Resultado 2 (entrega 30 de marzo de 2018):

Índice

1. Introducción

Con el presente informe se reporta los avances en varios ámbitos del proyecto y se entrega una página web con visualizaciones embebidas como prueba de concepto. En el capítulo 2 se muestra los avances acerca de la base de datos centralizada donde se incorpora las tres fuentes de datos mencionadas en conjunto (siniestros entregados por la AACH; redes sociales, por ejemplo Twitter; medios noticiosos en línea). Además se informa en el capítulo 3 el estado actual del desarrollo de modelos para la detección de patrones de robos sobre las fuentes individuales e integradas. En el capítulo 4 se describe una página web con tres diferentes tipos de visualizaciones embebidas como prueba de concepto para la visualización preliminar de los patrones encontrados en las tres fuentes antes mencionadas.

2. Base de Datos

A continuación se describe la implementación de los sistemas de administración de datos que sostiene el observatorio. Se destacan bases de datos relacionales para datos estructurados (SQL), así como un sistema para datos no estructurados (JSON), las que responden a las características específicas de las distintas fuentes y del uso de los datos. Primero se describe el entorno en el que existen las bases de datos para luego describir el acceso a los datos. Luego se describen las fuentes de información utilizadas para poblar la base de datos y finalmente se comenta sobre problemas y soluciones en la base de datos, fuentes de información y acceso.

2.1. Entorno de la base de datos

La arquitectura actual del observatorio responde al estado actual del desarrollo del proyecto de investigación. Esto se traduce en que las estructuras existentes tienen como objetivo el maximizar la flexibilidad tanto en la centralización y consolidación de información de múltiples fuentes de información como en el acceso esta información. La figura 1 presenta una vista esquemática del observatorio con sus principales módulos, fuentes de información y puntos de interacción con entidades externas (clientes y fuentes de información). Por último, cabe mencionar que los módulos para el minado de la información y aplicación de inteligencia artificial no están aún consideradas dado que se está en la etapa de investigación y exploración de datos.Sin embargo, para la componente de investigación, el acceso será a través de una interfaz (API) que permite acceder a la información de manera indexada (SOLR).

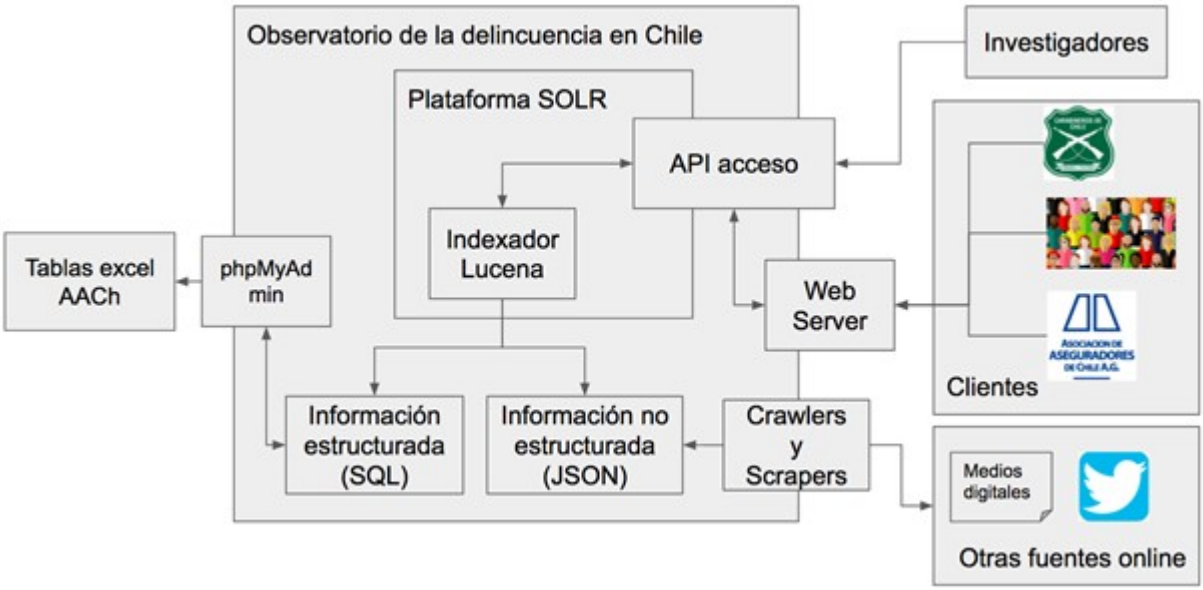


Figura 1. Arquitectura actual del observatorio. La punta de la flecha indica la fuente desde donde se obtiene la información.

Los módulos de la figura 1 se pueden agrupar en 3 capas: capa de adquisición de datos almacenamiento, capa de procesamiento y capa de presentación de la información. A continuación se presentará la capa de almacenamiento y sus interface.

2.2. Interfaces de acceso a la capa de Almacenamiento

En una vista general de la capa de almacenamiento se destacan bases de datos e interfaces de acceso a la información. A continuación se describen las interfaces de acceso a las bases datos.

Considerando la existencia de múltiples fuentes de información en distintos formatos y en diversas bases de datos (estructurada y no estructurada) es que se propone el uso de un de una plataforma de busqueda de informacion. De esta forma se hace transparente al usuario final la fuente y formato de la información, poniendo el foco en la búsqueda de la información adecuada.

Toda la información cargada al sistema se almacenará e indexará usando SOLR (http://lucene.apache.org/solr/). SOLR es una plataforma para la búsqueda de información almacenar grandes bases de y el acceso eficiente a los datos mediante un servicio web. Para lograr esto la información es indexada, tanto de las bases de datos estructurados como de los archivos con información no estructurada en formato json. La figura 2 presenta una muestra de las interfaces implementadas a través de esta plataforma web.

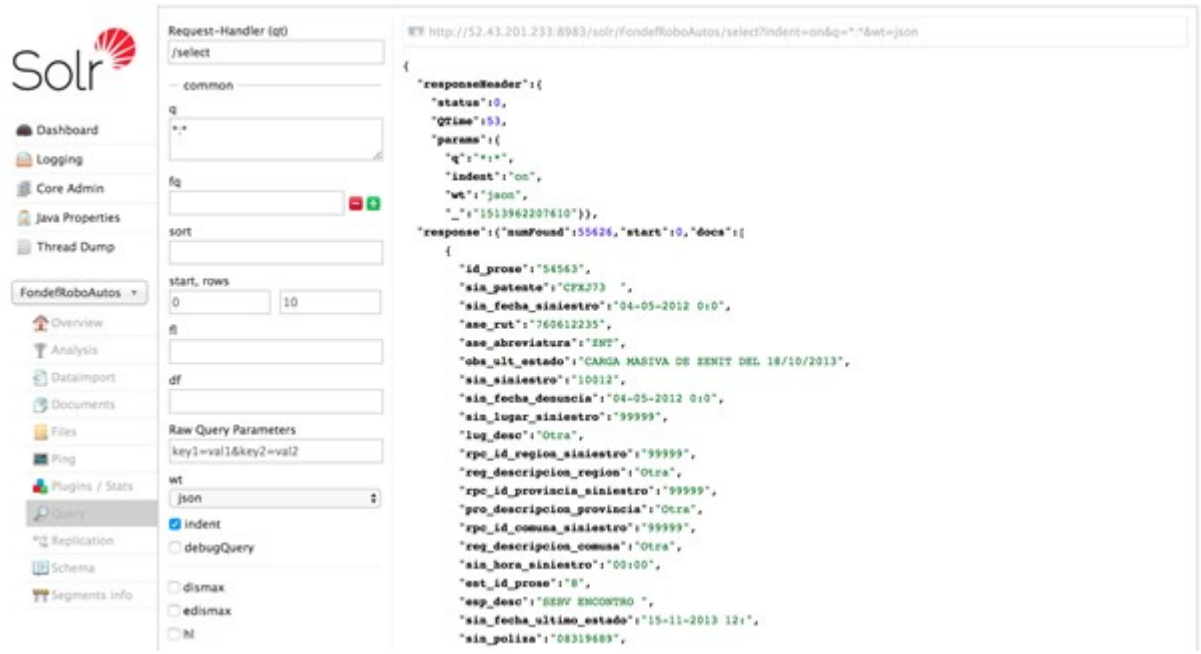


Figura 2. Interfaz de Solr, utilizada para la consulta de información no estructurada (texto libre de tweets y declaraciones de los robos) indexada en Lucene.

2.3. Descripción de los Datos

2.3.1. Asociación de Aseguradoras de Chile (AACh)

Esta base de datos incluye la información de siniestros denunciados por los clientes de las distintas compañias aseguradoras adheridas a la AACh.

1. Descripción: Consiste en una tabla con los datos entregados por la AACH con los siniestros denunciados por sus clientes. La tabla cuenta con 55.626 siniestros al día de hoy. La descripción completa de la tabla y sus atributos se encuentra en Anexo 1.
2. Adquisición: Esta información llega al observatorio en formato excel. Este excel es luego cargado manualmente a la base de datos en el observatorio. Una interfaz para la entrega automática de esta información está en planificación.
3. Almacenamiento: Se crea una tabla con los mismos campos del archivo excel que fue proveído por la AACH. Se insertan todos los datos del archivo en esta tabla. Esta tabla no requiere insertar/modificar los datos. Sólo modo lectura.

