PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
ESCUELA DE INGENIERIA

HERRAMIENTAS DE ANALÍTICA VISUAL PARA MODELOS DE TÓPICOS SOBRE COLECCIONES DE DOCUMENTOS

M. FERNANDA SEPÚLVEDA RAMÍREZ

Tesis presentada a la Dirección de Postgrado
como parte de los requisitos para optar al grado de
Magister en Ciencias de la Ingeniería

Profesor Supervisor:
DENIS PARRA

Santiago de Chile, Enero 2019

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
ESCUELA DE INGENIERIA

HERRAMIENTAS DE ANALÍTICA VISUAL PARA MODELOS DE TÓPICOS SOBRE COLECCIONES DE DOCUMENTOS

M. FERNANDA SEPÚLVEDA RAMÍREZ

Miembros del Comité:
DENIS PARRA
VALERÍA HERSKOVIC
RICHARD WEBER
MAURICIO LÓPEZ

Tesis presentada a la Dirección de Postgrado
como parte de los requisitos para optar al grado de
Magister en Ciencias de la Ingeniería

Santiago de Chile, Enero 2019

A quienes me han ayudado a
mantener la cordura que me queda
hasta el día de hoy.

AGRADECIMIENTOS

Agradezco a los miembros del proyecto Fondef ID 16I10222 por la motivación y la oportunidad de trabajar en el presente tema de tesis, además de permitirme ser parte de tan motivante proyecto. Agradezco al Departamento de Ciencias de la Computación de la Universidad Católica, por darme un segundo hogar y apoyarme casi como una familia en las buenas y en las malas de mi vida académica y a veces personal. Agradezco a mi profesor supervisor Denis Parra, por guiarme a lo largo de este proyecto y exigirme lo necesario para poder poder llegar al final de este camino.

Agradezco a los amigos que he formado a lo largo del camino, así como a mi hermano y madre, por apoyarme en distintas situaciones que me han permitido llegar hasta aquí. Agradezco a mi amiga Camila por ser tan buena amiga y un soporte emocional a lo largo de todo este camino. Finalmente agradezco a Felipe por estar siempre a mi lado, apoyarme y ser un pilar fundamental en mi vida y a lo largo de esta carrera.

INDICE GENERAL

INDICE DE FIGURAS

INDICE DE TABLAS

RESUMEN

En el presente trabajo de tesis se exploran herramientas para la visualización de tópicos localizados espacialmente, sobre un corpus de documentos de robos de vehículos en Chile, en el contexto del proyecto Fondef ID 16I10222, denominado "Observatorio digital de delincuencia en Chile", cuyo objetivo es consolidar la información recopilada por la Asociación de Aseguradoras de Chile (AACh) sobre robos de vehículos y con lo anterior realizar un sistema capaz de caracterizar los modi operandi de los delincuentes, así como su evolución, mediante técnicas de minería de datos. Debido a que las aseguradoras tienen datos con muchas dimensiones y carecen del conocimiento y capital humano para procesarlo, el aporte de este trabajo a la resolución de este problema es a través del estudio y desarrollo de herramientas que permitan la identificación de patrones de robos de vehículos, como por ejemplo los *portonazos*. La herramienta de analítica visual desarrollada permite analizar y descubrir patrones, usando métodos de aprendizaje de máquina no supervisado como modelos de tópicos, además visualizaciones interactivas para analítica visual. A partir de lo anterior se llevó a cabo la implementación de la herramienta con dos alternativas visuales: usando *small multiples* de gráficos de barras y por otra parte *small multiples* de gráficos de línea para representar series de tiempo. Ambas interfaces fueron sometidas a una evaluación con usuarios, donde se midió el desempeño en cuanto a tiempo, interacción y rendimiento de cada una al resolver múltiples tareas sobre tendencias, agregación y sobre información puntual. De la evaluación se descubrió que ambas interfaces estudiadas se obtienen un buen desempeño en cuanto a la resolución de las tareas propuestas, con excepción de la tarea enfocada en la comparación de distribuciones, donde la interfaz de barras logra un mejor desempeño, a costa de mayor número de interacciones.

Palabras Claves: análitica visual, visualización, modelos de tópicos, minería de datos, robos de vehículos

ABSTRACT

In the following thesis work, tools for the visualization of spatially located topics are explored, on a document corpus of vehicle theft documents in Chile, in the context of the project Fondef ID 16I10222 ”Digital Crime Observatory in Chile”, whose principal objective is to consolidate the information compiled by the Association of Insurers of Chile (AACh) on vehicle thefts and with the above make a system capable of characterizing the modi operandi of criminals, as well as its evolution, using data mining techniques. Because insurers have data with many dimensions and lack the knowledge and human capital to process it, the contribution of this work to the resolution of this problem is through the study and development of tools that allow the identification of vehicle theft patterns, such as *portonazos*. The developed visual analytics tool allows analyzing and discovering patterns, using unsupervised machine learning methods such as topic modeling, as well as interactive visualizations for visual analytics. From the above, the implementation of the tool was carried out with two visual alternatives: using *small multiples* of barcharts and on the other hand *small multiples* of linecharts. Both interfaces were tested on an evaluation with users, in which the performance was measured in: time, interaction and performance of each interface when solving multiple tasks on trends, aggregation and on specific information. From the evaluation it was discovered that both interfaces perform well in terms of the resolution of the proposed tasks, with the exception of performance when comparing distributions. In this type of tasks, the barchart interface achieved a better performance than the linechart interface, although the barchart interface implied a greater number of interactions in comparison to the linechart.

Keywords: data analysis, visualization, topic modeling, data mining, car theft

Capítulo 1. INTRODUCCIÓN

Los robos de vehículos son un tema de relevancia a nivel mundial debido las implicancias económicas que conlleva este tipo de delito, ya que este no afecta solo a los propietarios de vehículos, sino que afecta a la sociedad en general porque se ha vinculado a otros tipos de delitos, como el terrorismo y el tráfico de bienes y/o personas (Commission et al., 2004). Existen varios actores interesados que pueden tomar medidas para abordar este problema, desde los organismos encargados de hacer cumplir la ley hasta los fabricantes de automóviles y compañías de seguros. Todas estas instituciones recopilan datos que pueden analizarse para descubrir patrones de delitos de vehículos, pero varias veces carecen de herramientas especializadas para esta tarea.

Para contribuir a resolver este problema, bajo el proyecto Fondef ID 16I10222, denominado "Observatorio digital de delincuencia en Chile", se propone una herramienta interactiva que permite a los analistas de las compañías de seguros de automóviles descubrir patrones de robo de automóviles a partir de texto no estructurado proveniente de diferentes fuentes de información. El objetivo es consolidar la información recopilada por la Asociación de Aseguradoras de Chile (AACh) sobre robos de vehículos y poder realizar un sistema capaz de caracterizar tanto, los *modi operandi* y la evolución de los delincuentes, mediante técnicas de minería de datos.

El aporte de este trabajo a la resolución del problema anteriormente expresado es a través del estudio y desarrollo de herramientas que permitan la identificación de patrones de robos de vehículos, como por ejemplo los *portonazos*. El término *portonazos* es un término usado en el área de los robos de vehículos en Chile que implica el tipo de robo ocurrido fuera de una casa al momento de intentar ingresar por parte del dueño, abriendo el portón, es entonces cuando el vehículo es interceptado por varios individuos que aprovechándose de la situación se llevan finalmente el vehículo. Identificar este tipo de patrones es importante, pues permite tomar las medidas preventivas sobre los nuevos tipos de robo que surgen a través del tiempo y en distintas zonas geográficas.

Así, la herramienta a desarrollar tiene como fin principal (pero no único) ofrecer a analistas en el dominio de robos de vehículos un instrumento para poder inferir y encontrar nuevos patrones a partir de datos no estructurados. Para su evaluación, se usó de la base de datos de la AACh. En esta base de datos cada entrada equivale a un siniestro, es decir, a un vehículo robado. Cada siniestro, además es documentado con diversos datos, alcanzando así la base de datos un total de 70 atributos distintos. Entre estos se encuentra el relato escrito, el cual contiene información sobre cómo ocurrió la situación del robo. La herramienta a desarrollar logrará procesar el texto no estructurado con el fin de encontrar los patrones ya mencionados con mayor facilidad.

Para encontrar patrones se espera hacer uso de algoritmos de modelos de tópicos. Los modelos de tópico o modelamiento de tópicos (conocido en inglés como *topic modeling*) es una forma de encontrar estructuras en una colección de documentos (se denomina a la colección *corpus*). El modelamiento de tópicos es en un inicio un tipo de aprendizaje no supervisado que realiza agrupación de documentos (o como se denomina en la literatura *clusters*) en base a uno o varios tópicos. Se entiende por tópico una estructura semántica "abstracta" que caracteriza a un documento en base a su contenido. Por otra parte, cuando se habla de aprendizaje no supervisado, el modelo entrenado se clasifica en base a la propia estructura de los datos y cómo fue entrenado (es decir, los parámetros del modelo).

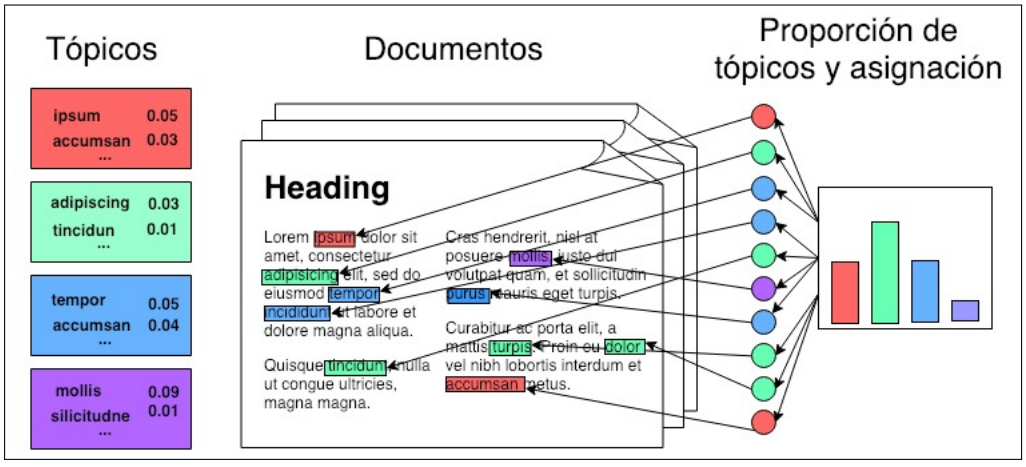


FIGURA 1.1. Figura ilustrativa del modelo probabilístico de topic modeling basada en el paper de David Blei (2012).

A continuación se describe un ejemplo práctico con uno de los algoritmos de modelamiento de tópicos llamado Latent Diritchlet Allocation (LDA) (Blei, 2012a), para una tarea que consiste en definir sobre un conjunto de historias qué palabras están asociadas a los tópicos "terror" y "comedia". Antes de ejecutar el algoritmo, a partir del conjunto de historias, lo que correspondería al corpus de documentos, se identifican las palabras distintas en cada documento para así crear el diccionario de palabras. Para la evaluación del modelo no se recibe información extra además de la misma estructura del documento (se dice que es no supervisado). Se usa el diccionario de palabras, que representa a todas las palabras que aparecen en el corpus de documentos, para que el algoritmo pueda determinar la probabilidad de cada palabra de pertenecer a cada tópico, y de cada documento de pertenecer a cada tópico. Con lo anterior se espera que el resultado del algoritmo asignará una probabilidad alta a "fantasma" de pertenecer a "terror" y "payaso" a "comedia", y así mismo con las historias, por ejemplo a "IT" de Stephen King de pertenecer a "terror" y a "El diario de Bridget Jones" de Helen Fielding de pertenecer a "comedia". El resultado del algoritmo es la la distribución de palabras para un tópico y distribución de tópicos para cada documento.

El problema de encontrar patrones utilizando modelos es que el análisis de los resultados puede resultar complejo, en especial cuando el corpus de documentos es muy grande. Entonces, es importante para poder identificar patrones en los robos, acompañar el conocimiento generado por el modelo con visualizaciones o herramientas analíticas que permitan sintetizar el resultado del modelo para facilitar su análisis. Este paso es el fundamental para la generación de conocimiento y el aporte real al área estudiada (Fayyad y Stolorz, 1997). Debido a la importancia de este punto es que en el presente trabajo se desarrolla la implementación de la visualización y su evaluación.

Al desarrollar una herramienta analítica de información interactiva, se requieren tomar decisiones de diseño acorde a las necesidades que se buscan resolver. Es por lo anterior que se usa la metodología de Tamara Munzner (Munzner, 2009) para así desarrollar la interfaz. Para realizar la visualización se sigue lo expuesto en el *framework* de la autora ya mencionada, y se implementa una visualización web con el fin de poder

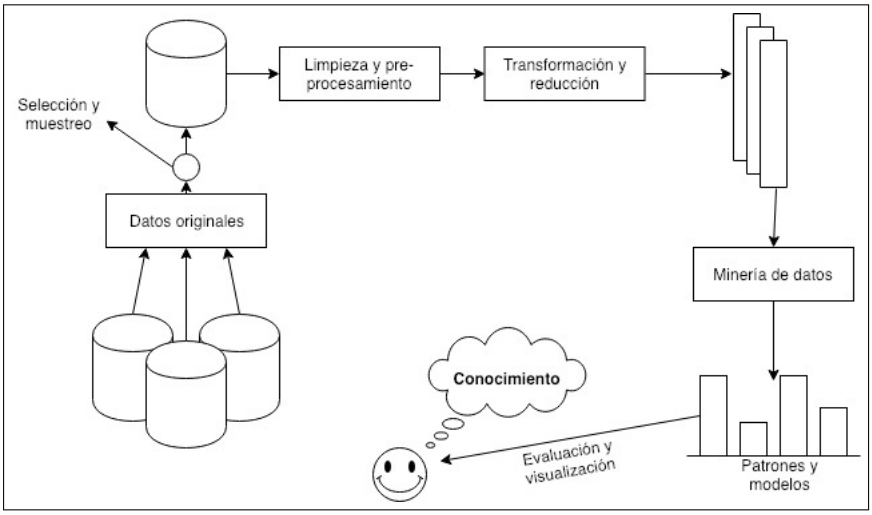


FIGURA 1.2. Proceso KDD para generar conocimiento de Fayad y Stolorz (1997).

probar y validar con usuarios reales la interfaz. La forma de validar la herramienta implementada es evaluando tareas relacionadas con la identificación de patrones de robos. En base a los resultados de la evaluación se obtiene el rendimiento, así como el tiempo por tarea e interacciones totales (cantidad de *clicks* y movimiento en la página) por cada tarea para todos los usuarios de la evaluación. Con estos resultados es posible comparar y determinar que tan efectiva fue una interfaz al momento de resolver una tarea.

Capítulo 2. MARCO TEÓRICO

Con el fin de dar un conocimiento previo y una base teórica al presente trabajo, es que a continuación se describen las teorías y el estado del arte actual en las dos áreas principales de este trabajo, que son modelamiento de tópicos y visualización de datos cómo área de conocimiento, y en particular, enfocada al modelamiento de tópicos.

2.1. Modelamiento de tópicos

Antes de describir en detalle lo que es modelamientos de tópicos, se describe su contexto. En los últimos años el aprendizaje de máquina, un área del campo de la inteligencia artificial, ha incrementado el interés en los investigadores en diferentes dominios (Brynjolfsson y Ms, 2017). El aprendizaje de máquina como su nombre indica es el conjunto de técnicas que permiten a un sistema computacional aprender, es decir, cada vez que cambie su estructura, programa o datos ya sea, debido al ingreso de algún input o a una respuesta a información externa, los cambios involucrarán una mejora en su rendimiento (Nilsson, 1996). Bajo esta área del conocimiento, surgen técnicas para el modelamiento de tópicos, que a grandes rasgos es una forma de analizar texto, cuyo fin es encontrar estructuras en una colección de documentos (*document corpus* o *text corpora*) con el objetivo de categorizarlos de forma semi-automática, o en terminología del área, hacer *clusters* o agrupación de documentos en base a uno o varios tópicos, logrando así también una reducción en la dimensionalidad del contenido de cada documento. En la Figura 1.1 se puede apreciar una ilustración de este proceso. Este tipo de técnicas es importante para el área de recuperación de información (*information retrieval*).

Si bien el concepto de modelar tópicos ha sido ampliamente estudiado en el ámbito de las ciencias sociales, no fue hasta años más recientes que, con el surgimiento de la ciencia de la computación, las tareas de clasificación se pudieron semi-automatizar. Una de las primeras aproximaciones al área y la cual define gran parte de los conceptos a tratar es *bag-of-words*, donde se asume que el orden de las palabras dentro de un documento no influye por lo que éste se ignora (Wallach, 2006). Siguiendo este planteamiento

sobre cómo estructurar el documento, se encuentra *Vector Space Model* para representar documentos (Manning, Raghavan, y Schütze, 2008), donde como su nombre lo índica, el documento se representa como una matriz de vectores. Estos conceptos son utilizados para métodos como *tf-idf* (*term frequency-inverse document frequency*) (Salton y McGill, 1986). En la técnica de *tf-idf*, se elige un vocabulario en base a las ocurrencias de las palabras (o como se refieren en el método, término) en los documentos. Luego por cada documento se mantiene la cuenta de la cantidad de ocurrencias de cada palabra del vocabulario. Luego de normalizar la frecuencia de las palabras (*term frecuency*) se compara con la frecuencia inversa de las palabras en todo el corpus del documento. El resultado de esta comparación es una matriz término-documento, donde cada columna contiene el valor *tf-idf*, valor que indica la relevancia de esa palabra en el documento. La técnica descrita es importante ya que fue la base para los motores de búsqueda (Baeza-Yates, Ribeiro, et al., 2011) y hasta no muchos años atrás fue popular en el área de recomendación de documentos (Pazzani y Billsus, 2007), esto en parte a que es una aproximación simple para explicar los métodos actuales de modelamiento de tópicos.

A partir del modelo *bag-of-words* basado en pesos de *tf-idf* surge *Latent Semantic Indexing* (LSI) (Deerwester, Dumais, Furnas, Landauer, y Harshman, 1990), método que a partir de la matriz término-documento utiliza la descomposición de valores singulares o *Singular-Value Decomposition* (SVD) sobre la matriz término-documento obtenida con *tf-idf*, para aproximarla a una combinación linear, y así identificar la varianza de la colección de documentos. Una extensión importante de esta aproximación es *probabilistic Latent Semantic Indexing* (pLSI) (Hofmann, 1999), donde cada palabra es modelada como una muestra de un *mixture model*, donde los componentes de la mezcla son variables multinomiales aleatorias, que pueden ser interpretadas como las representaciones de los tópicos o categorías. Entonces cada palabra es generada de un tópico, y un documento tiene entonces una mezcla de proporciones de pertenecer a cada tópico según las palabras en el documento, entonces se define la probabilidad de pertenecer al documento en base a esas proporciones. El problema de este modelo es que no se provee un modelo probabilístico a nivel de documento, si no que solo a nivel de palabras. Esto es un problema

según el teorema de Finetti (Diaconis y Freedman, 1990) el cual establece que cualquier colección de variables aleatorias intercambiables tiene una representación como una mezcla de distribuciones (*mixture distribution*). Por lo anterior pLSI al no considerar un modelo para los documentos, estos no son intercambiables entre sí. Bajo esta línea de pensamiento surge el modelo de *Latent Dirichlet Allocation* (LDA), que se explicará en detalle a continuación por ser la base para el resto de la lectura.

