualización son:

- Llama la atención de los usuarios y los invita a explorar la información gracias a su diseño interactivo.
- Permite identificar rápidamente los eventos más importantes gracias al tamaño de los marcadores, que es mayor si el evento tiene más *tweets* relacionados.
- Permite detectar rápidamente los distintos lugares afectados o involucrados en un evento.
- Permite hacer seguimiento de un evento en el tiempo y observar la duración de su presencia en *Twitter*.

- Permite hacer un análisis de impacto de un evento en otros lugares del mundo, gracias a la visualización que muestra el recuento de usuarios provenientes de cada país que comentaron sobre el tema.

3.2. Vistas de la Aplicación

La aplicación tiene dos vistas principales:

1. Vista del mapa de Chile: permite visualizar los eventos ocurridos en un día específico, distribuidos en las diferentes regiones del país.
2. Vista del planisferio: que permite visualizar eventos que involucran a otros países además de Chile.

Figura 3.1: Vista Principal de la aplicación en la que se muestran los eventos ocurridos el día 9 de Agosto del 2015 distribuidas en las diferentes regiones del país, y en se muestra la información relacionada al evento sobre el sistema frontal que afectó al Norte a Chile ese día.

La Figura 3.1 muestra la vista principal con el mapa de Chile. En la parte superior están los filtros que permiten al usuario personalizar su búsqueda. En la zona izquierda de la pantalla se encuentra el mapa donde se visualizan los eventos del día seleccionado. En la zona del centro de la pantalla se muestra la información asociada al evento seleccionado, como el número de *tweets*, número de menciones de los diferentes lugares mencionados en el texto de los *tweets*, la lista de titulares relacionados y el mapa del planisferio que visualiza la cantidad de usuarios por país que comentaron sobre el evento. En la zona derecha de la pantalla se muestran algunos *tweets* relacionados al evento.

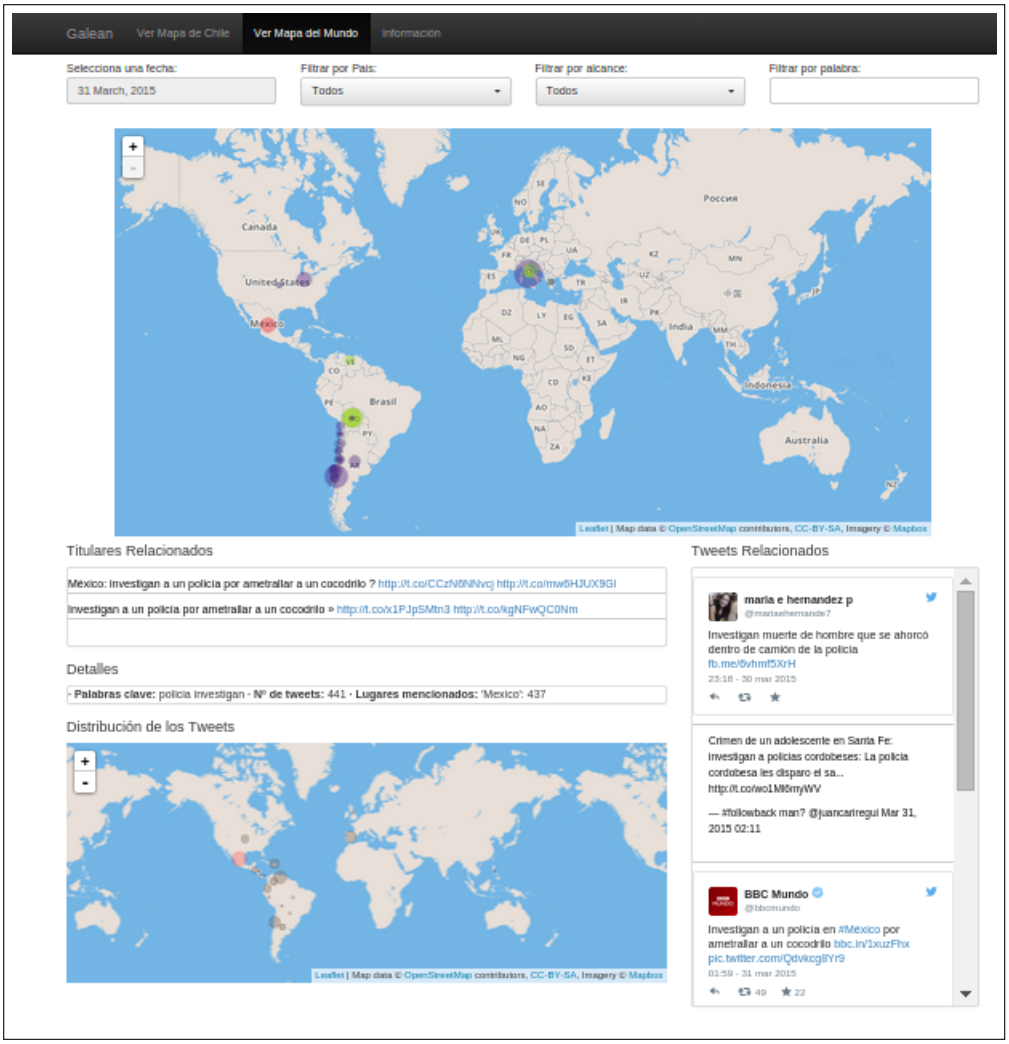


Figura 3.2: Vista Principal de la aplicación que muestra eventos de interés para Chile distribuidos en los distintos países del mundo el día 9 de Agosto del 2015, en el que se muestra información relacionada a un evento de un policía que intentó ametrallar a un cocodrilo en México.

La Figura 3.2 muestra la vista del planisferio completo. Esta vista se conforma por las mismas partes mencionadas para el caso anterior, pero usando otra distribución.

3.3. Interacción del usuario

La aplicación le brinda al usuario la posibilidad de navegar a través de la información realizando búsquedas fácilmente, ya que está construida pensando en que sea útil tanto para alguien que entra al sitio buscando información específica como para alguien que ingresa a navegar a través de los datos sin un destino claro.

Una búsqueda comienza cuando el usuario ingresa una fecha, aunque por defecto se inicializa con la última fecha para la que existen datos. Los eventos correspondientes a la fecha

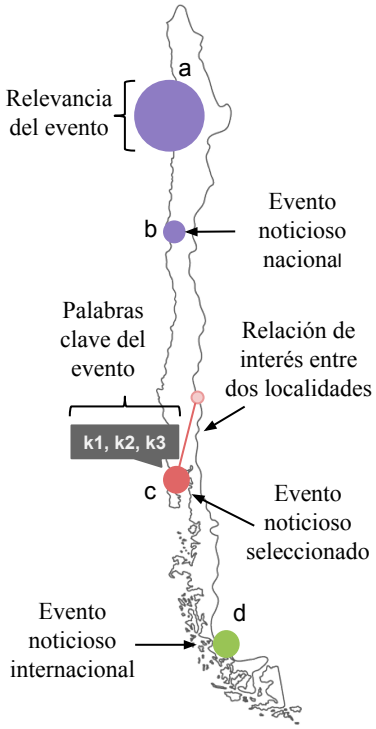


Figura 3.3: Simbología de la visualización de eventos noticiosos en la aplicación

seleccionada se visualizan en el mapa. Un evento puede involucrar a más de una localidad, pero el evento se ubica en las coordenadas correspondientes a la localidad más veces mencionadas en los *tweets* relacionados al evento.

Además de especificar la fecha el usuario puede utilizar algunos filtros para limitar la búsqueda. Uno de los filtros permite especificar la o las regiones de Chile involucradas en el evento y el otro es para especificar el alcance del suceso.

Las opciones del filtro por alcance varían dependiendo de si está utilizando la vista del mapa de Chile o si está utilizando la vista del planisferio. Para el caso de la vista de Chile se puede escoger entre visualizar eventos regionales (que involucran sólo a una región de Chile), visualizar eventos inter-regionales (que involucran a más de una región de Chile) o visualizar eventos nacionales (que involucran al país completo y no a alguna región en específico). Para el caso de la vista del planisferio el usuario puede filtrar para visualizar eventos nacionales (que involucran sólo a un país) o internacionales (que involucran a varios países).

Otro elemento que permite al usuario interactuar es el mapa. Como se mencionó antes, éste muestra los eventos detectados para el día seleccionado a través círculos, los que varían de tamaño dependiendo del número de *tweets* relacionados con el evento. El color de los círculos indica si son eventos que involucran sólo a Chile o si son eventos de carácter internacional, es decir, si se relacionan con algún otro país.

Al pasar el *mouse* por un círculo, éste cambia de color a rojo, si el evento involucra otros lugares, estos se muestran a través de círculos del mismo color y conectados a través de líneas al círculo principal. Al mismo tiempo las palabras claves del evento se muestran en

una bandera desplegable para que el usuario se pueda orientar respecto al tema del evento.

La Figura 3.3 muestra ejemplos de las especificaciones de la visualización. Cada uno de los círculos **a**, **b**, **c** y **d** representa un evento noticioso. Los círculos de color morado son eventos nacionales y el evento **d** de color verde es un evento internacional. El evento **c** se representa con un círculo de color rojo porque es un evento seleccionado, la conexión con el otro círculo muestra que hay otras localidades relacionadas con el evento y la bandera muestra las palabras clave que lo identifican.

Cuando el usuario hace *click* en un evento específico del mapa, en la pantalla aparece información asociada al evento. La información consiste en:

- Palabras clave que identifican al evento.
- Número de *tweets* relacionados.
- Nombres de lugares más mencionados en los *tweets* relacionados.
- Titulares relacionados con el evento.
- Un conjunto de *tweets* relacionados [1].
- Un mapa del mundo en el que se visualiza la distribución geográfica de los usuarios que *twittearon* en relación al evento.

[1] *Twitter* no permite publicar conjuntos de *tweets* en ningún formato sin un permiso especial, pero si permite mostrarlos embebidos en un sitio Web a partir del id del *tweets* utilizando la API de Twitter o publicar conjuntos de id's de *tweets* para que así la persona interesada pueda descargarlos[3].

Capítulo 4

Sistema de Recolección de Eventos

4.1. Metodología para modelar eventos noticiosos

En esta Sección se describe el *pipeline* para modelar eventos noticiosos utilizado por Mauricio Quezada en su tesis de magíster [11]. En el trabajo realizado por él, esta metodología fue utilizada de forma satisfactoria para la recolección de eventos de impacto de carácter internacional basado en *tweets* en idioma Inglés y el que utilizaba como fuente de noticias medios de comunicación internacionales como *CNN* y *The Breaking News* y *The Economist* .

En la Subsección 4.1.1 y la Subsección 4.1.2 se detallan los dos procesos principales que lo conforman.

4.1.1. Recuperación de documentos de interés periodístico

El objetivo de este primer proceso es emular un detector de eventos noticiosos, es decir, es un método que permite detectar noticias de interés periodístico. En este caso la detección se basa en identificar los principales tópicos sobre los que están publicando un conjunto seleccionado de medios de comunicación y luego recolectar información relacionada con esos tópicos. El diagrama presentado en la Figura 4.1 resume los pasos del proceso, los que se identifican con las letras **a**, **b**, **c** y **d**.

El proceso comienza recolectado titulares de las últimas noticias (**a**), los que se obtienen de las publicaciones en *Twitter* de un conjunto de cuentas oficiales seleccionadas de medios de comunicación.

Luego el texto de los titulares obtenidos es sometido a un proceso de limpieza, el cual remueve signos diacríticos, enlaces, menciones, *hashtags*, puntuaciones, espacios innecesarios y *stopwords*.

Una vez limpio el texto de los titulares, éstos se procesan para identificar tópicos, lo que

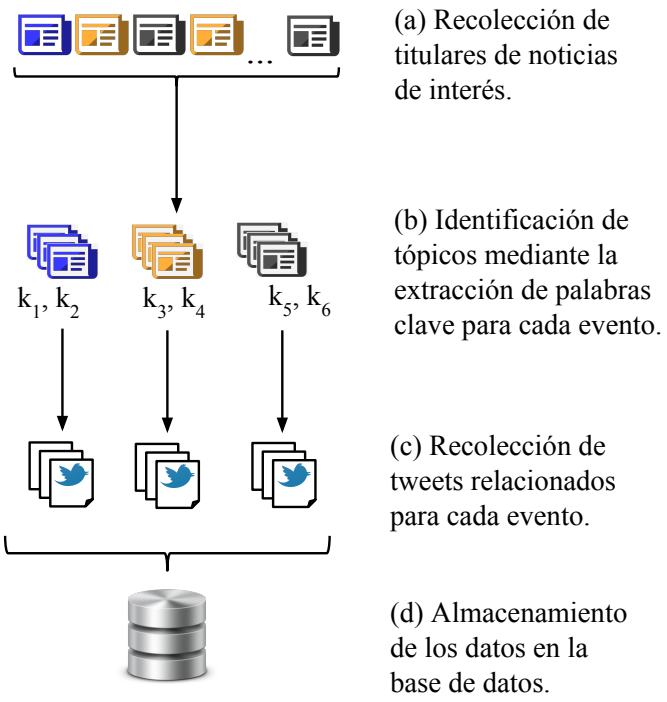


Figura 4.1: Diagrama de la metodología utilizada para la recuperación de documentos de interés periodístico.

se realiza mediante la extracción de palabras clave (**b**). El método de selección de palabras clave se basa en la búsqueda de términos frecuentes utilizando reglas de asociación sobre el conjunto total de titulares. Este proceso permite identificar pares de palabras clave comunes entre titulares de diferentes medios de comunicación, por lo que al finalizar, cada par de palabras clave representa un *evento* o tópico.