2.3.2. Twitter

1. Descripción: Con el fin de hacer análisis de la red social Twitter, se descargan tweets desde la API diariamente con patrones de búsqueda que tienen relación con el robo de vehículos. La descarga se realiza 2 veces al día y se extraen los tweets que en contenido hacen referencia a **@robados_chile** y/o **@autorobado_cl**. Además se descargan tweets con otras palabras claves relacionadas con el robo de vehículos en Chile.

2. Adquisición: En la descarga, mediante la API, se obtiene toda la metadata de los tweets (tweet id, fecha, tweet, geolocalización, etc.) e incluso se deja un registro de la consulta realizada para obtener dicho registro (atributo “query”, ver Anexo 1). Es importante destacar que algunas consultas podrían traer el mismo tweet más de una vez, sin embargo, se tomó en cuenta ese caso y se considera sólo uno (el que ya está en la base de datos).
3. Almacenamiento: Al día de hoy (jueves 15 de marzo) se cuenta con **1943** tweets cuyo desglose se adjunta a continuación:

Consulta a Twiter	Número de Tweets
@autorobado_cl	54
@robados_chile	1218
robado auto Chile	108
robado automovil Chile	1
robado patente	12
robado patente Chile	158
robaron auto Chile	33
robaron patente	5
robaron patente Chile	183
robo auto	37
robo auto Chile	11
robo automovil Chile	4
robo camion Chile	15
robo camioneta Chile	51
robo patente	10
robo patente Chile	43

Tabla 1. Nº de tweets extraídos mediante consultas

2.3.3. Noticias

1. Descripción: se extraen noticias con el fin de detectar patrones desde medios noticiosos.
2. Adquisición: Para llevar a cabo la extracción de noticias de robos de vehículos, se hicieron consultas al buscador google empleando palabras claves y de manera forzada a algunos portales de noticias. Por medio de técnicas de captura de contenido en páginas web (web scraping), se hizo la extracción de cada una de las noticias a los siguientes medios escritos: **Radio ADN, La Tercera, La Nación, Chilevisión noticias, Radio cooperativa, Radio Biobio, Publimetro.**
3. Almacenamiento: Cada noticia capturada fue almacenada en un archivo de formato no estructurado JSON cuyos atributos son: título de la noticia (texto), contenido de la noticia (texto), fecha de publicación (formato fecha año-mes-dia-hh-mm-ss), fuente de donde se extrajo la noticia (texto), URL (texto), si la noticia es usada o no (binario) y la codificación. Respecto al atributo si es usada o no la noticia, se hizo un

post-procesamiento de limpieza para asegurarnos que la noticia era efectivamente de robos. Dicho campo se marcará con Y, si la noticia es utilizada o de lo contrario, con N. Un ejemplo de una noticia o instancia capturada es la siguiente:


```
{
"encoding": "utf-8",
"title": "Roban vehiculo con ninos en su interior | Emol.com",
"content": "SANTIAGO.- Un operativo por parte de Carabineros se llevo a cabo este domingo despues de que cuatro sujetos realizaran un portonazo a una mujer ...y su familia en Villa Alemana.Los antisociales obligaron a la mujer a bajar del vehiculo y lograron robar una camioneta 4x4, el cual tenia a tres menores de edad en su interior, y escaparon por una carretera del sector.Carabineros encontraron a los ninos en un outlet de Curauma tras ser abandonados ahi por los delincuentes y fueron llevados a un centro asistencial para constatar lesiones.La camioneta fue encontrada enun sitio eriazo mientras que los sujetos se mantienen profugos.",
"source": "emol.com",
"date": "2017-12-11T07:10:55",
"url":
"http://www.emol.com/noticias/Nacional/2017/12/11/886776/Sujetos-roban-vehiculo-con-tres-menores-de-edad-en-su-interior.html",
"file_name": "news/2017-12-11T07:10:55_robo_AND_auto_emol.com.GoogleSearch",
"used": "Y"
}
```


Se recuperaron un total de 5913 noticias, de las cuales, con el post-procesamiento de limpieza, quedaron 3365 disponibles para análisis. El desglose se adjunta a continuación.

FUENTE	USADAS	NO USADAS
emol.cl	583	302
adnradio.cl	170	150
latercera.com	465	174
chvnoticias.cl	442	324
biobiochile.cl	776	217
cooperativa.cl	517	269
publimetro.cl	162	771
lanacion.cl	250	341

Tabla 2. Noticias descargadas mediante queries en google

2.4. Supuestos, dependencias y restricciones de desarrollo e implementación de las Bases de Datos

El desarrollo solo se ha visto restringido en cuanto a la disponibilidad de datos, esto debido a la dificultad de organizar múltiples instituciones para el traspaso de información. En consideración de esta dependencia sobre los datos, se asume que a medida que estos

estén disponibles serán incluidos en el observatorio. Se asume además que la estructura de los datos existentes al día de hoy, no cambiará de manera considerable. Más bien se irán agregando nuevos tipos de datos.

3. Modelos para detección de patrones/técnicas de extracción de patrones (BP, RW)

<texto introductorio a las diferentes técnicas utilizadas>

3.1. Modelos de Datos

Los datos de las diferentes fuentes de información: AACH, Twitter y noticias, pueden ser modelados de manera similar desde diferentes perspectivas con la finalidad de hacer minería de datos. Esto permite comparar de manera transparente la información de estas fuentes. Los datos están incompletos en diferentes campos, según la calidad de la fuente de datos y su completitud.

Modelamiento de eventos utilizando información enriquecida

Los eventos en el dataset son los robos reportados (independiente de la fuente), en este caso los datos que se representan son los eventos. En particular, un evento **e**, puede ser modelado en base a sus información enriquecida, como lo son su:

- **tipo de evento <AACH, Twitter, Noticia>,**
- **ubicación <longitud, latitud, comuna, región>,**
- **fecha <día de la semana, día del mes, mes, hora>,**
- **texto <tf-idf, n-gramas, word-embedding>,**
- **vehículo <patente, marca, modelo, color, avalúo>.**

Con eso un evento pasa a representarse por un vector multidimensional de metadatos.

Adicionalmente a esto se trabajó con modelos más sofisticados de los datos para extracción de patrones en robos en base a factores geográficos. Estos están descritos en la sección 3.3.

3.2. Extracción de patrones sobre la relación entre fuentes externas y los siniestros

Se desarrolló un modelo de minería de datos para comprobar la existencia de patrones que indiquen relación entre fuentes externas, tales como Twitter y medios noticiosos, con respecto a los datos de las aseguradoras.

El modelo desarrollado consistió en el proceso representado en la siguiente figura.

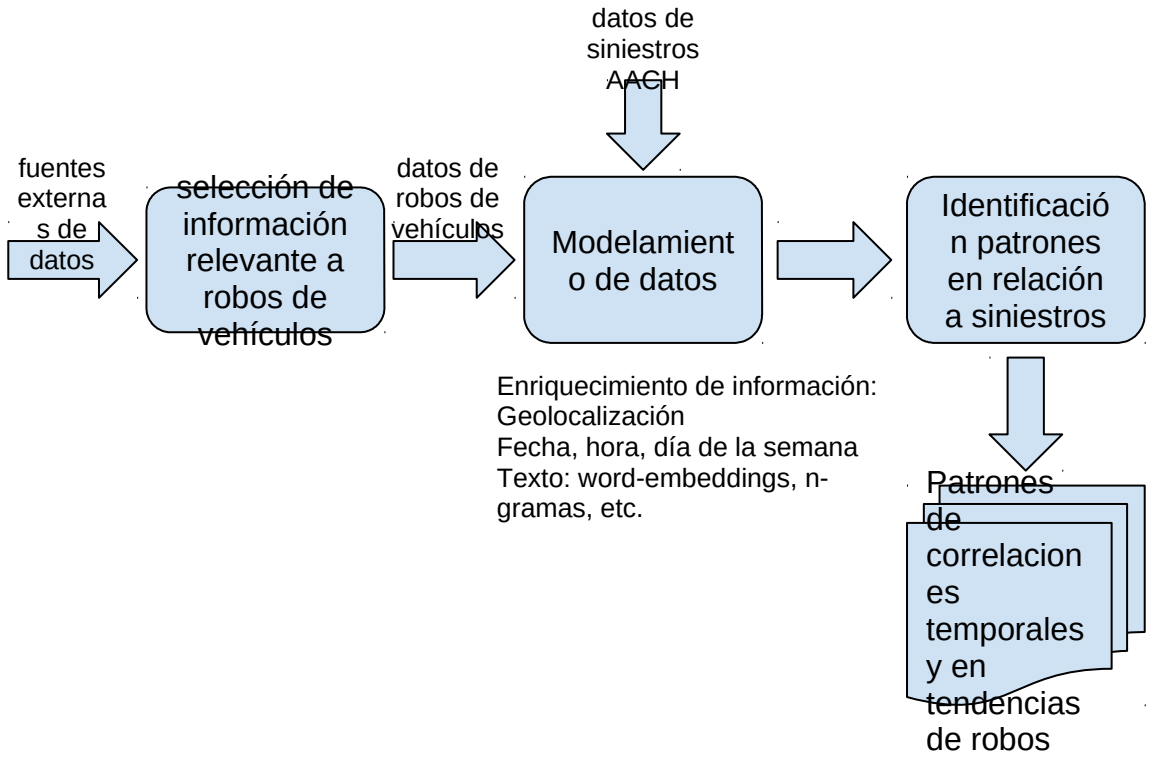


Figura 3. Proceso desarrollado para la búsqueda de patrones

Utilizando esta técnica se identifican claros patrones de relación temporal entre la fuente de datos de redes sociales, Twitter y los datos de los siniestros. Mostrando las mismas tendencias en cuanto a:

1. Volúmen de robos reportados en el tiempo, ya sea por días de la semana, por mes o por hora.
2. Volúmen de reportes por modelos y marcas de vehículos.

