Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento Ingeniería Industrial

Modelamiento y seguimiento de tópicos para detección de modus operandi en robos de autos

Alumno: Diego Garrido
Profesor: Richard Weber
Fecha: 20 de agosto de 2019
Santiago, Chile

Índice de Contenidos

Lista de Figuras

Lista de Tablas

Resumen

El presente trabajo utiliza modelos de tópicos para descubrimiento y seguimiento de temas o tópicos en el contexto de robo de vehículos, entendiendo un tópico como "un patrón repetitivo de términos co-currentes en un conjunto de documentos", en este caso, los documentos corresponden a los relatos de los robos de vehículos de la Asociación de Aseguradores de Chile (AACH). Un tópico en este contexto puede entenderse como un tipo de delito, por tanto, es de interés analizar estos tópicos descubiertos, pues su análisis podría darnos luces sobre las formas en que se comenten los delitos asociados a vehículos motorizados. No solo interesa descubrir estos tópicos, si no también analizar la evolución de estos tópicos en el tiempo, ya que esto nos permitiría estudiar la evolución de ciertos *modus operandi* para robo de vehículos.

1. Introducción

El aprendizaje no supervisado y semi-supervisado para procesamiento de texto es un área activa de investigación estos días. En muchas ocasiones el aprendizaje supervisado es inapropiado o imposible. Por ejemplo, en detección de comportamientos anormales es difícil predecir por adelantado que tipos de anomalías podrían ocurrir. Dentro de los métodos de aprendizaje no supervisado el modelamiento de tópicos es un enfoque prometedor para detección de comportamientos anormales [1]-[2]. Esto no solo permite dar advertencias sobre anomalías, además provee información sobre los patrones de comportamiento típicos.

El modelamiento de tópicos [3]-[5], es una herramienta estadística que busca encontrar los temas presentes en un conjunto de documentos (corpus), permitiendo organizar, buscar, indexar, explorar y comprender grandes colecciones de documentos. En minería de texto esto es asumido que los documentos no etiquetas pueden ser representados como una mezcla de tópicos, donde los tópicos son distribuciones sobre las palabras. Los tópicos son latentes y la inferencia en modelamiento de tópicos tiene por objetivos descubrirlos. En este sentido, los temas se pueden definir como "un patrón repetitivo de términos co-currentes en un corpus". Por ejemplo, se tiene el siguiente tópico, representado por sus cuatro palabras más probables, "salud", "médico", "paciente", "hospital", estas palabras sugieren el siguiente nombre para el tema: "Atención médica".

En los modelos convencionales de modelamiento de tópicos, las palabras dentro de un documento y los documentos son tratados como intercambiables, si bien tratar las palabras de forma intercambiable es una simplificación que es consistente con el objetivo de identificar los temas semánticos dentro de cada documento. Para muchas colecciones de interés, sin embargo, la suposición implícita de documentos intercambiables es inapropiada. Estos modelos funcionan bajo el supuesto que el corpus es estático, es decir, donde el número de tópicos y el tamaño del vocabularios son conocidos por adelantado. Además, la mayoría de estos fijan el número de tópicos y lo mantienen así a lo largo del corpus completo. Si bien lo anterior es apropiado para un corpus estático, esto toma suma importancia cuando analizamos conjunto de datos que varían en el tiempo (donde nuevos documentos se van generando día a día), sobre todo para los algoritmos en línea, donde es de interés hacer un seguimiento de los tópicos encontrados, donde de un periodo a otro se puede dar el nacimiento o muerte de un tópico.

En la literatura de modelamiento de tópicos se consideran dos tipos de modelos dinámicos. En el primer tipo la dinámica es asumida en la mezcla de tópicos [6]–[8]. En el segundo tipo la dinámica es asumida en los mismos tópicos [9]–[11], es decir, la distribución sobre las palabras de cada tópico cambia a través del tiempo. Algunos trabajos donde ambas tipos de dinámicas son consideradas [12], [13].

2. Marco teórico

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

2.1. Latent Dirichlet Allocation (LDA)

LDA es un modelo probabilístico generativo para el modelamiento de tópicos. Su principal característica, comparado con modelos como LSI, es que utiliza la distribución Dirichlet como prior para dos distribuciones multinomiales, con el objetivo de evitar sobreajuste.

El modelo recibe por entrada el corpus (colección de documentos), por ejemplo, los relatos de las denuncias de robo de vehículos y el número de tópicos a descubrir K, además de dos hiperparámetros, α y η para las distribuciones Dirichlet, que por defecto son 0.1 y 0.01 respectivamente. A partir del proceso de inferencia (el proceso generativo es mostrado en la Figura 2.1), LDA obtiene como salida la lista de tópicos latentes $\beta_{1:K}$, estas corresponden a una distribución de probabilidad sobre el vocabulario (palabras diferentes que sobrevivieron al procesamiento realizado sobre el corpus), así como las distribuciones de tópicos θ_d para cada documento d. Más formalmente:

Sean K tópicos, $\beta_{1:K}$ son distribuciones de probabilidad sobre un vocabulario fijo, dibujadas por una $Dirichlet(\eta)$. Para cada documento d del corpus D se asume que es dibujado por el siguiente proceso generativo:

1. Dibujar la mezcla de tópicos θ^d $\sim$Dirichlet(α).
2. Para cada palabra n:
   (a) Escoger la asignación del tópico Z $\sim$Mult(θ^d).
   (b) Escoger una palabra $W_{d,n}$ $\sim$Mult(β_z).