Del conjunto de pares de palabras clave, se seleccionan los pares más representativos dentro del conjunto. Esto permite hacer un filtro para considerar los eventos de más impacto y descartar tópicos menos importantes.

Una vez escogidas las palabras clave para cada evento, comienza la recolección de *tweets* relacionados (**c**). Para recolectar la información sobre los *eventos* se ejecutan en paralelo múltiples procesos de búsqueda de *tweets*. Para las búsquedas de *tweets* se utiliza como criterio que el *tweet* contenga las dos palabras clave que representan al evento.

La cantidad de consultas que se pueden realizar a la API REST de *Twitter* está limitada a un máximo de 450 cada 15 minutos para cada aplicación[5] registrada, por lo que las búsquedas se deben parcelar en intervalos de tiempo calculados para no sobrepasar el límite de velocidad.

La búsqueda comienza siendo en retrospectiva, es decir, comienza buscando *tweets* publicados antes de que el evento haya sido identificado a partir de los titulares. Esto permite obtener información sobre otros medios que lo publicaron con anterioridad o sobre usuarios que informaron o comentaron del tema previamente. Una vez recolectados los *tweets* anti-

guos, comienza un proceso de búsqueda permanente, con la finalidad de capturar los *tweets* que se van publicando durante el transcurso de tiempo en que se ejecute la recolección.

Debido a que la recepción de la información es más rápida de lo que es posible insertar registros en la base de datos, los *tweets* recolectados son escritos en archivos. Una vez finalizada la recolección para todos los eventos, el proceso principal se encarga de guardar los datos en la base de datos (**d**).

4.1.2. Identificación de noticias similares

El objetivo de este segundo proceso es agrupar eventos similares. Esto es necesario porque, si el proceso de recolección antes descrito se ejecuta para varias horas de un mismo día, generalmente ocurre que eventos que tratan el mismo tema son identificados más de una vez en horas diferentes. El diagrama de la Figura 4.2 presenta gráficamente el objetivo general del proceso de agrupación de noticias similares.

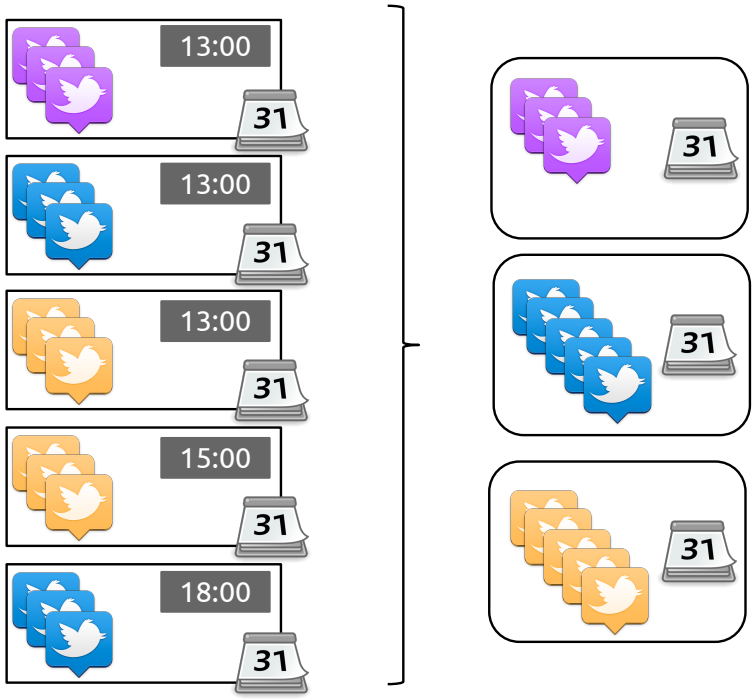


Figura 4.2: Descripción gráfica del proceso de identificación y agrupación de noticias similares. Los colores representan un tópico común.

Un ejemplo de casos como éste son los desastres naturales, ya que, en una primera instancia se informa del suceso y posterior a éste se informa sobre las consecuencias de la catástrofe, las reacciones de las personas, las medidas adoptadas por los representantes del estado, etc. Todas estas aristas de un mismo evento son abordadas a través de diferentes reportes periodísticos durante varias horas.

Para agrupar eventos similares se utiliza el algoritmo *union find*. En general este algoritmo

es utilizado para encontrar conectividad entre nodos de un grafo. Para este caso en particular las conexiones son las palabras clave repetidas en diferentes eventos. De esta forma todos los eventos que tienen palabras clave en común forman un conjunto.

Luego se aplica el algoritmo *tf-idf* sobre los conjuntos para identificar palabras de articulación. Éste algoritmo pondera las palabras clave considerando la cantidad de veces que se menciona el término dentro del conjunto de palabras clave que se analiza, pero al mismo tiempo calcula la frecuencia de utilización del término en todos los otros conjuntos de palabras de clave, determinando así, cuándo una palabra es en realidad una palabra comúnmente utilizada y no necesariamente se trata de un mismo evento. Como su nombre lo dice, las palabras identificadas como de articulación se utilizan para separar los conjuntos previamente formados.

4.2. Adaptación del sistema de recolección

El *pipeline* previamente descrito presentó resultados satisfactorios en su aplicación previa para el caso de eventos internacionales y con *tweets* en inglés, por lo que se consideró la base para la implementación de el sistema desarrollado en esta memoria.

En esta sección se detallan las principales modificaciones y extensiones realizadas a la metodología original durante la implementación de los procesos.

4.2.1. Fuentes de Noticias

Para la recolección de titulares noticieros se construyó una lista de cuentas de medios de comunicación tradicionales chilenos, las que fueron escogidas manualmente.

La selección de cuentas implicó un exhaustivo trabajo de búsqueda y seguimiento de cada una ellas para verificar que fueran utilizadas para publicar noticias actualizadas durante el día y que en su conjunto se abarcaran noticias de diferentes regiones de Chile.

Las cuentas pertenecen a medios de comunicación tradicionales. La mayor parte son periódicos que cuentan tanto con una versión impresa como digital, pero también hay cuentas de noticieros televisivos, medios radiales, periodistas e instituciones públicas como Carabineros de Chile y la Onemi. El listado completo de las cuentas utilizadas se incluye en los anexos6.4.

4.2.2. Proceso de limpieza de titulares

El proceso de limpieza de los titulares se aplica previo al procesamiento para la extracción de *keywords*.

En este paso se eliminan *stopwords*. Como los titulares se obtienen a partir de los *tweets*

Figura 4.3: Ejemplos de los resultados luego de la limpieza de titulares. Se destacan con rojo las palabras que fueron eliminadas por pertenecer al conjunto de palabras agregadas a la lista de *stopwords*.

publicados por los medios de comunicación, también se eliminan los *hashtags*, menciones y *links* que pueda contener el *tweet* y de esta forma obtener el titular lo más puro posible.

La lista de *stopwords* se obtiene del corpus que provee el paquete NLTK[1] disponible para Python. Además la lista fue complementada con un conjunto de palabras en español. Las palabras agregadas fueron identificadas como *stopwords* luego de inspeccionar los titulares recolectados durante un periodo de tiempo y notar que existen algunas que son frecuentemente utilizadas pero que no aportan información relevante sobre el tema del evento. La Figura 4.3 muestra ejemplos de *tweets* de noticias en los que se destacan algunas de las palabras que no aportan información y pertenecen al conjunto de palabras agregadas a la lista de *stopwords*. En general son frases utilizadas comúnmente por los medios de comunicación para destacar una noticia o palabras que son utilizadas como etiquetas para personalizar sus *tweets*. En la Tabla 4.1 se incluye la lista de palabras especiales que se agregan a las *stopwords*.

[1] `http://www.nltk.org/`

pre	dejar	hoy	asi	alla	fotos
san	quedo	dia	dan	cada	lunes
unas	ahora	dias	pide	mira	martes
unos	vengo	mes	estan	video	miercoles
dos	llego	ano	eran	foto	jueves
mil	dijo	minuto	sub	ultimo	viernes
paz	uso	hora	toda	vivo	sabado
mas	solo	horas	t13	senal	domingo
tras	sepa	hrs	ser	adn	opinion
deja	sigue	luego	dio	cnn	audio

Tabla 4.1: Palabras agregadas a las *stopwords* para ser utilizadas durante el procesamiento de limpieza del texto de los *tweets*

4.2.3. Filtros de búsqueda de *tweets*

La metodología del proceso original usaba como único criterio de búsqueda, que el *tweet* contenga las dos palabras claves que identifican al evento. En el proceso implementado para la versión chilena, a esa restricción se le suma que sean *tweets* escritos en español.

Esto se debe a que las palabras claves que identifican cada evento fueron extraídas de titulares escritos en español, y limitar la búsqueda de *tweets* escritos en español permite ubicar las palabras clave en un contexto más apropiado, disminuyendo el ruido en los datos.

4.2.4. Adición de sistema de registro

Para facilitar la identificación y tratamiento de errores, tanto presentes como futuros del sistema, se agrega un sistema de registro. Para ello se utiliza el paquete de software *logging* disponible para Python [2].

Las ventajas de contar con un sistema de registro son:

1. Permite identificar desde dónde y cuándo se ejecuta una llamada de registro.
2. Permite escribir el registro en archivos, a través de un *socket*, a través de la salida estándar, etc. al mismo tiempo.
3. Permite diferenciar lo que se registra dependiendo de la severidad y posteriormente facilita el proceso de filtrar la información.

Gracias a este sistema fue posible solucionar errores rápidamente y de forma efectiva.

[2] `http://docs.python.org/2/library/logging.html`

4.2.5. Almacenamiento de la información

La API de *Twitter* entrega la información de los *tweets* junto con la del usuario que publicó el *tweet* en formato JSON. [3].

Almacenamiento de eventos y titulares

Durante el proceso de recolección de titulares se obtienen *tweets* publicados por medios de noticias chilenos a través de las cuentas seleccionadas. El conjunto de *tweets* en formato JSON recolectados para cada hora se almacena en un archivo comprimido. Además se extrae el texto de los *tweets* que corresponde al titular noticioso y se escribe en un archivo. El archivo con los titulares se utiliza en el proceso que realiza la identificación de eventos.

Luego de procesar el texto de los *tweets* de los titulares para identificar eventos y extraer palabras clave, se almacenan los eventos identificados en la base de datos. Un evento se identifica por el par de palabras clave y la fecha y hora en que se detectó el evento. Luego se seleccionan los titulares que contienen las palabras clave del evento, y también son almacenados en la base de datos asociados al evento correspondiente. Cabe destacar que en este paso no se almacena el *tweet* completo del titular, si no, sólo el texto del *tweet* correspondiente al titular de la noticia.

Luego de almacenar lo necesario en la base de datos, el archivo de texto con los titulares noticiosos es eliminado.

Almacenamiento de usuarios y *tweets*

Al escribir los datos recolectados durante el proceso de recolección de *tweets* en la base de datos se separa la información del *tweet* y del usuario en dos tablas diferentes y los diferentes campos asociados a cada uno en columnas.

Si un *tweet* fue recolectado previamente y por lo tanto se encuentra registrado en la base de datos, entonces no se considera para el nuevo evento. Si se obtiene un *tweet* que no había sido recolectado previamente pero el usuario que lo publicó ya se encuentra registrado en la base de datos, entonces el usuario no se registra y el nuevo *tweet* se registra con una referencia al usuario que ya se encontraba previamente registrado.

Todos los usuarios nuevos se almacenan en la tabla de usuarios y también en la tabla de usuarios temporales, la que existe solamente para facilitar el proceso de geolocalización de usuarios descrito en el Capítulo 5. La tabla de usuarios temporales se vacía luego de ser procesada para geolocalizar los usuarios.

[3] `http://dev.twitter.com/rest/reference/get/search/tweets`

Almacenamiento de grupos de eventos o componentes

El proceso que agrupa eventos noticiosos similares almacena la información sobre los grupos identificados en la tabla *componentes*. Un componente o grupo se conforma por varios eventos y al igual que un evento se identifica por las palabras claves representativas y la fecha, que en este caso correspondiente al día sin detalle sobre la hora específica.

4.2.6. Ejecución periódica y continua

Para la automatizar la ejecución de los procesos se utiliza la herramienta *crontab* de Linux [4].

Para obtener información actualizada y de forma constante, el proceso de recolección está programado para ejecutarse al inicio de cada hora.

El proceso de agrupación de eventos se aplica sobre los datos recolectados durante un día completo. Está programado para ejecutarse posterior a la primera hora del día siguiente luego de que todos los datos recolectados para el día previo fueron almacenados.