Adicionalmente, se identifica un patrón secuencial que indica que los robos de vehículos reportados en la red social tienden a tener mejores tasas de recuperación en el tiempo.

Estos patrones fueron validados visualmente y en conjunto al experto de la AACH. También se calcularon índices de correlación entre robos reportados en diferentes meses para cada año. Esto último mostró una clara tendencia al aumento de la correlación en el tiempo entre redes sociales y los siniestros.

El modelo de minería de datos permitió encontrar los factores más correlacionados de la red social y de las noticias con la fuente de siniestros. Esto indica que tiene sentido utilizar reportes en tiempo real de las fuentes externas, como un estimador o complementario de lo que ocurrirá con los siniestros reportados a la compañía. Esto produce información complementaria, ya que permite a AACH obtener información sobre el parque automotriz no-asegurado que puede permitirle identificar nuevas oportunidades para el negocio.

3.3. Modelos para encontrar patrones espacio-temporales

Los patrones espacio-temporales encontrados en esta sección serán validados mediante la comparación de las distribuciones de distancias entre los puntos de robos de vehículos asegurados con puntos de robos simulados aleatoriamente. De esta forma se podrá determinar que los patrones identificados son significativos en cuanto a la relación espaciales entre puntos de robos de vehículos y lugares de atracción social. A continuación se describirán con mayor detalle esta propuesta.

3.3.1. Modelo de eventos en base a puntos de interés

Los robos de vehículos asegurados presentan patrones de comportamiento, los cual puede estar asociado a algunas características de la ciudad, como por ejemplo su cercanía a lugares de atracción social, restaurantes, centros comerciales, bares, bancos y escuelas. Para examinar esta idea, se propone caracterizar los puntos de robos de vehículos de acuerdo a su cercanía a lugares de atracción social. Esta caracterización consiste en modelar estos puntos de robos como vectores con sus coordenadas y con el número puntos de atracción social correspondientes a un radio de 200 metros. Este proceso se realizará para las categorías de restaurant, shopping mall y colegios. Para realizar esto se debió obtener las coordenadas de puntos de atracción social en un radio de 200 metros para cada punto de robo. Estas coordenadas fueron obtenidas mediante la herramienta de Google Places API Web Services, el cual permite consultar información sobre lugares en una amplia variedad de categorías, como establecimientos públicos, colegios, hospitales, restaurantes, ubicaciones geográficas, entre otros. El tipo de búsqueda que se realizó es por radar, es decir, permite buscar hasta 200 lugares por consulta. Estas frecuencias serán visualizadas en un histograma, en el cual se espera obtener gran cantidad de puntos de atracción social a distancias pequeñas. Sin embargo, es necesario plantearse la idea de que es posible que estos puntos de robos puedan estar cerca a puntos de atracción social debido a otros factores. En otras palabras, es necesario determinar que los lugares de atracción social sí son lugares de atracción de robos de vehículos. Para esto se plantea como hipótesis que existen puntos de atracción social donde es más probable que ocurran robos de vehículos.

Para evaluar esta hipótesis se propone modelar los puntos de atracción social como vectores con las coordenadas de estos puntos y las frecuencias de puntos de robos para un radio de 200, 400, 600, 800 y 1000 metros, donde la frecuencia para cada radio no es acumulativa. Este proceso se realizará para las categorías de puntos de atracción social ya mencionadas. Entonces, para verificar que estos puntos de atracción social realmente son puntos de atracción de robos se generarán coordenadas random dentro de la zona de interés, Santiago. Una vez obtenidas estas coordenadas aleatorias, se crearán los mismos vectores con las frecuencias de puntos de robos dentro de los radios ya descritos. Para crear estos vectores se utilizó el paquete RANN de R, el cual utiliza el algoritmo kd-tree para encontrar el número *p* de vecinos cercanos para cada punto en un conjunto de datos de entrada y un conjunto de datos de consulta. La ventaja de usar este algoritmo es que se ejecuta el tiempo O (M log M) siendo esto muy útil cuando los set de datos son de gran tamaño. Posteriormente, se compararan visualmente las distribuciones los vectores

obtenidos para las coordenadas reales y las aleatorias, para cada categoría de lugar de atracción, mediante un histograma. De esta forma podremos visualizar si los puntos de atracción social poseen una relación importante con los puntos de robos.
Finalmente, se aplicará clustering a estos vectores con la distribuciones para identificar regiones espaciales con un comportamiento similar en cuanto a la distribución de frecuencias de robos para los puntos de atracción social. Luego, se procederá a caracterizar estos cluster de acuerdo a las categorías de los lugares de atracción social e información asociada a los puntos de robos.

3.3.2. Modelo de eventos en base a texto, ubicación y tiempo

El uso de medios sociales en dispositivos con GPS provee una fuente de datos con información espacio-temporal. Servicios como Facebook Places, Twitter o Foursquare son ejemplos donde los usuarios pueden compartir noticias, opiniones, preferencias y actividades diarias con información espacio-temporal añadida. Esta fuente de información ha sido utilizada para estudiar fenómenos como patrones de movilidad (Luo et al. 2016) o actividades comunes en zonas urbanas (Zhang et al, 2017b), e incluso predecir eventos como actos criminales (Wang et al. 2012). Dado que entender actividades humanos complejas como actos criminales requiere combinar múltiples fuentes de información. Nuestro objetivo es descubrir patrones a partir de los datos compartidos en redes sociales que nos ayuden a un mejor entendimiento del proceso relacionado con el robo de autos. Los robos de autos son eventos donde la componente espacial (donde) y la temporal (cuando) tienen gran influencia. Para ello nos proponemos utilizar datos extraídos de la red social Twitter geolocalizados en Chile desde el 2013 hasta el 2017. Estos datos servirán para combinar los datos de ARCH con datos de Twitter y construir modelos que nos permitan un mejor entendimiento de los robos de autos en Chile mediante análisis exploratorio y predictivo.

Análisis exploratorio: El objetivo del análisis exploratorio es encontrar patrones espacio-temporales valiosos usando los datos de redes sociales. Con esto se podrían construir modelos que permitan, dado un punto geográfico o geográfico-temporal, entregar información descriptiva sobre las actividades que tienen lugar en esa ubicación.

Análisis predictivo: El objetivo del análisis predictivo es lograr predecir zonas geográficas y horarios donde es más probable que ocurran robos de autos. Para ello nos proponemos agregar información extraída de medios sociales a un modelo de predicción de actos criminales. Esta información servirá a AACH para alertar a sus clientes en las zonas detectadas.

Modelos: Una forma común para datos espacio-temporales de medios sociales es la tupla <timestamp, coordenadas, texto>. Considerando que cada elemento de la tupla tiene su propia forma de representación, no es trivial un modelo que agregue los tres tipos de datos de manera efectiva. **Embeddings:** el estado del arte para tareas como modelado de actividades en zonas urbanas (Zhang et al., 2017b) o detección de eventos localizados (Zhang et al., 2017a), se basa en representar los tres tipos de datos en un mismo espacio vectorial denso y de baja dimensional (modelo de embedding). Los vectores son calculados

a partir de patrones de co-ocurrencia en la tupla <T, L, Keyword>, donde T es una discretización del timestamp, L es una discretización de las coordenadas y Keyword es un término extraído del texto. El modelo de embedding podrá utilizarse para hacer consultas con cualquiera de las tres variables y recuperar las otras dos. Por ejemplo podremos consultar con la palabra "*portonazo*" y descubrir cuales son los horarios, zonas geográficas e incluso otras palabras que más co-ocurren con esta palabra de interes. También podrá consultarse el modelo con puntos de interés encontrados en los datos de AACH y descubrir qué información es compartida en los medios sociales en ese punto geográfico. Otras utilidades del embedding son que permitirá utilizar algoritmos de agrupamiento sobre esta representación para encontrar patrones además de que podrá ser agregada como variable de entrada al modelo de predicción de robos de autos. **Representación como documentos:** Para la representación como documentos nos proponemos utilizar las discretizaciones de las coordenadas y del timestamp como identificadores de documentos y la agregación del texto como el cuerpo. Luego utilizar técnicas de representación de textos como TF-IDF, modelado de tópicos (Blei et al., 2003) y embedding de texto (Mikolov et al., 2013b) para representar las celdas espacio-temporales a partir del contenido compartido en ellas. Esta representación podrá utilizarse para hacer consultas con cualquiera de las tres variables y recuperar las otras dos. Además permitirá utilizar algoritmos de agrupamiento sobre ella para encontrar patrones o ser agregada como variable de entrada al modelo de predicción de robos de autos.

Enriquecimiento de información: Se espera poder aumentar la cantidad de información de los modelos propuestos utilizando técnicas de "data augmentation" del estado del arte (Zhang et al., 2018) y a la vez propuestas por nosotros.