Latent Dirichlet Allocation es un modelo generativo estadístico introducido por David Blei, Andrew Ng y Michael Jordan (Blei, Ng, y Jordan, 2003). LDA es una técnica para descubrir tópicos de forma automática en un corpus de documentos. A cada documento se le asigna un tópico con cierta probabilidad. Un tópico es una distribución de probabilidad sobre las palabras de los documentos (donde al conjunto de palabras se le llama diccionario). De acuerdo con el modelo de tópicos probabilístico, la distribución sobre las palabras dentro del documento se expone de acuerdo a la siguiente ecuación:

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j) * P(z_i = j) \tag{2.1}$$

Donde w_i es la i-ésima palabra, T es la cantidad de tópicos, $P(w_i|z_i)$ es la probabilidad de la palabra w_i bajo el tópico z_j y finalmente $P(z_i = j)$ es la probabilidad de que el tópico z_j tenga la palabra w_i. Sea Φ la distribución multinomial sobre las palabras en el tópico y Θ la distribución multinomial para tópicos en los documentos. Los parámetros $\phi \in \Phi$ y $\theta \in \Theta$ indican la relevancia de la palabra para el tópico y la relevancia del tópico en el documento, respectivamente. El modelo generativo descrito por Blei introduce una distribución de Dirichlet con prior θ, un hiperparámetro α y con un parámetro β. En la figura 2.1 se muestra el modelo probabilístico de LDA.

El modelo de LDA descrito es considerado un tipo de modelo de aprendizaje no supervisado. Esto quiere decir que la clasificación de los tópicos no depende de una persona que etiquete los datos, si no que dependerá únicamente de la estructura de los datos. Por

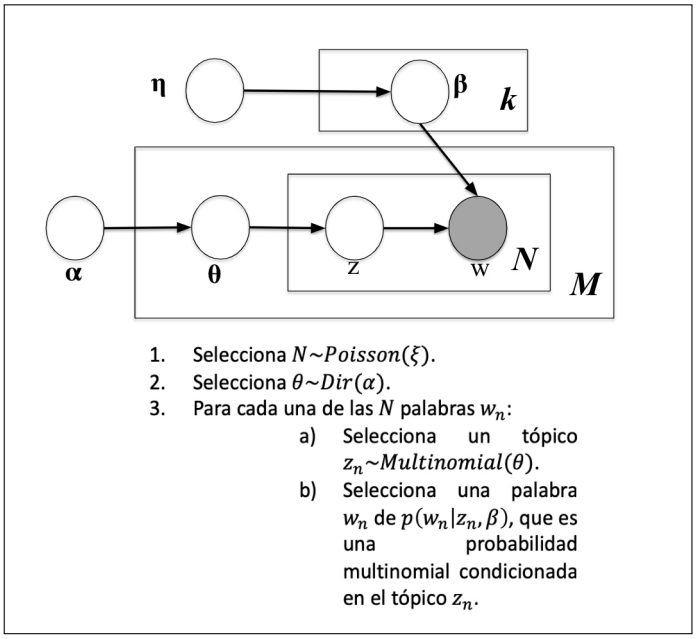


FIGURA 2.1. Modelo gráfico de LDA de David Blei (2003).

lo anterior LDA es un buen modelo para datos que no se encuentren previamente categorizados. Existen extensiones al modelo en dos ámbitos, el primero que incorporan más información al entrenamiento, y el segundo que añade supervisión. Entre las extensiones del modelo que añaden información extra (manteniendo así que sea no supervisado) está DTM (*Dynamic Topic Models*) (Blei y Lafferty, 2006) que añade la componente temporal al entrenamiento. Por otra parte entre las extensiones que añaden supervisión, es posible añadir *lexical priors* al modelo de LDA (Jagarlamudi, Daumé, y Udupa, 2012) (Andrzejewski, Zhu, y Craven, 2009), vale decir volver a entrenar el modelo actual integrando en el entrenamiento la interacción humana con las relaciones entre palabras y tópicos. Otra forma de integrar esta misma supervisión es añadiendo información sobre las relaciones entre palabras (Hu, Boyd-Graber, y Satinoff, 2011). Estas extensiones mejoran la eficiencia de la clasificación, sin embargo cabe destacar que el algoritmo de fondo usado en las extensiones mencionadas sigue la estructura básica de LDA ya explicada (Blei, 2012b). Fuera del ámbito de LDA existen soluciones para la clasificación de tópicos en distintas áreas del conocimiento, por ejemplo asociadas al procesamiento de

lenguaje natural (NLP) o al análisis de textos. Sin embargo este tipo de aproximaciones requieren de bases de conocimiento mayores (Mohr y Bogdanov, 2013) .

2.2. Visualización de datos

Los sistemas o interfaces de visualización proveen una representación visual de las fuentes de datos, diseñadas para ayudar a las personas a resolver tareas de forma más eficiente (Munzner, 2009). Por otra parte visualización se puede definir como la comunicación de información utilizando representaciones gráficas (Ward, Grinstein, y Keim, 2010). No se atribuye únicamente a un área del conocimiento, si no que es horizontal a todas las áreas que requieran comunicar información. La investigación en visualización ha reunido investigadores de ciencias de la computación, interfaces de usuario, psicología y percepción y estadísticas (Fayyad, Grinstein, y Wierse, 2002). Por lo anterior, la literatura en el área puede ser más específica dependiendo del campo al que se espera visualizar, por ejemplo una visualización financiera tendrá distintos objetivos que una visualización que busca analizar el comportamiento del caudal de un río, sin embargo, al momento de elegir cómo se codifica la información para visualizar cada una, se comparten las mismas decisiones de diseño. Frente a lo anterior, existen distintos *frameworks* o metodologías para diseñar interfaces.

Un *framework* para diseñar y analizar visualizaciones es el propuesto por Tamara Munzner (Munzner, 2009). En este se proponen cuatro capas anidadas cada una con el fin específico de validar distintas aristas de la situación a visualizar. Se muestra el *framework* en la Figura 2.2. En este *framework* el *output* de un nivel superior, es el *input* de los niveles inferiores.

Las descripción más detallada de las cuatro áreas es la siguiente:

- *domain situation*: se refiere al grupo de usuarios objetivo de la visualización, el dominio de su conocimiento e interés, sus preguntas y problemas, y los datos asociados. La salida de este nivel son las necesidades del usuario.

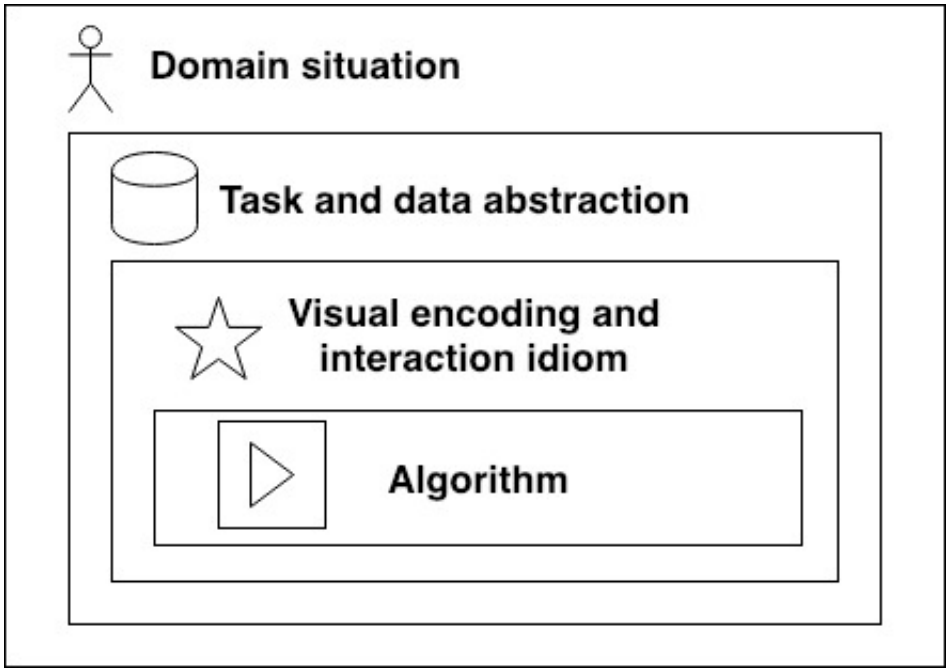


FIGURA 2.2. Framework de Tamara Munzner (2009).

- *task and data abstraction*: responde al *por qué* y *qué*. En esta capa se espera realizar un *mapping* entre los datos en su vocabulario especifico a una figura más abstracta y genérica con el fin de revelar las tareas asociadas. Es importante tener en cuenta el tipo de datos que se tiene y el formato, por ejemplo si es ordinal o nominal, y la escala de los datos. Por otra parte, una vez comprendido el set de datos se debe identificar el tipo de acción o tareas a realizar (responder el *por qué*). En la Figura 2.3 se describen las tareas genéricas clasificadas en 3 niveles: Analizar, Buscar y Consultar. Por otra parte también se definen los tipos de objetivo o meta que se buscan responder con estas tareas, especificados en la Figura 2.4.
- *visual encoding and interaction idiom*: responde al *cómo*. El idioma es cómo se representará la información. Depende de dos factores, el *visual encoding*, que es cómo exactamente lo verán los usuarios, y la interacción, que define cómo cambian los usuarios lo que ven. Algunas de las decisiones en esta capa son el tipo de Marca (*Marks*) como por ejemplo puntos, líneas y áreas, y los canales

visuales *Visual Channels* como la posición, la forma y el tamaño. Al momento de elegir los elementos del *visual encoding*, lo importante es lograr expresividad y efectividad.

- *algorithm*: cómo se conecta la interacción al *encoding visual* de forma eficiente.

Analizar		
Consumir		
Descubrir	Presentar	Disfrutar
Producir		
Anotar	Grabar	Derivar

Buscar		
	Objetivo conocido	**Objetivo desconocido**
Posición conocida	Buscar	Observar
Posición desconocida	Localizar	Explorar

Consultar		
Identificar	Comparar	Resumir

FIGURA 2.3. Niveles de acción: Analizar, Buscar y Consultar, según framework de Tamara Munzner (2009).

Lo importante de este *framework* es que una mala decisión en un nivel superior afectará a los niveles inferiores, logrando así exponer que el diseño de la visualización no necesariamente resolverá el problema que se busca atacar. Así mismo el *framework* propone cómo luego de identificar cada área validar la visualización por nivel.

Por otra parte, es importante recalcar la importancia de la visualización de datos en el área de *data science*. La visualización de datos es fundamental para la comprensión y el análisis de los datos. Un ejemplo claro es la famosa comparación de datos de Anscombe (Anscombe, 1973), donde se demuestra como cuatro sets de datos con medidas estadísticas idénticas se comportan de manera totalmente distintas al graficarlas en un *scatterplot*. De hecho, debido a la importancia de comprender los datos más allá del

FIGURA 2.4. Objetivos o metas de un usuario al realizar una acción framework de Tamara Munzner (2009).

valor, la visualización es parte del proceso KDD (Knowledge Discovery in Databases) (Fayyad, Piatetsky-Shapiro, y Smyth, 1996).

Luego bajo esta perspectiva, se pueden identificar dos grandes conjuntos de tipos de visualización en el ámbito de ciencias de la computación (Fayyad et al., 2002). La primera esta enfocada en el almacenamiento de datos y recuperación de información, y la segunda esta centrada en la noción del algoritmo, tal que permitan la detección o extracción de patrones y modelos estadísticos sobre los datos, esto es, relacionado al ámbito de reconocimiento de patrones, inteligencia artificial, aprendizaje de máquina y KDD (Fayyad et al., 2002). Es sobre este último grupo en que las herramientas de este trabajo se encontrarán enfocadas.

Capítulo 3. TRABAJO RELACIONADO

En la actualidad existen variadas herramientas de visualización de tópicos, entre ellas se encuentra la visualización con el algoritmo LDA de David Minmo (Yao, Mimno, y McCallum, 2009) basada en un sistema de visualización simple con los documentos pero efectiva en cuanto a su objetivo, donde se muestra la distribución de tópicos y la probabilidad de cada uno. Es una interfaz útil al momento de explorar como se ve en la Figura 3.1, sin embargo puede ser compleja a la hora de realizar tareas específicas, como aquellas tareas que implican análisis sobre los documentos, cuyo objetivo se encuentra fuera del ámbito de la exploración (Newman et al., 2010).

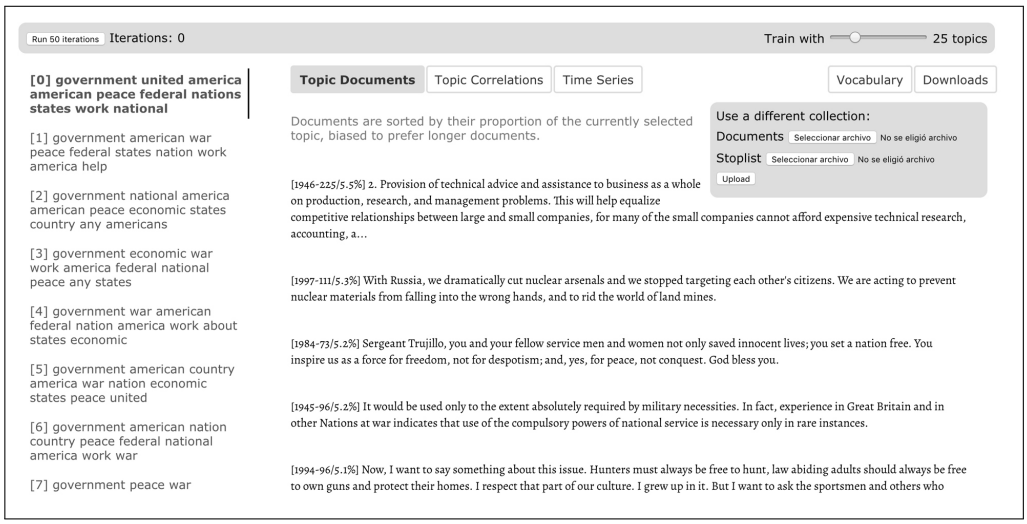


FIGURA 3.1. Interfaz de David Mimno (Yao et al., 2009) para explorar documentos.

Por otra parte, un *encoding* popular para visualizar la relación entre tópicos y las palabras asociadas es Termite (Chuang, Manning, y Heer, 2012), que es básicamente utilizar las ventajas de las matrices tópico-palabra. Otro trabajo popular es LDAvis (Sievert y Shirley, 2014) visualización que parte una vista general pero que permite la exploración en profundidad.

Otros trabajos relacionados, pero con el área de visualizaciones para crímenes, se encuentra una herramienta de visualización desarrollada bajo la autoridad del estado de Illinois sobre información criminal de este Estado (Block, 1995), dónde la visualización

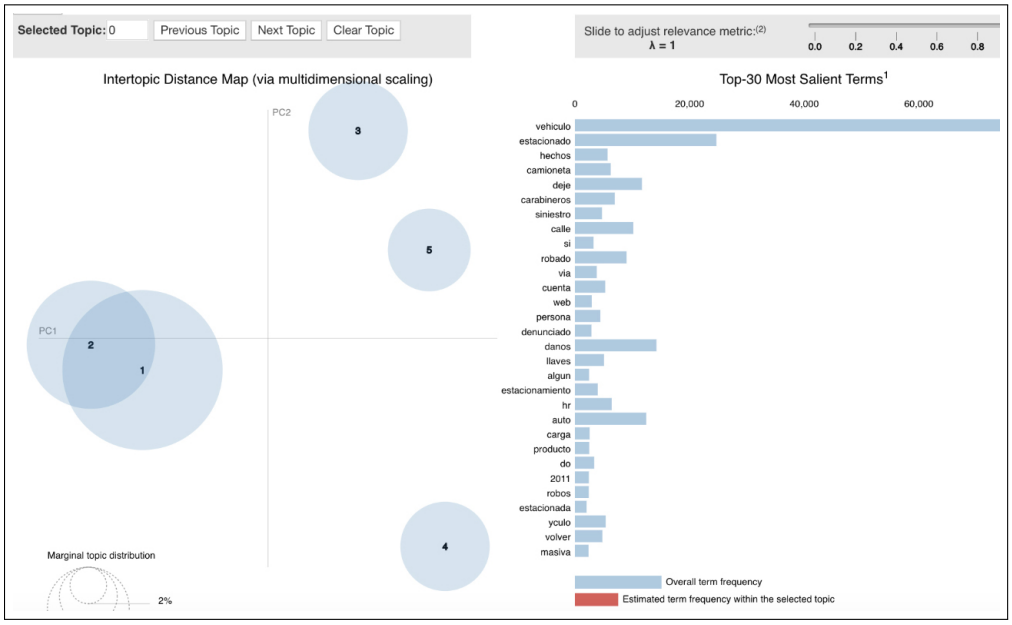


FIGURA 3.2. Ejemplo del encoding de LDAVis (Sievert y Shir-ley, 2014) sobre el set de datos de robo de vehículos.

consta de un mapa con los crímenes como puntos sobre el mismo. Otra visualización en el área de crímenes es CybercrimeIR (Chang, Ku, Wu, y Chiu, 2012), donde utilizan técnicas de análisis de texto como SVM, sin embargo la herramienta esta más enfocada al análisis que la visualización geo-espacial.

Explorando sobre otros *dashboard* de visualización, existe un análisis sobre textos históricos (Christoforidis, Heuwing, y Mandl, 2017) se evaluó el uso de *small-multiples* de *heatmaps* para las tareas más granulares permitían mayor eficiencia. Voila (N. Cao et al., 2018) es probablemente el trabajo más relacionado al implementar un sistema de monitoreo espacio-temporal. Sin embargo este sistema no se concentra en la información textual y en análisis de textos, si no que en estadísticas más cuantitativas y estructuradas.

Capítulo 4. OBJETIVOS

El objetivo de este trabajo es diseñar, implementar y evaluar una herramienta de visualización interactiva que permita agrupar documentos en tópicos utilizando algoritmos de *clustering* probabilístico y apoyando la interacción con información geográfica y temporal. Lo anterior con el fin de ofrecer a los analistas una herramienta para poder inferir y encontrar nuevos patrones en la estructura de los datos existentes. Se utiliza la base de datos de la Asociación de Aseguradoras de Chile (AACh), donde cada siniestro es una entrada en esta base de datos. El siniestro, además, es documentado con diversos datos, entre ellos el relato escrito de cómo ocurrió la situación del robo y el lugar donde ocurrió. Cada uno de estos relatos es un documento que se espera relacionar a la interfaz y serán la base para el modelamiento de tópicos.

Se busca responder a las siguientes preguntas de investigación:

- Sobre el modelo entrenado: (1) ¿LDA puede lograr una categorización con sentido de los documentos?
- Sobre la visualización: (2) ¿Cuál es la mejor forma de resolver tareas que impliquen tendencias en el tiempo? y (3) ¿Cuál es la mejor forma de resolver tareas que impliquen un análisis general agregado?

El objetivo general de este trabajo es que la herramienta desarrollada permita encontrar patrones en la estructura de los datos de forma más eficiente. Sobre la herramienta se plantean tres hipótesis.

H1: El modelamiento de tópicos utilizando herramientas de topic modeling como LDA permitirá encontrar tópicos reales y con significado.

H2: Visualizaciones como los gráficos de línea deberían tener un mejor rendimiento que otro tipo de gráficos, en este caso particular, en comparación a un gráfico de barras para tareas que impliquen tendencias en el tiempo.

H3: Visualizaciones como los gráficos de barras deberían tener un mejor rendimiento que otro tipo de gráficos, en este caso particular, en comparación a un

gráfico de líneas para tareas que impliquen información generalizada o agregada.

Capítulo 5. METODOLOGÍA

El primer paso en el desarrollo de este trabajo fue el procesamiento de los datos y análisis del mismo siguiendo la metodología KDD (*Knowledge Discovery in Databases*), que consta de cinco pasos: Selección de los datos, pre-procesamiento y limpieza, transformación, minería de datos y evaluación. Sobre este último punto, evaluación, se extiende el desarrollo en el siguiente capítulo, donde se trabajó el diseño de la visualización siguiendo el *framework* propuesto por Tamara Munzner.