4.3. Detalle de implementación

El modelo de recolección de eventos que se describe en este capítulo se conforma por dos de las principales componentes del sistema implementado. Una de las componentes se encarga de recolectar los datos sobre eventos noticiosos desde la red social *Twitter* e identificar eventos noticiosos para cada hora. La otra componente se encarga de modelar los datos agrupándolos en eventos diarios.

El primer paso del pipeline se encarga de recolectar titulares a partir del conjunto de cuentas de *Twitter* seleccionadas, limpiar el texto e identificar los pares de palabras claves que identificarán a cada evento noticioso. El algoritmo que se presenta a continuación resume este proceso.

```
def collect_headlines(news_accounts_file, date, lang='en'):
    ''' Collect headlines published by selected Twitter accounts
    and identify pairs of keywords for each event.
    '''

    one_hour_before = date - datetime.timedelta(hours=1)
    news_accounts = read_accounts(news_accounts_file)
    headlines, headlines_text, headlines_tweets =
        get_headlines_from_twitter(news_accounts, lang)
    keywords_pairs, groups = extract_event_keywords(headlines)
```

[4] http://help.ubuntu.com/community/CronHowto

```
    # Saving the data in files

    write_headlines(headlines, headlines_dir)
    write_headlines_text(headlines_text, headlines_dir)
    write_headlines_tweets(headlines_tweets, headlines_dir)
    write_keywords_scores(groups)

    # For each keywords pair create an event and save three
    # representative complete headlines
    for kw_pair in keywords_pairs:
        event_headlines = select_event_headlines(kw_pair,
                                                 headlines_text)
        with db.transaction():
            event_id = save_event(' '.join(kw_pair), date)
            for event_headline in event_headlines:
                save_headline(event_headline.encode('utf-8'),
                              event_id)

    return keywords_pairs
```

El segundo paso del *pipeline* corresponde al proceso de recolección de *tweets*. Este proceso utiliza los pares de palabras clave que retorna la función anterior y ejecuta en paralelo la búsqueda de *tweets* para cada uno de los eventos identificados. La búsqueda dura 50 minutos y los *tweets* recolectados son escritos en archivos de texto. Luego de que todos los procesos terminan su ejecución, los *tweets* son almacenados en la base de datos y asociados al evento correspondiente. El algoritmo a continuación resume este proceso.

```
def collect_tweets(keyword_pair_list, date, temp_dir):

    timeout = 3000  #secs = 50 min
    processes = []
    app_id = 0

    # Run a process to collect tweets during one hour
    # for each pair of keywords. Tweets are written in files
    for keyword_pair in keyword_pair_list:
        p = Process(target=search,
                    args=(app_id,
                          apps[app_id],
                          keyword_pair,
                          timeout,
                          date,
                          temp_dir))
        p.start()
        processes.append(p)
        app_id += 1
```

```
    # Join all processes
    for process in processes:
        process.join()

    # Store the information from text files generated in this
    # hour in database. One at a time.
    fdate = datetime.datetime.strftime(date, '%Y%m%d_%H%M%S')
    for fname in os.listdir(temp_dir):
        if fname.endswith(fdate + '.txt'):
            tweets = get_tweets_from_file(fname)
            keywords = get_keywords_from_filename(fname)

            save_to_database(keywords, tweets, date)

            # After save tweets in database the text file is removed.
            os.remove(temp_dir + '/' + fname)
```

Una vez al día se ejecuta el proceso que genera componentes conexas entre los eventos identificados durante el mismo día. Cada componente conexa representa un evento diario. La función que se presenta a continuación es la que realiza la tarea de generar las componentes conexas con los eventos identificados para un día especificado.

```
def generate_connected_components(day):
    ''' Generate connected components from events identified during
    a specific day
    Parameters: day is an object of type datetime
    '''

    keywords_by_hour, keyword_events_ids = __get_keywords_by_day(day)
    keyword_pairs = map(list, keyword_events_ids.keys())

    # Creates connected component using union find algorithm
    connected_components = create_connected_components(keyword_pairs)

    # Disconnect by articulation words using max tf-idf algorithm
    connected_components = disconnect_by_articulation_words(
                                connected_components,
                                keywords_by_hour,
                                keyword_events_ids)

    events = get_events_by_day(day)

    events_info = defaultdict(set)
    for event in events:
        keywords = frozenset(event.keywords.split())
        events_info[keywords].add(event.id)

    connected_components_data = list()
    for connected_component in connected_components:
```

```
        # Very large components are ignored
        if len(connected_component) >= 40:
            continue

        # Find events that make the connected component.
        ids_this_component = set()
        connected_component = frozenset(connected_component)
        connected_component_str = ' '.join(connected_component)
        for keyword_pair, ids in events_info.iteritems():
            if keyword_pair.issubset(connected_component):
                ids_this_component |= set(ids)
        connected_components_data.append((connected_component_str,
                                          list(ids_this_component)))

        # Save data
        c = save_component(keywords=connected_component_str,
                           date=day)
        if c:
            for e in list(ids_this_component):
                save_component_event(component_id=c,
                                     event_id=e)
        else:
            print 'Error, component doesnt exist'

    return connected_components_data
```

Capítulo 5

Geolocalización de datos

5.1. Descripción general

Una de las características importantes que describen a un evento es el lugar donde ocurre el suceso o los principales lugares involucrados. También es interesante conocer los lugares desde donde las personas están comentando el evento, ya que permite medir el interés y observar hasta que lugares del mundo se ha expandido.

Para obtener esa información se geolocalizan los eventos y los usuarios que publicaron *tweets* relacionados a los eventos. En la Sección 5.2 se describe la metodología utilizada para los procesos de geolocalización de eventos y usuarios y en la Sección 5.3 se detalla el proceso de implementación destacando lo que fue necesario modificar y/o extender a partir de la solución original.

5.2. Metodología para geolocalizar datos

El proceso de geolocalización consiste en añadir metadatos geográficos a partir del contenido de un documento. En esta sección se describe la metodología de geolocalización de datos extraídos desde *Twitter*, que es utilizada por Vanessa Peña en su trabajo de Tesis[10] y en la que se basó la implementación realizada en esta memoria.

En la Subsección 5.2.1 y la Subsección 5.2.2 se detallan la metodología utilizada para geolocalizar usuarios y eventos respectivamente.

5.2.1. Geolocalización de usuarios

Las cuentas de usuarios de *Twitter* tienen un campo de localización, en el cual, cada usuario puede o no escribir la localidad a la cual pertenece. El campo de localización permite

Figura 5.1: Localidades de algunos usuarios de *Twitter*. Los de la columna izquierda no pueden ser geolocalizados, mientras que los de la columna derecha si.

ingresar texto libre, por lo que no sigue un formato estándar ni es seguro que corresponda a una localidad válida.

La metodología utilizada para geolocalizar usuarios utiliza la información provista en ese campo. Las dificultades de este proceso radican en que no todos los usuarios tienen información en el campo de localización y, si la tienen, hay que darle formato a la información para poder identificar si se trata de un nombre de localidad válido.

Utilizando la metodología de geolocalización propuesta no es posible etiquetar geográficamente al total de los usuarios. A pesar de que la información geográfica no existe para todos, la información geográfica de los que usuarios que si la tienen continúa siendo valiosa, principalmente porque en general el conjunto de *usuarios* es lo suficientemente grande como para suponer que el porcentaje con localidad definida corresponde a una muestra representativa.

La Figura 5.1 muestra ejemplos de localidades de usuarios de *Twitter*. Como se puede observar las localidades de la columna del lado izquierdo no pueden ser geolocalizadas, ya que no corresponden a lugares reales, por otro lado, las localidades de la columna del lado derecho son lugares reales.

Los pasos que componen la metodología para geolocalizar un conjunto de usuarios son:

1. Limpieza del texto de localización: Usando expresiones regulares se identifica si el texto contiene coordenadas geográficas, si es el caso, se aplica un proceso de normalización de coordenadas que elimina los términos "UT:" y "i12T:" utilizados por algunos usuarios que tienen coordenadas como localización[1]; si el texto no contiene coordenadas geográficas, se normaliza eliminando espacios.
2. Resolución de coordenadas: Si el texto corresponde a coordenadas geográficas, se extrae la información de la latitud y longitud y se completa la información utilizando un índice geográfico. Para esto se utilizan algoritmos de proximidad que permiten identificar el lugar más cercano a las coordenadas obtenidas.
3. Resolución de localidades complejas: Para la resolución de las localidades que no contienen coordenadas geográficas se utiliza CLAVIN, un analizador de texto que permite identificar nombres de localidades a partir de texto no estructurado y resolverlos me-

[1]Algunos usuarios que utilizan *Twitter* en dispositivos como Iphone o Blackberry completan el campo de localización automáticamente con la información de coordenadas que les provee el servicio de GPS del dispositivo.

diante un diccionario geográfico construido con información obtenida desde GeoNames[2].

5.2.2. Geolocalización de eventos

La geolocalización de eventos tiene como finalidad identificar el lugar donde ocurre un evento o los lugares relacionados con el evento. El proceso de geolocalización se basa en el supuesto de que, si un evento está relacionado con una localidad, el nombre de esa localidad será mencionada en un cantidad considerable de *tweets* relacionados al evento.

El proceso de geolocalización de eventos toma todos los *tweets* relacionados con el evento, procesa el texto de los *tweets* aplicando un análisis semántico para extraer nombres de localidades y finalmente completa la información utilizando un índice geográfico.

Los pasos que componen la metogología propuesta son:

1. Extracción de localidades del texto de los *tweets*: Se utiliza el extractor de CLAVIN que permite identificar los lugares mencionados en los *tweets*.
2. Frecuencia y contextualización de localidades: Se procesa el conjunto de localidades extraídas para filtrar las más frecuentes. Este proceso también considera la frecuencia con que dos localidades localidades son mencionadas juntas en un mismo *tweet*, permitiendo contextualizar la información cuando el lugar mencionado coincide con un lugar existente en más de un país. Cuando se encuentran dos o más lugares de localidades mencionadas al mismo tiempo frecuentemente, éstas se unen separadas por comas como una misma localidad.
3. Resolución de localidades: Se utiliza el algoritmo de resolución de localidades de CLAVIN que completa la información geográfica de la o las localidades encontradas en los *tweets*.

5.3. Implementación de los sistemas de geolocalización

La metodología de geolocalización de datos descrita previamente fue utilizada como base para la implementación de los sistemas de geolocalización, pero se modificaron algunas partes para mejorar su funcionamiento para el caso de Chile.

A continuación se detallan las principales modificaciones realizadas durante la implementación en cada una de ellas.

[2] `http://www.geonames.org/`

5.3.1. Modificaciones a la metodología de geolocalización de usuarios

La metodología original se mantiene, excepto por una modificación en el tercer paso, en el que se resuelven localidades complejas. Las localidades complejas son las que no tienen coordenadas geográficas y en cambio tienen localidades ingresadas por el usuario en lenguaje natural.

La solución original descrita previamente utiliza CLAVIN para extraer nombres de localidades del texto del campo de localidad de los usuarios y luego los resuelve usando el índice geográfico. Mediante la realización de experimentos, se llegó a la conclusión de que era posible mejorar el método para el caso particular de geolocalización de usuarios de *Twitter*.

El campo de localidad de los usuarios de *Twitter* tiene un largo no mayor a 30 caracteres, lo que no permite al usuario ingresar frases complejas. La mayoría de los usuarios que ponen localidades válidas, indican nombre de ciudad y país separadas por comas o conjunciones.

El experimento realizado consistió en aplicar el algoritmo de análisis que realiza CLAVIN sobre una lista de lugares en Chile, los que se muestran en la Tabla 5.1. Los lugares están escritos usando el formato *Ciudad, País*. Como resultado, todas las localidades eran resueltas como lugares en Chile, lo que era suficientemente preciso en la solución original que buscaba posicionar usuarios de todo el mundo en sus respectivos países, pero sólo 32 % de los resultados incluía información precisa sobre la región de Chile en la que se encontraba la localidad.

Tabla 5.1: Lista de lugares en Chile usados para probar la geolocalización de usuarios chilenos.

Santiago, Chile	Quillota, Chile
Temuco, Chile	Chiloé, Chile
Arica, Chile	Ancud, Chile
Valdivia, Chile	Castro, Chile
Valparaíso, Chile	Tocopilla, Chile
Atacama, Chile	Antofagasta, Chile
San Bernardo, Chile	Copiapó, Chile
Concepción, Chile	Vallenar, Chile
Puerto Montt, Chile	Ovalle, Chile
Coquimbo, Chile	Illapel, Chile
La Serena, Chile	San Felipe, Chile
Viña del Mar, Chile	Rancagua, Chile
Puerto Varas, Chile	Curicó, Chile
Linares, Chile	Chillán, Chile
Iquique, Chile	Talcahuano, Chile
Talca, Chile	Coyhaique, Chile
Osorno, Chile	Punta Arenas, Chile

Luego se dividió el análisis que realiza CLAVIN en dos: el primer proceso de *extracción* de nombres de localidades en texto no estructurado y el segundo proceso de *resolución* de la localidad utilizando el diccionario geográfico. Se aplicaron ambos procesos sobre la lista de

localidades antes mencionada.