Evaluación:

- Se realizará una evaluación del enriquecimiento de datos midiendo la calidad de los modelos construidos. Para medir la calidad de los modelos construidos para representar datos espacio temporales evaluaremos la capacidad de predecir los elementos de la tupla <T, L, Keyword> cuando solo se conoce dos de ellos y queremos predecir el tercero.
- También evaluaremos el impacto de incluir datos extraídos de redes sociales para predecir zonas geográficas y horarios donde es más probable que ocurran robos de autos. Para ello nos proponemos agregar información extraída de medios sociales a un modelo de predicción de actos criminales. Esta evaluación se realizará con medidas clásicas para evaluar modelos de predicción como precisión, cobertura y F1.

3.3.3. Modelamiento y búsqueda de patrones en medios noticiosos

Para encontrar patrones de búsqueda en medios noticiosos fue necesario utilizar técnicas de minería de datos para el análisis de texto. Al hacer un análisis de bigramas (es decir, 2 palabras consecutivas) las noticias desde el 2013, se encontraron términos muy frecuentes. Dos de los patrones más recurrentes fueron "robo de", "robos de", "robos con", "robo con". Por lo tanto se hizo un análisis desde el 2013 al 2016 de las palabras que acompañaban a estos patrones. Se hizo además una comparación con la base de la AACH. El gráfico se adjunta a continuación.

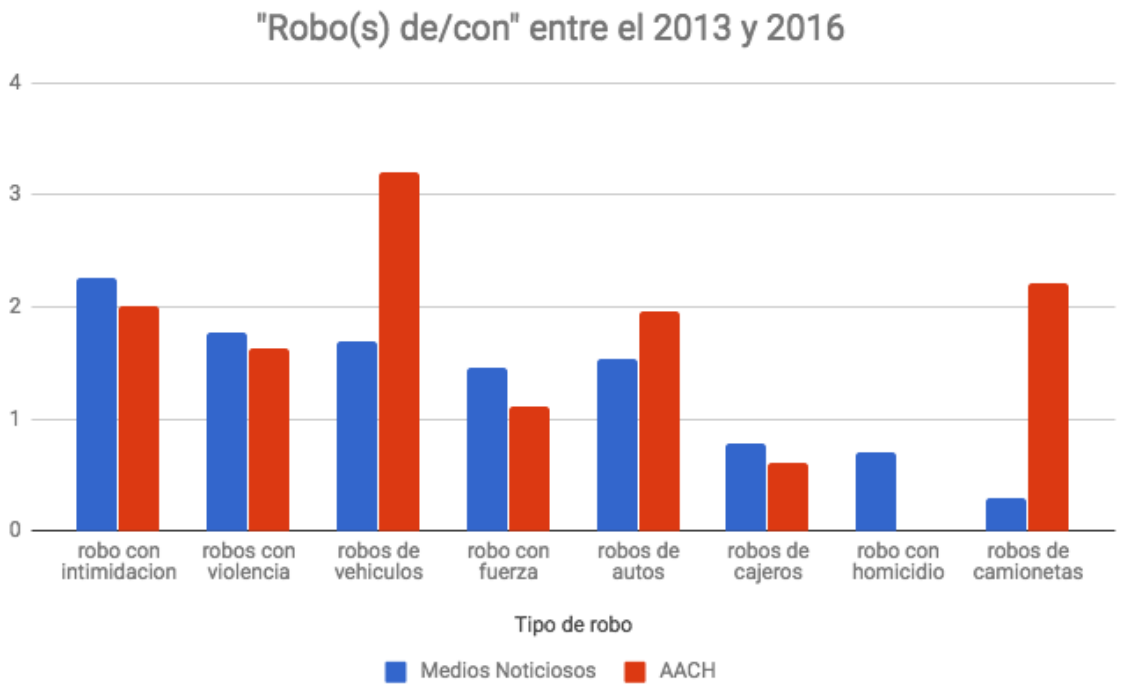


Figura 4. Distribución de los bigramas en los medios noticiosos y AACH

El gráfico nos indica que en general los medios noticiosos prefieren hablar con palabras más “sensacionalistas” como “robo con intimidación, robo con violencia que robo de “auto” o “automóviles”. De hecho, en los relatos de la AACH, las denuncias son precisas y se indica si el robo fue de un camión o automóvil. Sin embargo, en los medios noticiosos se habla simplemente de vehículos o autos. Además, se puede observar que el robo con homicidio no ocurre frecuentemente en los relatos de la AACH, pero sí aparece con frecuencia en la prensa.

Por otro lado, encontramos que al hacer un análisis de 3 gramas (3 palabras consecutivas) en medio noticiosos hay muchas noticias que mencionan lugares que tienen relación con robo de vehículos. En la siguiente nube de palabras se encuentran palabras frecuentes como “en la comuna”, “la comuna de”, “la intersección de”, “de la florida”, entre otros.

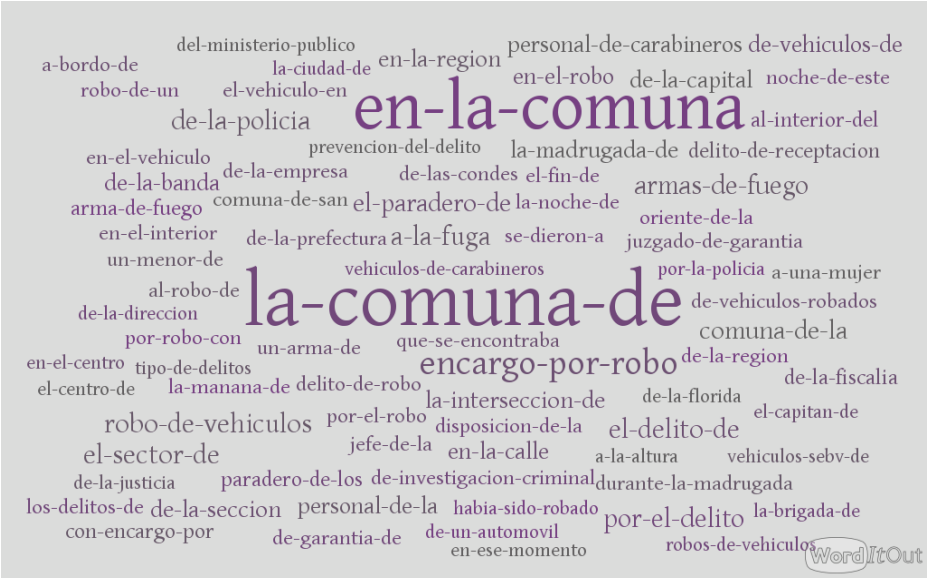


Figura 5. Nube de palabras de 3 gramas

Utilizamos una herramienta para extraer de manera automática los lugares que se hacen mención un texto (en una noticia). Esto también se puede cruzar con la base de datos proveída por la AACH y ver si existe algún tipo de relación para un futuro estudio. Para este proceso utilizamos la herramienta **Named Entity Recognition (NER)**[1] de la Universidad de Stanford. Esta herramienta es ampliamente utilizada para extraer entidades de los textos. Para este proceso fue necesario utilizar un diccionario en español puesto que por defecto está para idioma inglés. El proceso de obtener las localidades se muestra a continuación:

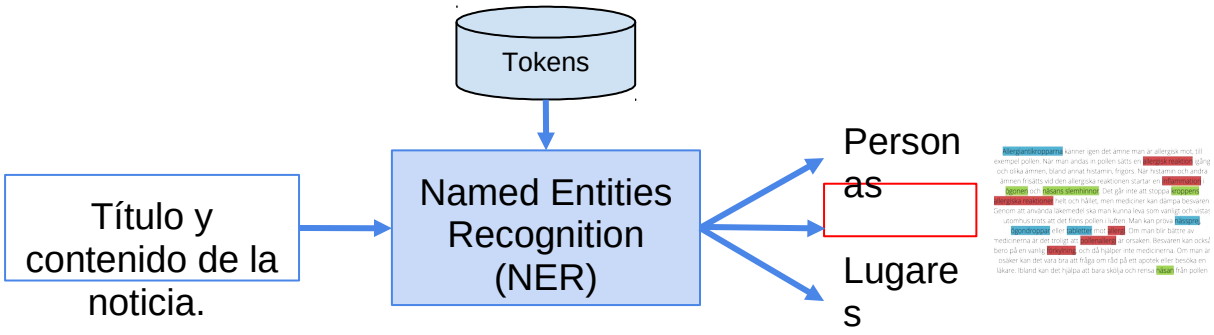


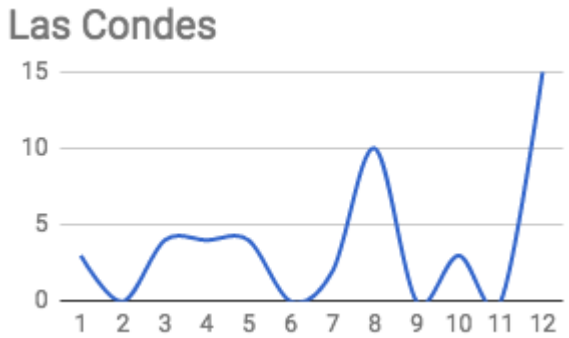
Figura 6. La noticia completa pasa por el NER, configurado para la detección de personas, organización y lugares, y además se carga el español para reconocer frases que hacen referencia a estas entidades. La salida corresponde al mismo texto pero con una anotación especial a las entidades encontrados. En nuestro caso sólo nos interesan los lugares.