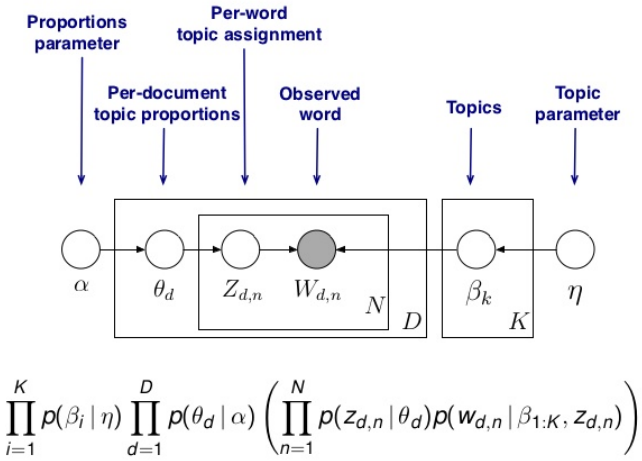

Figura 2.1: Representación gráfica de LDA. $\beta_{1:K}$ representa la distribución de probabilidad sobre el vocabulario de cada tópico y θ^d representa la mezcla de tópicos de cada documento.

2.2. Dynamic Topic Model (DTM)

Con respecto al modelamiento de tópicos dinámico, se tiene Dynamic Topic Model (Blei et al., 2006), de ahora en adelante DTM, corresponde a un modelo generativo que puede ser usado para analizar la evolución de tópicos no observados de una colección de documentos en el tiempo. Esta familia de modelos probabilísticos de series de tiempo corresponde a una extensión de LDA. En LDA tanto el orden en que aparecen las palabras en un documento como el orden en que aparecen los documentos en el corpus son ajenos al modelo. Mientras que se supone que las palabras son intercambiables, en DTM, el orden de los documentos juega un papel fundamental. Más precisamente, los documentos se agrupan por segmentos de tiempo donde si son intercambiables (por ejemplo, años, meses) y se supone que los documentos de cada grupo provienen de un conjunto de temas que evolucionó a partir del conjunto del segmento anterior.

DTM permite realizar un seguimiento de los K tópicos escogidos inicialmente (no incorpora nuevos temas), revelando como las palabras más relevantes de un tópico específico van cambiando en el tiempo y por ende hace un seguimiento de la trayectoria de las palabras en los tópicos. Este comportamiento lo puede capturar a través de la dependencia $\beta_{t,k}|\beta_{t-1,k}$ y $\alpha_t|\alpha_{t-1}$ que evolucionan con ruido gaussiano, la primera permite que la distribución de probabilidad sobre el vocabulario de cada tópico pueda cambiar de período a período, por ende las palabras más relevantes de un tópico podrían cambiar de un período a otro, mientras que la segunda permite que la mezcla de tópicos pueda cambiar de un período a otro, por ejemplo que un tópico se vuelva más o menos probable en el tiempo. El modelo gráfico de este proceso generativo es mostrado en la Figura 2.2.

El proceso generativo para el período t de un corpus secuencial sigue el siguiente proceso generativo:

1. Dibujar tópicos $\beta_t|\beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \sigma^2 I)$
2. Dibujar $\alpha_t|\alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$
3. Para cada documento d:
   (a) Dibujar $\eta_d \sim \mathcal{N}(\alpha_t, \delta^2 I)$
   (b) Para cada palabra n:
      i Escoger la asignación del tópico $Z \sim Mult(\pi(\eta_d))$
      ii Escoger una palabra $W_{t,d,n} \sim Mult(\pi(\beta_{t,z}))$

Notar que $\pi(n_d)$ es la parametrización para tener un simplex de dimensión K-1, es decir, la mezcla de tópicos para el documento d, de forma similar $\pi(\beta_{t,k})$ genera la distribución de probabilidad del tópico k en el segmento de tiempo t sobre el vocabulario, por ejemplo, $\pi(\beta_{k,t})_w = \frac{exp(\beta_{k,t,w})}{\sum_w exp(\beta_{k,t,w})}$.

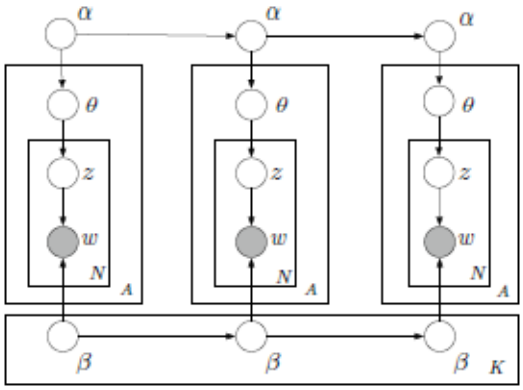

Figura 2.2: Representación gráfica de DTM para tres cortes en el tiempo. Cada tópico $\beta_{t,k}$ evoluciona sobre el tiempo junto a α_t para la mezcla de tópicos.

2.3. Métricas de evaluación

Uno de los grandes inconvenientes en el modelamiento de tópicos, es que el número de tópicos K es asumido a priori como conocido, por lo tanto, el curso de acción predominante en este caso, es encontrar el hiperparámetro K con mayor bondad de ajuste.