Al probar el algoritmo de *extracción* de nombre de localidades se observó que era en este paso en el cual el algoritmo no era capaz de extraer ambos nombres de localidades del texto. Las razones de la poca efectividad pueden ser muchas.

Sin embargo, al probar el algoritmo de *resolución* los resultados fueron satisfactorios aumentando a un 100 % de precisión con respecto a la región involucrada.

En base a los resultados obtenidos se desechó el algoritmo de *extracción*, siendo reemplazado por un algoritmo que procesa el texto del campo de localidad del usuario y lo intenta adaptar al formato "Ciudad, País", eliminando conectores como *en* o *de*, signos y puntuaciones no deseadas. Para esta nueva solución se asume que el campo de la localidad es relativamente similar al formato deseado. Luego el texto con formato es *resuelto* con CLAVIN.

5.3.2. Modificaciones a la metodología de geolocalización de eventos

La metodología original para geolocalizar eventos utiliza el extractor de nombres de localidades de CLAVIN. Al aplicar el proceso sobre los datos recolectados el proceso es capaz de identificar nombres de una gran cantidad de países, también cuando el lugar mencionado en el *tweet* corresponde a una capital o a ciudades conocidas de otros algunos países, pero la identificación de nombres de ciudades chilenas no es muy efectiva.

Algunos factores que pueden ser causantes de la poca efectividad son:

- Los *tweets* al estar limitados en cantidad de caracteres y ser publicados bajo un contexto que posiciona la noticia como chilena, muchas veces no mencionan a "Chile" explícitamente, sino que, se menciona directamente el nombre de la ciudad o región relacionada. Cuando esto ocurre, el proceso identifica algunos de los nombres de lugares, pero al buscar la localidad en el diccionario geográfico, la ubica en otros países, ya que en ellos también hay ciudades conocidas con los mismos nombres. Por ejemplo: Los Ángeles o San Felipe.
- El algoritmo que utiliza CLAVIN no está desarrollado pensando en ser aplicado en texto en español y al tener importantes diferencias semánticas con el idioma Inglés, los resultados no son equivalentes.
- El diccionario geográfico utilizado en la versión original no contiene algunos lugares en Chile.

Considerando los problemas encontrados se modifica el proceso original. El proceso de geolocalización finalmente implementado se conforma por los siguientes pasos:

1. Extracción de nombres de localidades candidatas: Se analiza cada palabra del texto de los *tweets* en busca de posibles nombres de ciudades, regiones o países. El criterio de selección es muy simple y se basa en verificar si la palabra comienza con mayúscula. El proceso mide la frecuencia con la que se menciona la localidad candidata dentro del conjunto de *tweets* y si la frecuencia no es lo suficientemente alta, la localidad

candidata se descarta. Es importante recalcar que el análisis realizado tanto en la solución original como en la modificación agregada, se trabaja bajo el supuesto de que, por la ley de los grandes números, la localidad más mencionada por los usuarios corresponde a la localidad más probablemente relacionada con el evento y por tanto el error corresponderá a un porcentaje menor dentro del conjunto completo de *tweets* analizados.

2. Resolución de localidades: Se utiliza un diccionario geográfico de localidades chilenas que contiene muchos nombres de ciudades y también uno de nombres de países (no incluye ciudades de otros países). Luego los lugares que se identifican como candidatos en el primer paso son buscados en ambos diccionarios geográficos. Por lo tanto, si el evento está relacionado a algún lugar en Chile, y el nombre de la ciudad o región es mencionado varias veces en los *tweets* relacionados al evento, entonces será geolocalizado en Chile y con información detallada de la ciudad. Si en cambio, está relacionado con otro país, también será geolocalizado, pero con información general (solo el país).
3. Verificación final utilizando el análisis original: Si luego de aplicar el paso anteriormente descrito no se obtiene la localidad del evento, el conjunto de *tweets* es analizado utilizando el proceso original basado en el uso de CLAVIN.

   Con este proceso se geolocalizan los eventos ocurridos en otros países cuando el nombre del país no fue mencionado explícitamente en los *tweets* con suficiente frecuencia, sin embargo, se menciona muchas veces el nombre de alguna ciudad dentro de él, por ejemplo, su capital.

Utilizando el proceso descrito, aumentan considerablemente los eventos correctamente geolocalizados en Chile y son pocos los eventos que llegan al tercer paso de verificación.

5.3.3. Posibles mejoras en la geolocalización de eventos

A pesar de las mejoras obtenidas, hay algunos problemas que escapan del trabajo de realizado en esta memoria ya que requieren un proceso de experimentación extenso. Estos problemas son:

- Tanto al utilizar el proceso de extracción original basado en CLAVIN como el proceso de búsqueda simple de localidades candidatas en el texto, existen errores para casos particulares como cuando existen nombres, apellidos o nombres de entidades que son iguales a lugares chilenos o de otras partes del mundo. Por ejemplo una noticia que tiene como protagonista a Jorge Valdivia, el futbolista, será ubicado en Valdivia, ya que su apellido se repetirá en muchos *tweets*, probablemente muchas más veces incluso que el lugar real donde pueda haber ocurrido el suceso.
- Hay equipos de fútbol en los que su nombre indica el lugar al que representan, por ejemplo *Unión La Calera*, *San Marcos de Arica*, *Curicó Unido*, etc. Estos casos al momento de ser geolocalizados tendrán al menos una etiqueta en la que se indicará que se relacionan con el lugar al que representan. En cambio hay otros casos como *Huachipato*, un club deportivo de la ciudad de Talcahuano, para el cual es imposible saber mediante el proceso de geolocalización que se relaciona con esa ciudad. Por último hay casos como el club deportivo de *Santiago Wanderers*, un club deportivo de Valparaíso,

que en su nombre menciona Santiago, texto que al ser geolocalizado quedaría ubicado incorrectamente en la capital de Chile, Santiago.

Capítulo 6

Exploración de los datos

6.1. Sumarización

El sistema recolector de datos se puso en marcha el día 30 de Octubre del 2014. La tabla 6.1 muestra la información respecto a la cantidad total de datos recolectadas hasta el día 17 de Agosto del 2015.

	Total Recolectados	Total Geolocalizados	Total chilenos
Eventos	17,850	9,895	7,545
Usuarios	8,847,648	3,716,602	260,807
Tweets	78,413,724	—	—

Tabla 6.1: Cantidad total de eventos, usuarios y *tweets* recolectados desde 30 de Octubre del 2014 hasta 17 de Agosto del 2015.

6.2. Distribución de eventos

Como los eventos son recolectados a partir de los titulares publicados por medios de comunicación chilenos, la mayor parte de ellos corresponden a eventos nacionales, pero también existe una parte importante de los eventos que tratan temas de carácter internacional. El gráfico de la figura 6.1 muestra que el 66 % de los eventos geolocalizados corresponde a eventos en Chile. El resto de los eventos se distribuyen en otros 136 países, dentro de los cuales predominan Estados Unidos, México, Argentina, Venezuela y Bolivia.

Los eventos geolocalizados en Chile, fueron divididos en dos categorías, eventos *nacionales* y eventos *regionales*.

A la primera categoría de eventos *nacionales* pertenecen los eventos que fueron geolocalizados en Chile, pero para los cuales no fue posible identificar nombres de lugares específicos dentro de Chile. En general son eventos que involucran al país entero, por ejemplos, renuncia

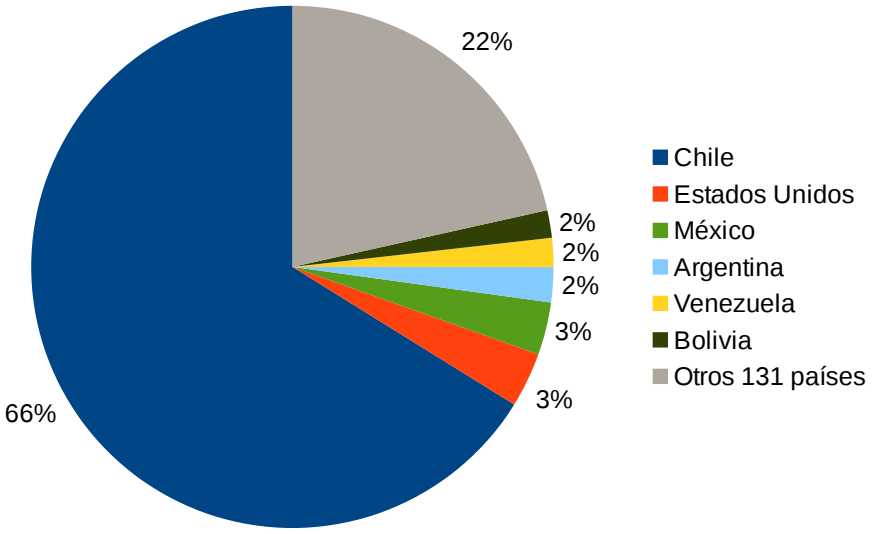


Figura 6.1: Distribución geográfica de los eventos geolocalizados almacenados en la base de datos

de un ministro de gobierno, aprobación de una nueva ley, etc. A esta categoría pertenecen 1239 eventos, equivalente al 16 % de los eventos geolocalizados en Chile.

A la segunda categoría de eventos *regionales* pertenecen los eventos que fueron geolocalizados en Chile y para los cuales fue posible identificar nombres de regiones o ciudades involucradas. Ejemplos de eventos pertenecientes a esta categoría son desastres naturales como incendios forestales, erupciones volcánicas o sismos, también encontramos en esta categoría eventos deportivos futbolísticos protagonizados por equipos regionales. A esta categoría pertenecen 6315 eventos, equivalentes al 83 % de los eventos geolocalizados en Chile.

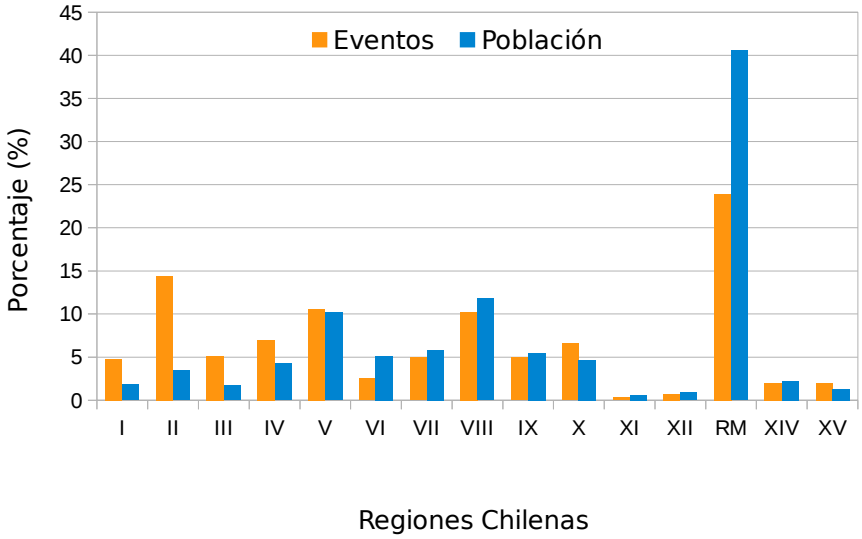


Figura 6.2: Distribución porcentual en regiones de eventos chilenos regionales y población chilena

El gráfico de la figura 6.2 muestra la distribución porcentual de los eventos regionales y la distribución porcentual de la población en cada región de Chile. El gráfico fue construido con la información recolectada entre el 1 de Noviembre del 2014 y el 30 de Abril del 2015.

Las barras color naranja muestran la distribución geográfica porcentual de los eventos regionales chilenos y las barras color celeste representan la distribución porcentual de la población en Chile. Como se puede observar, la mayor cantidad de eventos geolocalizados regionales están ubicados en la región metropolitana de Santiago, capital de Chile, lo que es esperable ya que Chile es un país centralizado en su capital y es en ella donde habitan la mayor cantidad de personas. Sin embargo, hay regiones en donde el porcentaje de eventos geolocalizados sobrepasa el porcentaje de la población en la región. Las regiones donde esto ocurre son las regiones I, II, III, IV y X. Analizando los datos se pudo ver que la mayor cantidad de eventos en esas regiones se debe a que esas regiones fueron afectadas por acontecimientos particulares durante el período en que se toma la muestra de datos. Durante el mes de Marzo del 2015 las regiones I, II, III, IV fueron afectadas por un temporal, que desbordó ríos, inundó terrenos poblados y dejó un número importante de personas desaparecidas y fallecidas. Por otro lado la region X fue afectada por erupciones volcánicas.

6.3. Distribución de usuarios

El sistema de recolección de *tweets* obtiene *tweets* escritos en español que contengan las palabras claves asociadas a los eventos. Por lo tanto, el espectro de usuarios que mantenemos en la base de datos abarca un conjunto amplio correspondiente a la población de habla hispana que comentó sobre algún evento identificado. Como se puede observar en el gráfico de la figura 6.3, sólo el 7 % de los usuarios geolocalizados están en Chile y la mayoría de los usuarios están geolocalizados en España, México o Argentina.