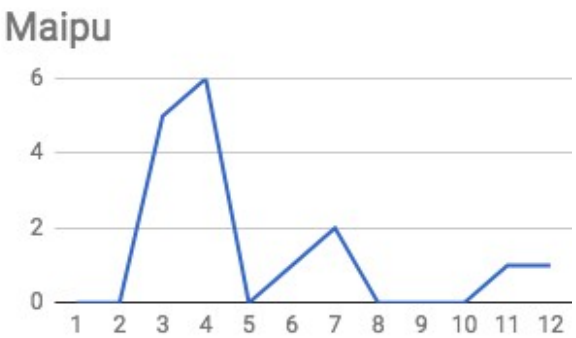
En medios noticiosos no es común que aparezcan las direcciones exactas de donde ocurren los robos, pero si el nombre de las comunas, ciudades o sectores.

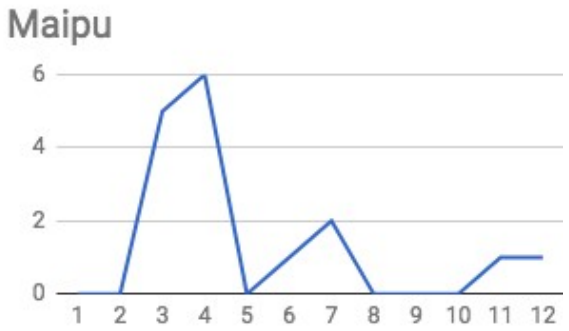
Para el siguiente análisis se consideraron todas las noticias del 2016. Se extrajeron las localidades y luego se registra la frecuencia de ocurrencias de la localidad en medios noticiosos en dichos meses y años indicados. Reducimos manualmente las localidades que

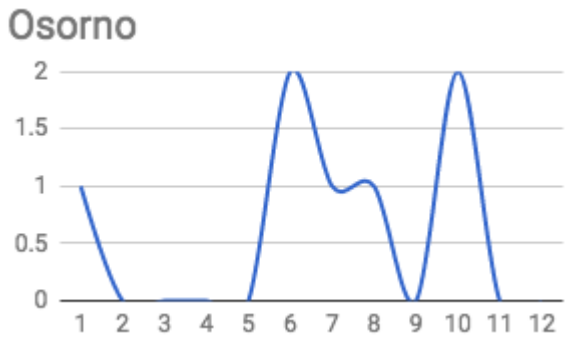
1 https://nlp.stanford.edu/software/CRF-NER.html

hacen referencia a un mismo lugar, como por ejemplo, “región metropolitana”, “santiago” y “gran santiago”. A continuación mostramos el resultado de 8 localidades de Chile que tuvieron mayores apariciones en medios noticiosos en el 2016.









En los gráficos se puede apreciar que comunas como San Bernardo tienen alzas en algunos meses, al igual que Maipú. Localidades como Quilicura o la misma comuna de Santiago, presentan mayor cantidad de ocurrencias de robos de vehículos. La comuna de "Las Condes" presenta un alza a mediados de año y a fin de año. Localidades al norte y sur (Arica y Osorno), presentan mucha varianza en sus datos. Esto último puede ser debido a que la prensa se concentra más en la zona central que lugares fuera de Santiago.

3.4. Modelamiento y seguimiento de tópicos en el tiempo

El modelamiento de tópicos (*topic modeling* en inglés) consiste en encontrar temas o tópicos que resumen una colección de documentos. Normalmente estos temas son "latentes", es decir, están implícitos en la colección de documentos --ya que no se han etiquetado explícitamente a los documentos como pertenecientes a ciertos temas--, y por lo tanto se deben descubrir a través de distintas técnicas que utilizan, entre otras cosas, la co-ocurrencia de palabras dentro de los documentos. Esta tarea es precisamente la que llevamos a cabo utilizando tanto los datos de la aseguradora (AACH) como los de Twitter: encontrar tópicos a partir de texto no estructurado y analizar si estos tópicos nos dan luces sobre formas en que se cometen delitos asociados a vehículos motorizados. El texto no estructurado corresponde a los relatos de los robos de vehículos (AACH) y a los mensajes posteados de forma pública en Twitter.

Algunas de las técnicas de modelamiento de tópicos están basadas en factorización matricial como LSI (latent semantic indexing) (Dumais et al, 2004) o NMF (Non-negative Matrix Factorization), (Xu et al, 2003) pero en este trabajo utilizaremos técnicas basadas en modelos probabilísticos generativos, en particular LDA (Latent Dirichlet Allocation) (Blei et al 2003) y una de sus extensiones, DTM (Dynamic Topic Modeling) (Blei et al, 2006).

Latent Dirichlet Allocation. LDA es un modelo probabilístico generativo para el modelamiento de tópicos. Su principal característica, comparado con modelos como LSI, es que utiliza la distribución Dirichlet como prior para dos distribuciones multinomiales, con el objetivo de evitar sobreajuste. Según (Steyvers, 2007), LDA especifica la siguiente distribución sobre palabras dentro de un documento

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j)$$

donde w_i es la palabra i-ésima, z_j es el tópico j-ésimo, T es la cantidad de temas, $P(w_i \mid z_i = j)$ es la probabilidad de la palabra w_i bajo el tópico z_j , y $P(z_i = j)$ es la probabilidad de que el tópico z_j haya sido muestreado para la palabra w_i .

Respecto a su utilización, el modelo recibe de entrada un corpus de documentos (por ejemplo, el texto de las denuncias de robo de vehículos) y el número de tópicos a descubrir (además de los hiper-parámetros α y β para las distribuciones Dirichlet; por defecto α=0.1 y β=0.01). A partir del proceso de inferencia, LDA obtiene como salida la lista de tópicos latentes (que corresponden a distribuciones de palabras), así como distribuciones de

tópicos por cada cada documento. Un ejemplo de estas salidas se puede ver en la Figura 7, donde se presenta parte de la distribución de palabras de tres tópicos y, para un documento específico que representa el relato de un robo, la distribución de tópicos con un gráfico de barras. La ilustración está basada en (Blei, 2012)

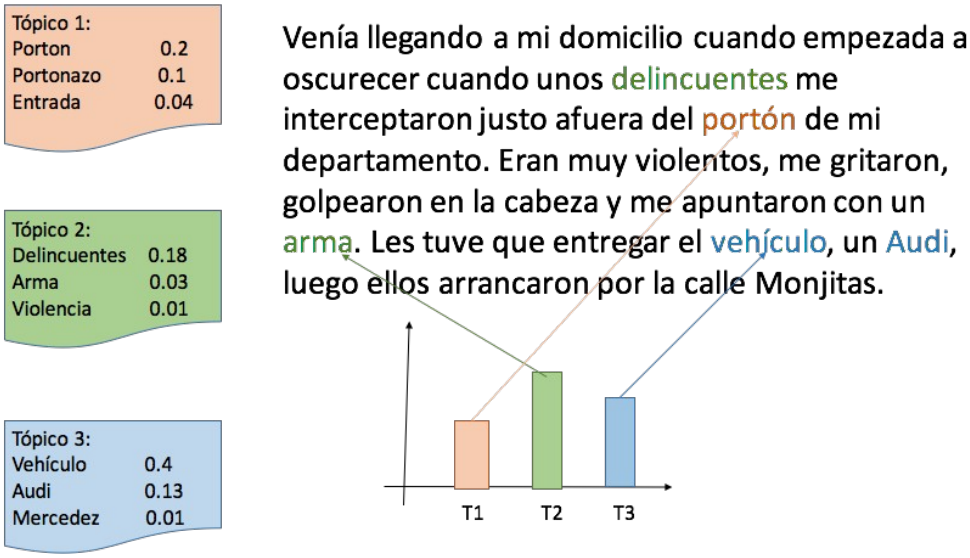


Figura 7. Resultado de ejecutar modelamiento de tópicos luego de utilizar LDA. Las figuras a la izquierda de distintos colores corresponde a tópicos (distribuciones de palabras, ordenadas según su probabilidad de pertenecer al tópico), y a la derecha se ve un ejemplo de un documento cuya distribución de tópicos se muestra en un gráfico bajo el texto. Ilustración basada en (Blei, 2012).

En la primera versión de nuestra herramienta de análisis, implementamos LDA para obtener tópicos, los cuales presentamos en detalles en la siguiente sección *4. Visualización Web: ITACaT*.

Dynamic Topic Modeling. Además de encontrar tópicos, nos interesa descubrir y analizar la evolución de estos tópicos en el tiempo, ya que esto nos permitiría estudiar la evolución de ciertos *modus operandi* para robo de vehículos. Para testear este tipo de análisis de tópicos en el tiempo, utilizamos DTM (Blei, 2006), una extensión de la popular técnica LDA (Blei, 2003). La salida de DTM es similar a la de DTM (una lista de tópicos, y la distribución de tópicos de cada documento), pero la principal diferencia entre DTM y LDA es que por cada ventana de tiempo, podemos observar cambios en la distribución de probabilidad de las palabras en un tópico. La Figura 8 muestra un ejemplo.