Algunas métricas usadas para estimar el número de tópicos (Griffiths and Steyvers, 2004; Cao et al., 2009; Deveaud et al., 2014; Arun et al., 2010; Blei et al., 2003; Wen Zhang et al., 2017) dado una serie de modelos de tópicos ajustados. Todos estos métodos requieren entrenar el modelo de tópicos múltiples veces sobre el mismo conjunto de datos para una serie de valores candidatos de K. En el *paper* original de LDA (Bleai et al., 2003), se sugiere calcular la *perplexity* (la inversa de la media geométrica de $logP(w_i's)$ en un conjunto de validación, valor que se debe minimizar. Griffiths and Steyvers (2004) [14], requieren el uso de *Gibbs sampler* para estimar los parametros de LDA,

y muestrear la posterior de *Gibbs sampler state* en intervalos regulares y escogen el K que maximiza la media armónica de la *log-likelihoods* muestreada. Cao et al. (2009) [15], estima la similitud coseno promedio entre las distribuciones de los tópicos $\overrightarrow{\phi_i}$, $\overrightarrow{\phi_j}$ $(i \neq j)$ y escoge el valor de K que minimiza esta cantidad. Arun et al. (2010) [16] minimiza la divergencia *Kullback-Liebler* simétrica entre los valores singulares de la representación matricial de las probabilidades de las palabras para cada tópico y la distribución del tópico dentro del corpus. Deveaud et al. (2014) [17] máximiza la distancia *Jensen-Shannon* promedio entre las distribuciones de tópicos $\overrightarrow{\phi_i}$, $\overrightarrow{\phi_j}$ $(i \neq j)$, muy similiar a Cao et al. (2009). Wen Zhang et al. (2017) [18] minimizar la suma de la entropía de los tópicos.

2.4. Interpretación de resultados

Los modelos de tópicos se caracterizan por tener un alto poder interpretativo, esto se debe a que la distribución de probabilidad de cada tópico sobre el vocabulario nos da una idea del tema al que pertenece este, por otro lado la mezcla de tópicos de cada documento muestra que tan importante es cada tópico en la generación de estos, como también dentro del corpus. En este sentido, las visualizaciones nos ayudan interpretar mejor los resultados de los modelos de tópicos, respondiendo las siguientes preguntas, ¿Cuál es el significado de cada tópico?, ¿Cuán predominante es cada tópico?, ¿Cómo se relacionan los tópicos entre sí?

Sievert, C., Shirley, K. (2014) , desarrollaron una herramienta de visualización para responder estas preguntas[1]. La herramienta a través de una visualización espacial responde la pregunta 2 y 3. Además para responder la pregunta 1 incorporan un gráfico de barras a la derecha del gráfico espacial que muestra las palabras más relevantes del tópico seleccionado dado un parámetro λ entre 0 y 1, entonces, la relevancia de la palabra w en el tópico k dado λ esta dada a través de la siguiente formula:

$$r(w, k|\lambda) = \lambda log(\phi_{k,w}) + (1 - \lambda) log(\frac{\phi_{k,w}}{p_w}), \lambda \in [0, 1]$$

Donde $\phi_{k,w}$ es la probabilidad de que el término w sea generado por el tópico k, p_w es la probabilidad de el término w en el corpus.

[1] `http://nbviewer.jupyter.org/github/bmabey/hacker_news_topic_modelling/blob/master/HN%20Topic%20Model%20Talk.ipynb#topic=3&lambda=0.46&term=`

3. Implementación

La base de datos utilizada corresponde a la base de datos de la AACH, *Robos_prose.csv*, donde se utilizaron los campos *sin_relato* el cual contiene el texto del relato de las víctimas de robo de vehículo y *sin_fecha_siniestro* que posee la fecha del robo, variable necesaria para segmentar el corpus en intervalos de tiempos con el objetivo de usar DTM. De la base de datos se consideraron los relatos desde el 2011 hasta el 2016, lo cual corresponde a 49.015 relatos.

El procesamiento realizado previó a utilizar DTM se resume en los siguientes pasos:

1. Aplicar un corrector ortográfico sencillo, a todos los relatos
2. Eliminar palabras que aportan poca información (*stopwords*)
3. Eliminar las palabras poco frecuentes con el fin de reducir el tamaño del vocabulario. Al eliminar *stopwords* el vocabulario posee 30.284 palabras, tras eliminar las palabras con una frecuencia menor a 10 el tamaño del vocabulario es de 5.431 palabras.
4. Discretizar el corpus en meses

Existen dos implementaciones disponibles para Python de DTM, DtmModel[2] y LdaSeqModel[3], la primera es un *wrapper* para Python de la implementación original hecha en C/C+ por David M. Blei y Sean M. Gerrish en 2015, por otro lado la segunda implementación es un esfuerzo por tener una implementación más pura en Python de la implementación original, por tanto es considerablemente más lenta que la primera (tarda casi 5 veces más), debido a esto se utilizó la primera implementación, además se ajusto el modelo para el siguiente conjunto de tópicos *K=2, . . . ,10* con el fin de encontrar el número óptimo de tópicos.

4. Resultados

Como se menciono en la sección anterior, se entrenó DTM sobre los relatos de la AACH desde el 2011 hasta el 2016, discretizando el corpus en meses, para *K=2, . . . ,10*, siendo el tiempo mínimo de entrenamiento de aproximadamente 25 minutos (*K=2*) y máximo de 60 minutos (*K=10*), siendo el tiempo total de ejecución de aproximadamente 7 horas. En la Figura 4.1 se observa que a medida que aumenta el número de tópicos a descubrir el tiempo de entrenamiento aumenta casi linealmente.

