Káñina, Rev. Artes y Letras, Univ. Costa Rica XL (Extraordinario): 33-47, 2016 / ISSN:2215-2636

EXTRACCIÓN DE TEMAS EMERGENTES EN MICROBLOGS UTILIZANDO MODELOS DE TEMAS Y DISCRIMINACIÓN DE BITÉRMINOS

EXTRACTION OF TRENDING TOPICS IN MICROBLOGS USING TOPIC MODELS AND DISCRIMINATION OF BITERMS

*Minor Quesada Grosso**

*Édgar Casasola Murillo***

*Antonio Leoni de León****

RESUMEN

La minería y explotación de datos contenidos en las redes sociales no solo ha sido foco de múltiples esfuerzos, sino que a pesar de los recursos y energía invertidos aún queda mucho por hacer dada su complejidad. Concretamente, el contenido de los textos publicados regularmente, en los sitios de *microblogs* (por ejemplo, en *Twitter.com*) puede ser utilizado para analizar tendencias. Estas últimas son marcadas por temas emergentes que se distinguen de los demás por un súbito y acelerado aumento de popularidad en periodos relativamente cortos, de un día o de unas cuantas horas. De este modo, el problema es extraer los temas sobre los cuáles se escribe e identificar cuáles de ellos son emergentes. Una solución reciente, conocida como *Bursty Biterm Topic Model (BBTM)* es un algoritmo que utiliza coocurrencias de palabras (bitérminos) para la identificación de temas emergentes y cuenta con un buen nivel de resultados en *Twitter*. Sin embargo, toma en cuenta todas las palabras, aún aquellas que no representan temas emergentes y por lo tanto, son menos útiles para identificarlos. De ahí, que esta investigación busca hacer una exploración inicial de la aplicación de una discriminación de los bitérminos utilizados por *BBTM* para modelar los temas emergentes.
Palabras clave: Temas emergentes, modelos de temas, detección de tendencias, redes sociales, procesamiento de leguaje temporal.

ABSTRACT

Mining and exploitation of data in social networks has not only been the focus of many efforts, but despite of the resources and energy invested, it still remains a lot of work given its complexity. Specifically, the content of the texts published regularly at sites of microblogs (as Twitter.com) can be used to analyze trends. The latter are marked by emerging topics that are distinguished from others by a sudden and accelerated

* Universidad de Costa Rica. Costa Rica. Correo electrónico: mqgrosso@gmail.com

** Universidad de Costa Rica. Escuela de Ciencias de la Computación, Programa de Posgrado en Computación e Informática y Centro de Investigaciones en Tecnologías de la Información y Comunicación (CITIC). Costa Rica. Correo electrónico: edgar.casasola@ucr.ac.cr

*** Universidad de Costa Rica. Profesor, Escuela de Filología, Lingüística y Literatura. Costa Rica. Correo electrónico: antonio.leoni@uc.ac.cr

Recepción: 15/1/2016. Aceptación: 16/3/2016.

increment of popularity. In this way, the problem is to extract the topics and identifying which of those topics are trending. A recent solution, known as *Bursty Biterm Topic Model* (*BBTM*) is an algorithm for identifying trending topics, with a good level of performance in Twitter. However, it takes into account all the words, including those that are not representative of the trending topics. For this reason, this investigation offers an initial exploration for the application of discrimination of biterms used by *BBTM* to modeling trending topics.
Key words: Trending topics detection, topic models, bursty topics, social media, natural language processing.

1. Introducción

Con la aparición de las redes sociales, tales como *Twitter*, múltiples esfuerzos se han realizado para lograr extraer y explotar la información contenida en ellas (Liu y Zhang 2012). El contenido de las publicaciones y comentarios que se realizan a cada momento en estos sitios puede ser utilizado para analizar tendencias en las poblaciones a nivel general. Estas tendencias son marcadas por temas emergentes que se distinguen de los demás por un súbito y acelerado índice de citas asociadas a un mismo tema; es decir, por un aumento repentino de popularidad en periodos relativamente cortos, de un día o de unas cuantas horas, por ejemplo (Wanner et al. 2014).

Estos análisis de tendencias son de gran importancia para investigadores, políticos y empresas, puesto que son una manifestación de los pensamientos, creencias, intenciones, opiniones y deseos de las personas. Por ejemplo, es posible seguir el hilo de noticias asociadas y observar su evolución con el paso del tiempo. También es posible saber cuales productos son populares en un momento determinado, además de las opiniones emitidas por los usuarios y clientes.

El problema de automatizar la identificación de estas tendencias consta de extraer los temas sobre los cuales se escribe e identificar cuáles de esos temas marcan tendencia. Una solución reciente, conocida como *Bursty Biterm Topic Model* (*BBTM*) (Yan et al. 2015), es un algoritmo para identificación de temas emergentes, con un buen nivel de resultados en Twitter. Sin embargo, este algoritmo trata a todos los vocablos[1] con la misma importancia, lo cual implica que debe procesar la mayor parte de los vocablos contenidos en cada comentario.