5.1. Selección y forma de los datos

Se decidió usar la base de datos de robo de vehículos de la AACh (Asociación de Aseguradoras de Chile) debido a la importancia de los patrones que podrían encontrarse, ya que los patrones en robos de vehículos podrían evidenciar modi operandis de las bandas delictuales especializadas a este tipo de atracos. Esta base de datos cuenta con un total de 55.626 entradas, donde cada entrada equivale a un siniestro, vale decir a un vehículo robado. Cada siniestro, además es documentado con diversos datos, alcanzando un tamaño de 70 atributos distintos, especificados en el Anexo A. De todos estos atributos, los que fueron seleccionados para trabajar con ellos son los siguientes:

- `id_prose`: identificador único de cada elemento en la base de datos. Cuenta con un total de 55.626 registros no vacíos.
- `sin_fecha`: fecha del siniestro, incluyendo día, mes y año. Cuenta con un total de 55.626 valores no vacíos.
- `sin_hora_siniestro`: hora del siniestro. Cuenta con un total de 55.626 valores no vacíos.
- `sin_direccion_siniestro`: dirección escrita del siniestro. Cuenta con un total de 46.819 valores no vacíos.
- `reg_descripcion_comuna`: comuna donde ocurrió el siniestro. Cuenta con un total de 55.624 valores no vacíos.

- `marc_desc`: marca del vehículo involucrado. Cuenta con un total de 55.626 valores no vacíos.
- `mod_desc`: modelo del vehículo involucrado. Cuenta con un total de 55.626 valores no vacíos.
- `tvc_desc`: tipo de vehículo según registro civil. Cuenta con un total de 55.626 valores no vacíos.
- `sin_relato`: es el relato escrito sobre cómo ocurrió la situación del robo. Cuenta con un total de 51.451 valores no vacío. En general es texto plano no estructurado registrado por miembros de las aseguradoras de acuerdo a cómo los clientes cuentan el detalle de cómo ocurrió el robo.

TABLA 5.1. Extracto de algunos elementos seleccionados de la base de datos.

id_prose	sin_fecha	sin_hora _siniestro	sin_direccion _siniestro	reg_descripcion _comuna	mar_desc	mod_desc	tve_desc	sin_relato
43563	$19-06-2012$	05:20	BENITO REBOLLADO AL FRENTE DE LA UNIVERSIDAD CATOLICA	Santiago	TOYOTA	YARIS SPORT XLI	AUTOMOVIL	YO DEJE EL VEHICULO A LAS 09:30 AM ESTACIONADO EN LA CALLE BENITO REBOLLADO AL FRENTE DE LA UNIVERSIDAD CATOLICA CAMPUS SAN JOAQUIN EN UN LUGAR HABILITADO PARA ESTACIONAR LUEGO SALGO DE LA UNIVERSIDAD A LAS 17:20 Y NO ENCUENTRO EN ESE MOMENTO DI AVISO AL SEGURIDAD DE LA UNIVERSIDAD PRO QUE ELLOS TIENEN CAMARAS LLAMARON A CARABINEROS LOS QUE LLEGARON HICIMOS LA DENUNCIA Y LUEGO ME DIRIGI AL CONTROL DE CAMARA DE SEGURIDAD DE LA UNIVERSIDAD AL REVISAR EL VIDEO PUDIMOS VER CUANDO LO HABIAN ROBADO
70770	$19-06-2012$	07:10	LOS PIRINEOS PROVIDENCIA SANTIAGO METROPOLITANA	Providencia	NISSAN	D22 TERRANO PICK UP AX	CAMIONETA	DEJE EL VEHICULO ESTACIONADO Y AL VOLVER NO ESTABA
76553	$29-09-2012$	07:10	EL TRANQUE/JOSE ALCALDE DELANO LO BARNECHEA	Providencia	SUBARU	NEW IMPREZA 1.5R 5D AWD2A	AUTOMOVIL	ESTABA ESPERANDO LA LUZ VERDE DEL SEMAFORO CUANDO SE ME CRUZO UN VH PÓR DELANTE Y OTRO VH ME CERRO POR EL LADO LUEGO SE BAJO UN INDIVIDUO ME ROMPIO EL VIDRIO ME AMENAZO Y ME TUVE QUE BAJAR DEL VH Y SE DIERON A LA FUGA CON MI VH

Una vez seleccionados los datos a utilizar, se prosiguió a revisar la relevancia de los datos. Con la fecha de los relatos se pudo determinar el horizonte de datos con valores

válidos. Como se puede ver la Figura 5.1, a partir del 2011 la cantidad de siniestros diaria es realmente considerable y comparables entre sí. Esto se debe a factores externos, que tienen que ver con la forma en que la aseguradora almacenaba sus datos antes de 2011. Por otra parte, no se posee suficiente información durante el año 2017, por lo que los pocos datos de ese año no son relevantes para poder analizarlos. Por lo anterior no se considerarán datos fuera de los años 2011 y 2016, reduciendo así la cantidad de datos a 51.967 ($6,6\,\%$ de reducción sobre el set de datos original).

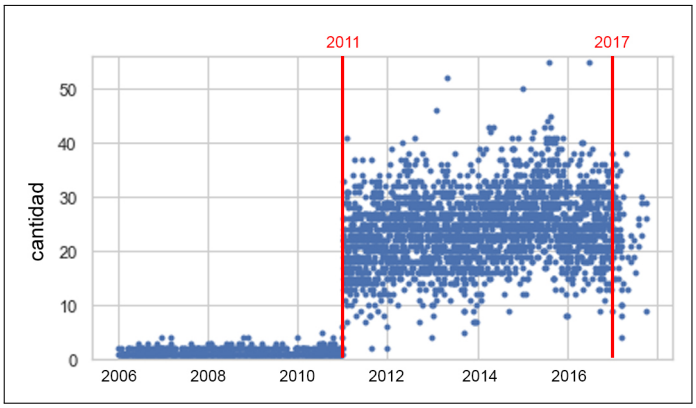


FIGURA 5.1. Scatterplot con la cantidad de robos por día según la base de datos de la AACh.

5.2. Pre-procesamiento y limpieza de los datos

Debido a que la gran mayoría de los datos son hechos reales, cuya información relevante para este trabajo son los relatos, se removieron solo aquellas entradas cuya información no permitiese el análisis de tópicos, vale decir, aquellos registros cuyos relatos de los siniestros estuviesen vacíos. Con lo anterior se redujo el tamaño del set de datos a 21.782 entradas, lo que implica una reducción del $58,1\,\%$ sobre el set de datos seleccionado. A pesar de la reducción en la cantidad de los datos, los datos restantes poseen mayor confiabilidad respecto al original, debido a que se tiene la información completa sobre el tipo de robo. Al realizar el análisis inicial del modelo de datos se evaluaron otros aspectos, que se desvían del foco de este trabajo. Estos se presentan en el Anexo B.

5.3. Transformación

Para poder trabajar los datos de forma eficiente, se debieron transformar dos de los atributos del set de datos original, que son el relato y la dirección descrita del lugar del siniestro.

5.3.1. Transformación de `sin_relato`

El texto original como se puede ver en el ejemplo de la Tabla 5.1 no está estructurado, posee faltas ortográficas, y además posee códigos internos o abreviaciones propias (tales como *VH* para vehículo). Antes de trabajar con el texto se seleccionó un conjunto aleatorio de 50 documentos (aproximadamente un $0,1\,\%$ del tamaño del set de datos), de los cuales se identificaron de forma manual las palabras con falta de ortografía y también las abreviaciones más comunes del documento. Con lo anterior se corrigió parte de las faltas ortográficas más evidentes, de forma manual. En esta revisión manual, también se identificaron códigos internos usados por las aseguradoras. Estos términos (como .ªlcohemiano") decidieron dejarse debido al significado que este tipo de palabras tienen por si mismas. Por otra parte horas como por ejemplo $4:00$ o $3hrs$ se removieron por ser información redundante, pero se mantuvieron números solos, dado que en la muestra hacían referencia a grupos, por ejemplo "4 personas.º "2 individuos". Luego uno de los primeros pasos para transformar el texto fue aplicar *stemming* al texto utilizando herramientas para el procesamiento de lenguaje natural de la librería NLTK Toolkit (Bird y Loper, 2004). *Stemming* hace referencia al proceso de convertir las palabras a su raíz, removiendo los sufijos, conjugaciones y plurales del final de las palabras. El *Stemmer* utilizado fue SnowballStemmer desarrollado y mantenido por Yoshiki Shibukawa (Shibukawa, 2009). El segundo paso, utilizando la misma herramienta mencionada, fue remover las *stopwords*. En este paso se eliminan todas las palabras que no aportan valor al significado del texto, como lo son los artículos (el, la, un, unos, etc.-). Se removió también la puntuación y los acentos para homogeneizar las palabras. Finalmente se aplicó

lemmatization, vale decir reemplazar las palabras por su forma más aceptada en el diccionario. De esta transformación se generó el atributo *relato*, que contiene el relato del siniestro limpio para poder ser procesado.

5.3.2. Transformación de `sin_direccion_siniestro`

Para transformar el texto escrito en coordenadas se utilizó el modelo descrito por (Quezada, Peña Araya, y Poblete, 2015). Se utilizó la API de Google para realizar la búsqueda de las direcciones y así obtener su latitud y longitud. Del resultado de la búsqueda, existió un total de 17.630 direcciones encontradas (reducción del $20,1\,\%$ sobre el set de datos procesado). Se realizó un *sampling* con 100 de los datos donde se obtuvo un nivel de acierto del $75\,\%$. De esta transformación se generaron los atributos *latitud* y *longitud*.

5.4. Modelamiento de tópicos

Luego de tener los datos preparados para procesarlos, se procede a determinar los valores para entrenar el modelo de tópicos de LDA sobre el set de datos. El primer punto es determinar la cantidad de tópicos. Para definir la cantidad de tópicos óptima que deben asignarse a cada set de datos, se utilizó el algoritmo lda-tuning de Nikita Murzintcev (Nikita, 2016), que usa las métricas de Griffiths (Griffiths, Jordan, Tenenbaum, y Blei, 2004), Cao et. al. (J. Cao, Xia, Li, Zhang, y Tang, 2009) y Deveaud (Deveaud, SanJuan, y Bellot, 2014). Con lo anterior se determinó que el número óptimo de tópicos para este set de datos es de 4 tópicos según la Figura 5.2.

Para entrenar el modelo se uso Gensim (Řehůřek y Sojka, 2010), una librería especializada en Python para realizar modelamiento de tópicos. Luego de determinar el diccionario para el set de datos y definir las matrices termino-documento y palabra-documento se entrenó de distintas formas el modelo variando los parámetros. Los parámetros finalmente escogidos para entrenar el modelo fueron los especificados en la Tabla 5.2.

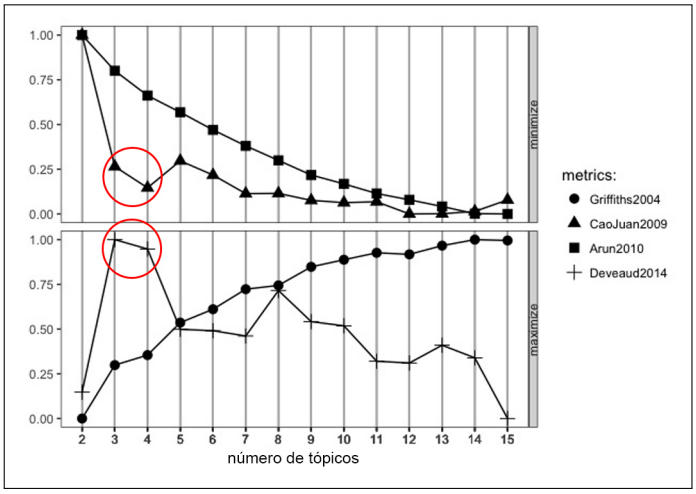


FIGURA 5.2. Cantidad óptima de tópicos según las distintas métricas que existen para determinar la cantidad.

TABLA 5.2. Parámetros considerados para el entrenamiento del modelo.

Parámetro	Valor	Explicación
chunksize	3000	Número de documento cargados en memoria cada momento y procesados en .Estep"de EM.
passes	20	Número de veces que se itera sobre todo el documento.
update_every	1	Número de chunks a procesar antes de moverse al "M step" de EM.
alpha	symmetric	Arreglo de 1 dimensión que representa la relación entre sparsity/uniformity. Al asumirse simétrica se le da más peso a los tópicos con palabras menos comunes en el corpus de documentos.
decay	0.5	Número que da peso al porcentaje previo del valor lambda que se perdió en comparación a la iteración anterior. Corresponde a Kappa (Hoffman, Blei, y Bach, 2010).
offset	1	Hiperparámetro que controla que tanto se debe demorar los primeros pasos de las primeras iteraciones.
eval_every	10	Cada cuántas actualizaciones se re-calcula el Log perplexity.
gamma_threshold	0.001	Cambio mínimo en el valor de gamma en cada iteración.
minimum_probability	0.01	Filtro para la probabilidad mínima de un tópico.

Lo anterior permitió entrenar el modelo, cuyo tiempo de entrenamiento en promedio fue de 15 minutos. Uno de los *output* del entrenamiento, además de obtener la probabilidad de cada documento de pertenecer a cada tópico, fueron las probabilidades de las palabras de pertenecer a cada documento. Esta última se ve reflejada en la Tabla 5.3. En

el capítulo de Resultados se muestra la interpretación de los usuarios y validación de la distribución generada por el modelo. Con los resultados del entrenamiento, se adjuntó al set de datos las columnas de probabilidad por tópico. Con la información extra añadida a la base de datos, se trabajó en la evaluación del resto del trabajo. Los atributos importantes en el set de datos generado son: las coordenadas geográficas (latitud, longitud), la comuna, el relato, la fecha y finalmente la probabilidad de pertenencia a cada uno de los 4 tópicos.

TABLA 5.3. Para cada tópico la probabilidad de aparición de las primeras 20 palabras.

Tópico 1	**Tópico 2**	**Tópico 3**	**Tópico 4**
robo (3,9 %)	auto (5,4 %)	vehículo (9,9 %)	vehículo (6,6 %)
total (3,3 %)	vehículo (4,1 %)	estacionado (5 %)	auto (2 %)
daños (2,6 %)	si (1,9 %)	deje (2,2 %)	dos (1 %)
circunstancia (1,9 %)	algún (1,5 %)	robado (1,6 %)	personas (1 %)
señalización (1,9 %)	llaves (1,3 %)	robo (1,3 %)	llaves (0,9 %)
presencia (1,9 %)	camioneta (1,3 %)	lugar (1,3 %)	casa (0,8 %)
4 (1,8 %)	estacionamiento (1,2 %)	automóvil (1,2 %)	roban (0,8 %)
alcohemiano (1,7 %)	robado (1,1 %)	auto (1,1 %)	llevaron (0,8 %)
carabinerosno (1,6 %)	cuenta (1,1 %)	encontraba (1,1 %)	calle (0,7 %)
hechos (1,1 %)	todas (0,9 %)	calle (1,1 %)	tipos (0,7 %)
vehículo (1 %)	seguridad (0,8 %)	hrs (1 %)	arma (0,7 %)
chapa (0,9 %)	comprobante (0,8 %)	domicilio (1 %)	iba (0,7 %)
externo (0,8 %)	carabineros (0,7 %)	volver (1 %)	domicilio (0,7 %)
puerta (0,8 %)	mecanismo (0,7 %)	carabineros (0,9 %)	pistola (0,6 %)
daños (0,7 %)	camion (0,6 %)	dia (0,9 %)	fuego (0,6 %)
web (0,6 %)	asegurado (0,6 %)	percato (0,8 %)	individuos (0,6 %)
call (0,6 %)	dia (0,6 %)	casa (0,8 %)	robo (0,6 %)
denunciado (0,6 %)	empresa (0,6 %)	había (0,8 %)	camioneta (0,6 %)
delantero (0,6 %)	denuncia (0,5 %)	salir (0,8 %)	bajan (0,6 %)
trasero (0,6 %)	patente (0,4 %)	dejo (0,1 %)	3 (0,5 %)

5.5. Evaluación preliminar

Se realizó una evaluación preliminar con los datos generados del modelo. Con el fin de explorar los resultados se creo una interfaz la cual se denominó ITACaT (*Interactive Tool for Analysis of Car Theft*). El diseño de la herramienta esta dividido en 4 componentes como se aprecia en la Figura 5.3. El componente principal (sección A) se establece

como un mapa que muestra los documentos. Cada documento es un punto en el mapa con un color según el tema asignado. La asignación a un tema se decide si el documento tiene más del 35 % de probabilidad de estar asignado a algún tópico. El mapa tiene una pequeña leyenda para relacionar el documento con el color respectivo del tópico asignado. Acerca de la interacción, cuando el analista interactúa con esta sección, puede ampliar y ver los documentos como puntos en un área geográfica. Se puede inspeccionar en detalle haciendo *click* en el documento de denuncia representado por cada círculo, donde aparecerá la información relacionada. La sección B está oculta por defecto, el usuario debe interactuar con el tópico deseado para ver esta vista. La vista detallada muestra la distribución como un mapa de calor, dada la intensidad de acuerdo con el porcentaje de relación con el tópico. Además, proporciona un gráfico de barras sobre las palabras más representativas del tema y la relación con el tópico. La sección C permite filtrar los documentos por diferentes atributos, entre ellos el año. La sección D ofrece una explicación resumida sobre el tópico seleccionado. Incluye el total de documentos asignados a estos temas y las 10 primeras palabras relacionadas, en orden de importancia.

La interfaz permitió validar el modelo y los parámetros finales del entrenamiento, debido a que se iteró sobre ella hasta que el resultado de la exploración tuviese sentido. Por otra parte también permitió dar cuenta de los problemas al intentar realizar tareas específicas sobre los datos. Uno de los principales problemas es que al intentar resolver tareas relacionadas a las tendencias o agregación de datos, es imposible de resolver con la vista de la sección A. Por otra parte, la sección B dificulta interpretar el movimiento de los tópicos, al verse el mapa de calor para cada tópico como una masa unida, como se puede apreciar en la Figura 5.4. Este último punto es el fundamental en las decisiones de diseño que se verán más adelante, pues en conjunto a lo visto en el trabajo relacionado, una división espacial con el fin de agregar los datos permitiría mejorar la forma en que se resuelven tareas. Tal como el plan cuadrante de carabineros de Chile divide Santiago por cuadrante asociado a cada comisaría, se pueden usar las divisiones ya existentes en Santiago, que son las comunas de la capital, para agregar la información.

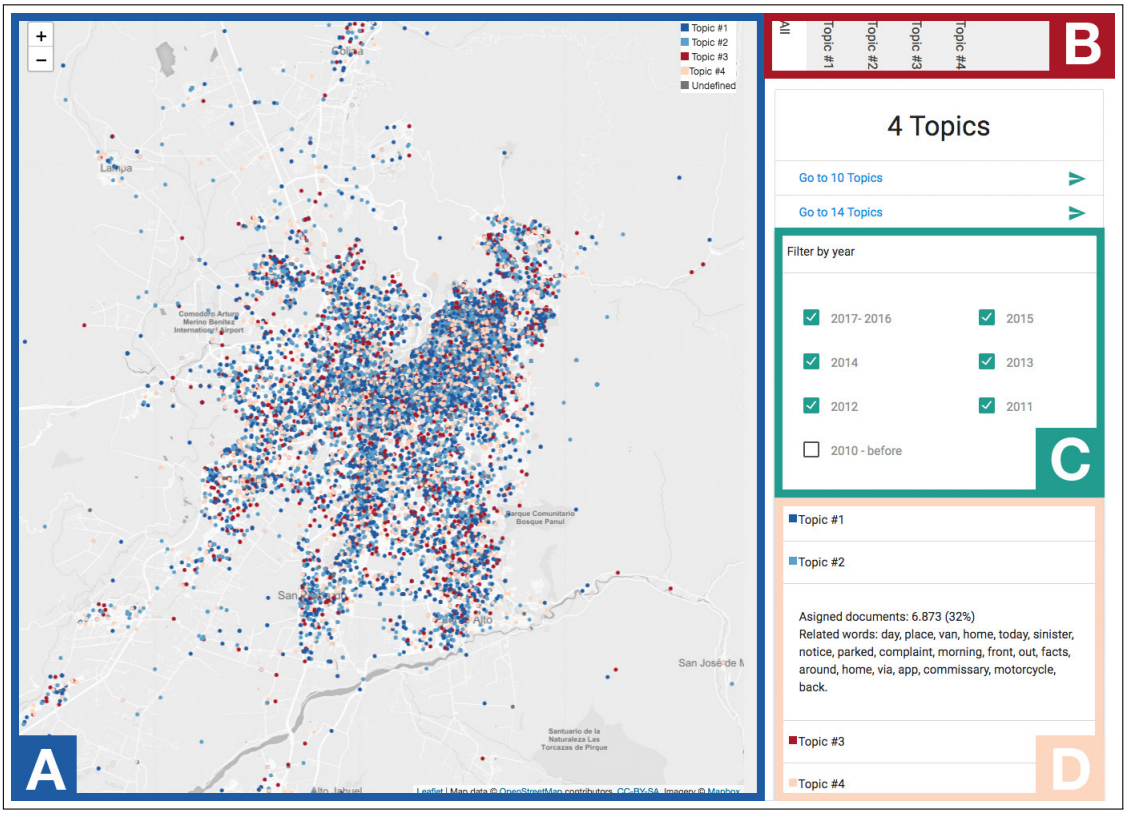


FIGURA 5.3. Interfaz ITACaT dividida por sus 4 secciones principales.