[2] `https://radimrehurek.com/gensim/models/dtmmodel.html`
[3] `https://radimrehurek.com/gensim/models/ldaseqmodel.html`

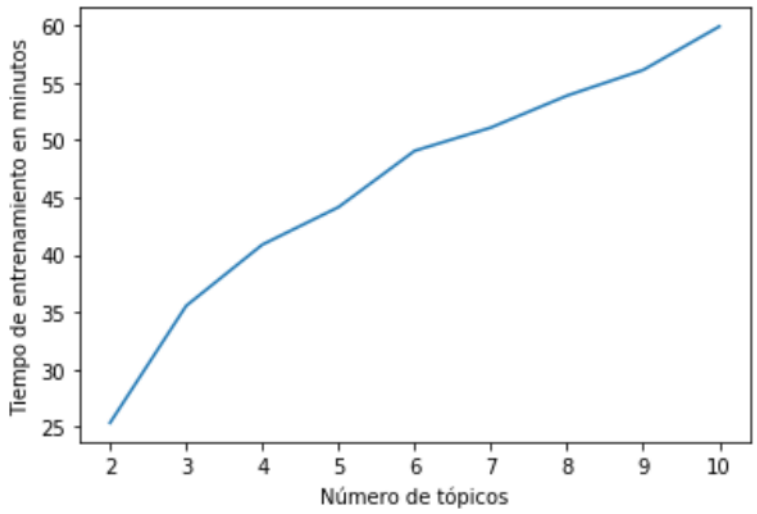

Figura 4.1: Tiempo de entrenamiento en minutos de DTM para *K=2, ... ,10*, se observa que el tiempo aumenta casi linealmente a medida que se incrementa el número de tópicos a descubrir.

Para encontrar el número óptimo de tópicos se utilizaron dos métricas, CaoJuan2009 (similitud coseno promedio entre tópicos) la cual se desea minimizar y Deveaud2014 (distancia *Jensen-Shannon* promedio entre tópicos) la cual se desea maximizar. En la figura 4.2 se observa que el número óptimo de tópicos es $K = 10$ donde CaoJuan2009 es mínima y Deveaud2014 es máxima.

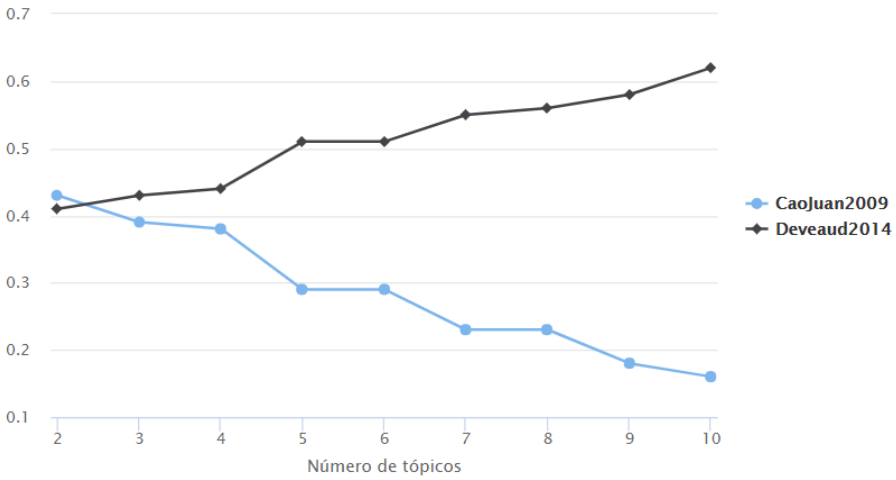

Figura 4.2: Resultados para distinto número de tópicos de CaoJuan2009 y Deveaud2014 para DTM.

Asignando cada documento d a su tópico más probable de haberlo generado usando θ_d se puede construir la serie mensual de cada tópico (véase Figura 4.3), de donde se observa que el tópico 1 y 9 son los más relevantes en el tiempo, en particular se tiene que el tópico 9 ha sido más o menos estable en el tiempo, pero por otro lado el tópico 1 ha sufrido un incremento importante en los últimos años, ya que desde el 2011 hasta el 2016 se ha prácticamente cuadriplicado (véase la Figura

4.4 para más detalle). Además se observa que el tópico 5 es un tópico que empezó a tomar relevancia a partir del 2013, probablemente se trata de un tópico que no existía desde el 2011.

