



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

IDENTIFICACIÓN DE CONTENIDO MULTIMEDIA RELEVANTE A PARTIR DE EVENTOS
UTILIZANDO SU INFORMACIÓN SOCIAL

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL EN COMPUTACIÓN

MAURICIO DANIEL QUEZADA VEAS

PROFESORA GUÍA:
BÁRBARA POBLETE LABRA

MIEMBROS DE LA COMISIÓN:
SERGIO OCHOA DELORENZI
MAURICIO MARÍN CAIHUAN

SANTIAGO DE CHILE
DICIEMBRE 2012

Resumen

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetur adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Una dedicatoria corta.

Agradecimientos

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Índice general

1. Introducción	1
1.1. Contexto y Motivación	2
1.2. Objetivos	3
1.2.1. Objetivo general	3
1.2.2. Objetivos específicos	4
1.3. Descripción general de la solución	4
2. Antecedentes	6
2.1. Twitter	6
2.2. Identificación automática de eventos	7
2.3. Clustering de documentos	7
2.3.1. Evaluación de clusterings	7
2.4. Resúmenes automáticos	7
2.4.1. Evaluación de resúmenes	7
2.5. Ranking de documentos	7
3. Especificación del Problema	8
3.1. Descripción detallada	8
3.2. Relevancia de una solución	8
3.3. Características de calidad	8
3.4. Criterios de aceptación	8
4. Descripción de la Solución	9
4.1. Modelo formal	9
4.2. Metodología de desarrollo	9
4.3. Metodología de obtencion del dataset	9
4.3.1. Recolección de eventos	9
4.3.2. Enriquecimiento de eventos	10
4.3.3. Identificación de documentos a partir de tweets	11
4.4. Desafíos técnicos	12
4.5. Restricciones de la API de Twitter	12
5. Conclusiones	13
5.1. Resumen del trabajo realizado	13
5.2. Objetivos alcanzados	13
5.3. Relevancia del trabajo realizado	13

5.4. Trabajo futuro	13
6. Anexos	14

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

Al igual que en el buffet de un restaurante, por mucho que se quisieran comer todos los platos favoritos, es imposible comer todo lo que uno quisiera por razones obvias. Una posibilidad es probar un poco de cada comida, para así saber qué es lo más delicioso y comer hasta hartarse.

Pero, ¿qué hacer si hay demasiados platos y no se conocen todos? de alguna manera hay que saber cuáles hay que probar, si el objetivo es comer lo mejor posible. Un amigo puede recomendar una u otra comida, lo cual puede servir para orientarse. Entonces se pueden escoger pequeñas muestras de acuerdo a las recomendaciones.

Complicando más el escenario, qué pasa si este restaurant tiene además música en vivo, y por alguna razón, se tiene el privilegio de escoger qué escuchar. En este caso, ya no es posible “probar” un poco de cada tipo existente, no sólo por la cantidad, sino porque no es posible juzgar un grupo musical por una canción o un extracto de ella. Si se quiere tener la mejor velada, pudiendo disfrutar de cada uno de los panoramas que ofrece, es necesario tener algo de información para poder escoger.

Pasando a un contexto diferente, supóngase que este gran buffet es la Web y los distintos platos corresponden a contenido publicado en ella. Por lo tanto, dada la gran cantidad de información disponible, se hace necesario poder encontrar lo más atractivo de acuerdo a la preferencia del usuario o de los usuarios. Nótese que se está haciendo otra suposición importante con esta analogía, y es que se está considerando que la información es íntegramente para ser *consumida*, y no, por ejemplo, para generar más contenido, conocimiento, o para ser utilizada por máquinas, etc. Dentro de este contexto se plantea la pregunta de cómo seleccionar el contenido más atractivo dentro de todo lo que hay disponible en un momento dado.

Siguiendo el razonamiento de la analogía, una manera de poder seleccionar sólo el contenido más “atractivo” (de acuerdo a las preferencias del usuario), es probar un poco de cada uno. O bien, generar un *resumen* de cada documento y presentarlo al usuario.