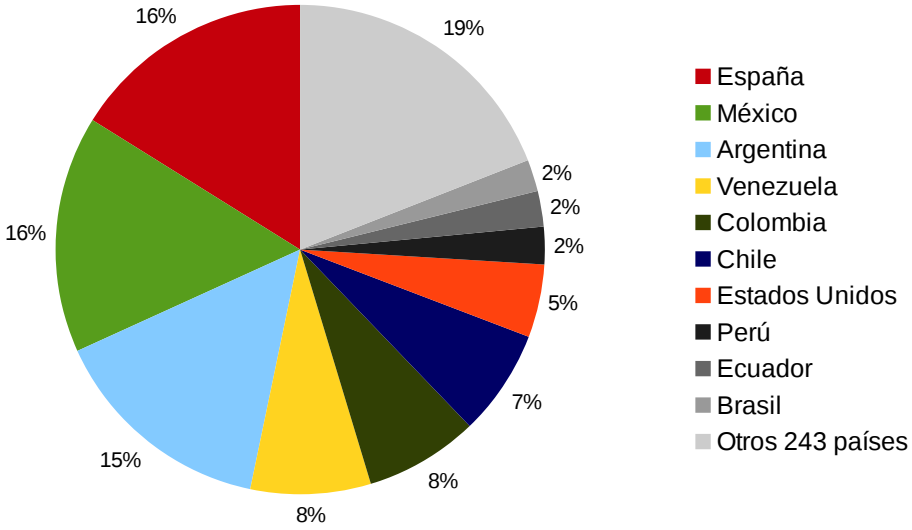


Figura 6.3: Distribución geográfica de los usuarios geolocalizados almacenados en la base de datos

Esto no quiere decir que los usuarios de España, México o Argentina comenten mucho

sobre noticias chilenas, ya que basta con que el usuario comente sobre un evento para que quede guardado en la base de datos. Al analizar los datos se pudo observar, por ejemplo, que muchos usuarios españoles comentaron sobre los ataques terroristas que sufrió Francia en el mes de Enero del 2015, un evento que fue publicado por los medios de comunicación nacionales chilenos y por lo tanto forma parte del registro.

Al igual que los eventos, los usuarios también pueden ser divididos en usuarios *nacionales* y usuarios *regionales*, ya que algunos tienen como localidad "Chile" y otros tienen información respecto a la región, comuna o ciudad en la que viven.

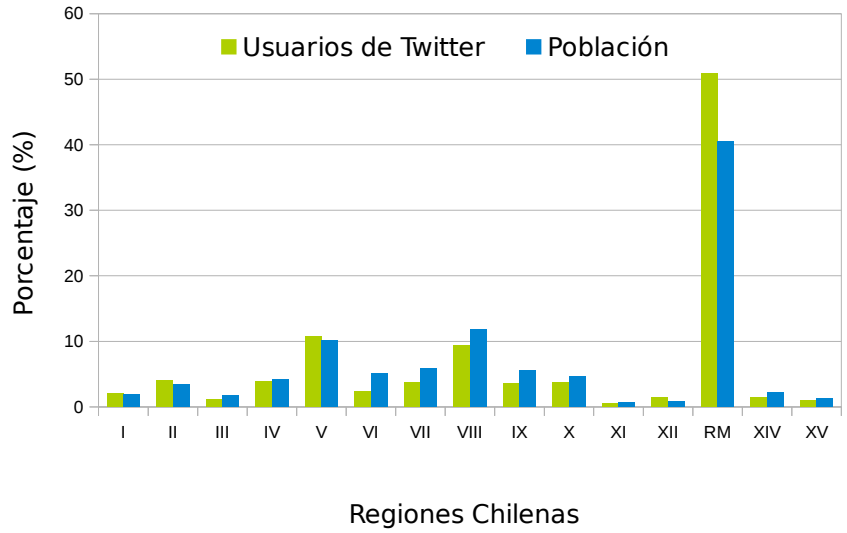


Figura 6.4: Distribución porcentual en regiones de usuarios chilenos en regiones y población chilena

El gráfico de la figura 6.4 muestra con barras verdes la distribución porcentual en regiones de los usuarios de *Twitter* chilenos que tienen información detallada respecto al lugar de Chile en el que habitan. Con barras celestes se muestra la distribución porcentual de la población en las regiones.

Como se puede observar el porcentaje de usuarios de *Twitter* y el porcentaje de la población en cada región están correlacionados. Lo que puede significar que durante el período hemos recolectado cuentas de usuarios chilenos que podrían representar una muestra proporcional a la población en cada región de Chile.

6.4. Caracterización de los eventos noticiosos chilenos

Varios meses posterior a la puesta en marcha del sistema de recolección de datos se realizó un análisis sobre los eventos identificados entre el 1 de Noviembre del 2014 y el 30 de Abril del 2015.

Se midió cuales fueron los eventos más comentados por usuarios chilenos durante ese

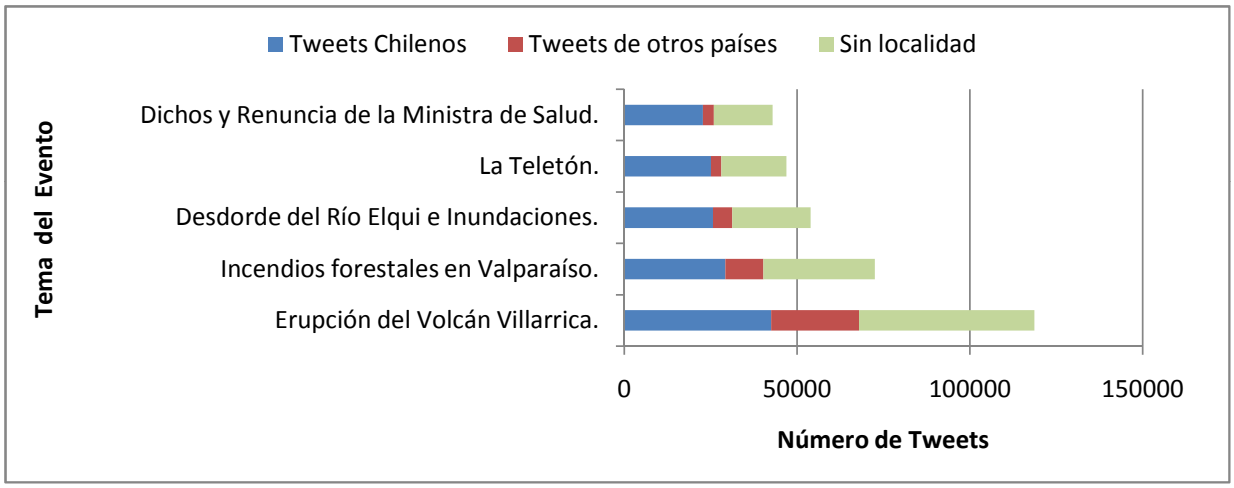


Figura 6.5: Gráfico de los eventos recolectados entre el 1 de Noviembre del 2014 y el 30 de Abril del 2015 que fueron más comentados por usuarios chilenos.

periodo. El gráfico de la figura 6.5 muestra la información obtenida. Lo primero que podemos observar es que todos los eventos son eventos que ocurrieron en Chile, a pesar de que, durante el periodo considerado el 27 % de los eventos geolocalizados eran eventos ocurridos en otros países.

También se puede observar, que el evento televisivo Teletón, causó interés similar a otros eventos de alto impacto como los incendios que afectaron a Valparaíso o el desborde del río Elqui.

Si se amplía el ranking a los 10 eventos más comentados por chilenos, el que se presenta en la tabla 6.2, se puede ver que otro evento televisivo logra posicionarse entre los más comentados, el Festival de Viña.

Ranking	*tweets* chilenos	Tópico del evento
1	42.483	Erupción del Volcán Villarrica
2	29.253	Incendio Forestal en Valparaíso
3	25.684	Desborde del Río Elqui a causa de las lluvias el día 26/03/2015
4	25.050	La Teletón
5	22.723	Dichos y posterior renuncia de la ministra de salud
6	20.249	Desborde del Río Elqui a causa de las lluvias el día 27/03/2015
7	18.976	Festival de Viña el día 25/02/2015
8	18.589	Desaparecidos en Diego de Almagro por inundaciones
9	16.958	Formalización principales involucrados Caso Penta
10	16.942	Festival de Viña el día 26/02/2015

Tabla 6.2: Tabla de los eventos más comentados por usuarios chilenos identificados durante el periodo entre el 1 de Noviembre del 2014 al 30 de Abril del 2015.

En el gráfico también se puede ver el número de *tweets* asociados a cada evento y los colores indican si corresponden a *tweets* publicados por usuarios chilenos (azul), usuarios

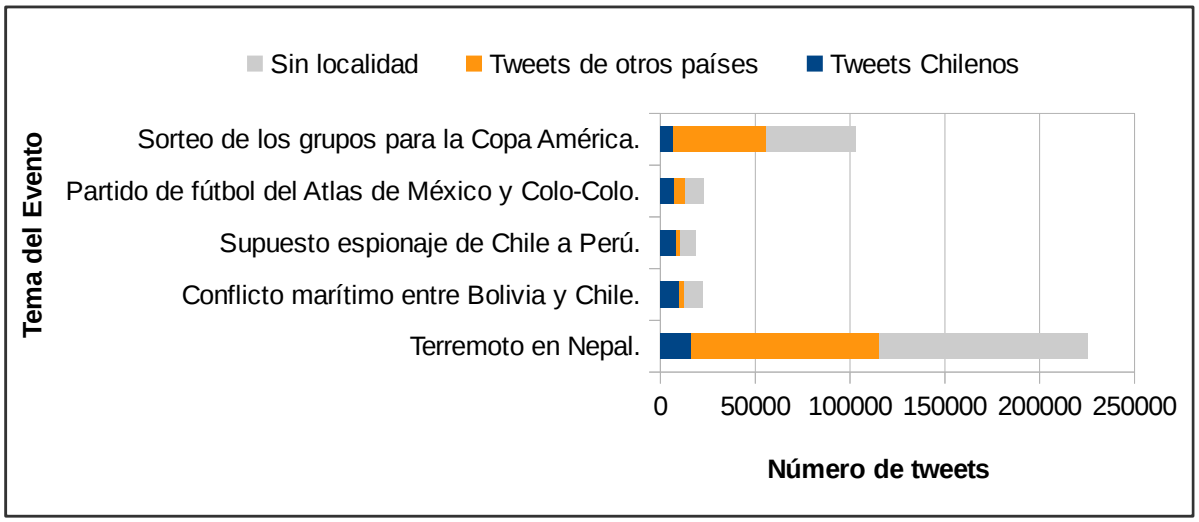


Figura 6.6: Gráfico de los eventos geolocalizados en otros países recolectados entre el 1 de Noviembre del 2014 y el 30 de Abril del 2015 que fueron más comentados por usuarios chilenos.

de otros países (rojo) o si no se sabe la localidad de los usuarios (verde). Todos los eventos fueron comentados por usuarios de otros países. Podrían ser chilenos viviendo en el extranjero, parientes o amigos de personas chilenas, personas que se interesan en Chile sin tener relación con el país, etc. Sin embargo, en base a la información que se tiene, no se puede saber qué relación tienen esos usuarios con Chile.

Al medir cuales fueron los eventos internacionales más comentados por chilenos se obtuvieron los resultados presentados en el gráfico de la figura 6.6. El primer lugar corresponde al terremoto en Nepal, el cual no sólo fue comentado por usuarios chilenos, si no que por usuarios provenientes de muchos otros países. El segundo y tercer lugar son eventos relacionados con Bolivia y Perú, dos países cercanos a Chile y ambos eventos también involucran a Chile. El cuarto y quinto lugar lo ocupan eventos relacionados con fútbol, un deporte que se sabe es muy popular en Chile. El sorteo de los grupos para la Copa América también fue muy comentado por usuarios de otros países.

Al analizar cuáles fueron los países que acumularon mayor número de *tweets* de usuarios chilenos en relación a sus eventos, descartando a Chile que es por lejos el más *twitteado* por usuarios chilenos, se obtiene el resultado que se muestra en la figura 6.7. El él se puede observar que los países que, en el tiempo, llamaron más la atención de los usuarios chilenos fueron: México, Argentina, Reino Unido, Nepal y Francia.

Al analizar los datos para cada uno de ellos por separado, se pudo observar que en relación a México, un tema muy *twitteado* por los usuarios chilenos fue el caso de los estudiantes mexicanos desaparecidos. Además fue un tema muy persistente, ya que a pesar de que fue un hecho acontecido antes del periodo para el cual se realiza el análisis, los nuevos descubrimientos del caso siguieron publicándose y comentándose hasta muchos meses después. Otros eventos comentados por usuarios chilenos, estaban relacionados a partidos de fútbol que tenían como protagonistas a equipos mexicanos y chilenos.