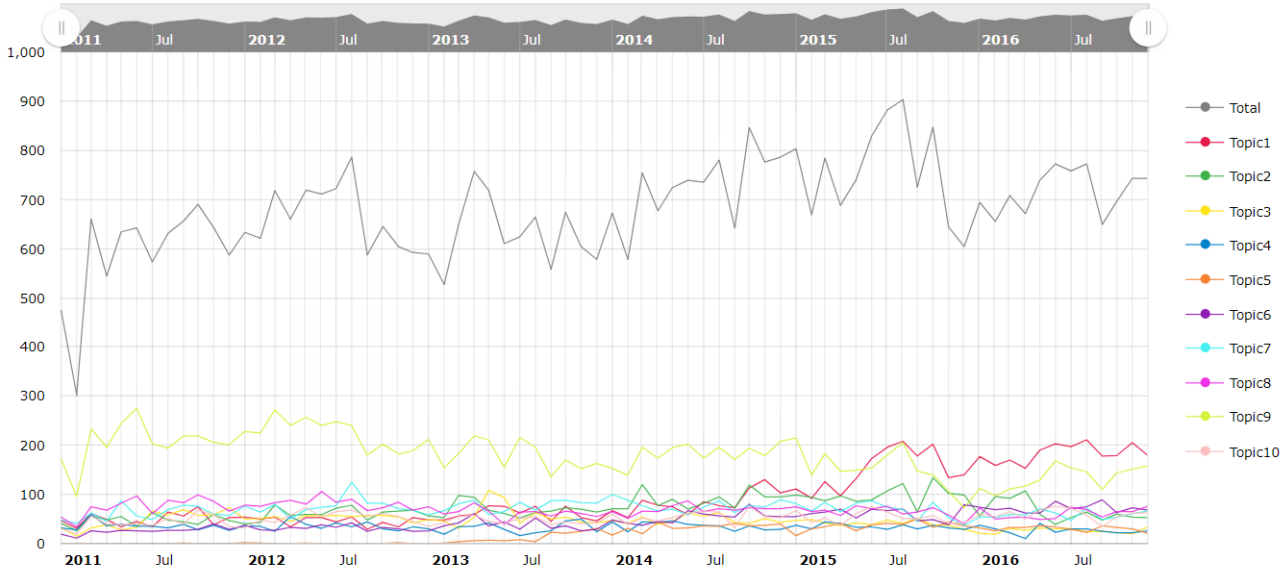

Figura 4.3: Serie mensual de los 10 tópicos desde el 2011 hasta el 2016.

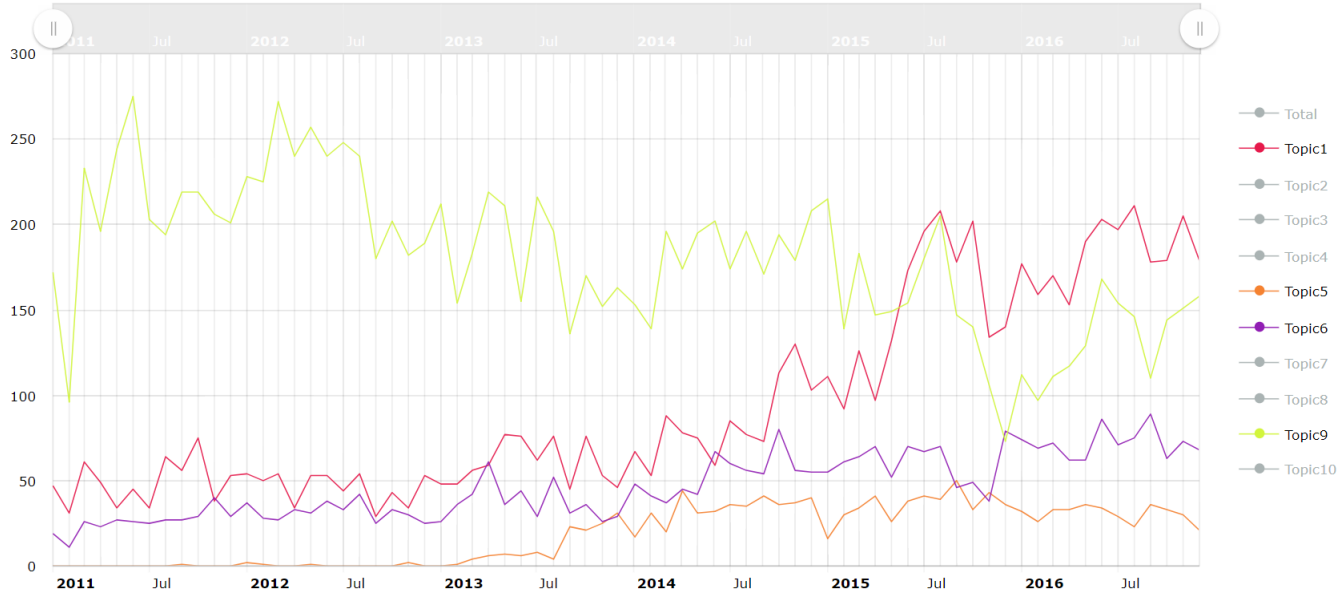

Figura 4.4: Serie mensual de los tópicos 1, 5, 6 y 9 desde el 2011 hasta el 2016.

Las siguientes visualizaciones se obtuvieron usando la libería LDAvis[4] analizando el més de enero del 2011, el gráfico de la izquierda es un mapa de distancia intertópico obtenido a través de escalamiento multidimensional considerando la distribución marginal de tópicos empleando los dos primeros componentes principales que se obtienen mediante PCA, el gráfico de barras del lado derecho corresponden a las Top-30 palabras más relevantes del tópico seleccionado.

En la Figura 4.5 lado izquierda se observa que los tópicos 1, 5 y 6 están bastante lejos del resto, esto se debe a que la mayoría de los tópicos hablan de delitos donde el dueño no se encontraba presente en el momento del suceso. Además, se tiene que las palabras más relevantes del tópico 1 parecen describir un robo con violencia, pues entre estas se hallan las palabras “arma”, “armas”, “fuego”, “pistola”.