Sin embargo, tal como se dijo antes, no es posible hacer lo anterior para contenido que no puede ser resumido directamente para ser consumido (la música como tal, o bien, vídeos o imágenes). En esta perspectiva, sólo las recomendaciones pueden ayudar a determinar lo más conveniente

de acuerdo al usuario.

Este trabajo consistió en el desarrollo de una primera aproximación que permite generar resúmenes automáticos de eventos bien definidos a partir de los documentos Web que hablan de éstos. Los documentos son considerados *multimedios*: texto, imágenes, vídeos, sonidos; es decir, no necesariamente texto. El contenido es filtrado o seleccionado de acuerdo a indicadores sociales: los objetos más *tomados en cuenta* en la Web son considerados más importantes.

El sistema implementado consideró dos tipos de eventos: noticias y conciertos musicales. Para obtenerlos, se utilizó el servicio de Google News¹ y Last.fm². Para medir la relevancia de los documentos y obtener los mismos se utilizó la red social Twitter³, que provee una *Application Programming Interface* (o API) para realizar búsquedas y obtener información sobre los *tweets* o mensajes cortos que publican los usuarios de la red.

La estructura de este informe es como sigue: en este capítulo se comenta el contexto dentro del cual se desarrolló este sistema, las contribuciones realizadas, los objetivos y una descripción general de la solución; en el siguiente capítulo se discute el estado del arte y el marco teórico del cual se desprende este trabajo; el capítulo 3 describe más en detalle el problema a resolver, su relevancia y sus dificultades, para luego, en el capítulo 4, describir la solución implementada, más un par de casos de estudio sobre los resultados obtenidos, para finalizar con las conclusiones de este trabajo en el capítulo 5.

1.1. Contexto y Motivación

La tasa de crecimiento de la cantidad de datos en la Web, y en particular, en las *redes sociales online* (OSN, *Online Social Networks*), es de tal magnitud que se vuelve necesario encontrar formas de filtrar y buscar sólo la información relevante dentro de todas las fuentes que hablan del mismo tópico o tema.

En el contexto de las redes sociales online, cada día se publican millones de *actualizaciones de estado* (mensajes breves sobre el estado actual del usuario) con respecto a distintos tópicos, ya sean conversacionales, personales o sobre algún evento en particular⁴. Además, el auge de los teléfonos inteligentes o *smartphones* con mayor capacidad de procesamiento e integrados con todo tipo de sensores (cámaras fotográficas, de vídeo, acelerómetro, barómetro, osciloscopio, etc.), hace posible el generar aun más información y e incluso en tiempo real sobre lo que acontece en el mundo, en internet, o bien sobre el estado particular de cada usuario.

Este aumento y evolución de la generación de datos no sólo influye en la riqueza en la variedad de éstos, sino también en el comportamiento de los usuarios a lo largo del tiempo. Actualmente, una gran parte de los usuarios valora más el contenido de tipo multimedia (imágenes y videos) en las redes sociales online⁵. Se hace entonces necesario encontrar formas para satisfacer estas

¹<http://news.google.com>

²<http://last.fm>

³<http://twitter.com>

⁴Pear Analytics. Twitter Study <http://es.scribd.com/doc/18548460/Pear-Analytics-Twitter-Study-August-2009>

⁵The Rise of Visual Social Media <http://www.fastcompany.com/3000794/rise-visual-social-media>. En el artículo

necesidades de los usuarios, las cuales ya han sido abordadas en parte, como la generación de resúmenes automáticos orientado a motores de búsqueda, o la determinación de la relevancia tanto de documentos en la Web como de actualizaciones de estado en las redes sociales.

