Cristobal Dotte, Vicente Oyanedel y Sebastián Santana
Introducción a la Minería de Datos
Primavera 2016

Desafío 1

I. Introducción

En Chile, así como en otros países en vías de desarrollo existen múltiples estructuras esenciales que son sísmicamente vulnerables, por lo que es prácticamente dificil reforzarlas todas antes de la ocurrencia de un futuro terremoto. Es por esta razón, que un sistema que alerte a los ocupantes de estructuras vulnerables, que aún no han sido reforzadas, pocos segundos antes del arribo de las ondas sísmicas, podría evitar un desenlace fatal. Sería ideal que los seres humanos puedan informar de los sismos manualmente, ya que los seres humanos pueden proporcionar información detallada y precisa sobre ocurrencia de los sismos rápidamente, así como de sus consecuencias. Sin embargo, debido al significativo retraso que existe en el procesamiento de la información reportada por las personas (Por ejemplo, por Twitter), se han desarrollado técnicas para detectar rápidamente eventos sismicoss.

Un error en el que se podría incurrir fácilmente sería considerar que toda la información presente en Twitter que alerte sobre un sismo sea genuina y represente una alerta, por tanto uno de los problemas que se debe sortear tiene que ver con cómo **verificar la veracidad de la información**. Por tanto, a través de minería de datos en tiempo real se pueden obtener resultados erróneos por problemas en los datos, lo que haría pensar que un sistema de este estilo no tiene ventajas versus los actualues sistemas de detección automatizada de sismos, basadas en sensores fisicos. Sin embargo, este proceso de detección automatizada es sumamente costoso, ya que requiere de la instalación de aparatos y de su debida mantención. Así, existe un gran incentivo a desarrollar técnicas efectivas en la detección de sismos, ya que su proceso puede ser replicado a otro tipo de alertas de emergencia que no posean tanto desarrollo en sistemas de detección

II. Problematica central: “Clasificación de tweets de alerta en tiempo real”

Nuestro objetivo será desarrollar un clasificador que, dado un aviso (En nuestro caso, un Tweet) sea capaz de identificarlo como una ‘Alerta en tiempo real’ o no; si existe un evento sísmico en desarrollo. Este clasificador deberá determinar la existencia del suceso a partir de datos que sean publicados en twitter.

En primer lugar, para producir un data-set de entrenamiento, se piensa aprovechar la existencia de datos sobre los eventos telúricos en Chile suministrados por el Centro Sismológico Nacional[1], los cuales cuentas con el registro de la fecha y hora exacta de los sismos. En esta parte se asumirá que todos los tweets con palabras claves como: temblor, tembló, temblo, terremoto, temblando, sismo; que sean emitidos en un rango de tiempo muy cercana a un sismo, serán etiquetados como una alerta de sismo. Se utilizará este método, ya que permite clasificar una gran cantidad de datos de manera simple y económica, tanto en tiempo como recursos económicos.