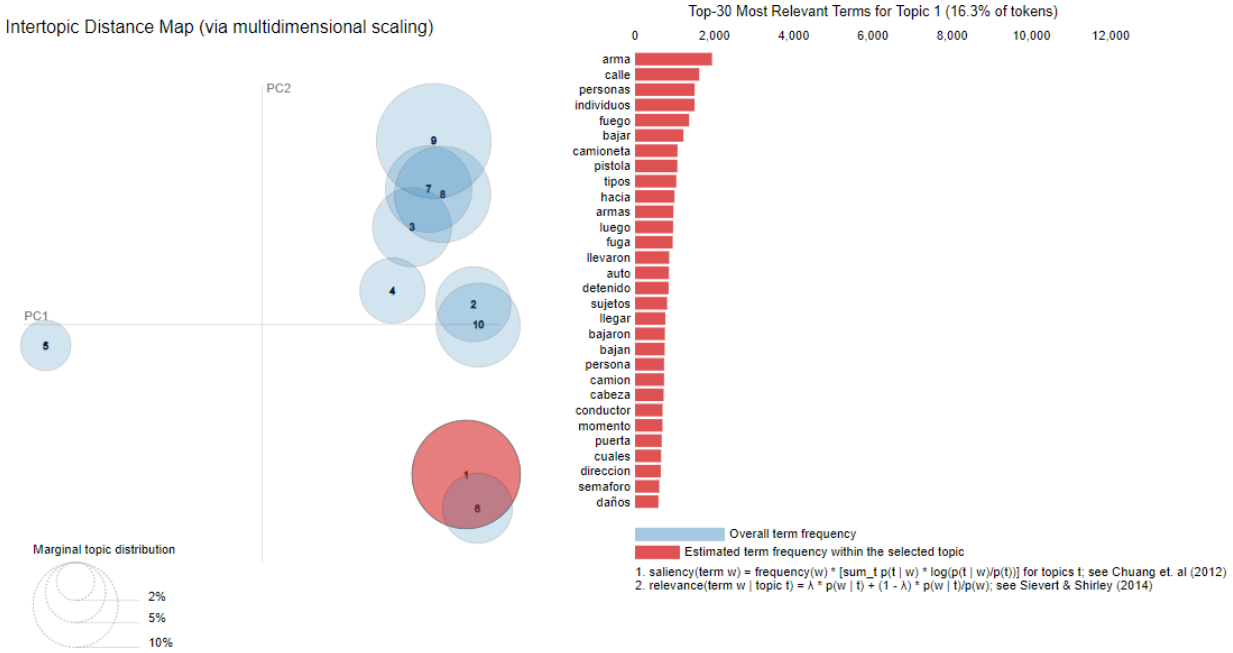

Figura 4.5: Lado izquierdo, distancia intertópico vía escalamiento multidimencional considerando la distribución marginal de los tópicos empleando el primer y segundo componente principal de PCA (PC1 y PC2). Lado Derecho, top-30 palabras más relevantes del tópico 1.

Tras seleccionar la palabra más relevante del tópico 1, la palabra “arma” se observa en el lado de izquierdo de la Figura 4.6 que el tópico 1 y 6 aumentaron de tamaño y el resto disminuyó hasta casi desaparecer, esto se debe a que la visualización captura la probabilidad de que un tópico genere la palabra seleccionada, en este caso, los tópicos 1 y 6 tienen una probabilidad significativa de generar la palabra “arma”.