Surge como motivación el poder identificar y extraer contenido relevante de la Web, a partir de eventos, y además avanzar un paso más arriba en el nivel de abstracción: considerar los documentos no por su contenido textual, lo que permite abarcar imágenes, vídeos, sonidos y multimedios. Algunas aplicaciones directas de esto son, entre otras:

- Ayudar al trabajo periodístico mediante una colección de contenido multimedia relacionado a un evento noticioso. Por ejemplo, la versión online de Radio Biobío⁶ frecuentemente publica breves artículos sobre sucesos que tienen impacto en las redes sociales, mostrando un pequeño conjunto de mensajes con comentarios de la gente⁷. Una aplicación directa involucraría considerar además contenido multimedia, y organizar este contenido de acuerdo a la relevancia que tiene dentro de las redes.
- Enriquecer la búsqueda en la Web a través de contenido multimedia. Una persona buscando información sobre un concierto podría obtener imágenes y vídeos de éste fácilmente una vez identificado el concierto.
- Siguiendo lo anterior, un grupo musical podría obtener toda la información multimedia asociada a su concierto, tanto para sus fans como para ellos mismos, potenciando su popularidad.
- Poder distinguir entre eventos similares rápidamente. Por ejemplo, un usuario que desee obtener información sobre “Gaza”, puede referirse tanto a la banda de música como al conflicto en Israel. El poder distinguir rápidamente mediante una imagen o un vídeo acelera mucho el proceso. *Una imagen vale más que mil palabras.*

El sistema implementado es una primera aproximación que puede satisfacer los ejemplos mencionados.

1.2. Objetivos

1.2.1. Objetivo general

El objetivo principal de este trabajo fue el de poder evaluar e implementar en la práctica un sistema de extracción de contenido multimedia basado en la información social asociada a este contenido.

se menciona un estudio sobre comportamientos y preferencias de los usuarios en las redes sociales hecho por ROI Research: http://www.slideshare.net/performics_us/performics-life-on-demand-2012-summary-deck

⁶<http://www.biobiochile.cl/>

⁷Como muestra: <http://www.biobiochile.cl/2012/12/01/aporte-de-lustrabotas-de-santiago-a-la-teleton-provoca-admiracion-en-redes-sociales.shtml>, y <http://www.biobiochile.cl/2012/12/01/rechazo-provocan-condicionamientos-de-compra-de-ripley-y-unimarc-para-donar-a-la-teleton.shtml>

1.2.2. Objetivos específicos

- Abstraerse del problema de identificación de eventos a partir de documentos Web, llevando a cabo una metodología de obtención de datos simple.
- Implementar un modelo de *clustering* para separar los documentos en *subtópicos* de cada evento, sin considerar el tipo de contenido de estos documentos.
- Analizar la efectividad del sistema implementado, evaluando casos de estudio.

1.3. Descripción general de la solución

Este trabajo puede considerarse como un punto de partida para el desarrollo de un modelo de recuperación de contenido multimedia, similar a lo que corresponde a la generación de resúmenes automáticos para múltiples documentos. En particular, se implementó un sistema que permite considerar distintas estrategias para continuar desarrollando en el futuro. Además:

- Se llevó a cabo una metodología para la obtención de documentos y enriquecerlos con datos obtenidos de fuentes sociales;
- Se implementó un procedimiento que separa estos documentos en *clusters*, *sin considerar su contenido*. Sólo se utilizó la información social asociada; y
- Se implementó además una forma de *rankear* u ordenar los resultados de acuerdo a *relevancia*, siendo ésta medida de acuerdo a la información social asociada a los documentos generados.

El sistema implementado puede dividirse en tres componentes principales:

1. La que obtiene descripciones de eventos a partir de fuentes de éstos en la Web, enriqueciéndolos con información social;
2. Otra componente que procesa y separa los documentos a partir de la información social; genera *objetos Web* y los separa en subtópicos de cada evento, respectivamente; y
3. La componente que entrega los k documentos más relevantes por cada evento obtenido, basándose en los subtópicos identificados.

Se utilizaron las API de Google News como de Last.fm para la obtención de eventos: noticias y conciertos, respectivamente. Para el enriquecimiento de los eventos se utilizó la información social que provee Twitter y su API de búsqueda de *tweets*. De la misma forma, se consideraron los metadatos de los mismos mensajes para medir la relevancia de los documentos generados.

Un documento es identificado por la URL que lo ubica en la Web. El contenido no es más que la concatenación de los tweets que mencionan al documento. Se realizó una limpieza y preprocesamiento de los datos, quitando las *stopwords* y realizando *stemming* sobre el contenido en texto. Luego, se aplicó *tf-idf* sobre los documentos, representándolos como vectores en el *space vector model*. Para identificar los subtópicos de un evento se utilizó el algoritmo de clustering *k-means* sobre los vectores.