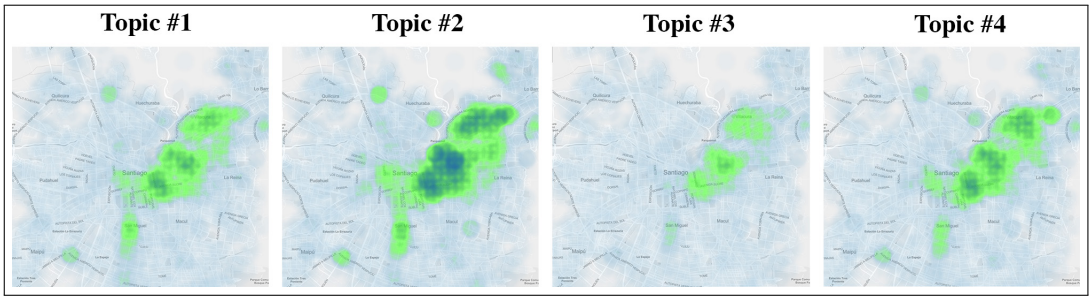


FIGURA 5.4. Mapa de calor de los tópicos sobre santiago según la visualización de ITACaT.

Con el fin de evaluar las hipótesis planteadas, los siguientes pasos son diseñar e implementar herramientas que permitan la validación de los puntos mencionados. Según la sección 4, para **H1** se evaluará con una encuesta a los usuarios, mostrando la distribución de palabras entregada por el modelo y validando que la interpretación de los usuarios

sobre el tópico, es coherente con el contenido de los documentos con mayor probabilidad de pertenecer al tópico. Para **H2** y **H3** por otra parte se evaluará con el diseño e implementación de las herramientas, y será validado a través de un estudio de usuario.

Capítulo 6. DISEÑO E IMPLEMENTACIÓN

Para tomar las decisiones de diseño se siguió el Framework de Tamara Munzner explicado en el Capítulo 2. Al justificar las decisiones por capa, el final del análisis se concluye resumiendo al responder las preguntas fundamentales: ¿Por qué?, ¿Qué? y ¿Cómo?.

6.1. Dominio o situación

Los datos seleccionados como ya se han mencionados son los relacionados a robos de vehículos y la distribución de probabilidad de cada tópico. En base a lo interior el interés principal es identificar patrones interesantes en los robos de vehículos, a partir de los modi operandi de los delitos. El usuario necesita una forma eficiente de comparar robos y tópicos de los robos entre zonas geográficas, como las comunas, además se deben identificar las tendencias en los tópicos ya sea de forma espacial o temporal, esto es identificar que tópicos han aumentado y disminuido en el tiempo, o bien en comparación a otras comunas. Finalmente la interfaz debe también dar información agregada sobre cómo se comporta en general cada sector.

6.2. Tareas y abstracción de datos

En base a las necesidades del usuario, se espera que las acciones que permita la visualización sean:

- Analizar y consumir los datos: permitiendo descubrir patrones con el objetivo de comprender los atributos, que en este caso serían los tópicos. Además se espera que se presente también la información geográfica con el fin de comprender la distribución de los robos por cada sector geográfico.
- Consultar los datos: permitiendo identificar y comparar los datos con el objetivo de encontrar tendencias.

Los datos se encuentran actualmente en formato relacional, vale decir, cada siniestro se encuentra tabulados por sus atributos: coordenadas, relato, fecha y probabilidad de

pertenencia a cada uno de los 4 tópicos. Esta forma inicial permite una buena agregación de los datos, así como transformaciones simples como agrupaciones por año o por tópico de los documentos.

6.3. Idioma

Para lograr el análisis y consumo de los datos se definirá el idioma a utilizar por la visualización con el fin de atender a las tareas descritas con anterioridad.

6.3.1. Visual encodings

- Alternativas de *encodings*: existe una gran cantidad *encodings* visuales, sin embargo dependiendo del tipo de dato y la necesidad, algunos son más relevantes de estudiar que otros. A continuación se revisan algunas de las alternativas existentes.
  1. *Choropleth Map*: Originalmente se contaba con la información geográfica como eventos (información ordinal de las coordenadas). Sin embargo como se planteó en el Capítulo 5 con la evaluación preliminar, este tipo de *encoding*, mapas de puntos, es útil para explorar, pero al tener tareas de comparación es difícil determinar por ejemplo ¨qué comuna es la que tiene más robos¨, si no se puede observar de forma agregada para cada comuna. En este caso resulta conveniente reorganizar la información y agruparla dependiendo del área geográfica en que se encuentra, y así poder usar un *encoding* como el *choropleth map* de la Figura 6.1, cuya información ordinal es por área. En el *choropleth map* presentado las marcas son las comunas pertenecientes a la conurbación de Santiago, excluyendo a las comunas satélite, obteniendo así un total de 35 áreas representando las comunas que se pueden ver en el Anexo C. Por otra parte, los canales escogidos por el *choropleth map* son los colores, para presentar la cantidad de robos por comuna, la paleta escogida es lineal y secuencial para atributos ordenados, donde la escala ocurre

por la saturación y luminosidad del color. La paleta escogida finalmente fue estudiada con la herramienta Color Brewer (Harrower y Brewer, 2003).

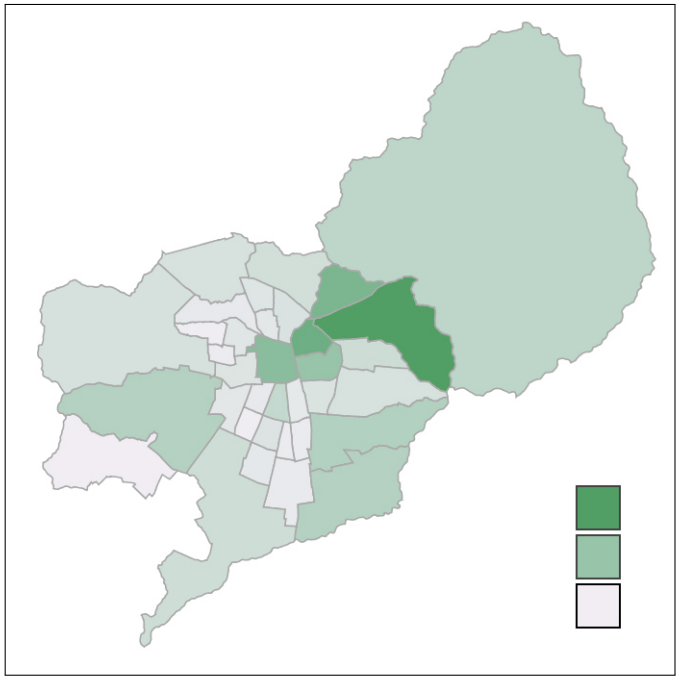


FIGURA 6.1. Choropleth encoding sobre la conurbación de Santiago, donde el encoding del color implica la cantidad de robos para esa comuna.

2. Gráfico de barra y gráfico de línea: debido a las tareas de comparación en tendencias e información agregada los gráficos de barra y de línea tienden a ser una buena opción a priori para este tipo de tareas (Saket, Endert, y Demiralp, 2018). Dado que se tiene la probabilidad de pertenencia a cada tópico para todos los documentos y la información geográfica por comuna, una forma de sintetizar o abstraer estos datos es agrupándolos por comuna. Frente a este es necesario hacer una disminución de la dimensionalidad de los datos manteniendo las probabilidades de pertenencia a los tópicos. La forma de disminuir la dimensionalidad agrupando los relatos por comuna y agregándolos por promedio. Luego las probabilidades de los tópicos se normalizan para seguir manteniendo la suma de las probabilidades igual a 1.

Con esta disminución de la dimensionalidad, se pueden comparar distribuciones por comuna para cada año distinto. En la Figura 6.2 se puede ver el resultado de esta disminución y visualización respectiva.

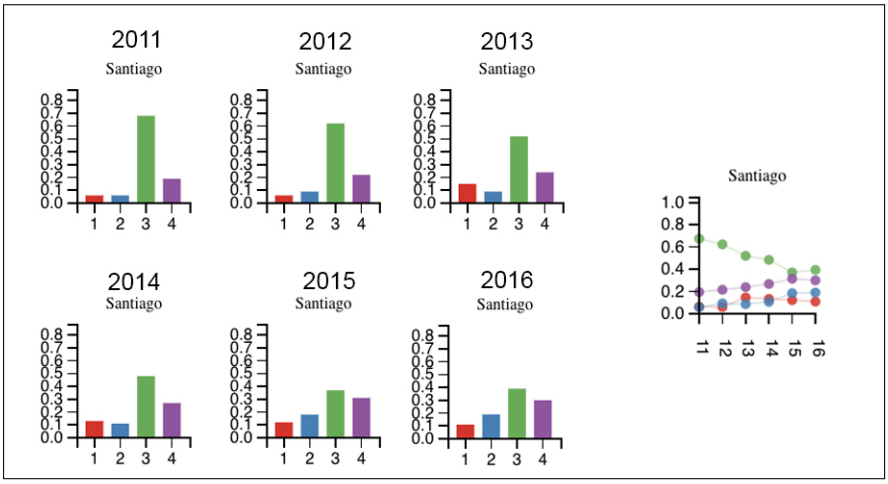


FIGURA 6.2. Distribución de tópicos para la comuna de Santiago entre los años 2011 a 2016, en la izquierda como gráfico de barra y a la derecha como gráfico de línea.

- *Facet*: *facet* en la literatura de visualización es entendido como el conjunto de decisiones de división de la visualización ya sea en múltiples vistas o capas (Munzner, 2009). Un tipo de *facet* son los *small multiples*. Los *small multiples* permiten la exploración de forma más eficiente con menos pasos sobre grandes cantidades de datos. Así mismo permiten una buena perspectiva de los datos (van den Elzen y van Wijk, 2013) así como una forma de agrupar de forma más eficiente la información en vez de sobre cargar un solo *encoding* (Heer, Bostock, Ogievetsky, et al., 2010).

Finalmente las interacciones estudiadas son la relación entre comuna y gráfico, tal que al realizar *hover* (vale decir, exaltación del elemento al pasar el cursor sobre este) sobre una comuna el color de esta cambie como también aparezca el nombre de la comuna. Para facilitar la comparación, también se puede seleccionar una comuna y dejarla marcada. La interacción sobre los gráficos también incluye *hover* sobre las barras o líneas

con el fin de ver la probabilidad exacta asociada. Finalmente, para el caso del gráficos de barra debido a que, a diferencia del gráfico de línea, para mostrar los años se necesita N veces más espacio (N según la cantidad de años a mostrar), se añadió a la interfaz con gráficos de barra un filtro por año.

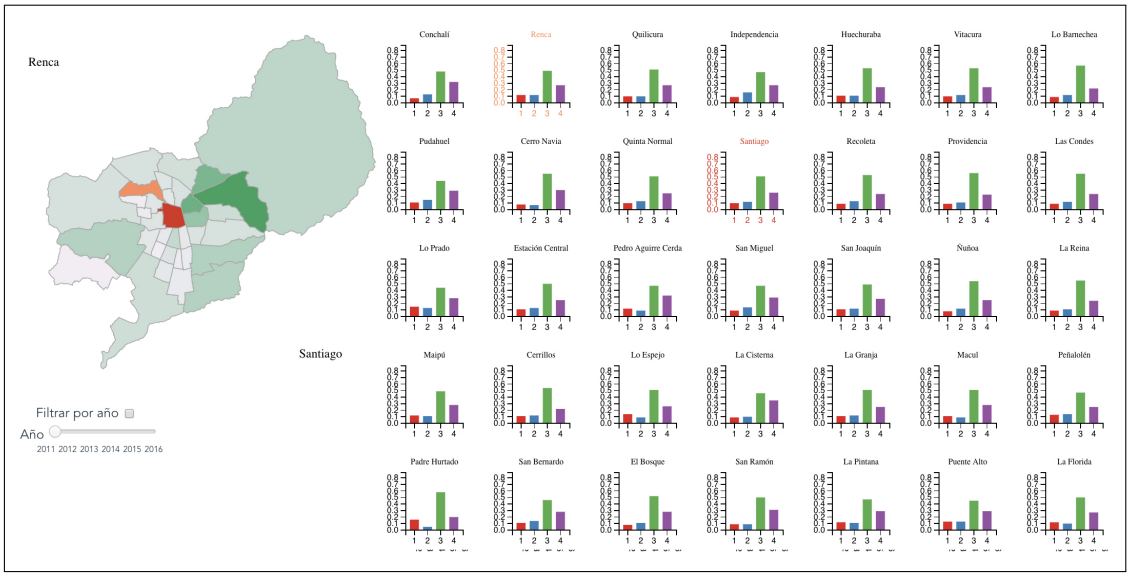


FIGURA 6.3. Implementación de la visualización descrita con small multiples de gráficos de barra.

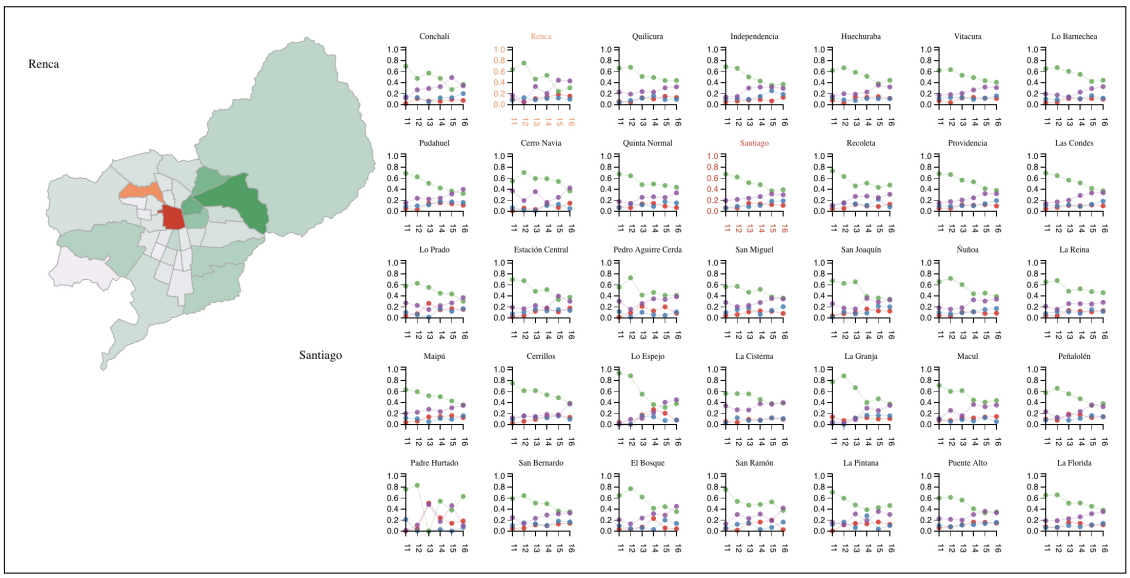


FIGURA 6.4. Implementación de la visualización descrita con small multiples del gráfico de líneas.

6.4. Implementación

Finalmente, a modo de resumen se presenta la respuesta a cada una de las preguntas fundamentales del *framework*:

¿Por qué?: La visualización debe permitir al usuario comparar tópicos por zonas geográficas de forma global, comparar tendencias de los tópicos en el tiempo e identificar los tópicos asociados a cada zona geográfica.

¿Qué?: se deben en un inicio visualizar los datos de crímenes de vehículos que incluyen datos no estructurados como el texto pero que han sido procesados en probabilidades de acuerdo al modelamiento de tópicos, datos geográficos y datos temporales. Debido a que la gran cantidad de delitos se concentran en Santiago, serán localizados solo en esta área.

¿Cómo?: Vamos a usar distintos canales, partiendo por la información geográfica cuya marca será la vista en un mapa delimitado por áreas geográficas que en esta caso serán las comunas. En el mapa los *encoding* serán en primer lugar el color con una escala lineal secuencial con el fin de demostrar cuales comunas se cometen más delitos. Por otra parte por cada comuna existirá un gráfico de barras o un gráfico de líneas.

Con lo anterior ya definido, se realizó la implementación del sistema descrito. Las herramientas utilizadas fueron principalmente D3.js.

Capítulo 7. ESTUDIO DE USUARIO

Con el fin de validar el diseño de la implementación según el *framework* seguido en la Figura 7.4 y probar las hipótesis de este trabajo, se realizó un estudio de usuario. Dado que el fin es probar las hipótesis mencionadas en el Capítulo 4, el estudio de laboratorio tuvo el fin de medir el tiempo por interfaz así como los errores, y así determinar que interfaz permitió a los usuarios llevar a cabo las tareas especificadas. Debido a que el objetivo es comparar las dos interfaces, se determinó el número de usuarios de prueba utilizando el software GPower con el cual se realizó el Power Analysis (Susanne, Edgar, Buchner, y Faul, 2007) utilizando los parámetros especificados en la Figura 7.1.

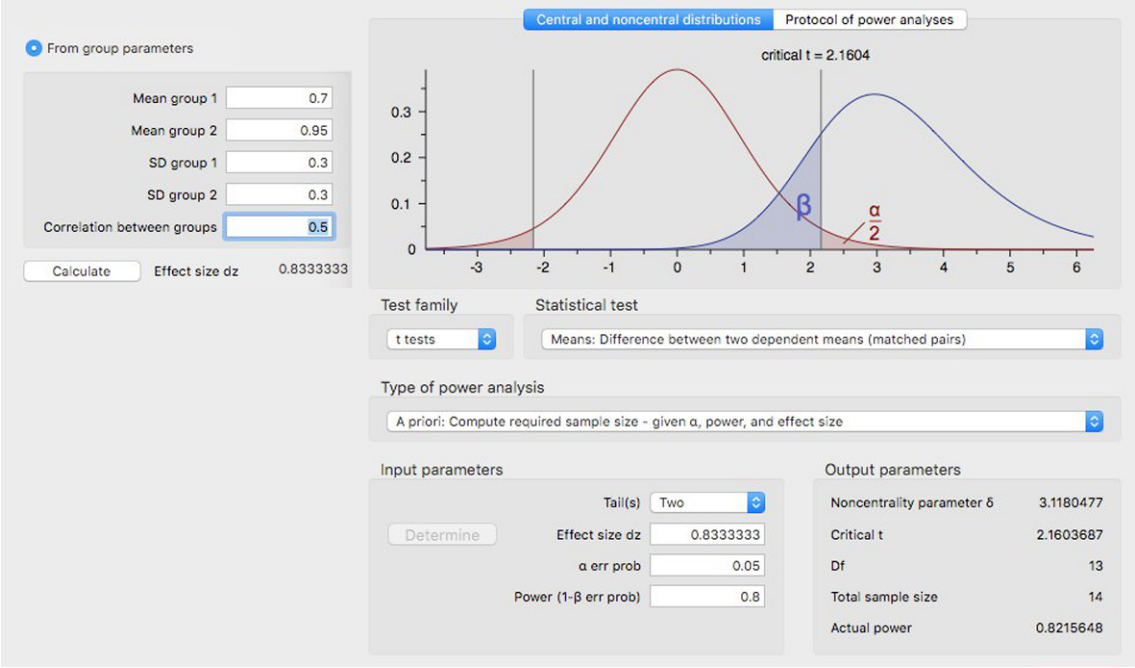


FIGURA 7.1. Parámetros utilizados en GPower para realizar el Power Analysis.