[4] `https://github.com/bmabey/pyLDAvis`

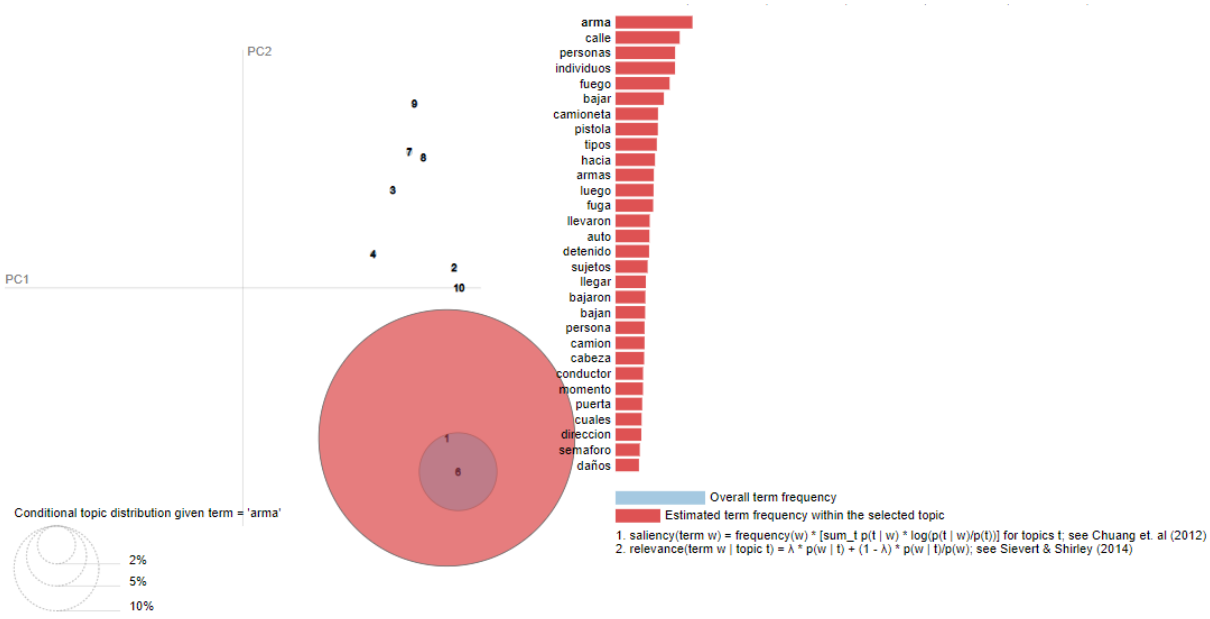

Figura 4.6: Lado izquierdo, distribución condicional sobre los tópicos de la palabra “arma”. Lado Derecho, top-30 palabras más relevantes del tópico 1.

En la Figura 4.7 al seleccionar la palabra “porton” del tópico 6 (el resultado es similar si se selecciona la palabra “domicilio”) se observa que el tópico 1 tiene probabilidad casi nula de generar esa palabra, sumado a las otras palabras relevantes del tópico 6 como “domicilio”, “porton” y teniendo en cuenta que una de las palabras más relevantes del tópico 1 es “calle”, es razonable pensar que el tópico 1 habla de asaltos realizados en plena calle y el tópico 6 de asaltos ocurridos cerca del domicilio de la víctima (conocidos como “portonazo”).

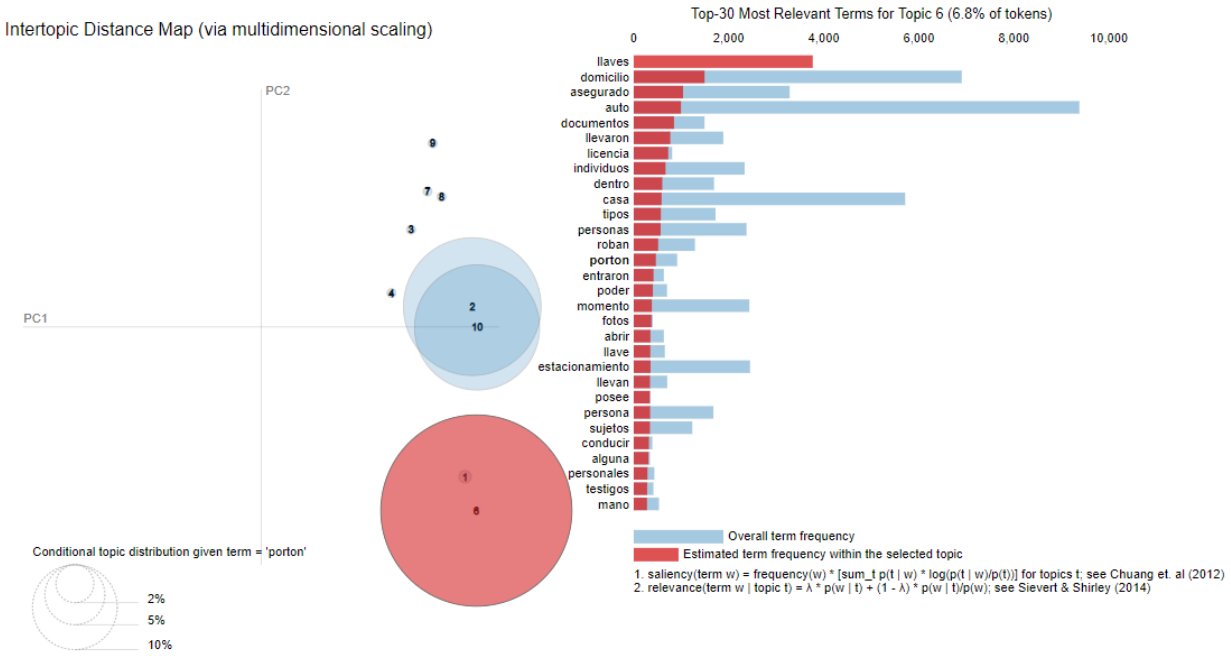

Figura 4.7: Lado izquierdo, distribución condicional sobre los tópicos de la palabra "porton". Lado Derecho, top-30 palabras más relevantes del tópico 6.

5. Conclusiones y Trabajo futuro

Del análisis preliminar se concluye que los robos con violencia han ido en aumento desde el 2011 y que hasta el 2016 se han prácticamente cuadriplicado. De los robos con violencia se identifican dos tipos, los robos con violencia realizados en plena calle y cerca del domicilio de la víctima ("portonazo").

LDAvis es una potente herramienta de visualización para entender los tópicos, pero tiene la desventaja de que solo genera visualizaciones de resultados por segmento de tiempo, por lo que un potencial usuario debería inspeccionar decenas de visualizaciones para tener un mejor entendimiento del fenómeno (debido a la discretización del corpus y estimaciones de las distribuciones de los tópicos para cada mes), en este sentido se podría incorporar una herramienta que levante alertas cuando un tópico sufre un cambio significativo en su distribución de un período a otro aprovechando los resultados que arroja el modelo (por ejemplo, si la divergencia *Jensen-Shannon* de un tópico con respecto al período anterior supera cierto valor).