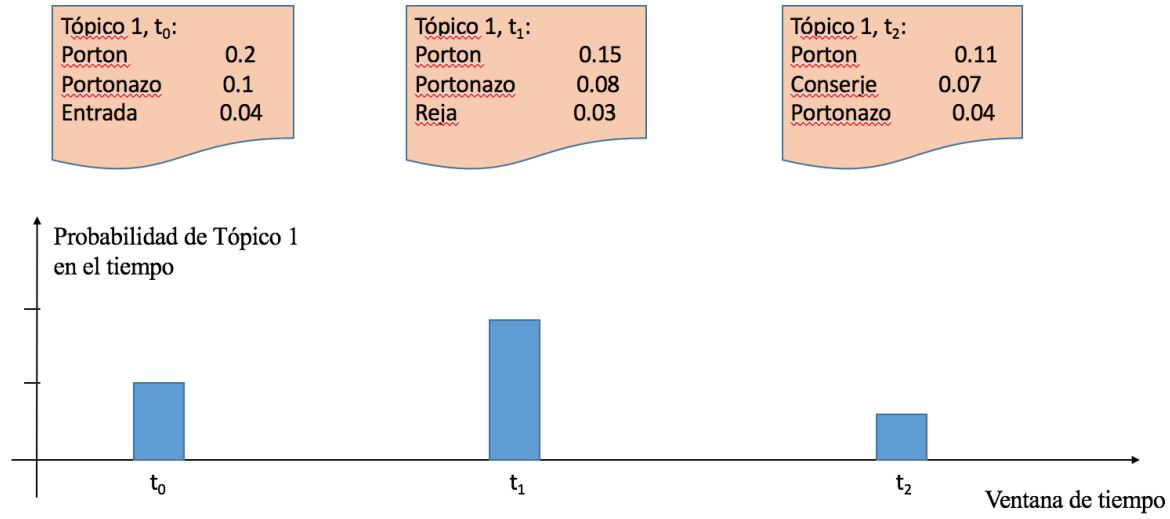


Figura 8. Ejemplo de evolución de un tópico usando DTM. En la parte superior se ve cómo cambia la probabilidad de las palabras del tópico en el tiempo (to, t1, t2) y abajo se muestra la probabilidad del tópico en el dataset.

En la práctica, DTM funciona de manera similar a LDA, pero considera secuencias de tópicos en lugar de tópicos estáticos. LDA no asume diferencias temporales entre los documentos. Sin embargo, esta suposición no se cumple cuando se analizan colecciones que pueden abarcar años, como las actas de congresos durante años consecutivos (Blei, 2012). En tales casos, queremos que un número fijo de tópicos evolucione con el tiempo, que es lo que proporciona DTM. En lugar de una distribución única sobre palabras, un tema ahora es una *secuencia de distribuciones sobre palabras*. Formalmente, el hiper-parámetro Dirichlet β_k se convierte en $\beta_{t,k}$ en DTM, donde el parámetro evoluciona con ruido gaussiano en función del estado anterior (Blei, 2006):

$$\beta_{t,k}|\beta_{t-1,k} \sim \mathcal{N}(\beta_{t-1,k}, \sigma^2 I),$$

donde *t* es un segmento de tiempo y *k* es un tópico. De forma similar, las proporciones de tópicos θ específicos del documento se extraen ahora de una distribución de Dirichlet con el hiper-parámetro α_t, y la estructura secuencial entre los modelos viene dada por (Blei, 2006)

$$\alpha_t|\alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I).$$

Al encadenar tópicos y distribuciones de proporción de tópicos, se obtiene una colección secuencial de modelos de tópicos, con el tópico *k*-ésimo en el segmento de tiempo *t* evolucionando desde el tema k-ésimo en el momento de tiempo *t-1*.

Evaluación: Tanto para LDA como para DTM, un componente importante en el modelado de tópicos corresponde a identificar un número apropiado de tópicos, ya que este valor es ingresado por el usuario. La evaluación de esta parte del proceso se realiza probando varios

números de tópicos y observando su comportamiento en base a varios índices[2]:

- Índices de (Arun et al., 2010) y (Cao et al, 2009), que se deben minimizar, e
- Índices de (Deveaud et al., 2014) y (Griffiths et al., 2004), que se deben minimizar.

Como ejemplo, para una muestra de los relatos de robos de autos de la base de datos de AACH entre los años 2015-2016, obtenemos el siguiente resultado de los índices mencionados en la Figura 9. Esto nos indica que el número óptimo de tópicos podría ser 4 o 14, lo que se puede posteriormente validar al revisarlo con expertos.

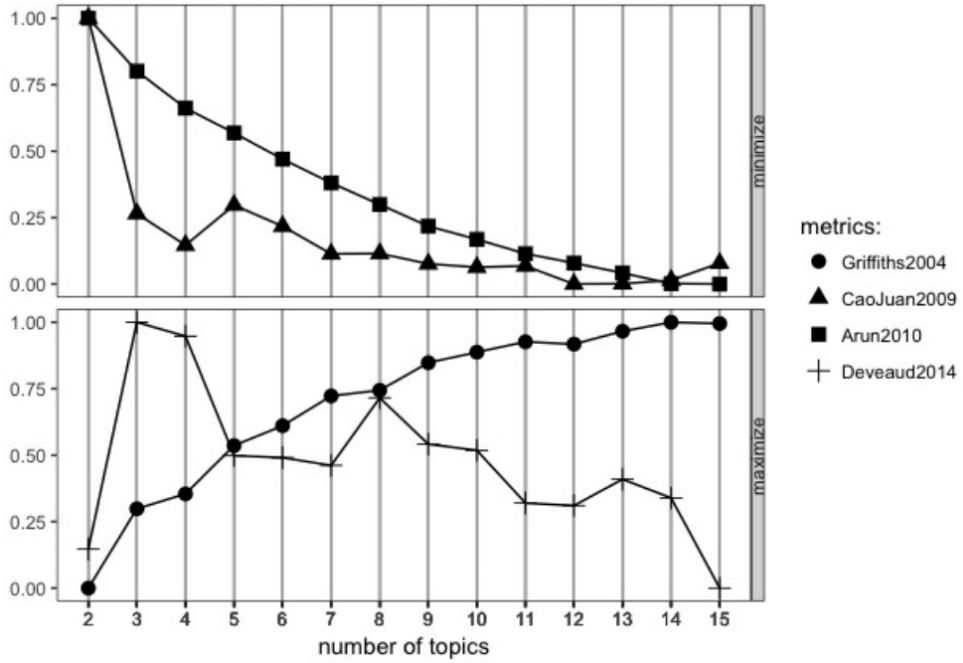


Figura 9. Resultado de análisis de distinto número de tópicos sobre una muestra (2015-2016) del dataset de la AACH.

Por otro lado, para evaluar los tópicos desde un punto de vista de coherencia y semántica, utilizaremos 2 analistas de la AACH a quienes les solicitaremos que evalúen la calidad de los tópicos encontrados utilizando como metodología la descrita en (Cheng et al, 2009). Esta metodología va en la dirección de solicitar a los usuarios información sobre *word intrusion* (palabras erróneamente incluidas en un tópico) y *topic intrusion* (tópicos erróneamente asignados a un documento).

3.5. Modelo de tópicos semi-supervisados

Un aspecto que deseamos incluir en este proyecto es utilizar el conocimiento de los analistas de compañías de seguro para mejorar nuestros modelos, ya que ellos son expertos en el dominio del problema (robo de vehículos). Por un lado, es posible ver el proceso de minería de datos en una sola dirección, donde el modelo recibe datos de entrada y produce patrones de salida para ser consumidos. Por otro lado podemos ver este proceso de forma iterativa donde el experto, al analizar los resultados, puede proveer información que mejore el modelo y este se mejora incrementalmente, continuando el ciclo.

2 https://cran.r-project.org/package=ldatuning

Esto es conocido como *Human-in-the-loop ML*, es decir, usar una persona en el proceso de aprendizaje del modelo.

En este contexto, el modelo (Jagarlamudi et al, 2012) presenta una opción factible para implementar esta funcionalidad. En este artículo, a diferencia de LDA que es completamente no-supervisado, los autores presentan una versión semi-supervisada que permite entregar información parcial sobre los tópicos para mejorar el modelo. Por ejemplo, un analista podría etiquetar un número limitado de tweets (del orden de decenas) indicando que pertenecen o no a ciertos tópicos, y el modelo debería actualizarse en función de dicha información. El usuario podría seguir indicando feedback en varias iteraciones, implementando de esta forma el proceso de "human-in-the-loop" LDA.

Evaluación: Crearemos una metodología en base a la descrita en (Cheng et al, 2009) para medir los niveles de *word intrusion* y *topic intrusion*, pero a diferencia de LDA, la analizaremos en función del feedback que provee el usuario (analista de la AACH)

3.6. Modelo de Angel

Evaluación:

4. Interfaz Web de Herramienta de Reporte y Análisis

La visualización actual es una aplicación web la cual agrupa distintos análisis que se han realizado a lo largo del proyecto con distintos tipos de visualización. La aplicación web se encuentra protegida con usuario y contraseña dentro del sistema actual con el fin de controlar el acceso a usuarios autorizados, como se observa en Figura 4.1 (a). Una vez que el usuario se encuentra autenticado, se presenta un menú lateral, Figura 4.1 (b), que cuenta con las distintas opciones de visualización: Informe mensual requerimiento de AACH, 3D Maps (implementados con la API Uber DeckGL) e ITACat: Sistema espacial de tópicos.