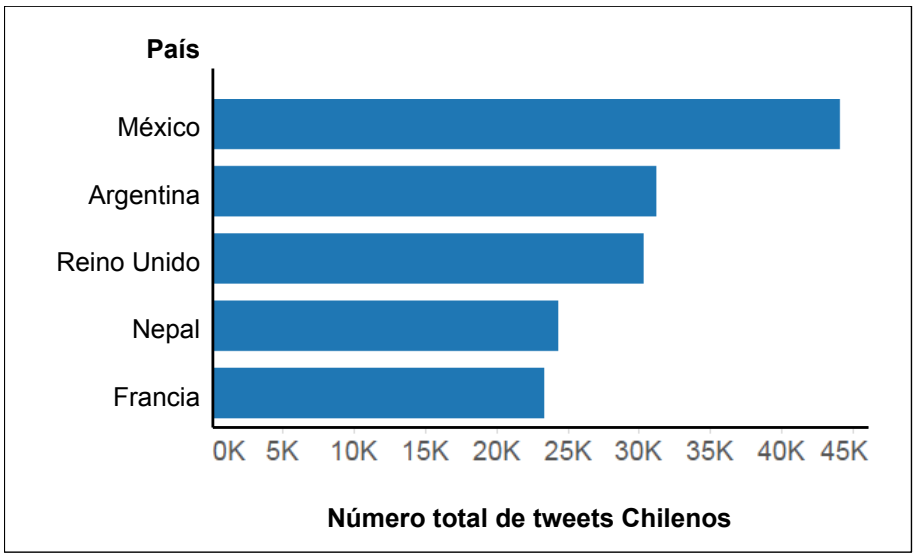


Figura 6.7: Gráfico de los países que acumulan mayor número de *tweets* de usuarios chilenos en relación a sus eventos dentro del periodo entre el 1 de Noviembre del 2014 y el 30 de Abril del 2015.

En relación a Argentina un evento muy comentado fue el caso de la inesperada muerte del Fiscal Nisman, un tema muy polémico. También hay eventos localizados en Argentina que se identificaron durante la erupción del volcán en el sur de Chile el cual afectó a algunos pueblos al otro lado de la cordillera.

En relación a Nepal, la atención se debe principalmente al evento correspondiente al terremoto que afectó a ese país y que dejó gente desaparecida y fallecida.

En relación a Francia se identifica un alza importante en el número de *tweets* de usuarios chilenos durante el mes de Enero. Los eventos que llamaron la atención de los chilenos estaban relacionados con los atentados terroristas en contra de Charlie Hebdo, un semanario satírico Francés, de parte de sujetos pertenecientes a la rama de Al Qaeda en Yemen, que asumió la responsabilidad por el ataque. Este evento no sólo llamó la atención de los usuarios chilenos, si no también, de usuarios de mucha otras partes del mundo.

Finalmente, en relación al Reino Unido, no fue posible identificar un evento específico de gran impacto que ocasionara un gran número de *tweets* de los usuarios. Sin embargo, como se muestra en la nube de etiquetas de la Figura 6.8 construida con las palabras claves de los eventos relacionados con el Reino Unido, se puedo observar que algunas de las palabras más repetidas fueron *Manchester*, *City*, *Arsenal*, *liverpool*, *Champions*, *League*, *Pellegrini*, *Alexis*, *Sanchez*, entre otras. Claramente el tópico más frecuente que relaciona a este país con Chile, es el fútbol. Y una razón importante puede ser que un jugador chileno y un entrenador chileno forman parte de importantes equipos de fútbol de ese país.

A partir de los análisis realizados se puede decir que, a pesar de que la prensa chilena informa sobre eventos ocurridos en otros países, ya que parte importante de los eventos recolectados involucran a otros países, los usuarios chilenos se interesan más por los eventos que ocurren en Chile.

Figura 6.8: Nube de etiquetas construida con las palabras claves de los eventos relacionados con el Reino Unido que son identificados a partir de los titulares de los medios de comunicación chilenos.

Los eventos ocurridos en otros países que llaman más la atención de los usuarios chilenos son eventos de alto impacto o son están directamente relacionados al fútbol.

Conclusión y Trabajo Futuro

Este trabajo consistió en la implementación de un sistema que permite recolectar, almacenar y visualizar contenido Web de carácter noticioso generado en la red social *Twitter* en Chile. En particular:

1. Se implementó un proceso de recolección de información sobre eventos noticiosos en *Twitter* utilizando una adaptación de la metodología utilizada por Mauricio Quezada en su trabajo de tesis de magíster [11]. El proceso de recolección se puso en marcha el 10 de Octubre del 2014, por lo que hasta la fecha se cuenta con una base de datos que contiene más de 17.800 eventos noticiosos de interés.
2. Se implementó un proceso de geolocalización de datos obtenidos de *Twitter*, en particular para la geolocalización de usuarios y eventos noticiosos. Para la implementación se utiliza como base el algoritmo diseñado por Vanessa Peña durante su trabajo de tesis de doctorado [10], el cual se modifica para adaptarse de mejor manera a las necesidades de este proyecto. Los procesos de geolocalización se utilizaron para geolocalizar los datos almacenados y se ejecuta periódicamente de forma automática para realizar la geolocalización de los nuevos datos recolectados.
3. Se implementó una primera versión de la aplicación Web que permite visualizar los eventos noticiosos y explorar los datos aplicando filtros. La visualización utilizada también está basada en una visualización presentada por V. Peña-Araya, M. Quezada y Barbara Poblete en su artículo *Galean: Visualization of Geolocated News Events from Social Media* [9] recientemente publicado.
4. Se realiza un análisis exploratorio de los datos recolectados durante el periodo entre el 1 de Noviembre del 2014 y el 30 de Abril del 2015.

Pese a que durante el desarrollo de la memoria se cumplieron casi a cabalidad los objetivos planteados existen aspectos en los cuales se podría continuar trabajando.

Como se mencionó en el Capítulo 5, con la metodología utilizada para geolocalizar eventos en ocasiones se producen errores, ya sea porque el evento es localizado en el lugar incorrecto, o porque no se logra localizar, por ejemplo, para eventos que tratan sobre personas públicamente conocidas que tienen nombres de localidades o cuando en los *tweets* relacionados al evento no se menciona un lugar con suficiente frecuencia.

En relación a la aplicación Web, al ser una primera versión, pueden realizarse muchas mejoras que extiendan el trabajo realizado y de esa forma ofrecer más o mejores opciones visualización y exploración de datos a los usuarios de la aplicación.

En cuanto a los objetivos:

El objetivo general fue cumplido, se logra desarrollar un sistema que recolecta, almacena y visualiza información de carácter noticioso generado en la red social *Twitter* en Chile.

- El primer objetivo fue cumplido. Se logró poner en marcha un sistema de recolección periódica de datos orientada a la detección de eventos de interés, el cual utiliza como fuente los titulares noticieros publicados por medios de comunicación chilenos y en base al número de *tweets* relacionados se puede medir el interés generado en las personas que utilizan la red social *Twitter*.
- El segundo objetivo fue cumplido de forma parcial. La información se almacena continuamente en una base de datos relacional, la que permite realizar consultas para efectuar análisis de los datos. No se logra resolver en completitud el problema de cómo dar acceso a la totalidad de los datos, ya que las normas de *Twitter* prohíben compartir datos extraídos en cualquier formato digital que permita hacer uso de ellos computacionalmente, es decir, formatos digitales como bases de datos, listas indexadas, json, etc. Por ahora usuarios podrían acceder sólo a la información que se provee mediante la aplicación Web desarrollada.
- El tercer objetivo fue cumplido. En la primera versión de la aplicación Web se presenta una visualización que permite observar los eventos posicionados en el mapa y utilizando otras variables como el color o el tamaño de los marcadores es posible identificar el impacto y el alcance de un evento. También se muestra información como titulares relacionados, *tweets* relacionados, etc.
- El cuarto objetivo fue cumplido. El diseño de la solución propuesta permite extraer módulos específicos de la solución para ser utilizados en otros proyectos. Además la aplicación Web utiliza un *framework* simple que facilita extender la solución agregando más funcionalidades sin mayores dificultades.

Como se mencionó en la introducción, se espera que el software desarrollado pueda formar parte de los servicios de almacenamiento digital de la Biblioteca Nacional de Chile. El trabajo aquí realizado representa un punto de partida para continuar avanzando en temas relacionados con el archivo digital público de la información contenida en redes sociales y su acercamiento a las personas.

Glosario

tweet: Mensaje publicado en Twitter compuesto por un máximo de 140 caracteres.

retweet: Mensaje publicado en Twitter que hace referencia a otro *tweet* publicado previamente por otro usuario.

twittear: Acción de publicar un *tweet* en *Twitter*.

mención: En *Twitter* es una palabra que tiene la forma @nombredeusuario y que permite interactuar con otros usuarios.

hashtag: En *Twitter* es una palabra que va precedida del símbolo # y que permite diferenciar, destacar y agrupar por una palabra o tópico especifico.

Bibliografía

[1] Actualización sobre el archivo de *Twitter* de la biblioteca del congreso de estados unidos. `http://blogs.loc.gov/loc/2013/01/update-on-the-twitter-archive-at-the-library-of-congress/`. Consultado el 1 de Octubre del 2014.

[2] Alexa top 500 global sites. `http://www.alexa.com/topsites`. Consultado el 1 de Agosto del 2015.

[3] Contrato y política de *Twitter* para elaboradores. `https://dev.twitter.com/es/overview/terms/agreement-and-policy`. Consultado el 1 de Agosto del 2015.

[4] La biblioteca del congreso de estados unidos adquiere el archivo completo de *Twitter*. `http://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/`. Consultado el 1 de Octubre del 2014.

[5] Límite de solicitudes a la api rest de *Twitter*. `http://dev.twitter.com/rest/reference/get/search/tweets`. Consultado el 1 de Agosto del 2015.

[6] Tratamientos del *Tweets*. `https://dev.twitter.com/overview/terms/display-requirements`. Consultado el 1 de Agosto del 2015.

[7] Helen HOCKX-YU. Archiving social media in the context of non-print legal deposit. Lyon, France, 2014. IFLA WLIC 2014.

[8] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

[9] Vanessa Peña Araya, Mauricio Quezada, and Barbara Poblete. Galean: Visualization of geolocated news events from social media. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 1041–1042, New York, NY, USA, 2015. ACM.

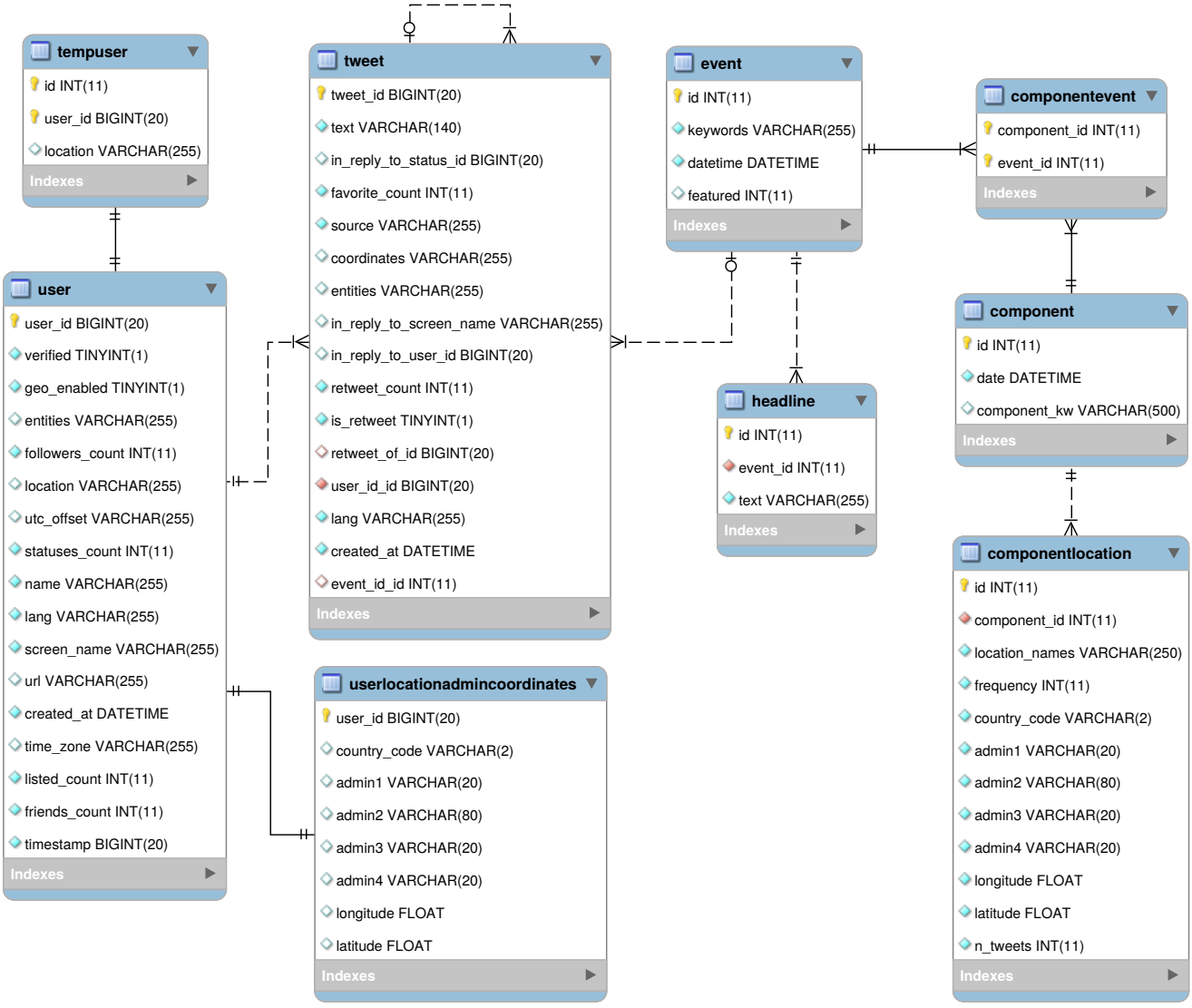
[10] Vanessa Peña. Multidimensional modeling and visualization of online social network events.