La mayoría de los algoritmos de tópicos que abordan los documentos a lo largo del tiempo usan el mismo número de tópicos en todo el tiempo, por ende si surgen nuevos tópicos en el tiempo estos quedaran clasificados dentro de tópicos existentes desde el principio. Es razonable pensar que en el fenómeno de robo de autos aparezcan nuevos tópicos o modus operandi, o desaparezcan otros. En esta dirección existen dos líneas generales de trabajo:

1. Dividir el corpus en épocas, entrenar un modelo de tópicos estático en cada época, en base a

métricas definir el número óptimo de tópicos en cada época y luego hacen la conexión entre los tópicos de una época con la anterior [20].

2. Versiones dinámicas de Hierarchical Dirichlet Process (HDP)[21], versiones dinámicas de HDP se tienen en [22, 23], la base de estos modelos esta el proceso Dirichlet, el cuál permite que los documentos sean generados por mezclas de tópicos de diferente tamaño, es decir, el número de tópicos no es fijo y es encontrado a partir del corpus.

En un trabajo futuro sería interesante probar una versión dinámica de HDP debido a que modelan de forma natural el dinámismo en el número de tópicos en el tiempo permitiendo la aparición y muerte de tópicos o modus operandi en este contexto. Actualmente existe una implementación hecha en C+ por Olga Isupova y Danil Kuzin[5] 2016 a partir de una implementación escrita en C+ por Chong Wang 2010 de HDP.

[5] `https://github.com/OlgaIsupova/dynamic-hdp`

Referencias

[1] H. Jeong, Y. Yoo, K. M. Yi, and J. Y. Choi, "Two-stage online inference model for traffic pattern analysis and anomaly detection," Machine Vision and Applications, vol. 25, no. 6, pp. 1501–1517, 2014.

[2] J. Varadarajan and J. Odobez, "Topic models for scene analysis and abnormality detection," in Proceedings of the 2009 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops), Sept 2009, pp. 1338–1345.

[3] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009, pp. 935–942.

[4] T. Hofmann, "Probabilistic latent semantic indexing," in Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '99. New York, NY, USA: ACM, 1999, pp. 50–57.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, Mar. 2003.

[6] T. Hospedales, S. Gong, and T. Xiang, "Video behaviour mining using a dynamic topic model," International Journal of Computer Vision, vol. 98, no. 3, pp. 303–323, 2012.

[7] D. Kuettel, M. Breitenstein, L. Van Gool, and V. Ferrari, "What's going on? Discovering spatio-temporal dependencies in dynamic scenes," in Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2010, pp. 1951–1958.

[8] I. Pruteanu-Malinici, L. Ren, J. Paisley, E. Wang, and L. Carin, "Hierarchical Bayesian modeling of topics in time-stamped documents," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 6, pp. 996–1011, June 2010.

[9] C. Wang, D. Blei, and D. Heckerman, "Continuous time dynamic topic models," in Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-08). Corvallis, Oregon: AUAI Press, 2008, pp. 579–586.

[10] X. Fu, J. Li, K. Yang, L. Cui, and L. Yang, "Dynamic online HDP model for discovering evolutionary topics from Chinese social texts," Neurocomputing, vol. 171, pp. 412–424, 2016.

[11] C. Chen, N. Ding, and W. Buntine, "Dependent hierarchical normalized random measures for dynamic topic modeling," in Proceedings of the 29th International Conference on Machine Learning (ICML-12), ser. ICML '12, J. Langford and J. Pineau, Eds. New York, NY, USA: Omnipress, July 2012, pp. 895–902.

[12] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in Proceedings of the 23rd

International Conference on Machine Learning, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 113–120.

[13] A. Ahmed and E. Xing, "Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream," in Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10). Corvallis, Oregon: AUAI Press, 2010, pp. 20–29.

[14] T. L. Griffiths and M. Steyvers, "Finding scientific topics," Proceedings of the National academy of Sciences, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[15] J. Cao, T. Xia, J. Li, Y. Zhang, and S. Tang, "A density-based method for adaptive lda model selection," Neurocomputing, vol. 72, no. 7, pp. 1775–1781, 2009.

[16] R. Deveaud, E. SanJuan, and P. Bellot, "Accurate and effective latent concept modeling for ad hoc information retrieval," Document numérique, vol. 17, no. 1, pp. 61–84, 2014.

[17] R. Arun, V. Suresh, C. V. Madhavan, and M. N. Murthy, "On finding the natural number of topics with latent dirichlet allocation: Some observations," in Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 2010, pp. 391–402

[18] Zhang, W., Cui, Y., Yoshida, T. (2017). En-LDA: an novel approach to automatic bug report assignment with entropy optimized latent dirichlet allocation. Entropy, 19(5), 173.

[19] Sievert, C., Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces (pp. 63-70).

[20] Wilson, A. T., Robinson, D. G. (2011). Tracking Topic Birth and Death in LDA. Sandia National Laboratories.

[21] Teh, Y. W., Jordan, M. I., Beal, M. J., Blei, D. M. (2005). Sharing clusters among related groups: Hierarchical Dirichlet processes. In Advances in neural information processing systems (pp. 1385-1392).

[22] Ahmed, A., Xing, E. P. (2012). Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. arXiv preprint arXiv:1203.3463.

[23] Isupova, O., Kuzin, D., Mihaylova, L. (2016, July). Dynamic Hierarchical Dirichlet Process for abnormal behaviour detection in video. In Information Fusion (FUSION), 2016 19th International Conference on (pp. 750-757). IEEE.