La aplicación se encuentra implementada en el framework Ruby on Rails (RoR) debido a las ventajas que ofrece RoR en su arquitectura MVC, que permite separar la lógica de las vistas sin necesidad de usar dos sistemas distintos. Además de ser el servidor para las visualizaciones, la ventaja que ofrecen las librerías otorga flexibilidad para extender de forma ordenada las nuevas necesidades y *features* del proyecto.

Para ingresar a la aplicación se pueden ocupar las siguientes credenciales:

Correo: test@test.com
Contraseña: 123456

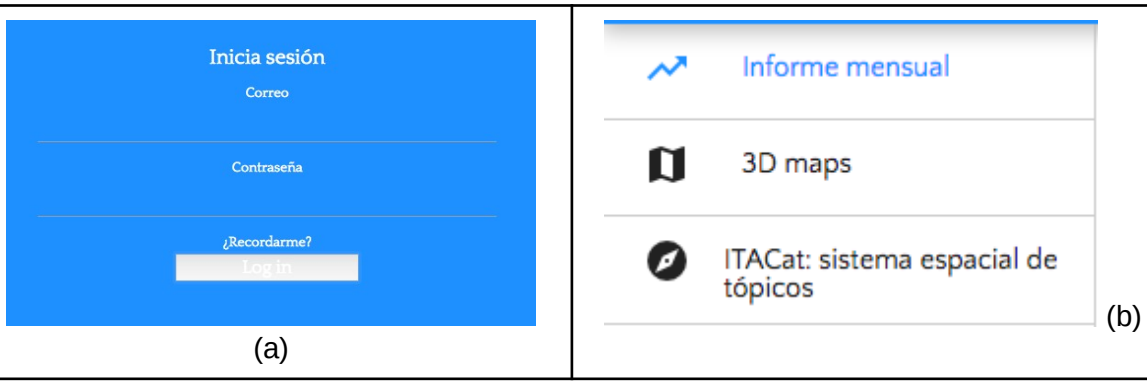


Figura 10. Interfaz del sistema para (a) autenticación y (b) menú principal para acceder a reportes y herramientas de análisis.

4.1. Visualizaciones Interactivas Basadas en Reporte AACh

En esta sección se presentan gráficos correspondientes a los informes mensuales realizados por la AACH, uno de ellos se ve en la Figura 11. Estos gráficos son interactivos, lo que permiten visualizar los datos de manera más fácil con el cursor y en otros casos poder mostrar o esconder otros valores. Están hechos con Chart.js, un framework en javascript especializado en visualizaciones de datos. La ventaja de este framework por sobre otros es que sus visualizaciones de gráficos (charts) son pre-compilados, y estos además acogen distintos estándares de visualización de información. Otra herramienta similar es D3.js, popular y flexible pero no orientada exclusivamente a presentar gráficos. Sin embargo, trabajar con D3.js en esta visualización hubiese requerido re-escribir código ya optimizado en Charts.js. La única desventaja general de Chart.js frente a D3.js es que el primero renderiza como *canvas* mientras que el segundo como *SVG*, haciéndolo un poco menos eficiente en el navegador. A pesar de esto, como se dibuja un número de gráficos en el orden de decenas, no es un problema en cuanto a la renderización de imágenes.

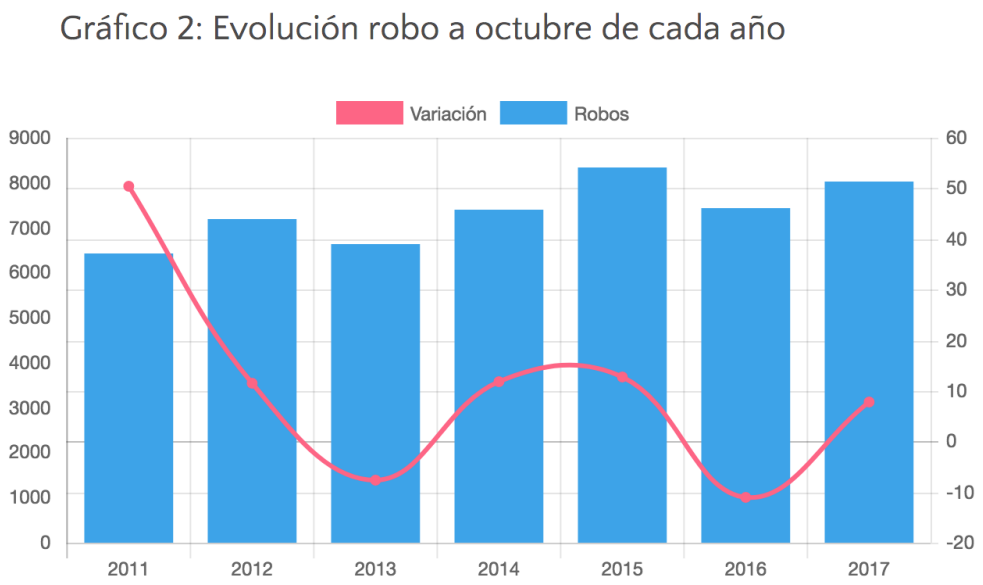


Figura 11. Gráfico interactivo basado en reporte mensual que realiza AACh.

4.2. Visualización Espacial de Robo de Vehículos

Esta sección de la herramienta cuenta con dos mapas interactivos: uno para explorar histogramas de robos en base a diferentes filtros, y otro para comparar patrones en distintos momentos. El mapa interactivo 1 (Figura 12) muestra un mapa de Santiago con robos de distintos años de las categorías automóviles y Station-Wagon. El mapa permite que se cambien valores de exploración, tales como el año, los cuatrimestres a cada año y la categoría de vehículo. Además, se puede regular el tamaño de las barras del histograma espacial para visualizar a distintos niveles de granularidad.

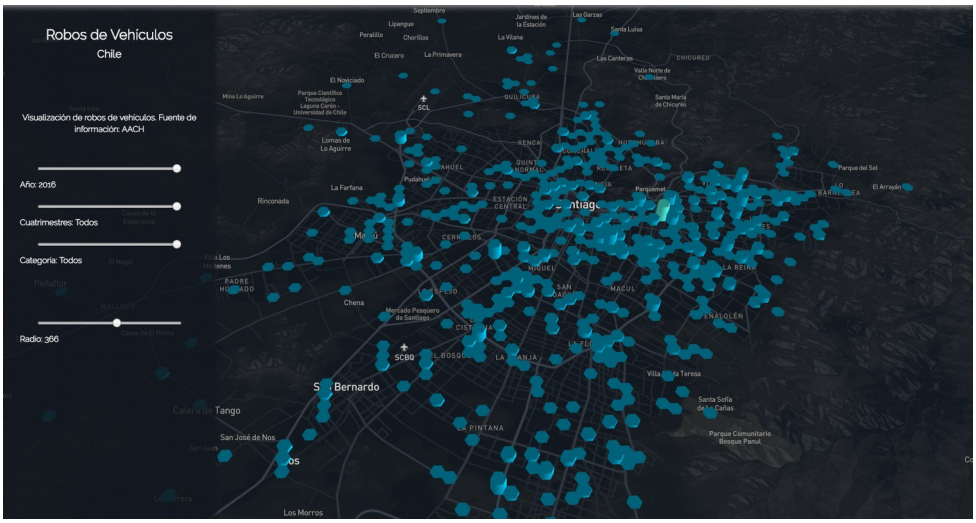


Figura 12. Visualización 3D que presenta histograma espacial de robo de vehículos.

El mapa comparativo (Figura 13) muestra 3 mapas de Santiago de forma simultánea. El primer mapa corresponde al total de robos. El segundo mapa corresponde a los robos ocurridos entre las 08:00 y 20:00 horas. El tercer mapa corresponde a los robos ocurridos entre las 20:00 y 08:00 horas. Así, se dispone de una forma de analizar los robos que son efectuados en el día o en la noche, y se visualizan los sectores más vulnerables. Además, el mapa tiene la opción de ver los datos en 3D o planos. Se destaca que se pueden generar estos gráficos agregándolas en base a otras variables (días de semana vs. fin de semana, marca 1 vs. marca 2, etc.), por lo que las posibilidades de análisis son amplias.

Ambas visualizaciones se encuentran implementadas en React y utilizan la librería DeckGL de Uber. La ventaja de DeckGL es la calidad de renderizado de imágenes y cómo usa los datos en tiempo real para ajustar la visualización. Sin embargo, necesita más recursos de memoria RAM para poder realizar las visualizaciones. La memoria necesaria dependerá de la cantidad de *layers* (capas) que posee la visualización. La fluidez de la visualización es de alrededor 60 *fps* (frames per second), por lo que se debe considerar un poder de procesamiento adecuado de CPU o GPU. Finalmente cabe destacar que DeckGL está construido sobre WebGL, una popular librería de renderizado 2D y 3D escrita en javascript, cuyas visualizaciones son embebidas en html a través de tags de tipo canvas.