Para el análisis se basó en la hipótesis tal que nuestra expectativa es que una interfaz se comporte mejor que la otra considerando una tarea de comparación estándar. Esperamos que la proporción de respuestas correctas sea de $0{,}7$ a $0{,}95$ (la media de cada grupo) con una desviación estándar el $0{,}3$ cada uno, y una correlación de $0{,}5$.

Por otra parte los t-test a realizarse serían *matched pairs*, con una cantidad de 2 colas, a un nivel $\alpha = 0{,}05$ de significancia y un poder estadístico de $0{,}8$. En base a esto el tamaño de efecto dz esperado sería de $0{,}833$. Todos estos valores fueron el *input* del *power analysis*, con el cual se obtuvo que el tamaño mínimo del grupo a estudiar sería de 14 personas. Para aumentar el poder estadístico según el gráfico de la Figura 7.2 a un $95\,\%$ por ejemplo, el tamaño de la muestra sería de aproximadamente 20 personas.

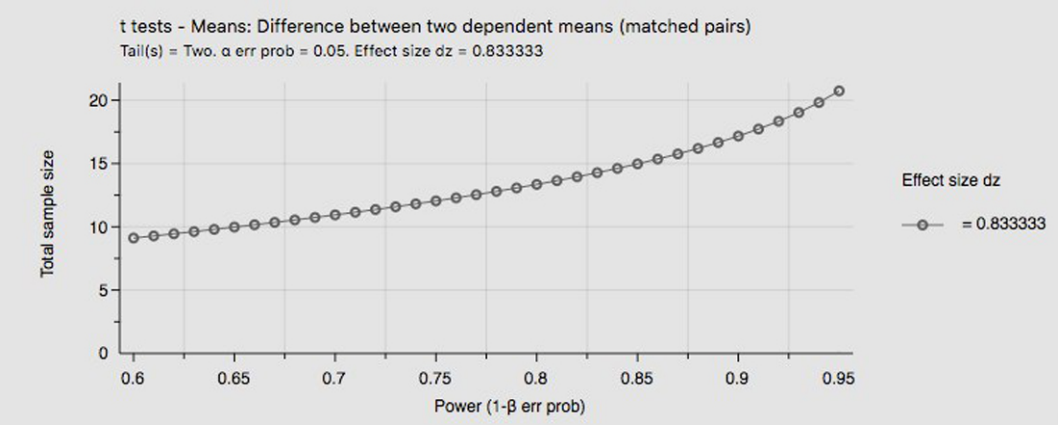


FIGURA 7.2. Poder estadístico de acuerdo a distintos tamaños de muestra por GPower.

Se seleccionó un grupo de 22 personas, estudiantes universitarios de ingeniería computación, eléctrica, hidráulica e industrial entre los 19 y 26 años. El estudio se realizó a través de una página web adaptada para la resolución de las tareas como se ve en la Figura 7.3. Se adaptaron 2 páginas cuya única diferencia fue si empezaban con el *encoding* de un gráfico de barras o de líneas, construidas en Vue.js para facilitar la interacción y la captura de información. Ambas interfaces tenían en común los elementos de la Figura 7.3, en (1) se encuentra la tarea actual, en (2) se encuentra el espacio destinado a responder y en (3) podían dar *feedback* o presentar problemas con el software de Hotjar. Por último cada página capturaba por cada usuario el número de *clicks* y el tiempo entre que una persona entraba a la página de una tarea y en que presionaba terminar tarea, todos estos datos asociados al usuario.

Se describe el procedimiento del estudio a continuación:

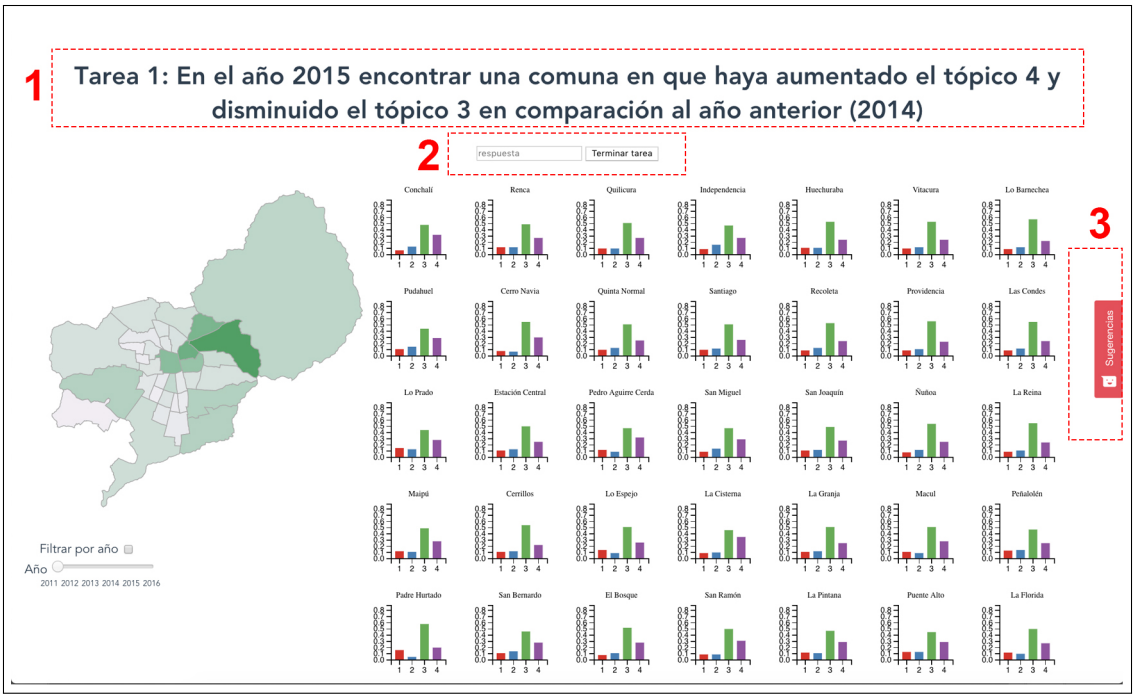


FIGURA 7.3. Interfaz del estudio de usuario. En (1) se describe la tarea actual, en (2) el usuario responde, y en (3) el usuario registra sus dudas o comentarios.

1. Motivación del estudio y el contexto: Se les explicó el contexto a los usuarios según el Anexo D con detalles sobre el objetivo de la implementación.
2. Evaluación inicial de los tópicos: Se le presentaron a los usuarios los tópicos del modelo entrenado y sus palabras asociadas. La primera encuesta que debieron responder fue su interpretación de cada uno de los tópicos.
3. Uso de la interfaz: Lo siguiente fue explicarles la interfaz a utilizar. Se les explicó que en total debían responder 12 preguntas, 6 asociadas a una visualización de gráfico de barras y 6 visualizaciones de gráfico de líneas. Todas las tareas fueron distintas, sin embargo los objetivos del tipo de tarea se repitieron para poder comparar las interfaces. Las tareas se realizaron por interfaz, vale decir, cada usuario partió analizando una de las dos interfaces, que implicó resolver 6 tareas seguidas, de dificultad incremental, sobre una interfaz. Se controló que se repartieran de forma pareja la cantidad de usuarios que partía evaluando el la interfaz de gráfico de barras y los que partieron evaluando la interfaz de gráfico de líneas.

4. Resolución de tareas: Los usuarios debieron responder un total de 6 tareas que apuntan a probar la efectividad de la interfaz desarrollada y a su vez comparar cuál tipo de small multiple, si el de líneas o barras se desempeña mejor y en qué tipo de tareas. En la Tabla 7.1 se describen las tareas de la interfaz.
5. Encuesta de carga cognitiva: Luego de realizar las 6 tareas de la interfaz que le tocó a cada usuario, debieron responder el test de carga cognitiva NASA-TLX antes de avanzar a la siguiente interfaz.
6. Probar segunda interfaz: se repiten los dos puntos anteriores, pero con la interfaz que no les tocó en un inicio.

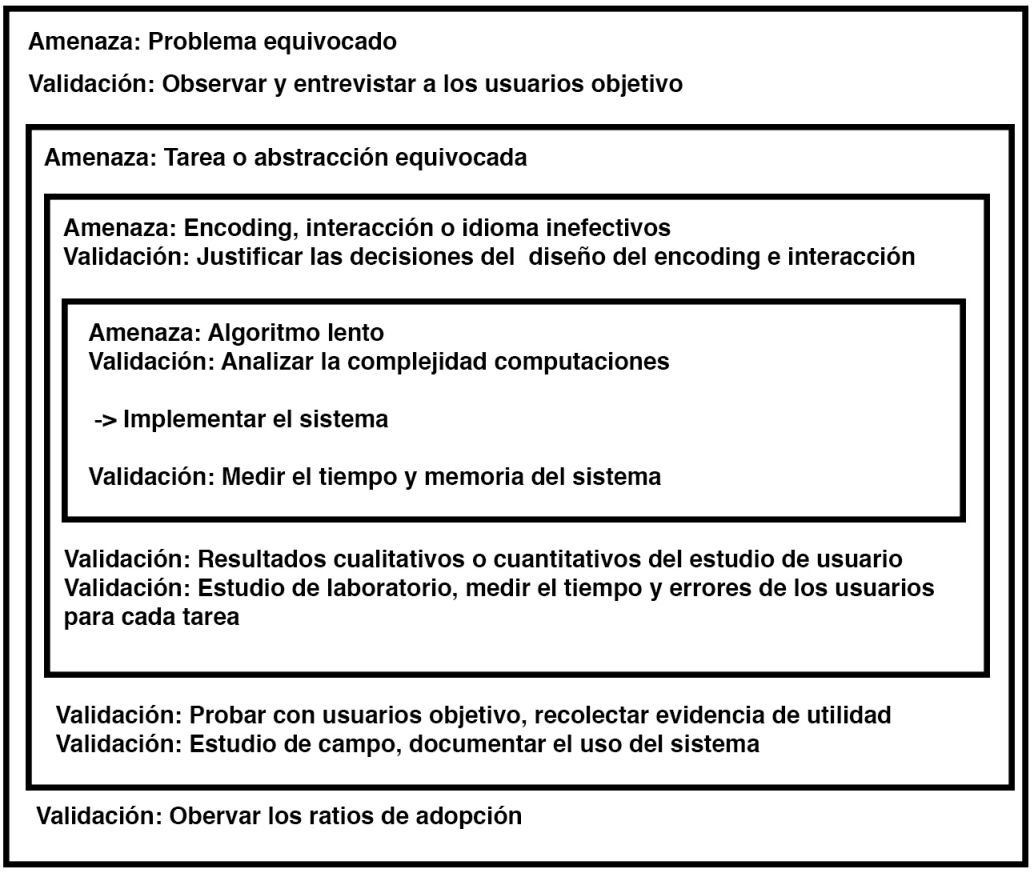


FIGURA 7.4. Validaciones según el Framework de Tamara Munzner (2009).

TABLA 7.1. Descripción de las tareas utilizadas en el estudio de usuario.

Número Tarea	Objetivo	Descripción para gráfico de barras	Descripción para gráfico de líneas
1	Comparación, tendencias en los tópicos, uso de información temporal.	En el año 2015 encontrar una comuna en que haya aumentado el tópico 4 y disminuido el tópico 3 en comparación al año anterior (2014).	En el año 2013 encontrar una comuna en que haya aumentado el tópico 1 y disminuido el tópico 4 en comparación al año anterior (2012).
2	Comparación, cambios en el tiempo, tendencias en el tiempo.	Encontrar una comuna en que para todos los años el tópico 1 sea menor al tópico 2.	Encontrar una comuna en que para todos los años excluyendo el 2011 el tópico 2 sea menor al tópico 1.
3	Comparación, agregación, uso de información geo-espacial.	Identificar una comuna que en comparación a Santiago, tenga en promedio mayor probabilidad del tópico 1.	Identificar una comuna que en comparación a Santiago, tenga en promedio mayor probabilidad del tópico 4.
4	Identificación de outliers, uso de información temporal.	En el año en 2014 identificar una comuna en que el tópico 2 sea el segundo mayor en relevancia.	En el año en 2015 identificar una comuna en que el tópico 4 sea el mayor en relevancia.
5	Identificar distribuciones, diferencias entre los datos.	Identifique las 2 comunas que tengan el patrón (distribución) de tópicos más distinto.	Identifique las 2 comunas que tengan el patrón (distribución) de tópicos más distinto.
6	Relacionar con información externa, información agregada.	Cual es la comuna más relacionada al tópico de delitos de armas.	Cual es la comuna más relacionada al tópico de delitos con daños.

Los resultados se midieron en el comportamiento del usuario (cantidad de *clicks* y segundos) y rendimiento (porcentaje de respuestas correctas) como se muestra en la Tabla 8.1, y en el gráfico correspondiente en la Figura 8.1, donde la *Interface* 1 corresponde al gráfico de barras e *Interface* 2 corresponde al gráfico de líneas.

TABLA 8.1. Resultados del rendimiento de las pruebas de usuario. Cantidades promedio y en paréntesis la desviación estándar.

Tarea	Clicks promedio		Segundos promedio		Porcentaje de correctas	
	Barras	Líneas	Barras	Líneas	Barras	Líneas
1	14,5	1,5	187,7	261,7	0,783 (+- 0,42)	0,696 (+- 0,47)
2	12,5	1,9	288,5	295,1	0,818 (+- 0,39)	0,714 (+- 0,46)
3	9,6	2,2	405,3	442,0	1,000 (+- 0,00)	0,909 (+- 2,94)
4	4,3	1,4	477,9	505,1	0,727 (+- 0,45)	1,000 (+- 0,00)
5	5,0	1,5	613,9	624,7	-	-
6	5,2	1,2	734,1	713,8	0,682 (+- 0,48)	0,818 (+- 0,39)

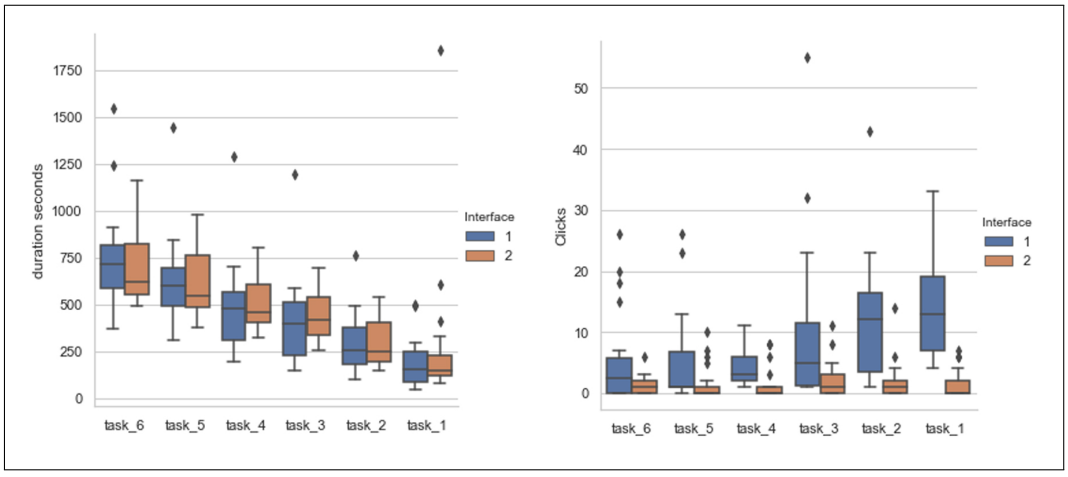


FIGURA 8.1. Cantidad de segundos y clicks por tarea respectivamente, donde la Interface 1 corresponde al gráfico de barras e Interface 2 corresponde al gráfico de líneas.

Para determinar si existe diferencia estadística entre los resultados obtenidos, y así determinar cual interfaz se comportó mejor, se realizó un t-test pareado a las respuestas y el resultado se muestra en la Tabla 8.2, donde se encuentran marcadas aquellas con p-value menor a $0,05$. Para la cantidad de *clicks*, la hipótesis nula para cada tarea es

que existe una diferencia significativa entre ambas interfaces. Por lo anterior se acepta la hipótesis nula, con cual se puede afirmar que la diferencia entre la interfaz de lineas y la de barras es significativa, y en este caso, es la interfaz de barras la que implica mayor cantidad de *clicks* o interacciones. Para la duración la hipótesis nula es que existe diferencia significativa en el tiempo que tomó terminar una tarea en cada interfaz. Del resultado de los test no se puede aceptar ni rechazar esta hipótesis, por lo que no hay evidencia para decir que una interfaz implicó más tiempo que la otra en las 6 tareas evaluadas. Finalmente sobre el rendimiento, la hipótesis nula es que para las tareas 1, 2, 3, 4 y 6 una interfaz tuvo mejor rendimiento que la otra. Para las tareas 1, 2, 3 y 6 no existe evidencia suficiente que permita determinar si se acepta o rechaza la hipótesis. Para el caso de la tarea 4, se puede aceptar la hipótesis nula, por lo cual es posible afirmar que una interfaz se comportó mejor que la otra, en este caso la interfaz de líneas se comportó mejor que la interfaz de barras.

En cuanto a las tarea 5 donde el objetivo era identificar el par de elementos con la distribución más distinta se procedió a evaluar los resultados de forma distinta. Para determinar si una respuesta fue correcta o no, se realizó el *ranking* del grado de similitud entre distribuciones de las comunas. Para lo anterior se utilizó la métrica de divergencia de Jensen-Shannon (James, Ellison, y Crutchfield, 2018). Esta divergencia es una medida simétrica y suavizada de la divergencia de Kullback–Leibler (Kullback y Leibler, 1951), donde su valor es siempre finito para variables aleatorias finitas. Cuantifica en palabras simples que tan distintas son dos o más distribuciones. Su forma básica es como se ve en la Ecuación 8.1 (1) para las distribuciones X e Y, y en (2) su generalización para un número arbitrario de variables aleatorias con peso, donde H es la entropía.

$$\begin{aligned} (1) JSD(X||Y) &= H(\frac{X+Y}{2}) - \frac{H(X)+H(Y)}{2} \\ (2) JSD(X_{0:n}) &= H(\Sigma w_i X_i - \Sigma(w_i H(X_i))) \end{aligned} \tag{8.1}$$

TABLA 8.2. Resultados t-test pareado. El p-value destacado para los clicks es mayor para el grafico de barras, en la tarea 4 el valor es mayor para el gráfico de líneas.