Para el ranking de los documentos se decidió usar una ponderación simple sobre una serie de indicadores que dependen de los tweets y de las URLs de cada evento.

Entre las herramientas utilizadas, se usó lenguaje de programación Python, varias librerías para el manejo de datos (tales como `nltk`, `scipy`, `scikit-learn`, por nombrar las más importantes), el sistema de almacenamiento Redis, entre otras herramientas que son mencionadas en la descripción detallada de la solución.

Capítulo 2

Antecedentes

2.1. Twitter

Twitter es una red social online que permite conectar a personas mediante la comunicación de mensajes cortos, rápidos y frecuentes. Estos mensajes son publicados en el perfil del usuario que los emite, pueden ser vistos directamente por los seguidores de este usuario o ser vistos directamente en el perfil o buscándolos mediante una funcionalidad que provee el servicio. Además, un usuario puede *seguir* a otros para poder ver en su *timeline* los mensajes de todos a quienes sigue.

FIGURA TWITTER

Estos mensajes, o *tweets*, pueden además *mencionar* a otros usuarios, mediante la convención “=@usuario [texto]=” indica que se está mencionando a la persona con el nombre “usuario”. Adicionalmente, existen varias convenciones o costumbres que han surgido a lo largo del tiempo en esta red desde sus inicios el año 2007:

- Respuestas o *replies*: son mensajes del tipo @usuario [texto], que ocurren usualmente en una conversación entre dos usuarios.
- Menciones o *mentions*: un poco más general a una respuesta, el nombre del usuario mencionado puede estar en cualquier parte del mensaje. La diferencia semántica es que no se le habla “directamente” al usuario mencionado, como en una respuesta, sino que sólo es mencionado por si el mensaje es de su interés o no.
- *Retweets*: son mensajes del tipo RT @usuario: [texto]. Ocurren cuando se quiere compartir el mensaje de otro usuario, o citarlo para mencionarlo en el mismo mensaje.
- *Hashtags*: son palabras precedidas por el caracter #, que indican un identificador a cierto evento o suceso dentro o fuera de la red. Suelen usarse para categorizar de cierta forma un tópico, pero son libres de usarse como los usuarios quieran.
- Mensaje simple: un mensaje sin menciones ni hashtags.

Ejemplos:

- Mensaje simple: Jason Funk disipa patitos;

- Respuesta: @jason estoy de acuerdo con lo que dices;
- Mención: creo que @jason es una cumbre de sabiduría;
- Retweet: RT @jason: Jason Funk disipa patitos; y
- Hashtag: Estoy escribiendo mi memoria #dcc #summarization

Estos mensajes están limitados a 140 caracteres de extensión. Sumando esto a la integración de la red con otros servicios y dispositivos, y a la cantidad de mensajes publicados cada minuto, permite utilizar esta red como una gran fuente de datos.

Twitter además provee varios servicios adicionales, como por ejemplo, un servicio de acortamiento de URLs, para permitir incluir una URL larga sin perjudicar la cantidad de caracteres restantes para el mensaje; un servicio de alojamiento de fotos y vídeos, para hacer más sencilla la publicación de mensajes multimedia desde dispositivos móviles; un servicio de búsqueda que permite buscar una cantidad determinada de tweets sobre un término de búsqueda o un hashtag.

2.2. Identificación automática de eventos

2.3. Clustering de documentos

2.3.1. Evaluación de clusterings

2.4. Resúmenes automáticos

2.4.1. Evaluación de resúmenes

2.5. Ranking de documentos

Capítulo 3

Especificación del Problema

- 3.1. Descripción detallada
- 3.2. Relevancia de una solución
- 3.3. Características de calidad
- 3.4. Criterios de aceptación

Capítulo 4

Descripción de la Solución

4.1. Modelo formal

4.2. Metodología de desarrollo

4.3. Metodología de obtencion del dataset

Se describe a continuación el proceso diseñado para la obtención de datos para alimentar al sistema implementado.