La investigación Xia et al. (2015) sugiere que es posible discriminar estos términos, eliminando aquellos menos importantes. De esta manera, es posible obtener los términos claves para identificar cada tema. Uno de los beneficios obtenidos al aplicar esta discriminación es la reducción en la cantidad de términos por procesar y por lo tanto, una disminución del procesamiento necesario.

BBTM basa su funcionamiento en los modelos de temas (*Topic models*), una técnica ampliamente utilizada para la extracción de temas en colecciones de textos (Liu 2012). Un modelo de temas es un modelo probabilístico que encuentra términos relacionados entre sí, que identifican los temas contenidos en una colección de textos.

El tiempo y memoria que le toma a *BBTM* obtener sus resultados depende directamente de la cantidad de términos utilizados. Por lo tanto, la importancia de la discriminación de términos radica en una disminución de tiempo y memoria.

Este artículo se enfoca en realizar una evaluación de la combinación de esta discriminación de términos con *BBTM*, como mecanismo para reducir la cantidad de bitérminos necesarios para identificar temas emergentes en Twitter. Para realizar esta discriminación, se propone crear un grafo a partir de la coocurrencia de los términos y aplicar el método introducido en Shetty y Adibi

(2005) para encontrar nodos importantes en un grafo. La idea sería entonces que *BBTM* se ejecute principalmente sobre aquellos términos que indiquen mayor importancia.

El resto del artículo está organizado de la siguiente forma: se inicia con un resumen del trabajo relacionado. Luego, se describe brevemente el algoritmo *BBTM*. Seguidamente, se introduce el método utilizado para la discriminación de términos. Después, se presentan los experimentos y sus resultados. Finalmente, las conclusiones.

2. Trabajo relacionado

2.1. Modelos de temas para textos largos

BBTM forman parte de la familia de métodos llamados modelos de temas (*topic models*). Estos métodos explotan la estructura semántica que se encuentra implícita en los textos para modelar los temas contenidos en esos textos.

Los modelos de temas fueron creados originalmente para extraer temas de textos largos, tales como artículos o noticias. Por ejemplo, *Latent Dirichlet Allocation* (*LDA*) (Blei, Ng, y Jordan 2003), es un modelo de tema que se ha usado ampliamente, debido a su capacidad de ser extendido para agregar nuevas funcionalidades.

2.2. Modelos de temas para textos cortos

El principal problema que presentan los textos cortos[1] presentes en los *microblogs*[2] es la poca densidad de palabras o esparcidad (*sparcity*). La esparcidad provoca que los métodos para textos largos no hallen suficiente concurrencia de palabras para encontrar similitud entre textos e identificar su tema (Phan, Nguyen, y Horiguchi 2008). Por tanto, se han tenido que buscar soluciones diferentes para abordar el procesamiento de textos cortos.

Una de las aproximaciones para tratar con textos cortos ha sido la utilización de datos externos para enriquecer su interpretación. Por ejemplo en Phan et al. (2008) se utiliza colecciones de documentos externos para aprender temas de ellos, utilizando LDA y luego, utiliza esos temas para ayudar a clasificar textos cortos. En Jin et al. (2011) se propone *Dual Latent Dirichlet Allocation Model* (*DLDA*), una versión de *LDA* que aprende temas de colecciones de documentos largos y de colecciones de textos cortos conjuntamente, permitiendo aprovechar la información presente en los textos largos para clasificar los textos cortos.

Biterm Topic Model (*BTM*) presentado en (Cheng et al. 2014) logra tener buenos resultados procesando textos cortos, definiendo un modelo capaz de abordar el problema sin necesidad de preprocesamientos. Consecuentemente, otros trabajos se han basado en *BTM* extendiéndolo para abarcar distintas tareas. Por ejemplo, en (Zhu et al. 2015) se modela la evolución de temas a través del tiempo en *microblogs* (como *Twitter*) utilizando *BTM*.

2.3. Extracción de termas emergentes

En Xia et al. (2015) clasifica noticias en Twitter extendiendo a *BTM* al agregar discriminación de términos para utilizar sólo aquellos que son más representativos de cada noticia. El modelo resultante es *Discriminative Biterm Topic Model* (*d-BTM*).

Bursty Biterm Topic Model (*BBTM*) (Yan et al. 2015), es un modelo para identificar temas emergentes, también basado en BTM, que utiliza datos sobre la popularidad explosiva (*burstiness*) de los bitérminos como información durante el proceso de modelaje de temas.

Por otra parte, el tema de encontrar temas emergentes ha sido abordado de diferentes maneras. Por ejemplo en (Mathioudakis y Koudas 2010) se propone *TwitterMonitor*, un sistema que detecta temas emergentes Twitter identificando palabras claves emergentes y agrupándolas. En (Cataldi, Di Caro, y Schifanella 2010) también se detecta temas emergentes en Twitter, pero modelando el ciclo de vida de los términos para determinar aquellos que son frecuentes en un intervalo específico de tiempo. Por último en (Li, Sun, y Datta 2012) se crea *Twevent*, un sistema para detectar eventos en Twitter, identificando segmentos de

Tweets que sean frecuentes en una ventana de tiempo específica.