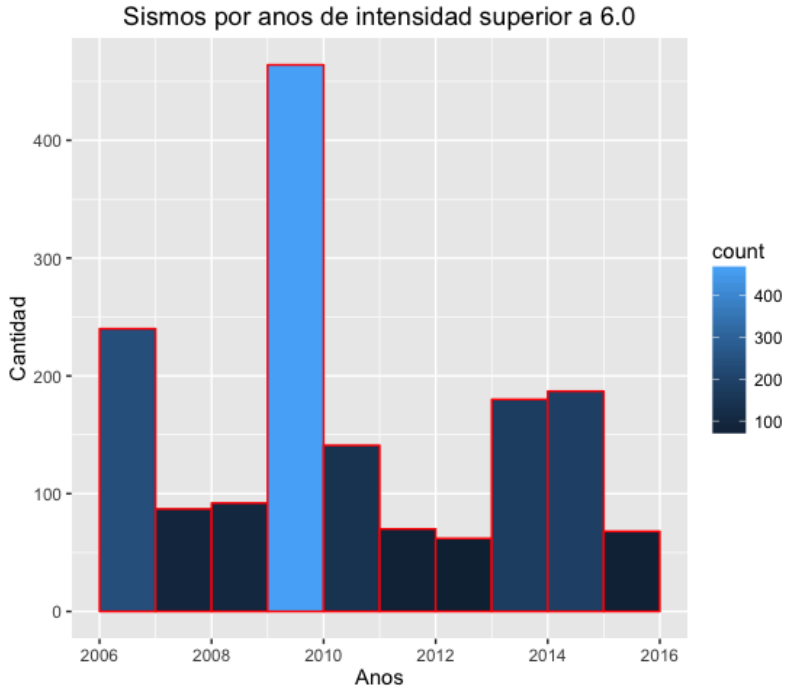
[1] http://sismologia.cl/

Proyectos anteriores[2] muestran un desempeño que difiere significativamente según la técnica que se utilice en la clasificación de estos tweets. Motivados por el interés de aprender sobre la técnica de aprendizaje *Deep Learning*, se piensa utilizar esta técnica para entrenar un clasificador que probablemente opere de forma semi-supervisada. Lo anterior ya que siempre existirán fuentes de información confiables (como el Centro Sismológico Nacional), los que brindarán datos clasificados. Además, para nuestra implementación tomaremos en cuenta las conclusiones que llegaron los proyectos pasados para así evitar cometer los mismos errores y aprovechar los avances que hayan logrado.

Se espera probar el desempeño de distintos métodos de clasificación, una vez obtenidos los datos etiquetados, y finalmente optar por el clasificador que produzca mejores resultados.

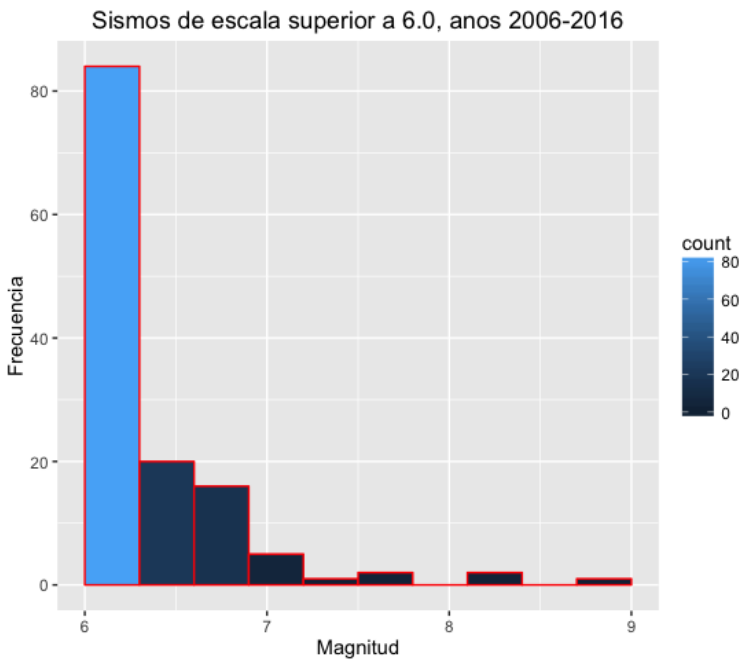
III. Descripción de los datos y exploración inicial

En primer lugar, fue necesario establecer un set de datos donde tuviésemos la certeza que se hablaba de un evento sísmico, de esta manera podríamos etiquetar este set de datos y utilizarlo para entrenar nuestro clasificador. La primera limitación fue que la API pública de Twitter no permite descargar datos antiguos, por lo que se utilizó una implementación en Python que permite buscar tweets anteriores a través de búsqueda avanzada de Twitter en navegadores[3]. Además, se obtuvieron los datos de todos los sismos registrados en Chile desde el año 2006, los cuales se pueden apreciar en el grafico siguiente.



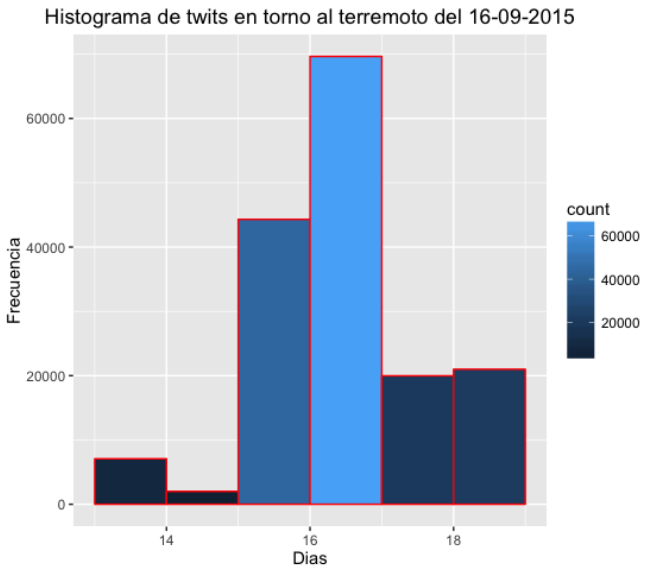
[2] https://users.dcc.uchile.cl/terremotos/