	Resultado t-test		
Tarea	**Clicks**	**Duración**	**Correctas**
1	$\boldsymbol{p-value = 7,369e-05}$ $Z = 3{,}964$ Barra: $M = 14{,}5$; $SD = 8{,}6$ Línea: $M = 1{,}5$; $SD = 2{,}3$	$p-value = 0{,}289$ $Z = -1{,}061$ Barra: $M = 187{,}7$; $SD = 124{,}3$ Línea: $M = 261{,}7$; $SD = 368{,}8$	$p-value = 0{,}527$ $Z = 0{,}632$ Barra: $M = 0{,}78$; $SD = 0{,}42$ Línea: $M = 0{,}70$; $SD = 0{,}47$
2	$\boldsymbol{p-value = 4,100e-4}$ $Z = 3{,}496$ Barra: $M = 12{,}5$; $SD = 9{,}7$ Línea: $M = 1{,}9$; $SD = 3{,}1$	$p-value = 0{,}859$ $Z = -0{,}178$ Barra: $M = 288$; $SD = 156{,}2$ Línea: $M = 295$; $SD = 125{,}0$	$p-value = 0{,}317$ $Z = 1$ Barra: $M = 0{,}818$; $SD = 0{,}39$ Línea: $M = 0{,}714$; $SD = 0{,}46$
3	$\boldsymbol{p-value = 0{,}0243}$ $Z = 2{,}253$ Barra: $M = 9{,}6$; $SD = 13{,}0$ Línea: $M = 2{,}2$; $SD = 2{,}9$	$p-value = 0{,}508$ $Z = -0{,}663$ Barra: $M = 405{,}3$; $SD = 225{,}9$ Línea: $M = 441{,}9$; $SD = 138{,}2$	$p-value = 0{,}157$ $Z = 1{,}414$ Barra: $M = 1$; $SD = 0$ Línea: $M = 0{,}909$; $SD = 2{,}94$
4	$\boldsymbol{p-value = 9,890e-4}$ $Z = 3{,}292$ Barra: $M = 4{,}3$; $SD = 3{,}0$ Línea: $M = 1{,}36$; $SD = 2{,}55$	$p-value = 0{,}624$ $Z = -0{,}490$ Barra: $M = 477{,}8$; $SD = 232{,}3$ Línea: $M = 505{,}1$; $SD = 144{,}2$	$\boldsymbol{p-value = 0{,}0143}$ $Z = -2{,}449$ Barra: $M = 0{,}727$; $SD = 0{,}45$ Línea: $M = 1$; $SD = 0$
5	$\boldsymbol{p-value = 0{,}0415}$ $Z = 2{,}038$ Barra: $M = 4{,}9$; $SD = 7{,}3$ Línea: $M = 1{,}5$; $SD = 2{,}8$	$p-value = 0{,}869$ $Z = -0{,}165$ Barra: $M = 613{,}9$; $SD = 239{,}1$ Línea: $M = 624{,}7$; $SD = 197{,}9$	-
6	$\boldsymbol{p-value = 0{,}0274}$ $Z = 0{,}271$ Barra: $M = 5{,}2$; $SD = 7{,}5$ Línea: $M = 1{,}2$; $SD = 1{,}5$	$p-value = 0{,}787$ $Z = 0{,}271$ Barra: $M = 734{,}1$; $SD = 271{,}2$ Línea: $M = 713{,}8$; $SD = 208{,}2$	$p-value = 0{,}180$ $Z = -1{,}341$ Barra: $M = 0{,}682$; $SD = 0{,}48$ Línea: $M = 0{,}818$; $SD = 0{,}39$

La ventaja de esta métrica es que es simétrica, por lo que las respuestas de pares de valores por ejemplo ”Colina-Lo Espejo” tendrá igual valor que ”Lo Espejo-Colina”. Con el valor de la divergencia calculado con anterioridad, se ordenó en un *ranking* la lista de pares de mayor a menor valor de divergencia, tal que el par con mayor valor de divergencia (los más distintos) fuesen el primer par de la lista. De esta forma también se eliminaron los duplicados. Con la posición del fueron marcadas los pares que respondieron los usuarios. Mientras más arriba en el *ranking* se encuentra la respuesta del usuario, se considera mejor. La distribución de las respuestas correctas se encuentra en la Figura 8.2.

Por otra parte se tabuló el promedio, la desviación estándar y el valor del test-t de los resultados del test NASA-TLX de carga cognitiva (Hart y Staveland, 1988), y se ven sus resultados en la Tabla 8.3.

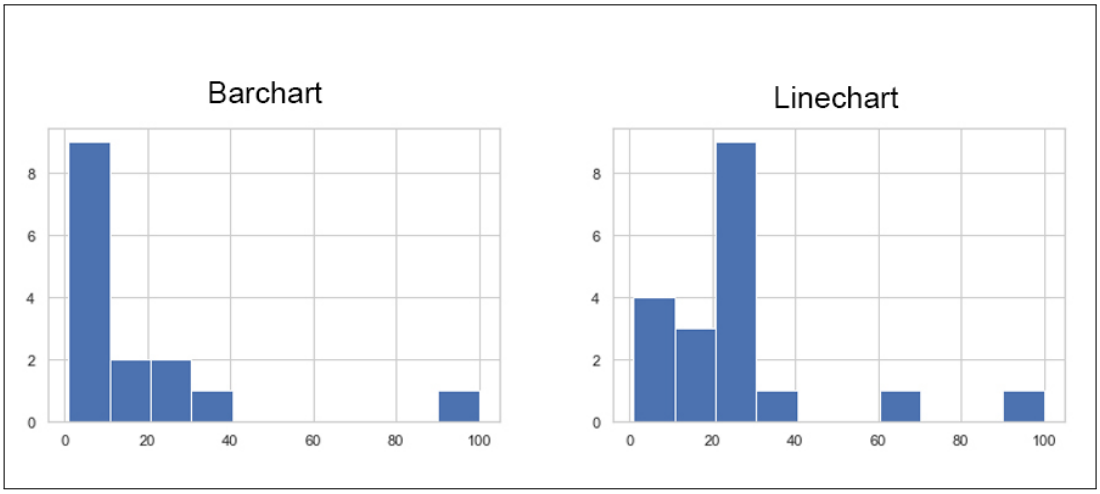


FIGURA 8.2. Histograma de respuestas correctas para la Tarea 5. Para la interfaz del gráfico de barras la media = 15, desviación estándar = 25 y mediana = 5 mientras que para la interfaz del gráfico de líneas son media = 24, desviación estándar = 23 y mediana = 21.

TABLA 8.3. Resultados promedio de la percepción de usuario según el test TLX. No se observan diferencias significativas en niguna de las dimensiones de este test entre las dos interfaces.

	Barras		**Líneas**		
	Promedio	Desviación Estándar	Promedio	Desviación Estándar	t-value
Exigencia mental	5,8	2,1	5,9	2,4	0,879
Exigencia Física	3,9	2,5	3,5	2,2	0,175
Exigencias Temporales	4,8	2,3	5,1	2,8	0,732
Rendimiento	7,8	1,7	7,3	1,9	0,210
Esfuerzo	6,6	2,1	6,4	2,4	0,475
Nivel de frustración	5,2	2,4	5,9	2,3	0,394

Por último los resultados de la interacción de la interfaz se muestran a continuación. En la Figura 8.3 se muestra un Scrollmap utilizando la herramienta CrazyEgg (Cunningham y Robertson, 2014), donde se puede ver que la distribución de la atención del usuario fue similar en ambos casos. En la Figura 8.4 se muestra un mapa de calor, el cual esta hecho con las coordenadas x e y de los clicks sobre cada interfaz. En este se muestran las zonas de mayor intensidad con un gráfico KDE (Muller et al., 1984), se sobre puso la cantidad de clicks en zonas de 25 píxeles. En este gráfico se puede observar

como en una de las interfaces la interacción se concentra solo en un sector mientras que en la otra se esparce a través de toda la interfaz.

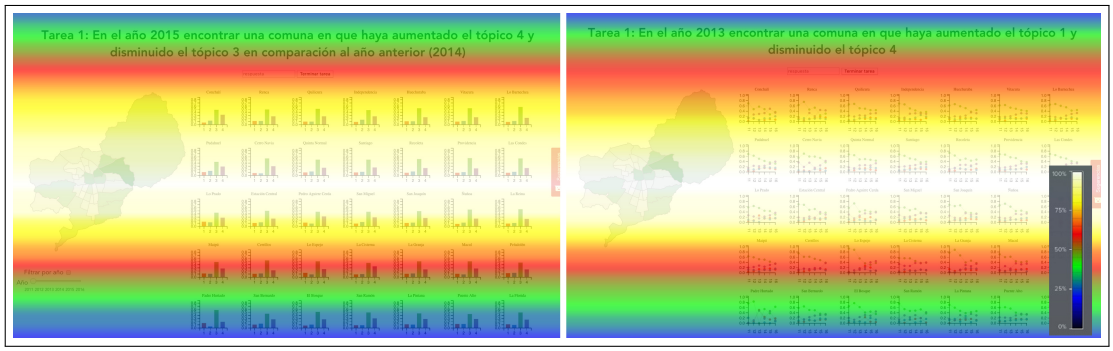


FIGURA 8.3. Atención por área en cada una de las interfaces. Se ve que en general ambas interfaces tienden a concentrar la atención de forma similar.

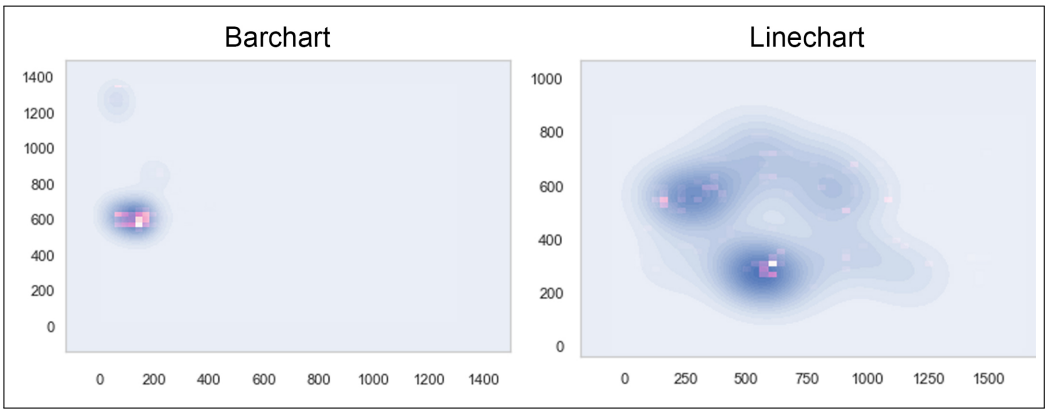


FIGURA 8.4. Mapa de calor de los clicks en la interfaz. Se ve la dispersión del uso de cada interfaz en comparación en cuánto a la cantidad de clicks realizados.

De manera aparte, también se presentan los resultados de los tópicos identificados por los usuarios, en la Figura 8.5, donde se corrobora que los usuarios en general pudieron encontrar sentido a la agrupación de las palabras.

De los resultados expuestos, se analizan las hipótesis supuestas en la Sección 4 de objetivos. Sobre la primera hipótesis, es posible validar que la categorización de tópicos realizada por el algoritmo fue reconocida por los usuarios, y que los tópicos tienen sentido con respecto a los documentos. En la Figura 8.5 se realiza el análisis para cada tópico.

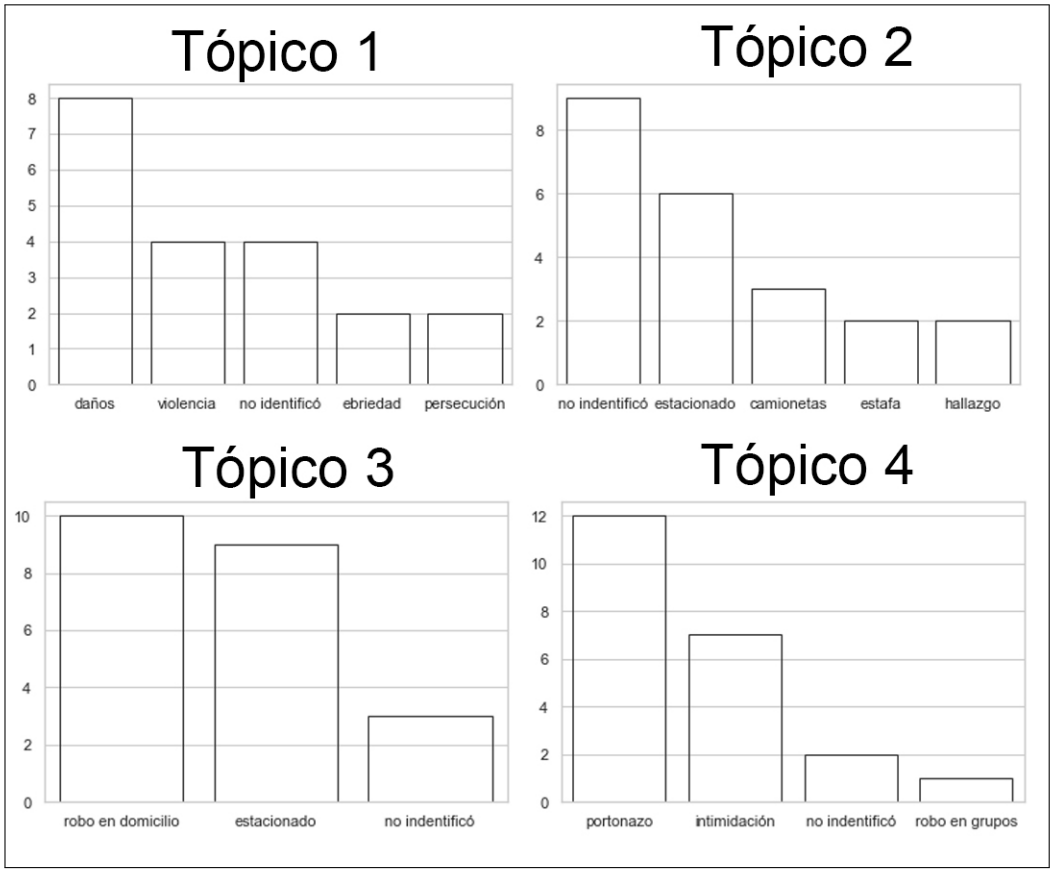


FIGURA 8.5. Tópicos identificados por los usuarios.

Para el tópico 1 un ejemplo de documento es el id 61378 con un $0{,}8$ de probabilidad a este es el siguiente:

> siniestro denunciado vía web antes poder hechos robo vehículo aparición inmediata con la chapa del conductor quebrada, puerta lateral derecha abollada, parabrisas trizado y daños parte inferior.

Se puede corroborar que el documento asociado y el tópico 1 identificado por los usuarios hacen sentido entre sí, pues ambos incluyen daños como parte de su clasificación. Otra validación importante es para el tópico 4, en que la mayoría de los usuarios identificaron la palabra portonazo. Al identificar un documento con una alta probabilidad de pertenecer, en este caso el documento id 60775 con $0{,}9$ de probabilidad dice así:

> tenia vehículo detenido mientras abría el portón dejando la puerta abierta y las llave puestas bloqueo mi bmw blanco cuando por el asiento del copiloto subió al auto intente ir hacia el vehículo pero hizo marcha atrás dándose a la fuga

Se puede ver como el modi operandis de este tipo de robo es efectivamente el que por los medios de comunicación chilena tilda como ”portonazos”. Este análisis se repitió para cada tópico, seleccionando aleatoriamente 10 documentos para cada tópico, donde el criterio para la selección fue si la probabilidad de pertenencia a ese tópico era mayor a $0{,}8$.

Con estos resultados y la validación realizada se puede corroborar la Hipótesis 1, en que la categorización realizada por LDA tiene sentido y es posible que se identifique por los usuarios.

Respecto a la segunda hipótesis, donde se esperaba que las tareas relacionadas con el gráfico de líneas se comportaran mejor en tareas que implicaran tendencias, en base a los resultados, las tareas asociadas con tendencias son las tareas 1 y 2. De acuerdo a la Tabla 8.2 la interfaz con los gráficos de barra logró un porcentaje levemente mayor de respuestas correctas, sin embargo al observar la Tabla 8.2, no se puede corroborar o negar la hipótesis de que una se comporte mejor que la otra. Por lo anterior, se puede concluir que ambas interfaces pueden funcionar de forma similar para resolver tareas que impliquen identificación de tendencias.

Sobre la tercera hipótesis, donde se esperaba que los gráficos de barra se desempeñaran mejor que los gráficos de línea para análisis generales agregados, las tareas 5 y 6 correspondían a este tipo de tareas. En cuánto a la tarea 5, como se ve en la Figura 8.2, en el gráfico de barra se obtuvieron la mayor cantidad de aciertos respecto a la distribución más distinta. De hecho, considerando el *ranking* de las 15 de distribuciones más distintas como correctas (*ranking*@15), el valor del t-test en que se comparan las respuestas correctas e incorrectas para cada interfaz es de $0{,}016$, con lo que se puede validar la hipótesis de que el gráfico de barra funciona mejor en la tarea especifica de

identificar a nivel general diferencias entre las distribuciones. Esta hipótesis es solo validada en particular para la tarea de diferencias distribuciones distintas, ya que para la tarea 6 también de agregación, según los resultados no se puede aceptar o negar la hipótesis de que la interfaz de gráfico de barra funcione mejor que la de gráfico de línea, incluso como promedio, la interfaz de líneas muestra mejores resultados.

Capítulo 9. CONCLUSIONES

Las herramientas de analítica visual para robos de vehículos deben obedecer al objetivo del estudio. Si el fin es encontrar patrones en el comportamiento de los robos, según el *framework* estudiado es más eficiente al momento de comparar, dividir el espacio en áreas de interés en vez de usar una herramienta exploratoria. En el caso de Santiago una división existente, la cual se aprovechó para la implementación de esta herramienta es la comuna. Al tener áreas comparables, la identificación de patrones o tendencias, así como la relación entre áreas y tipos de robos es más fácil de identificar. Tareas que son imposibles en interfaces que tienen toda la información disponible, como la herramienta ITACaT mencionada en la sección de metodología, son abordables al usar una división espacial. Un *encoding* que facilita la forma de comparar la información entre áreas son los *small multiples*. El tipo de glifo que contiene el *small multiple* depende del objetivo de la tarea a realizar.

Un de las conclusiones de este estudio, es la evidencia de que el gráfico de barra requiere más interacción del usuario, esto es por la Tabla 8.1 en que es evidente que la interfaz de barras requirió muchos más *clicks* que la interfaz de líneas. En la Figura 8.4 se puede ver que el gráfico de barra concentra todos sus *clicks* en un punto, que coincide con la posición del filtro por año para esta interfaz. En comparación el gráfico de barras requiere más interacción que el gráfico de líneas al no poder incluir todos sus años al mismo tiempo, sin embargo, los usuarios no consideraron que alguna interfaz requiriera más exigencia física o esfuerzo al usar que la otra, según sus respuestas en el test de carga cognitiva en la Tabla 8.3, no se puede determinar, aunque en ambos casos el promedio fue levemente mayor para el gráfico de barra, la diferencia no es lo suficientemente significativa. Esto quiere decir que a pesar de que el usuario tenga que interactuar más con la interfaz, esto no pareció ser un problema para los usuarios.

Para concluir de acuerdo a la construcción y las herramientas estudiadas, para la exploración es útil tener herramientas que permitan ver ”*the big picture*”de forma generalizada, sin embargo, cuando se requieren realizar tareas, agrupar la información de

acuerdo a divisiones abstractas o reales, como lo son las comunas, permite resolver tareas puntuales de mejor forma, pues permite comparar. La forma de comparar la información puede ser con distintos *encoding*, pero los estudiados particularmente fueron los gráficos de barra y de línea. Ambos poseen sus beneficios y desventajas, sin embargo ambos fueron útiles por si solos para resolver la mayoría las tareas encomendadas, por lo que para tareas generales podrían ser usados sin discriminación. La consideración particular de este estudio es que si es necesario hacer discriminación entre distribuciones, es mejor usar un gráfico de barras frente a un gráfico de líneas.

Sobre el trabajo futuro relacionado con el área es posible ver 2 posibles líneas de mejora:

1. Mejora en el modelo: actualmente se entrena solo una vez el modelo. Por lo anterior incluir la idea de ”human-in-the-loop”podría mejorar la asignación de tópicos. Esto podría permitir al analista encontrar una mejor distribución de tópicos o bien aumentar o disminuir la cantidad tópicos a evaluar en tiempo real. En la Figura 9.1 se puede ver una interfaz con algunas de las funcionalidades que debería tener esta interfaz, en (1) permitir los parámetros del modelo, en (2) la cantidad de tópicos y presets necesarios, y en (3) el resultado de la evaluación.
2. Mejorar exploración: el modelo evaluado no aprovecha la información espacial como lo hacen las visualizaciones que permiten la exploración. Una de las soluciones sería implementar una visualización de los documentos sobre la interfaz ya estudiada con el fin de evaluar si este tipo de funcionalidad mejoraría o no la eficiencia.