Las etapas de generación del dataset son las siguientes:

- Recolección de eventos (noticias y conciertos);
- Enriquecimiento de los eventos existentes mediante tweets; e
- Identificación de documentos a partir de los tweets por cada evento.

Se recolectaron datos (eventos y tweets) desde el 19 de noviembre de 2012 hasta XXXXXXXXXXXX todos los días desde la medianoche hasta que el proceso termina exitosamente.

4.3.1. Recolección de eventos

Se consideraron dos tipos de eventos para el sistema: noticias y conciertos musicales. Los conciertos incluyen festivales de varios artistas.

- Noticias Para obtener las noticias, se utilizó el servicio de Google News¹. Existe una API (en proceso de obsolescencia, pero funcional a la fecha de este trabajo) que permite obtener no sólo los titulares y breve descripción de cada noticia, sino también un conjunto de entre 4-10

¹<http://news.google.com>

noticias relacionadas de otras fuentes. Esto sirvió para alimentar los términos de búsqueda para la etapa siguiente. Se guardaron los siguientes datos de una noticia:

- Título,
 - Descripción,
 - URL de la fuente, y
 - Titulares de las noticias relacionadas.
- Concierdos Utilizando el servicio de Last.fm para obtener los conciertos y festivales de una ubicación en particular², se obtuvieron los conciertos y festivales de las siguientes ubicaciones:
 - Santiago, Chile;
 - Londres, Inglaterra;
 - Glastonbury, Inglaterra;
 - Las Vegas, Nevada, EE.UU.; y
 - Estocolmo, Suecia.
 - Título del evento (concierto o festival);
 - Artistas que participan; y
 - Fechas de inicio y término (esta última no siempre está como dato).
- Además de otros datos descriptivos, como la ubicación, descripción breve, sitio web de la banda o festival, etc.

Cada vez que se obtienen los eventos se vuelven a obtener los conciertos, pero sólo agregando los nuevos. Las noticias siempre son nuevas, aun así por implementación no se consideraron los repetidos.

4.3.2. Enriquecimiento de eventos

Se obtuvieron tweets utilizando el servicio de búsqueda que provee Twitter en su API³. El objetivo es enriquecer los eventos con la información social que hay en la Web sobre éstos.

Para cada uno de los eventos obtenidos en la fase anterior, se utilizaron los términos de búsqueda asociados a ellos: los titulares de las noticias relacionadas y los nombres de los artistas para los eventos noticiosos y musicales, respectivamente.

- Para las noticias, se hace una búsqueda en Twitter de los titulares al mismo tiempo en que se obtienen de Google News, y nuevamente al día siguiente, es decir, 2 búsquedas por cada titular de un evento. Se quitan las tildes y caracteres no ASCII y las stopwords, para evitar problemas con la implementación y no hacer calce de stopwords en la búsqueda de Twitter, respectivamente.
- Para los conciertos y festivales, se utilizaron los nombres de los artistas y del evento como términos de búsqueda. De acuerdo a la información asociada al evento, se busca por una mayor cantidad de días:

²<http://www.lastfm.es/api/show/geo.getEvents>

³<https://dev.twitter.com/docs/api/1.1/get/search/tweets>

- Se busca desde un día antes de inicio del evento;
- Si está presente la fecha de término del evento, se busca cada día dentro del intervalo “fecha de inicio” a “fecha de término” hasta tres días terminado el evento.
- Si no está presente la fecha de término (por ejemplo, un concierto o un festival de un día), se busca hasta tres días pasada la fecha de inicio.

4.3.3. Identificación de documentos a partir de tweets

Luego de obtener los tweets asociados a cada evento, el siguiente paso fue generar los documentos que fueron usados para la generación de los resúmenes. Nuevamente, el modelo consistió en que cada documento se modeló como un vector de palabras, donde el identificador del documento es una URL, y sus componentes corresponden al contenido de los tweets que tienen esa URL en el texto del mensaje.