[3] https://github.com/Jefferson-Henrique/GetOldTweets-python

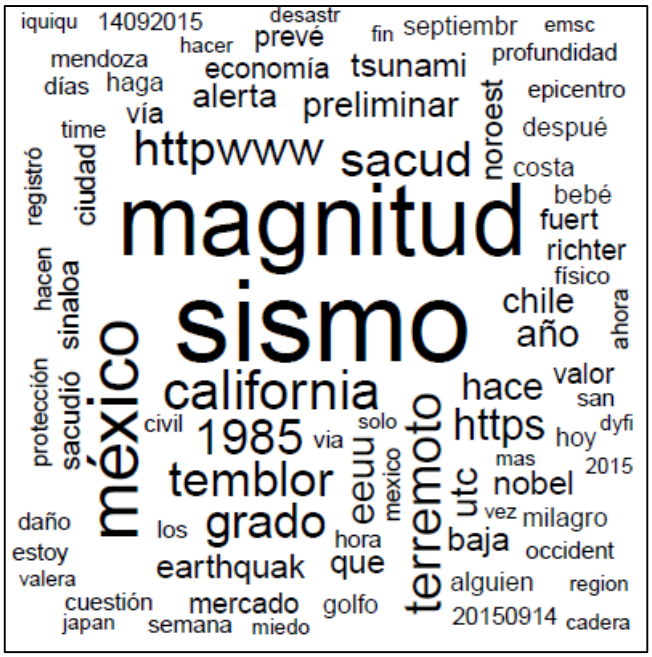


En el gráfico de la izquierda se pueden ver la frecuencia de sismos en Chile según su magnitud en los últimos 10 años. Es importante, para la fase de exploración, elegir un evento suficientemente "Fuerte" para asegurar una reacción notable a través de las redes sociales.

Para la exploración inicial se decidió utilizar el terremoto del 16 de Septiembre del 2015, el cual tuvo una magnitud de 8.3 Richter; dado que es un terremoto reciente y se supone que mientras más reciente sea el evento, más actividad habrá en Twitter (Dado el aumento en el uso de éste medio en los últimos años).
Se extrajeron los Twits con alguna de las palabras claves de los días 13, 14, 15, 16, 17, 18 y 19 de Septiembre del 2015, para explorar las palabras más mencionadas, conocer la información que proveen y ver su distribución al pasar los días. Con éstos datos se obtuvieron conclusiones notables y visualizaciones:



Como la intuición diría, los días posteriores al evento (Después del 16) hubo un aumento notable en la cantidad de Twits con las palabras claves. Aunque es destacable que los días previos poseen una frecuencia no despreciable de Twits. Además los días posteriores continúa una tendencia a seguir hablando del evento en cuestión.



Ésta nube de palabras corresponde a las palabras más frecuentes en los Twits **previos** al evento en cuestión. Puede verse como aparecen las palabras México, 1985 y California que recuerdan eventos pasados ocurridos en fechas similares. Además de otras alertas que pudiesen haberse filtrado en las consultas, dado que aparecen alertas de tsunami, profundidad, etc.

A la derecha corresponde a los días **posteriores** al evento aparecen palabras que notan claramente la referencia al evento en cuestión. Palabras como Chile, tsunami, muerto… Aunque siguen apareciendo referencias al terremoto de México y otras menciones que no se relacionan directamente con el terremoto del 16.

Del análisis de los datos, se desprenden las siguientes conclusiones preliminares:

1. Los días previos al evento telúrico se mencionan las palabras claves en contextos no-sísmicos, por ejemplo, como citas ‘poéticas’, canciones, declaraciones, etc. (No alertas)
2. También, en los días previos, aparecieron twits que si bien eran “Alertas” hablaban de otro evento en otra área geográfica.
3. El día del evento se produjo gran cantidad de twits que se pueden etiquetar como ‘Alertas’, aunque no todas lo son. Además, la cantidad de twits con las palabras clave, claramente aumentó.
4. Los días posteriores presentan una gran cantidad de twits con las palabras claves, aunque la mayoría no representan ‘Alertas’ como tal.

Estos puntos se toman en cuenta para pulir nuestro método de etiquetado de twits, dado que tenemos clara la importancia de él buen etiquetado en el desempeño de los clasificadores, en general.