[11] Mauricio Daniel Quezada Veas. Identification and characterization of high impact news

events on twitter. Master's thesis, Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas, Departamento de Ciencias de la Computación, Santiago, Chile, 2014.

Anexo A: Base de datos

Figura 6.9: Esquema entidad relación de la base de datos utilizada para almacenar la información almacenada y procesada.



Anexo B: Fuentes de noticias Chilenas

La Tabla 6.3, Tabla 6.4, Tabla 6.5 y Tabla 6.6 muestran la información de las cuentas de *Twitter* de medios noticiosos Chilenos utilizadas como fuentes de noticias para la recolección de datos.

Medio de Comunicación	Tipo	Cuenta en *Twitter*	Zona de Chile
Chilevisión Noticias	Televisivo	@chv_noticias	Nacional
24 Horas TVN	Televisivo	@24horastvn	Nacional
Ahora Noticias MEGA	Televisivo	@ahnoticiasmega	Nacional
Tele 13	Televisivo	@t13	Nacional
CNN Chile	Televisivo	@cnnchile	Nacional
ADN Radio Chile	Radial	@adnradiochile	Nacional
Radio Bio-Bio	Radial	@biobio	Nacional
Radio Cooperativa	Radial	@cooperativa	Nacional
Mauricio Bustamante	Periodista	@tv_mauricio	Nacional
Salvador Schwartzmann	Periodista	@s_schwartzmann	Nacional
Portal Terra	Digital	@terrachile	Nacional
Carabineros de Chile	Institución	@carabdechile	Nacional
Onemi Chile	Institución	@onemichile	Nacional
El Mostrador	Periodístico	@elmostrador	Nacional
La Cuarta	Periodístico	@lacuarta	Nacional
La Nación	Periodístico	@nacioncl	Nacional
La Segunda	Periodístico	@la_segunda	Nacional
La Tercera	Periodístico	@latercera	Nacional
Las Últimas Noticias	Periodístico	@lun	Nacional
The Clinic	Periodístico	@thecliniccl	Nacional
Diario la Hora	Periodístico	@diariolahora	Nacional
Publimetro	Periodístico	@publimetrochile	Nacional
Emol	Periodístico	@emol	Nacional

Tabla 6.3: Listado de cuentas de medios de comunicación Nacionales utilizadas para la recolección de titulares noticieros.

Medio de Comunicación	Tipo	Cuenta en *Twitter*	Zona de Chile
	Periodístico	@diarioatacama	Región de Atacama
	Periodístico	@chañarcillo	Región de Atacama
	Periodístico	@diarioelnortino	Iquique
	Periodístico	@ellongino	Región de Tarapacá
	Periodístico	@elboyaldia	Iquique y Tarapacá
	Periodístico	@eldia_cl	La Serena
	Periodístico	@mercurioafta	Antofagasta
	Periodístico	@mercuriocalama	Calama
	Periodístico	@elmorrocotudo	Arica y Parinacota
	Periodístico	@elnortero	Antofagasta y Calama
	Periodístico	@elobservatodo	La Serena y Coquimbo
	Periodístico	@elovallino	Ovalle
	Periodístico	@serenaycoquimbo	Serena y Coquimbo
	Periodístico	@estrella_antofa	Antofagasta
	Periodístico	@estrelladearica	Arica
	Periodístico	@laestrellaiqq	Iquique
	Periodístico	@estrella_loa	Valle del Loa
	Periodístico	@estrella_toco	Tocopilla

Tabla 6.4: Listado de cuentas de medios de comunicación del Norte de Chile utilizadas para la recolección de titulares noticieros.

Medio de Comunicación	Tipo	Cuenta en *Twitter*	Zona de Chile
	Periodístico	@diarioivregion	Región de Valparaíso
	Periodístico	@elaconcagua	Valle de Aconcagua
	Periodístico	@el_amaule	Curicó y Talca
	Periodístico	@diarioelcentro	Región del Maule
	Periodístico	@el_ciudadano	Santiago
	Periodístico	@diariolabrador	Melipilla
	Periodístico	@diariolider	San Antonio
	Periodístico	@elmartutino	Valparaíso y Viña
	Periodístico	@soyvalparaiso	Valparaíso
El Observador de Quillota	Periodístico	@eo_enlinea	Quillota
	Periodístico	@elparadiario14	La Florida
	Periodístico	@elrancaguino	Rancagua
	Periodístico	@elrancahuaso	Rancagua
	Periodístico	@laprensacurico	Curicó
	Periodístico	@ultimahoracl	Rancagua

Tabla 6.5: Listado de cuentas de medios de comunicación del Centro de Chile utilizadas para la recolección de titulares noticieros.

Medio de Comunicación	Tipo	Cuenta en *Twitter*	Zona de Chile
	Periodístico	@noticiasmalleco	Malleco
	Periodístico	@cronicachillan	Chillán
	Periodístico	@diariodeaysen	Aysen
	Periodístico	@diarioconce	Concepción
	Periodístico	@australtemuco	Región de La Araucanía
	Periodístico	@soyvaldiviacl	Región de Los Ríos
	Periodístico	@austral_osorno	Osorno
	Periodístico	@elconcecuente	Región del BioBío
	Periodístico	@ddivisadero	Coyhaique
	Periodístico	@diarioelgong	Temuco
	Periodístico	@diario_eha	Puerto Varas
	Periodístico	@informadordig	Cuenca del Río Imperial
	Periodístico	@ellanquihue	Puerto Montt
	Periodístico	@elmagallanews	Punta Arenas y Magallanes
	Periodístico	@elnaveghable	Valdivia y Los Ríos
	Periodístico	@elpatagonicocl	Punta Arenas
	Periodístico	@pinguinodiario	Punta Arenas
	Periodístico	@elrepuertero	Puerto Montt y Los Lagos
	Periodístico	@elsurcl	Concepción, Región del BioBío
	Periodístico	@elvacanudo	Osorno y Los Lagos
	Periodístico	@ladiscusioncl	Chillán
	Periodístico	@estrellachiloe	Chiloé
	Periodístico	@estrellaconce	Concepción
	Periodístico	@laopinon	Temuco y Araucanía
	Periodístico	@laprensaustral	Magallanes
	Periodístico	@latribunacl	Los Angeles
	Periodístico	@lanalhue	Cañete
	Periodístico	@diariolasnotic	Malleco
	Periodístico	@TribunaC	Colchagua

Tabla 6.6: Listado de cuentas de medios de comunicación del Sur de Chile utilizadas para la recolección de titulares noticieros.

Anexo C: Artículo publicado

Durante el período de realización de la memoria, se realizó una caracterización de eventos noticiosos Chilenos extraídos de Twitter. Los resultados se publicaron en el Workshop TAIA'15 realizado durante la conferencia SIGIR 2015 en Santiago. El artículo se titula *"Spatio and Temporal Characterization of Chilean News Events in Social Media"*. El artículo se adjunta en las siguientes cuatro páginas.

Spatio and Temporal Characterization of Chilean News Events in Social Media

Jazmine Maldonado
Department of Computer Science
University of Chile
Santiago, Chile
jamaldonadof@ug.uchile.cl

Vanessa Peña-Araya
Department of Computer Science
University of Chile
Santiago, Chile
vpena@dcc.uchile.cl

Barbara Poblete
Department of Computer Science
University of Chile
Santiago, Chile
bpoblete@dcc.uchile.cl

ABSTRACT

Online Social Networks play a leading role in news consumption. As a consequence, most newspapers and other media use these platforms to promote their content. However, the geographic bias in the media, in addition to the demographic bias in Online Social Networks can lead to inaccurate and incomplete view of the news in a country. Being aware of these two kinds of bias in news published in Online Social Networks is useful to understand the context in which events develop. We selected Chile as a case study to observe these problems. Chile is a country with a high degree of participation in Online Social Networks and suffers from both issues: media covers mostly news from Santiago, its capital, and most of Online Social Networks users are located in this city. We built a dataset of Chilean news headlines extracted from Twitter.

We conducted a characterization of news and messages which comment them. We focus on the geographical and temporal features of news. In this paper we present the results of this analysis in addition to the description of the dataset. Our findings show that as expected, news and Twitter users are mainly concentrated in Chile's capital. In addition, users in Chile focus on local news paying little attention to international events. We observed that a considerable number of users discussing Chilean news are located outside of Chile. We conclude that users in Chile are subject to bias in news media coverage of information, which privileges news from the largest cities.

Keywords

Geo-temporal analysis, event coverage in Social Media, case study

1. INTRODUCTION

Online social networks have become an important source of information. Their growth has allowed for regular citizens and not only traditional news media to inform when an important event happens. Indeed, it is not uncommon that the first information about breaking news is published by a regular user instead of a journalist. However, even if the involvement of users in Online Social Networks has helped decrease the bias in information coverage introduced by news media, by no means are these platforms geographically representative source of information. We assume the task of characterizing news extracted from Online Social Network by the geographical distribution of the location they mention, in addition to that of the Online Social Network users that share them. We believe that this analysis will help understand the context in which news events develop, in addition to unveiling biases in information coverage.

Chile is one of the top ten countries using Twitter [1]. As Chile is a very centralized country, the geographical distribution of Twitter users and news visibility is not homogeneous and mainly focused in Santiago, its capital. To observe the characteristics of the geographical behaviour of Twitter users and news over time, we built a dataset by gathering news headlines from Chilean newspapers' Twitter accounts. For each news topic extracted from the headlines we retrieve its keywords and the tweets [2] that comment about it. In this work we present a case study of this dataset.

2. RELATED WORK

This case study is based on an extension of the work of Vanessa Peña-Araya, Mauricio Quezada and Barbara Poblete [2] which presents spatio-temporal event models.

The work of Graells-Garrido and Lalmas [1] covers how the physical centralization of Chilean population affects the participation of Twitter users. Although we also cover centralization, our work is broader as it covers other issues and considers behaviour of the Chilean media.

The work developed by Yom-Tov and Diaz [4] analyzes how users participate in an event considering the geographical distance to the venue and considering whether they have relationships with people in the venue. In our dataset we observe both national and international participation in the events, with different percentage of participation depending on whether they are international or national events, but the analysis we made is more general and is presented as a comparison between the media coverage and the number of people who comment about the events.

[1] http://www.forbes.com/sites/victorlipman/2014/05/24/top-twitter-trends-what-countries-are-most-active-whos-most-popular/

[2] tweets: messages posted in Twitter composed of maximum 140 characters

TAIA'15, August 13, 2015, Santiago, Chile.

3. EVENT EXTRACTION AND DATASET CONSTRUCTION

To build the Chilean news dataset, we conduct a 2 steps process. We first collect news events from Chilean Twitter accounts of media entities. The second step is to geolocate where events occur and also the users who comment them.

News event collection is done by periodically retrieving tweets from a set of selected twitter news accounts from Chile. The set of accounts was manually curated and complete including accounts from online newspapers in regions of Chile. It also includes radio accounts, TV news accounts, institutions and some journalist accounts. Periodically, we extract representative terms for each news item and perform a search to retrieve related messages from the public Twitter API. This methodology was created by collaborators for another project, see Acknowledgements Section.

In the second step of the process users and events are geotagged. To geolocate users, the system takes the user location field in their Twitter profile and resolves using CLAVIN [3]. On the other hand to geolocate an event we extract the locations that are mentioned in an event's headlines and comments. To obtain locations with higher-level of administrative divisions than country or region, we use two sources of geographical locations. First of all, the system searches for text matches in a dedicated list of Chilean cities or country names. If no location is matched or the number of mentions of those that were indeed found are too small, the system uses CLAVIN. Even when both methods are used, some events are not possible to geotag. This happens because there were not locations mentioned in tweets related to those events or because the places mentioned are cities which cannot be identified using CLAVIN.

In the set of events that can not geotagged there are events that do not mention names of cities but mention other information like the names of the local soccer teams like "Colo-Colo", names of some known monument like "La Moneda", names of people who are international figures like "Maradona", etc. Some of these cases were identified and they are considered in the process of geotag an event, but using our approach they must be identified manually so they are exceptions.