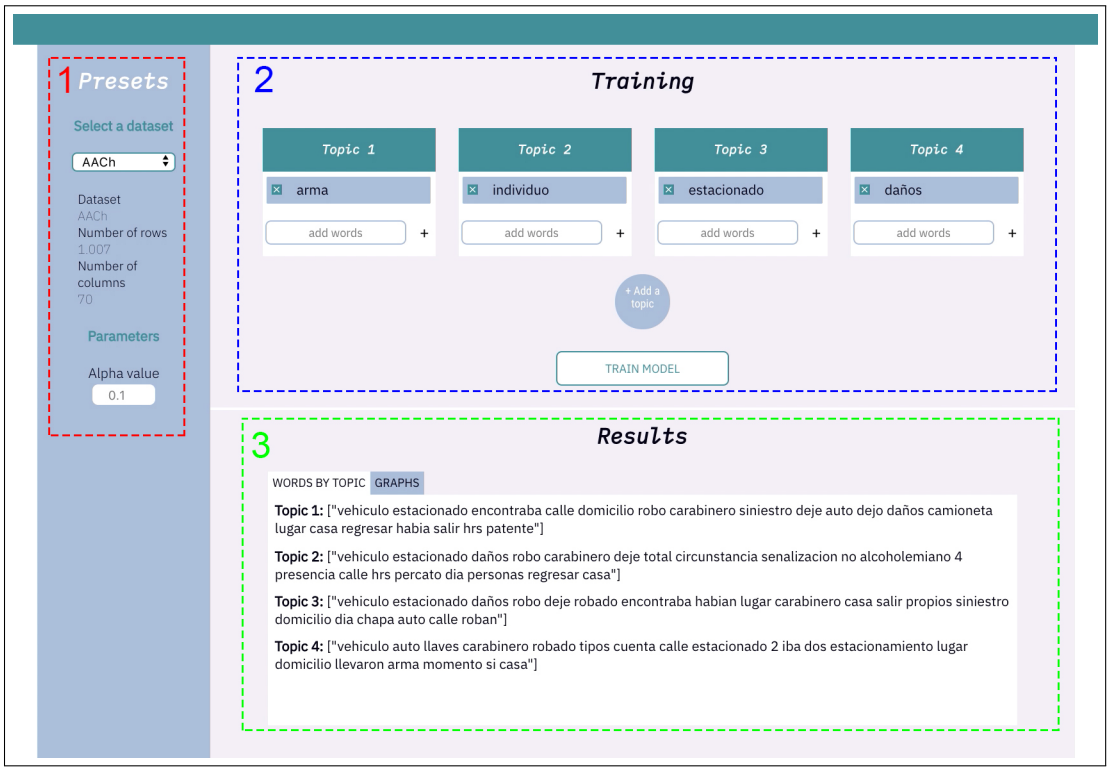


FIGURA 9.1. Visualización para entrenar tópicos de forma interactiva.

BIBLIOGRAFIA

Andrzejewski, D., Zhu, X., y Craven, M. (2009). Incorporating domain knowledge into topic modeling via dirichlet forest priors. En *Proceedings of the 26th annual international conference on machine learning* (pp. 25–32). New York, NY, USA: ACM. Descargado de http://doi.acm.org/10.1145/1553374.1553378 doi: 10.1145/1553374.1553378

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*(1), 17-21. Descargado de https://www.tandfonline.com/doi/abs/10.1080/00031305.1973.10478966 doi: 10.1080/00031305.1973.10478966

Baeza-Yates, R., Ribeiro, B. d. A. N., y cols. (2011). *Modern information retrieval*. New York: ACM Press; Harlow, England: Addison-Wesley,.

Bird, S., y Loper, E. (2004). Nltk: The natural language toolkit. En *Proceedings of the acl 2004 on interactive poster and demonstration sessions*. Stroudsburg, PA, USA: Association for Computational Linguistics. Descargado de http://dx.doi.org/10.3115/1219044.1219075 doi: 10.3115/1219044.1219075

Blei, D. M. (2012a). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Blei, D. M. (2012b, abril). Probabilistic topic models. *Commun. ACM*, *55*(4), 77–84. Descargado de http://doi.acm.org/10.1145/2133806.2133826 doi: 10.1145/2133806.2133826

Blei, D. M., y Lafferty, J. D. (2006). Dynamic topic models. En *Proceedings of the 23rd international conference on machine learning* (pp. 113–120). New

York, NY, USA: ACM. Descargado de http://doi.acm.org/10.1145/1143844.1143859 doi: 10.1145/1143844.1143859

Blei, D. M., Ng, A. Y., y Jordan, M. I. (2003, marzo). Latent dirichlet allocation. *J. Mach. Learn. Res.*, *3*, 993–1022. Descargado de http://dl.acm.org/citation.cfm?id=944919.944937

Block, C. R. (1995). Stac hot-spot areas: A statistical tool for law enforcement decisions. En *Crime analysis through computer mapping. washington, dc: Police executive research forum* (pp. 15–32).

Brynjolfsson, E., y Ms, A. (2017, 8). What's driving the machine learning explosion? *Harvard Business Review*, *1704*(4), 201-213.

Cao, J., Xia, T., Li, J., Zhang, Y., y Tang, S. (2009). A density-based method for adaptive lda model selection. *Neurocomputing*, *72*(7-9), 1775–1781.

Cao, N., Lin, C., Zhu, Q., Lin, Y.-R., Teng, X., y Wen, X. (2018). Voila: Visual anomaly detection and monitoring with streaming spatiotemporal data. *IEEE transactions on visualization and computer graphics*, *24*(1), 23–33.

Chang, W., Ku, Y., Wu, S., y Chiu, C. (2012). Cybercrimeir–a technological perspective to fight cybercrime. En *Pacific-asia workshop on intelligence and security informatics* (pp. 36–44).

Christoforidis, A., Heuwing, B., y Mandl, T. (2017). Visualising topics in document collections.

Chuang, J., Manning, C. D., y Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. En *Proceedings of the international working conference on advanced visual interfaces* (pp. 74–77).

Commission, A. C. J., y cols. (2004). Arizona auto theft study. *Phoenix, AZ: Statistical Analysis Center Publication*.

Cunningham, H., y Robertson, J. (2014). Crazy egg. *Journal of the Canadian Health Libraries Association/Journal de l'Association des bibliothèques de la santé du Canada*, *34*(2), 123–126.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., y Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391-407. doi: 10.1002/(SICI)1097-4571(199009)41:6⟨391::AID-ASI1⟩3.0.CO;2-9

Deveaud, R., SanJuan, E., y Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, *17*(1), 61–84.

Diaconis, P., y Freedman, D. A. (1990). Cauchy's equation and de finetti's theorem. *Scandinavian Journal of Statistics*, *17*(3), 235–249. Descargado de `http://www.jstor.org/stable/4616171`

Fayyad, U., Grinstein, G. G., y Wierse, A. (Eds.). (2002). *Information visualization in data mining and knowledge discovery*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Fayyad, U., Piatetsky-Shapiro, G., y Smyth, P. (1996, noviembre). The kdd process for extracting useful knowledge from volumes of data. *Commun. ACM*, *39*(11), 27–34. Descargado de `http://doi.acm.org/10.1145/240455.240464` doi: 10.1145/240455.240464

Fayyad, U., y Stolorz, P. (1997). Data mining and kdd: Promise and challenges. *Future generation computer systems*, *13*(2-3), 99–115.

Griffiths, T. L., Jordan, M. I., Tenenbaum, J. B., y Blei, D. M. (2004). Hierarchical topic models and the nested chinese restaurant process. En *Advances in neural information processing systems* (pp. 17–24).

Harrower, M., y Brewer, C. A. (2003). Colorbrewer.org: An online tool for selecting colour schemes for maps. *The Cartographic Journal*, *40*(1), 27-37. Descargado de https://www.tandfonline.com/doi/abs/10.1179/000870403235002042 doi: 10.1179/000870403235002042

Hart, S. G., y Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. En P. A. Hancock y N. Meshkati (Eds.), *Human mental workload* (Vol. 52, p. 139 - 183). North-Holland. Descargado de http://www.sciencedirect.com/science/article/pii/S0166411508623869 doi: https://doi.org/10.1016/S0166-4115(08)62386-9

Heer, J., Bostock, M., Ogievetsky, V., y cols. (2010). A tour through the visualization zoo. *Commun. Acm*, *53*(6), 59–67.

Hoffman, M. D., Blei, D. M., y Bach, F. (2010). Online learning for latent dirichlet allocation. En *Proceedings of the 23rd international conference on neural information processing systems - volume 1* (pp. 856–864). USA: Curran Associates Inc. Descargado de http://dl.acm.org/citation.cfm?id=2997189.2997285

Hofmann, T. (1999). Probabilistic latent semantic analysis. En *Proceedings of the fifteenth conference on uncertainty in artificial intelligence* (pp. 289–296). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. Descargado de http://dl.acm.org/citation.cfm?id=2073796.2073829

Hu, Y., Boyd-Graber, J., y Satinoff, B. (2011). Interactive topic modeling. En *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies - volume 1* (pp. 248–257). Stroudsburg, PA, USA: Association for Computational Linguistics. Descargado de http://dl.acm.org/citation.cfm?id=2002472.2002505

Jagarlamudi, J., Daumé, H., III, y Udupa, R. (2012). Incorporating lexical priors into topic models. En *Proceedings of the 13th conference of the european chapter of the association for computational linguistics* (pp. 204–213). Stroudsburg, PA, USA: Association for Computational Linguistics. Descargado de http://dl.acm.org/citation.cfm?id=2380816.2380844

James, R. G., Ellison, C. J., y Crutchfield, J. P. (2018). dit: a Python package for discrete information theory. *The Journal of Open Source Software*, *3*(25), 738. doi: https://doi.org/10.21105/joss.00738

Kullback, S., y Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, *22*(1), 79–86.

Manning, C. D., Raghavan, P., y Schütze, H. (2008). Scoring, term weighting and the vector space model. *Introduction to information retrieval*, *100*, 2–4.

Mohr, J. W., y Bogdanov, P. (2013). Introduction—topic models: What they are and why they matter. *Poetics*, *41*(6), 545 - 569. Descargado de http://www.sciencedirect.com/science/article/pii/S0304422X13000685 (Topic Models and the Cultural Sciences) doi: https://doi.org/10.1016/j.poetic.2013.10.001

Muller, H.-G., y cols. (1984). Smooth optimum kernel estimators of densities, regression curves and modes. *The Annals of Statistics*, *12*(2), 766–774.

Munzner, T. (2009, noviembre). A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, *15*(6), 921–928. Descargado de http://dx.doi.org/10.1109/TVCG.2009.111 doi: 10.1109/TVCG.2009.111

Newman, D., Baldwin, T., Cavedon, L., Huang, E., Karimi, S., Martinez, D., ... Zobel, J. (2010). Visualizing search results and document collections using topic

maps. *Web Semantics: Science, Services and Agents on the World Wide Web*, *8*(2-3), 169–175.

Nikita, M. (2016). Select number of topics for lda model. *Retrieved from*.

Nilsson, N. J. (1996). *Introduction to Machine Learning (An Early Draft to a proposed textbook)*. Descargado de `http://ai.stanford.edu/~{}nilsson/mlbook.html`

Pazzani, M. J., y Billsus, D. (2007). Content-based recommendation systems. En *The adaptive web* (pp. 325–341). Springer.

Quezada, M., Peña Araya, V., y Poblete, B. (2015). Location-aware model for news events in social media. En *Proceedings of the 38th international acm sigir conference on research and development in information retrieval* (pp. 935–938). New York, NY, USA: ACM. Descargado de `http://doi.acm.org/10.1145/2766462.2767815` doi: 10.1145/2766462.2767815

Řehůřek, R., y Sojka, P. (2010, 22 de mayo). Software Framework for Topic Modelling with Large Corpora. En *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA. (`http://is.muni.cz/publication/884893/en`)

Saket, B., Endert, A., y Demiralp, C. (2018). Task-based effectiveness of basic visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 1-1. doi: 10.1109/TVCG.2018.2829750

Salton, G., y McGill, M. J. (1986). *Introduction to modern information retrieval*. New York, NY, USA: McGraw-Hill, Inc.

Shibukawa, Y. (2009). Python package snowballstemmer 1.2.1: https://perma.cc/xpj6-jnmf. 2015.

Sievert, C., y Shirley, K. (2014). Ldavis: A method for visualizing and interpreting topics. En *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63–70).

Susanne, M., Edgar, E., Buchner, A., y Faul, F. (2007, 09). A short tutorial of gpower. *Tutorials in Quantitative Methods for Psychology*, *3*. doi: 10.20982/tqmp .03.2.p051

van den Elzen, S., y van Wijk, J. J. (2013). Small multiples, large singles: A new approach for visual data exploration. En *Computer graphics forum* (Vol. 32, pp. 191–200).

Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. En *Proceedings of the 23rd international conference on machine learning* (pp. 977–984).

Ward, M., Grinstein, G., y Keim, D. (2010). *Interactive data visualization: Foundations, techniques, and applications*. Natick, MA, USA: A. K. Peters, Ltd.

Yao, L., Mimno, D., y McCallum, A. (2009). Efficient methods for topic model inference on streaming document collections. En *Proceedings of the 15th acm sigkdd international conference on knowledge discovery and data mining* (pp. 937–946).

ANEXO A. DETALLES COMPLETOS DE LA BASE DE DATOS DE LA AACH

A continuación se presentan los atributos de la base de datos suministrada por la AACh, sin incluir los atributos con información personal, sumando un total de 70 atributos. Por otra parte se especifica además de su identificador interno, el tipo dato (ordinal, nominal, temporal), la cantidad de elementos no vacíos y la descripción de ese atributo.

TABLA A.1. Resumen y descripción de los atributos de la base de datos.

Nombre	Tipo de dato	Filas no vacias	Descripción
id_prose	ordinal	55.626	*ID única transacción.*
sin_patente	nominal	55.626	*Patente vehículo siniestrado.*
sin_fecha_siniestro	fecha	55.626	*Fecha (día, mes y año) en donde ocurrió el siniestro de robo.*
ase_abreviatura	nominal	55.626	*Abreviatura de aseguradora en la cual se registra el robo.*
obs_ult_estado	nominal	3.179	*Validación interna.*
sin_siniestro	ordinal	55.626	*Número de siniestro interno de la compañía. Cada número es independiente.*
sin_fecha_denuncia	fecha	55.626	*Fecha denuncia siniestro.*
sin_lugar_siniestro	ordinal	55.626	*ID Tipo de lugar en donde ocurrió el siniestro.*
lug_desc	nominal	55.626	*Tipo de lugar en donde ocurrió el siniestro.*
rpc_id_region_siniestro	ordinal	55.626	*ID Región en donde ocurrió el siniestro.*
reg_descripcion_region	nominal	55.624	*Región en donde ocurrió el siniestro.*
rpc_id_provincia_siniestro	ordinal	55.626	*ID Provincia en donde ocurrió el siniestro*
pro_descripcion_provincia	nominal	55.624	*Provincia en donde ocurrió el siniestro*
rpc_id_comuna_siniestro	ordinal	55.626	*ID Comuna en donde ocurrió el siniestro*
reg_descripcion_comuna	nominal	55.624	*Comuna en donde ocurrió el siniestro*
sin_direccion_siniestro	nominal	46.819	*Relato de dirección del siniestro.*
sin_hora_siniestro	hora	55.626	*Hora de ocurrencia del siniestro.*
est_id_prose	ordinal	55.626	*Validación interna.*
esp_desc	nominal	55.626	*Validación interna.*
sin_fecha_ultimo_estado	fecha	55.626	*Validación interna.*
sin_poliza	nominal	50.793	*Número de poliza.*
sin_item	ordinal	48.875	*Validación interna.*
sin_check_registro_civil	nominal	55.626	*Validación interna.*
sin_concurrencia	nominal	55.626	*Validación interna.*
sin_ind_caso_raro	nominal	55.626	*Validación interna.*
sin_glosa_caso_raro	nominal	2	*Validación interna.*
tve_id_tipo_vehiculo	ordinal	55.626	*ID Tipo de vehículo según registro civil*
tve_desc	nominal	55.626	*Tipo de vehículo según registro civil*
mar_id_vehiculo	ordinal	55.626	*Identificación interna de la marca del vehículo siniestrado.*
mar_desc	nominal	55.626	*Marca del vehículo involucrado en el robo.*
mod_id_vechiculo	ordinal	55.626	*ID del modelo asociado al vehículo siniestrado.*
mod_desc	nominal	55.626	*Modelo del vehículo siniestrado.*
sin_ano_vechiculo	ordinal	55.626	*Año de fabricación otorgado por el registro civil.*
sin_motor_vehiculo	nominal	53.456	*Motor asociado al vehículo robado.*
sin_chasis_vehiculo	nominal	53.200	*Chasis del vehículo asegurado*
cve_id_vehiculo	ordinal	55.626	*ID del color del vehículo.*
cve_desc	nominal	55.626	*Color del vehículo asegurado.*
sin_valor_comercial_veh	ordinal	55.626	*Valor comercial del vehículo, sin detalle de las unidades utilizadas.*
sin_dispositivos_vehiculo	nominal	55.626	*Validación interna.*
sin_dispositivos_desc	nominal	188	*Validación interna.*
rpc_id_region_asegurado	ordinal	55.626	*ID Región en donde ocurrió el siniestro.*
reg_descripcion_region_asegurado	nominal	55.626	*Región de residencia de la persona que contrató el seguro.*
rpc_id_provincia_asegurado	ordinal	55.626	*ID Comuna de residencia de la persona que contrató el seguro.*
pro_descripcion_provincia_asegurado	nominal	55.626	*Comuna de residencia de la persona que contrató el seguro.*
rpc_id_comuna_asegurado	ordinal	55.626	*ID Comuna de residencia de la persona que contrató el seguro.*
reg_descripcion_comuna_asegurado	nominal	55.626	*Comuna de residencia de la persona que contrató el seguro.*
rpc_id_region_conductor	ordinal	55.626	*ID Región de residencia de la persona que conducía el auto al momento del siniestro.*
reg_descripcion_region_conductor	nominal	55.626	*Región de residencia de la persona que conducía el auto al momento del siniestro.*
rpc_id_provincia_conductor	ordinal	55.626	*ID Provincia de residencia de la persona que conducía el auto al momento del siniestro.*
pro_descripcion_provincia_conductor	nominal	55.626	*Provincia de residencia de la persona que conducía el auto al momento del siniestro.*
rpc_id_comuna_conducto	ordinal	55.626	*ID Comuna de residencia de la persona que conducía el auto al momento del siniestro.*
reg_descripcion_comuna_conductor	nominal	55.626	*Comuna de residencia de la persona que conducía el auto al momento del siniestro.*
sin_comisaria	nominal	45.121	*Relato de comisaría en donde se constató el heho.*
sin_fecha_parte	nominal	55.626	*Fecha del parte registrado en comisaría.*
sin_tribunal_fiscalia	nominal	3.061	*Tribunal y/o fiscalía que adscribe al caso siniestrado (recibido por carabineros).*
sin_fecha_citacion	nominal	55.626	*Validación interna.*
sin_relato	nominal	51.451	*Relato del siniestro.*
car_id	ordinal	55.626	*Registro del vehículo asegurado reportado.*
cre_id	ordinal	55.626	*Validación interna.*
sin_activo	nominal	55.626	*Validación interna.*
sin_cargado_bizagi	nominal	55.626	*Validación interna.*
id_PROSE_BPM_ROBO	ordinal	55.626	*Validación interna.*
UltimoOrden	nominal	49.258	*Validación interna.*
FechaOrden	nominal	49.258	*Validación interna.*
UsuarioBPM	nominal	627	*Validación interna.*
CargadoBPM	nominal	55.626	*Validación interna.*
PrimeraFechaValidacionProse	fecha	55.624	*Validación interna.*
EncontradoOtraVia	nominal	55.626	*Informa si la información del hallazgo fue distinta a carabineros.*
tareasBizagi	nominal	48.791	*Validación interna.*
FechaCarga	nominal	55.626	*Fecha en que se agregó al sistema (de la aseguradora) el robo vehicular*

ANEXO B. ANÁLISIS EXPLORATORIO DE LOS DATOS

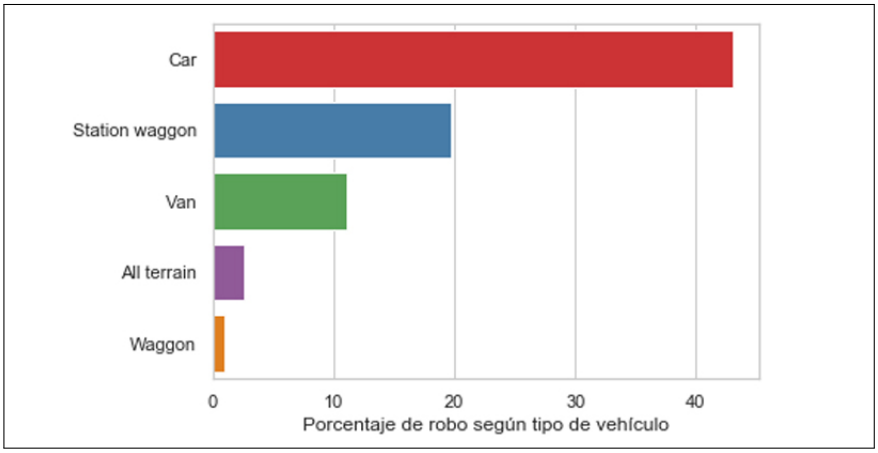


FIGURA B.1. Porcentaje de robo de vehículos por tipo.