El caso en el que un tweet no tenía ninguna URL en su contenido fue abordado de la siguiente forma: la URL asociada es una tal que representa al mismo tweet (utilizando el servicio de Twitter), y el contenido de ese documento es el mismo tweet, de forma de no dejar el tweet sin ser representado.

Este proceso fue abordado recorriendo todos los eventos del dataset, observando todos los tweets asociados a cada evento, extrayendo la URL si es que hay alguna y guardando el documento con el nuevo tweet. Se marcan los tweets observados para no tener que repetir el proceso, ya que es intensivo en conexión a la red.

Dada la condición breve de los mensajes publicados en la red social, muchos de los usuarios y/o servicios que publican mensajes con una URL en su interior suelen utilizar *acortadores (url shorteners)* para los enlaces, y así no utilizar mucho espacio dentro de un mensaje. Otra ventaja que ofrecen es que algunos servicios como bit.ly dan estadísticas sobre los visitantes a estos enlaces (y así saber quiénes vienen de cierta red social u otra, por ejemplo). Twitter, a su vez, actualmente también ofrece acortamiento de URLs por defecto. Esto suele producir que un enlace acortado se resuelva a otro enlace también acortado, por lo que es necesario resolver la URL completa para evitar duplicados o *pseudo-duplicados* (en el caso en que dos URLs sintácticamente distintas apunten al mismo recurso). EN LA FIGURA.....

FIGURA DE LINKS CORTOS

Por lo anterior, una vez identificada la URL del texto de un tweet, se resuelve su URL completa (que puede ya serlo de antemano), lo que consume recursos de ancho de banda y tiempo.

4.4. Desafíos técnicos

4.5. Restricciones de la API de Twitter

La API de búsqueda de Twitter permite obtener tweets de acuerdo a un término de búsqueda. Se utilizó este servicio para enriquecer los eventos con información social utilizando como términos de búsqueda tanto los títulos de las noticias como los nombres de los artistas para las noticias y los conciertos, respectivamente.

Funciona de la siguiente forma: cada vez que se hace un request a la URL dada por el servicio, éste retorna a lo más 100 tweets por página, con un máximo de 15 páginas (indicando en el request qué página queremos consultar), dando como total hasta 1500 tweets por búsqueda. Existirán términos de búsqueda que no presenten ningún resultado (ya sea por estar mal escritos o simplemente que no sean un tópico de discusión), o por el contrario, que se generen más tweets que los retornados por la búsqueda por cada ventana de tiempo que demore ésta (por ejemplo, un *trending topic* o tópico que sea muy mencionado en la red social).

Existe una limitación de uso de este servicio: sólo es posible hacer hasta 180 requests por cada 15 minutos, o 1 request cada 5 segundos. Además, sólo retorna tweets de hasta 7 días de antigüedad, y sus resultados no son necesariamente en tiempo real y su estabilidad varía de acuerdo a factores externos.

Los tweets retornados vienen en formato JSON (*Javascript Simple Object Notation*), e incluyen varios metadatos sobre el tweet aparte de los principales, como autor, fecha, contenido. Algunos de estos metadatos son:

- Cantidad de *retweets* hechos hasta la fecha;
- Si posee alguna URL o *hashtag* en el texto;
- Si es una *mención* a otro usuario;
- La ubicación de donde se envió el tweet;
- etc.

Además incluye datos sobre el autor, como por ejemplo:

- Si la cuenta está *verificada*;
- La cantidad de seguidores del usuario;
- Cantidad de amigos (seguidores que también lo siguen);
- Cantidad de tweets;
- Su descripción, y si incluye alguna URL, etc;
- Ubicación (dada por el mismo usuario);
- Fecha de creación de la cuenta;
- etc.

Capítulo 5

Conclusiones

- 5.1. Resumen del trabajo realizado
- 5.2. Objetivos alcanzados
- 5.3. Relevancia del trabajo realizado
- 5.4. Trabajo futuro

Capítulo 6

Anexos

Bibliografía

- [1] Ioannis Karatzas. and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, Berlin, 2nd edition, 2000.
- [2] Philip Protter. *Stochastic Integration and Differential Equations*. Springer, 1990.
- [3] Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*. Number 293 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin [u.a.], 3. ed edition, 1999.