As a result of using this process over a period of six months, we constructed a dataset of Chilean interest events which we plan to share with the community for research purposes. The dataset is composed of events, and each one contains: (i) the date when it was detected; (ii) its most representative keywords; (iii) one or more locations related to it (where it happened, countries involved, etc); and (iv) a set of tweets IDs commenting about it. In addition, each tweet ID is geolocated, when possible.

4. CHILEAN NEWS EVENTS

In this section we analyze the Chilean news dataset focused on its geographical and temporal characteristics. We first give an overview of the dataset (subsection 4.1). In subsection 4.2 we analyze the geographical distribution of Chilean news and Chilean Twitter users. We finally inspect the interest that Chilean media and Twitter users have in international countries in subsection 4.3.

4.1 Dataset Overview

The dataset is composed of events collected over a period of six months, starting from November 1, 2014 to April 30, 2015. Table 1 provides an overview of dataset in terms of the number of events and Twitter users gathered. It also gives the number and percentages of events and users that were geotagged (one or more locations were assigned). From the total of 6507 geotagged events, 4740 of them (72.8%) are news concerning Chile, and 1767 (27.2%) are news about other countries. For the purpose of this paper, we call the first kind of news *national events*. On the other hand, news events concerning countries outside Chile are referred to as *international events*.

	Users	Events	Tweets
geotagged	2,179,255(44%)	6,507(60%)	14,121,553(51%)
non-geotagged	2,796,443(56%)	4,415(40%)	13,618,515(49%)
Total	4,975,698	10,922	27,740,068

Table 1: General information and totals of events and users of the dataset

4.2 Geographical distribution of Chilean news and Twitter users

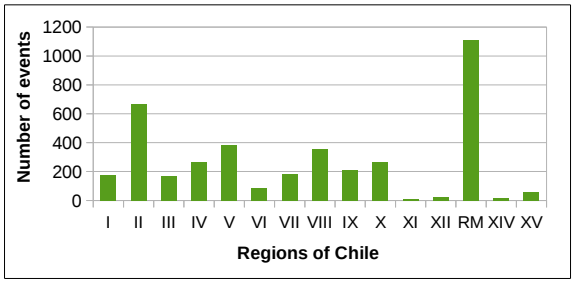


Figure 1: Number of events detected for each region of Chile

To analyze the geographical distribution of Chilean events it is important to know that Chile is divided in to 15 *regions*. Regions are the first-level administrative division of Chile and are named by a Roman numeral. We divide events in two categories: (i) events regarding specific Chilean locations and (ii) events concerning the whole nation. Events in category (i) occur when at the moment of geotagging the events, was possible to find names of cities or regions of Chile a considerable number of times into the set of tweets associated with the event. In the case of events in category (ii) only was possible to find "Chile" a lot of times but not mentions of cities or regions. For the first category we found events like natural disasters, regional initiatives, accidents, etc. For events in which the whole country is involved, we found news such as political topics, a new law approvals, TV show topics, national soccer matches, among others. From the total 4740 Chilean events in the dataset, 3450 of them are events associated with specific locations in Chile and 1290 are general Chilean events.

Figure 1 displays the distribution of news events related with each region of Chile. The figure only represents events concerning a particular Chilean region, category (i), and it does not include events belonging to (ii) event category.

Nevertheless, although we consider headlines for all regions of Chile, most events refer to the capital of the country Santiago which is located in the Metropolitan Region (RM). Region II also has many events, mainly because on March 25 a large temporary affected the area leaving a large number of dead and missing persons. Furthermore, the regions XI and XII of Chile, have very few events that mention them.

From this information, one can ask whether the concentration of news events in the capital is because no newsworthy events occurs in other regions of the country, or if instead it is the product of bias in information coverage by the media. From the data we observe that the national media tends to inform only about big disasters in smaller localities, but little about everyday problems of people there. However, even low impact news in the capital, such as malfunction of Santiago's subway trains, are covered by national TV news.

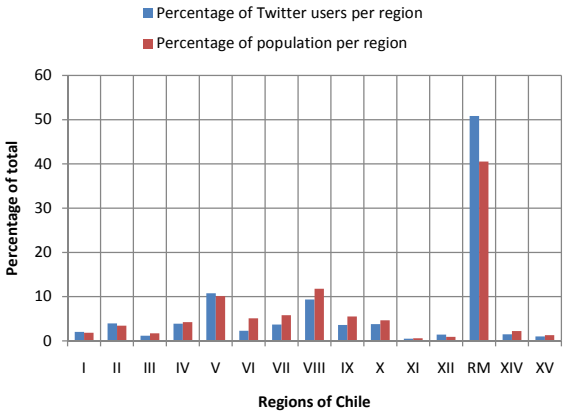


Figure 2: Percentage of number of Twitter users and population of each region of Chile

Regarding geolocated Twitter users, from the total of users who comment on news published in Chile (national and international), 8,4% of them are in Chile. The rest consists of users from other countries, among them, there are some Spanish-speaking countries with a considerable amount of users as Spain, Mexico, Argentina, Venezuela and Colombia, which together represent 62% of users in the dataset. The remaining 30% is made up of users located in 245 other countries. The user geotagging process obtains the information from Twitter user location field, so is unknown if the geolocated Twitter users effectively live in the country that they say they live, or if they are Chileans that live in those countries. As for the geographical distribution of Chilean users, 27.8% of them only put "Chile" in their location field, and the remaining 72.2% have "Chile" in addition to a region name or a city name.

The Figure 2 shows the distribution of Chilean users that specified in which part of Chile they are from and the population of each region of Chile. In the same way of events distribution, the concentration of users is mainly in the capital, Santiago, and the difference between users distribution is even more drastic than event distribution. But we also can observe the distribution of Twitter users is very similar to population in each region. The *centralization* is a known problem of Chile in which most of the population and resources are located in Santiago and somehow this condition is reproduced in the number of Twitter users in the dataset we build. About the relation between the number of tweets and users per region, there are in average 18 tweets per user in each region with an standard deviation of 1,8.

4.3 Geopolitical interest of Chilean media and Twitter users

We present the geopolitical interest in international countries by Chilean media and Chilean Twitter users. As we previously mentioned, 27.2% of news events are international events, involving other countries than Chile. This means that, even if the seeds of events are Chilean media entities, a considerable part of headlines are international news, showing the Chilean media interest in external countries.

Of all collected data, the events with more tweets are international events or related with other countries that are not Chile, for example, the Oscar Awards, the Nepal earthquake, Charlie Hebdo Shooting, etc. As expected, people that comment about them are from different Spanish-speaking countries and the most are not from Chile. The international events with more interest of Chilean people are related with nearby countries, have less than 6,000 Chilean tweets and are not in the top 30 of most tweeted by Chileans.

Although Chilean media and users have interest in international events, Chilean users are by far more interested in Chilean news. Table 2 shows the top five events most tweeted by Chilean people.

Description	Location	Chilean Tweets
Villarrica Volcano Eruption	IX - La Araucanía	42,483
Forest Fire in Valparaíso	V - Valparaíso	29,253
Elqui River Overflow	Chile	25,684
Chile's Teleton	Chile	25,050
Resignation of the Health Minister Helia Molina	Metropolitan Region of Santiago	22,723

Table 2: Most commented events by users in Chile

Locations of the events were originally in Spanish, and they were translated to English for the purpose of this paper. Even though the dataset includes several international events tweeted by Chilean users, the top five events most twitted by Chileans occur in Chile. About Chilean engagement, in average, we observe that users in Chile tweet aprox. 2 tweets per event, in those with the most activity.

Figure 3 shows the number of tweets of users of other countries in events with highest participation of Chilean people. Although the Chilean participation is higher than in other events, users in other countries also have an important presence in these national events.

To inspect which countries are relevant for Chilean media and Twitter users, Figure 4 shows the top ten countries found in events. On the left side it shows the amount of events concerning each country, and on the right side it shows the distribution of events per month. As can be observed, the interest of media is similar to the interest of Chilean people. The only exception are Nepal and Uruguay, which appear to be more interesting for Chilean people but were not covered much by media. In the upper-right side of the visualization, we can see that the most tweeted events in Nepal are mainly in April, the month when the Nepal

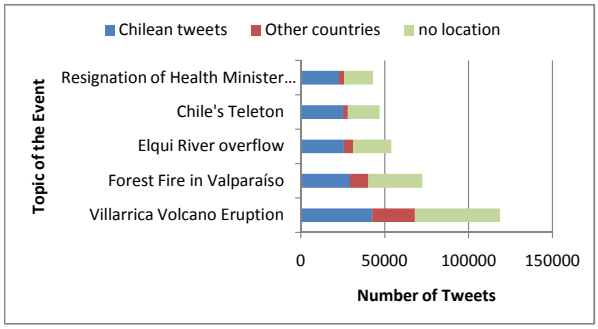


Figure 3: User participation in the events with more tweets by users in Chile

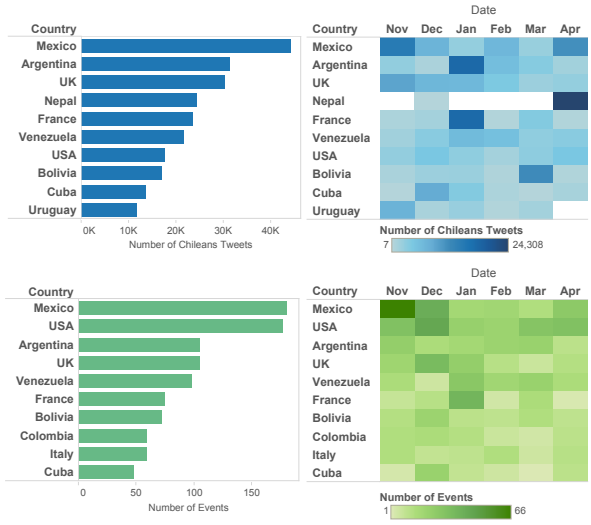


Figure 4: Top Ten countries that concentrated the most interest for Chilean users (in Blue) and for Chilean Media (in Green) during the period Nov/2014 to Apr/2015.

Earthquake happened and after the Everest avalanche. In the case of Uruguay, we can see that events are more evenly distributed over time, so they are probably not the same event but several different events involving the country. A look at event's keywords, showed us that the most tweeted Uruguayan events are related with soccer matches between Chile and Uruguay in November.

Regarding time, we can observe that Chilean media has a more even coverage of countries over time, than the interest of Chilean users. Indeed, there are periods of time when Chileans did not tweet about certain countries, like Nepal, to only pay attention when important news happens. Mexico and USA appear to be covered by Media and Twitters users in a similar way. Analysing the keywords of Mexican events, we observe that the news of the Mexican's students disappearance was extensively covered by Chilean media and also very commented by Chileans users. It is important to notice that, even if the event happened before we started our collection, people and media still continued to talk about it, being as it was a very persistent topic over time. Different is the case of nations like United Kingdom, France and Italy. By inspecting events keywords involving these countries, we observe that the most frequents events are related with soccer. This could be due to the fact that several Chilean soccer players are members of important European teams. Only France has an important spike in January, the month when a series of terrorist attacks occurred.

5. CONCLUSIONS

We have presented our preliminary findings of Chilean news on Twitter. For this we have built a dataset of events from Chilean news media on Twitter and gathered the tweets that discuss them. By analyzing events by their geographical and temporal characteristics, we have three main conclusions. The first one is that even if Chilean media has interest in international events, Chilean Twitter users are more interested in local news. In fact, Chilean users only focus on international news for extremely high-impact events (e.g. terrorist attacks and natural disasters) or for soccer matches. Our second finding is that Twitter users from countries outside Chile are interested in Chilean news: a considerable number of them commented about Chilean events, particularly for natural disasters. As we mention before, a large part of users geotagged with countries outside of Chile and tweet in Chilean events could very well be Chileans living abroad. Finally, we were able to confirm that Chilean media displays from geographical bias in news coverage as most of news were from Santiago, Chile's capital. This is most likely driven by commercial interests moved by audience distribution in the country.

6. ACKNOWLEDGMENTS

We acknowledge Janani Kalyanam from UCSD and Mauricio Quezada for designing and allowing us to use their news event collection procedure.

The authors were partially supported by FONDECYT Grant 11121511 and the Millennium Nucleus Center for Semantic Web Research under Grant NC120004.

References

[1] E. Graells-Garrido and M. Lalmas. Balancing diversity to counter-measure geographical centralization in microblogging platforms. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14, pages 231–236, New York, NY, USA, 2014. ACM.

[2] M. Quezada, V. Peña Araya, and B. Poblete. Location-aware model for news events in social media. In *SIGIR*, 2015, Santiago, Chile.

[3] B. Technologies. CLAVIN: Cartographic Location And Vicinity INdexer. `http://clavin.bericotechnologies.com/`, 2012–2013.

[4] E. Yom-Tov and F. Diaz. Out of sight, not out of mind: On the effect of social and physical detachment on information need. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 385–394, New York, NY, USA, 2011. ACM.