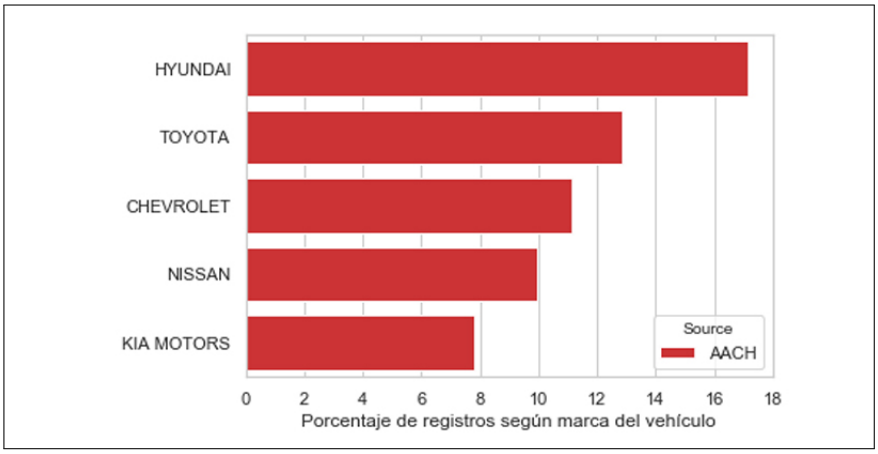
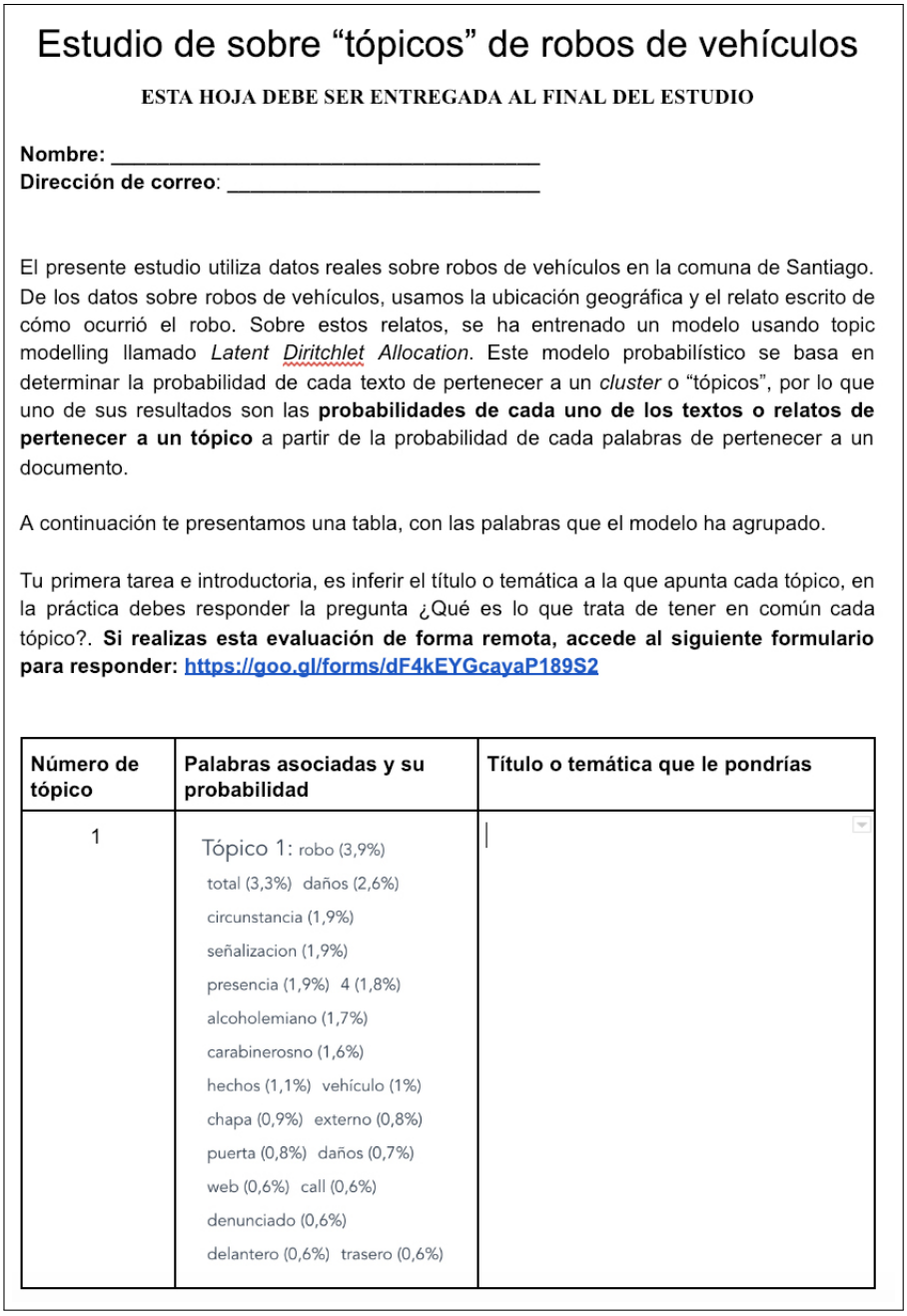


FIGURA B.2. Porcentaje de robo de vehículos por marca.

ANEXO C. COMUNAS DE SANTIAGO

Cerrillos La Reina Pudahuel Cerro Navia Las Condes Quilicura Conchalí Lo Barnechea Quinta Normal El Bosque Lo Espejo Recoleta Estación Central Lo Prado Renca Huechuraba Macul San Miguel Independencia Maipú San Joaquín La Cisterna Ñuñoa San Ramón La Florida Pedro Aguirre Cerda Santiago La Pintana Peñalolén Vitacura La Granja Providencia

ANEXO D. INFORMACIÓN EN ESTUDIO DE USUARIO



Estudio de sobre “tópicos” de robos de vehículos

ESTA HOJA DEBE SER ENTREGADA AL FINAL DEL ESTUDIO

Nombre: _____
Dirección de correo: _____

El presente estudio utiliza datos reales sobre robos de vehículos en la comuna de Santiago. De los datos sobre robos de vehículos, usamos la ubicación geográfica y el relato escrito de cómo ocurrió el robo. Sobre estos relatos, se ha entrenado un modelo usando topic modelling llamado *Latent Diritchlet Allocation*. Este modelo probabilístico se basa en determinar la probabilidad de cada texto de pertenecer a un *cluster* o “tópicos”, por lo que uno de sus resultados son las **probabilidades de cada uno de los textos o relatos de pertenecer a un tópico** a partir de la probabilidad de cada palabras de pertenecer a un documento.

A continuación te presentamos una tabla, con las palabras que el modelo ha agrupado.

Tu primera tarea e introductoria, es inferir el título o temática a la que apunta cada tópico, en la práctica debes responder la pregunta ¿Qué es lo que trata de tener en común cada tópico?. **Si realizas esta evaluación de forma remota, accede al siguiente formulario para responder:** https://goo.gl/forms/dF4kEYGcayaP189S2

Número de tópico	Palabras asociadas y su probabilidad	Título o temática que le pondrías
1	Tópico 1: robo (3,9%) total (3,3%) daños (2,6%) circunstancia (1,9%) señalizacion (1,9%) presencia (1,9%) 4 (1,8%) alcoholemiano (1,7%) carabinerosno (1,6%) hechos (1,1%) vehículo (1%) chapa (0,9%) externo (0,8%) puerta (0,8%) daños (0,7%) web (0,6%) call (0,6%) denunciado (0,6%) delantero (0,6%) trasero (0,6%)	

FIGURA D.1. Estudio de usuario.

2	Tópico 2: auto (5,4%) vehículo (4,1%) si (1,9%) algún (1,5%) llaves (1,3%) camioneta (1,3%) estacionamiento (1,2%) robado (1,1%) cuenta (1,1%) todas (0,9%) seguridad (0,8%) comprobante (0,8%) carabineros (0,7%) mecanismo (0,7%) camion (0,6%) asegurado (0,6%) día (0,6%) empresa (0,6%) denuncia (0,5%) patente (0,4%)	
3	Tópico 3: vehículo (9,9%) estacionado (5%) deje (2,2%) robado (1,6%) robo (1,3%) lugar (1,3%) automovil (1,2%) auto (1,1%) encontraba (1,1%) calle (1,1%) hrs (1%) domicilio (1%) volver (1%) carabineros (0,9%) día (0,9%) percato (0,8%) casa (0,8%) habia (0,8%) salir (0,8%) dejo (0,8%)	
4	Tópico 4: vehículo (6,6%) auto (2%) dos (1%) personas (1%) llaves (0,9%) casa (0,8%) roban (0,8%) llevaron (0,8%) calle (0,7%) tipos (0,7%) arma (0,7%) iba (0,7%) domicilio (0,7%) pistola (0,6%) fuego (0,6%) individuos (0,6%) robo (0,6%) camioneta (0,6%) bajan (0,6%) 3 (0,5%)	

FIGURA D.2. Estudio de usuario.

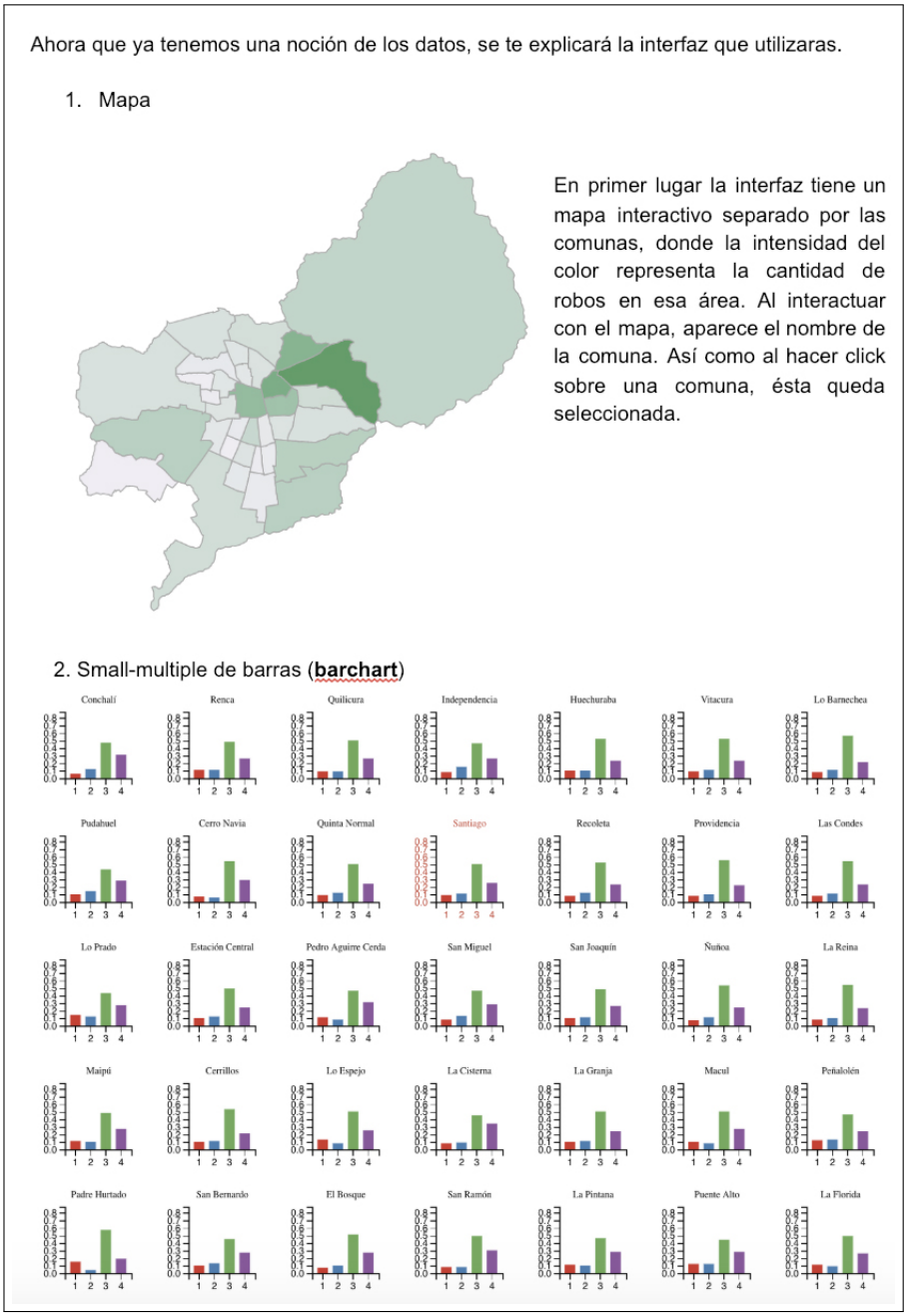
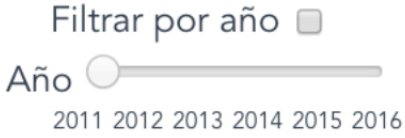


FIGURA D.3. Estudio de usuario.

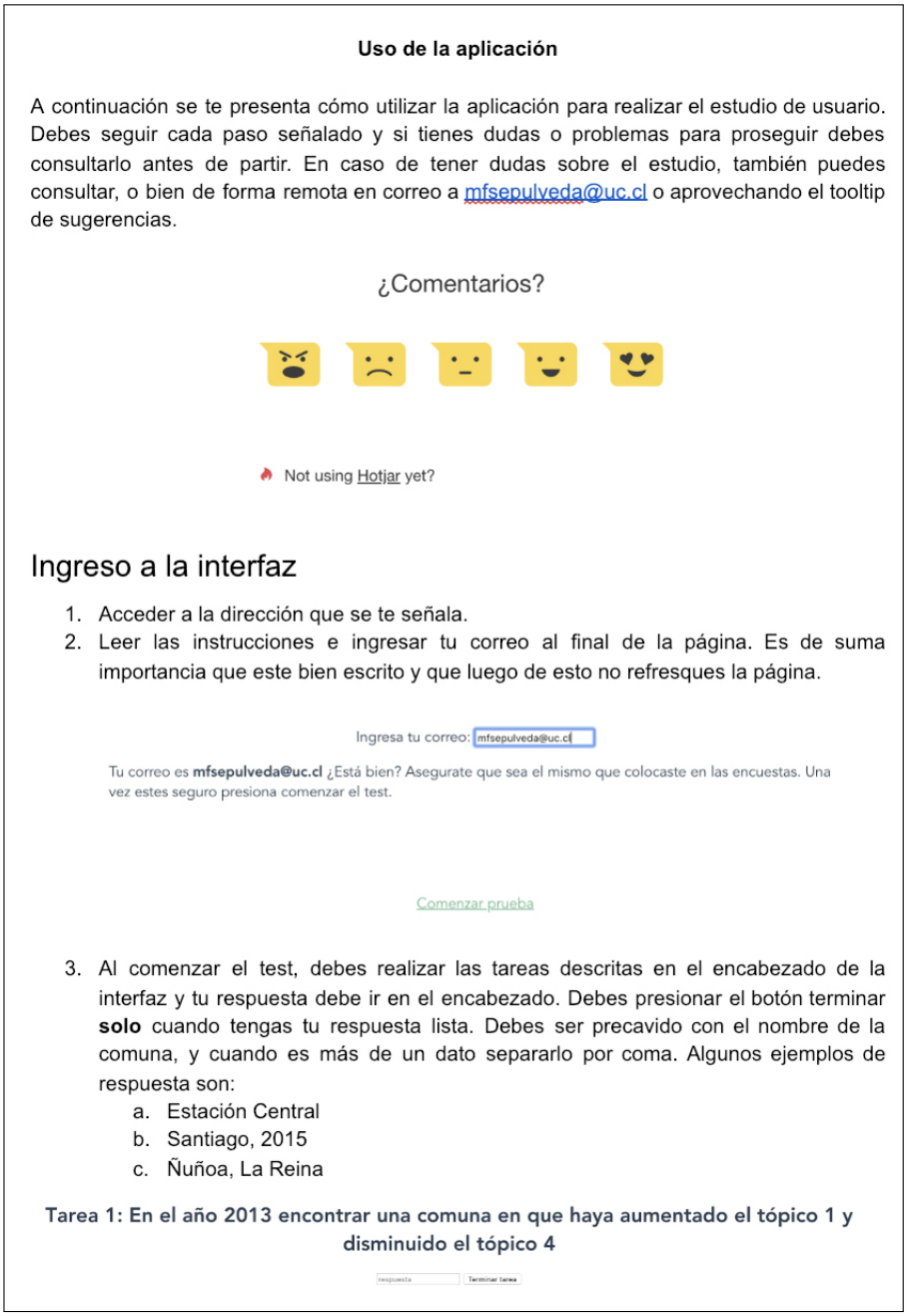
Por cada comuna existe un gráfico asociado. En el eje x se encuentra el número de tópico y en el eje y la probabilidad de ese tópico en el mapa. Además esta interfaz cuenta con un filtro por año, pues la vista principal es el agregado de todos los años. Debes habilitar el checkbox de filtro por año para utilizar el slider.



3. Small-multiple de líneas (**linechart**)

Cada línea representa un tópico. En el eje x se tienen los años, mientras que en el eje y se tiene la probabilidad del tópico para ese año.

FIGURA D.4. Estudio de usuario.



Uso de la aplicación

A continuación se te presenta cómo utilizar la aplicación para realizar el estudio de usuario. Debes seguir cada paso señalado y si tienes dudas o problemas para proseguir debes consultarlo antes de partir. En caso de tener dudas sobre el estudio, también puedes consultar, o bien de forma remota en correo a mfsepulveda@uc.cl o aprovechando el tooltip de sugerencias.

¿Comentarios?

Not using Hotjar yet?

Ingreso a la interfaz

1. Acceder a la dirección que se te señala.
2. Leer las instrucciones e ingresar tu correo al final de la página. Es de suma importancia que este bien escrito y que luego de esto no refresques la página.

Ingresa tu correo: mfsepulveda@uc.cl

Tu correo es **mfsepulveda@uc.cl** ¿Está bien? Asegurate que sea el mismo que colocaste en las encuestas. Una vez estes seguro presiona comenzar el test.

Comenzar prueba

3. Al comenzar el test, debes realizar las tareas descritas en el encabezado de la interfaz y tu respuesta debe ir en el encabezado. Debes presionar el botón terminar **solo** cuando tengas tu respuesta lista. Debes ser precavido con el nombre de la comuna, y cuando es más de un dato separarlo por coma. Algunos ejemplos de respuesta son:
    a. Estación Central
    b. Santiago, 2015
    c. Ñuñoa, La Reina

Tarea 1: En el año 2013 encontrar una comuna en que haya aumentado el tópico 1 y disminuido el tópico 4

FIGURA D.5. Estudio de usuario.