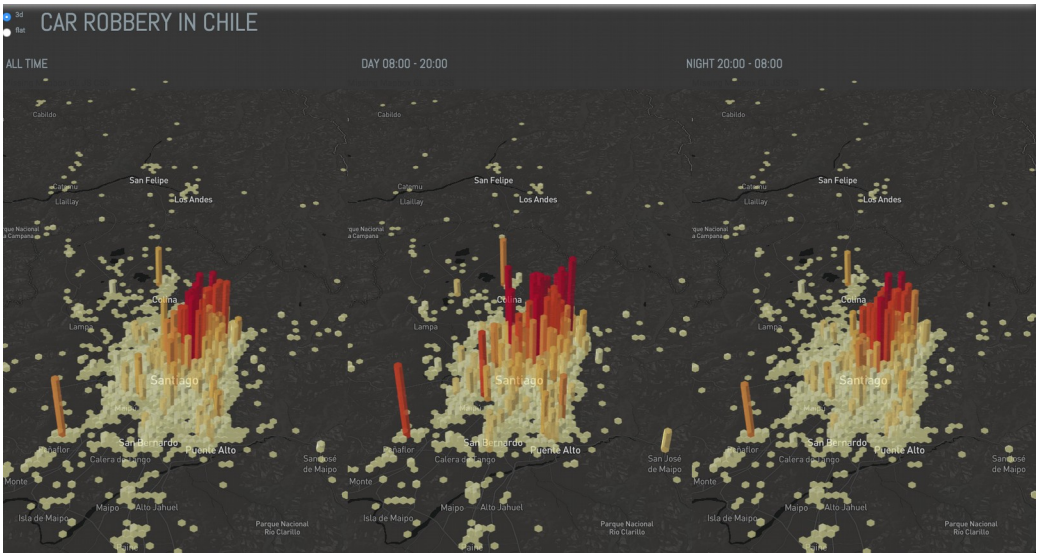


Figura 13. Mapa comparativo de robo de vehículos presentados como histograma 3D espacial

4.3. Visualización ITACaT

La visualización ITACaT es un sistema que permite a los usuarios explorar un mapa para ver los relatos que han sido geolocalizados, además de los tópicos de estos relatos obtenidos a través de procesarlos con *LDA*. Cada relato contiene información sobre cómo ocurrió el robo, pero su nivel de detalle depende de la información provista por la persona que realizó la denuncia. Utilizando LDA, cada relato se asigna al tópico con mayor probabilidad, ya que como se explicó en la Sección 3, un relato puede estar asignado a varios tópicos a la vez con distinta probabilidad. La visualización parte inicialmente con 4 tópicos, pero presenta una vista también para 10 y 14 tópicos. Estos números de tópicos no son arbitrarios, se obtuvieron en base a los índices de Griffiths, Cao, Arund y Devaud, indicados en la Sección 3.

En la visualización, se le asigna un color a cada tópico, y el relato se encuentra ubicado en el mapa y visualizado como un círculo.

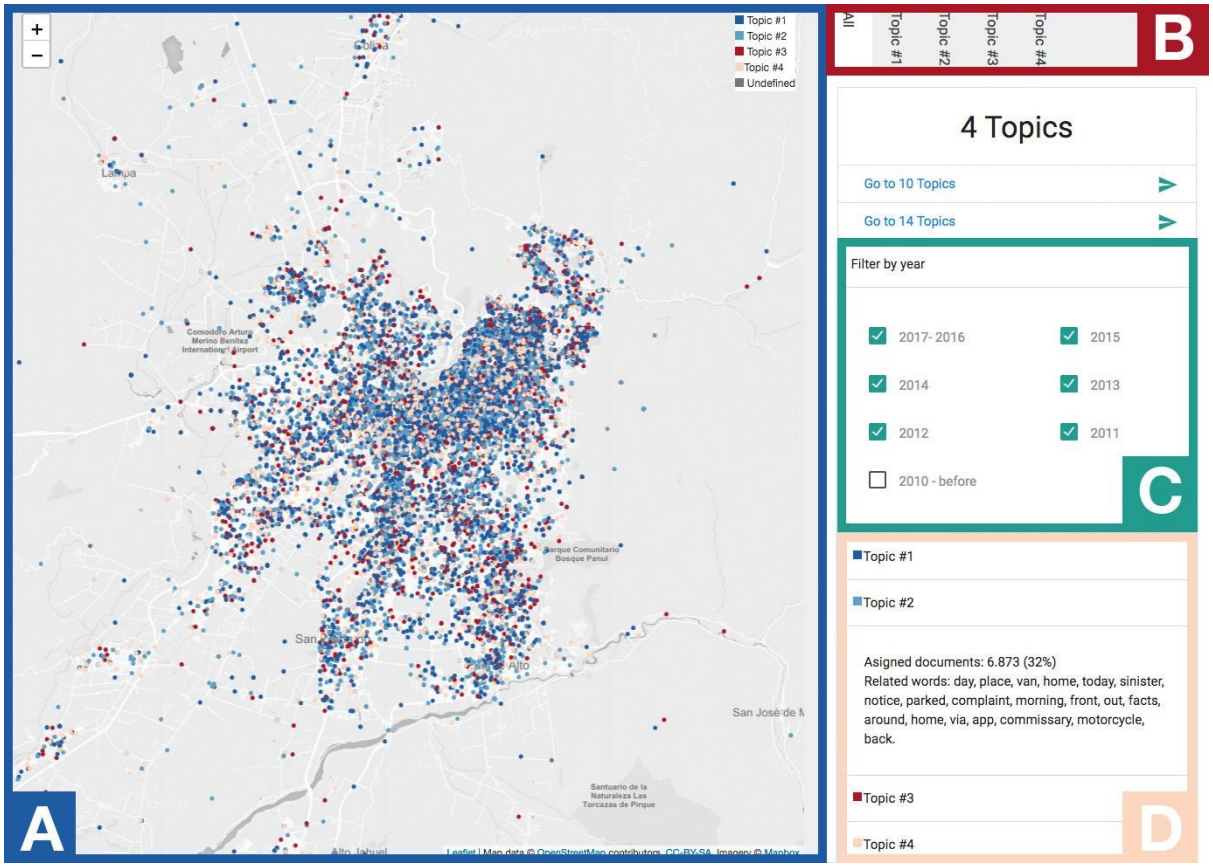


Figura 13. ITACaT con sus secciones. La sección A presenta la vista principal con los relatos geolocalizados, la sección B permite explorar en detalle la distribución de cada tópico, la sección C permite filtrar los relatos y la sección D es una forma de ver rápidamente el contenido por tópico, es decir, la lista de palabras más representativas.

El sistema utiliza la librería Leaflet[3] para incluir elementos en el mapa y D3.js[4] para los gráficos. Así mismo, los frames del mapa son extraídos usando la API de Mapbox[5]. Este conjunto de herramientas tienen la ventaja de funcionar bien entre sí dado que todas son librerías ejecutadas desde el lado del cliente, especializadas en visualización.

5. Conclusiones

6. Literatura

[1] Dumais, S. T. (2004). Latent semantic analysis. Annual review of information science and technology, 38(1), 188-230.

[2] Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 267-273). ACM.

3 http://leafletjs.com/
4 https://d3js.org/
5 https://www.mapbox.com/

[3] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[4] Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In Proceedings of the 23rd international conference on Machine Learning (pp. 113-120). ACM.

[5] Blei,D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.

[6] Arun,R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp.391-402). Springer, Berlin, Heidelberg.

[7] Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. Neurocomputing, 72(7-9), 1775-1781.

[8] Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. Document numérique, 17(1) 61-84.

[9] Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National academy of Sciences, 101(suppy1), 5228-5235

[10] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In Advances in neural information processing systems(pp.288-296)

[11] Jagarlamudi, J., Daumé III, H., & Udupa, R. (2012). Incorporating lexical priors into topic models. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (pp. 204-213). Association for Computational Linguistics.

[12] Luo, F., Cao, G., Mulligan, K., Li, X. (2016): Explore spatiotemporal and demographic characteristics of human mobility via Twitter: A case study of Chicago. Applied Geography 70, 11-25.

[13] Mikolov,T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119.

[14] Zhang, C., Zhang, K., Yuan, Q., Peng, H., Zheng, Y., Hanratty, T., Wang, S., and Han, J. (2017b). Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning. In Proceedings of the 26th International Conference on World Wide Web, pages 361–370. International World Wide Web Conferences Steering Committee.

[15] Wang, X., Gerber, M. S., Brown, D. E. (2012): Automatic Crime Prediction Using Events Extracted from Twitter Posts. In: Yang S.J., Greenberg A.M., Endsley M. (eds) Social Computing, Behavioral - Cultural Modeling and Prediction. SBP 2012. Lecture Notes in Computer Science, vol 7227. Springer, Berlin, Heidelberg, 231-238

[16] Zhang, C., Zhang, K., Yuan, Q., Peng, H., Zheng, Y., Hanratty, T., Wang, S., and Han, J. (2017b). Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning. In Proceedings of the 26th International Conference on World Wide Web, pages 361–370. International World Wide Web Conferences Steering Committee.

[17] Zhang, C., Liu, L., Lei, D., Yuan, Q., Zhuang, H., Hanratty, T., and Han, J. (2017a). Triovecevent: 17/18 Embedding-based online local event detection in geo-tagged tweet streams. In Proceedings of the 23rd ACM SIGKDD International

Conference on Knowledge Discovery and Data Mining, pages 595–604. ACM.

[18] Zhang, C., Liu, M., Liu, Z., Yang, C., Zhang, L., and Han, J. (2018). Spatiotemporal activity modeling under data scarcity: A graph-regularized crossmodal embedding approach.