En comparación con los métodos anteriores para detectar temas emergentes, *BBTM* presenta las siguientes ventajas:

a) Al estar basado en *BTM* logra modelar de manera efectiva textos cortos superando el problema de baja densidad de palabras.
b) Al incorporar la información sobre la popularidad repentina de los términos, logra identificar temas emergentes de manera eficiente sin necesidad de heurísticas ni posprocesamiento.

A continuación se presenta una explicación breve del funcionamiento general del algoritmo *BBTM*.

3. Bursty Biterm Topic Model

BBTM fue creado para procesar textos cortos (Cheng et al. 2014). Esto quiere decir que extrae los temas, sean estos emergentes o no, de una colección de textos cortos, por ejemplo una serie de comentarios de *Twitter*. En la Figura 1 se muestran los pasos que sigue *BBTM* para obtener temas emergentes.

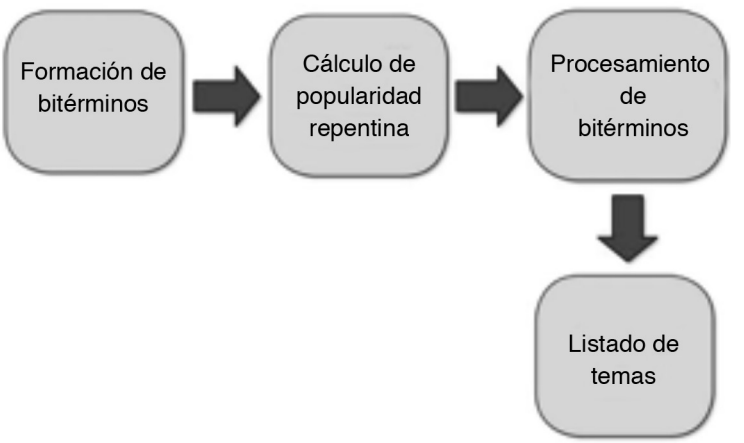


FIGURA 1

Pasos de BBTM para obtener temas emergentes a partir de una colección de textos cortos.

Como primer paso *BBTM* forma los bitérminos al tomar cada término en un texto y combinarla con todas las restantes palabras en el mismo texto. Un ejemplo de este proceso es mostrado en la Figura 2.

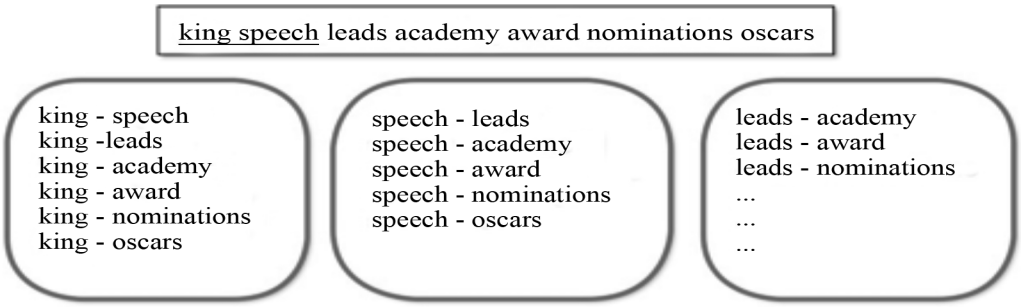


FIGURA 2

Formación de bitérminos a partir de una publicación de Twitter.

Para *BBTM*, cada conjunto de textos cortos está formado por una mezcla de temas, y cada tema es una distribución de probabilidad sobre las palabras. Por ello, una vez constituidos todos los bitérminos del conjunto de texto, se procede a encontrar la distribución de probabilidad que produce cada tema. En la Figura 3 se observa cómo distintas publicaciones de *Twitter* están asignadas a un conjunto de temas y como existe un conjunto de términos claves que identifican esos temas.

FIGURA 3

Para *BBTM* en una colección de publicaciones o documentos existen varios temas y cada uno de estos temas tiene un conjunto de términos claves que lo identifican.

Para encontrar la relación entre temas y términos, es necesaria una cantidad suficiente de muestras de estos términos. En los textos cortos se presenta el problema de escasez de términos en un mismo documento, impidiendo que los modelos de temas tradicionales funcionen adecuadamente. *BBTM* utiliza las siguientes dos estrategias para superar este problema (Cheng et al. 2014):

1. Utilizar bitérminos en lugar de términos. Esto permite que el modelo trabaje con la concurrencia de dos palabras a la vez, en lugar de sólo una.
2. Modela los temas tratando la colección completa de textos como si fuera un sólo documento. Esto solventa el problema que produce la poca cantidad de palabras en cada texto dentro de los métodos estadísticos.

Para descubrir la distribución probabilística de palabras para cada tema, *BBTM*, utiliza una técnica denominada Muestreo de *Gibbs*. Una técnica estocástica que permite aproximar la distribución de los bitérminos en cada tema. Entre más bitérminos se tengan que procesar, el muestreo de *Gibbs* tendrá más procesamiento que realizar. Por esta razón es importante reducir la cantidad de bitérminos. En la Figura 4 se encuentra un esquema simple de este paso.

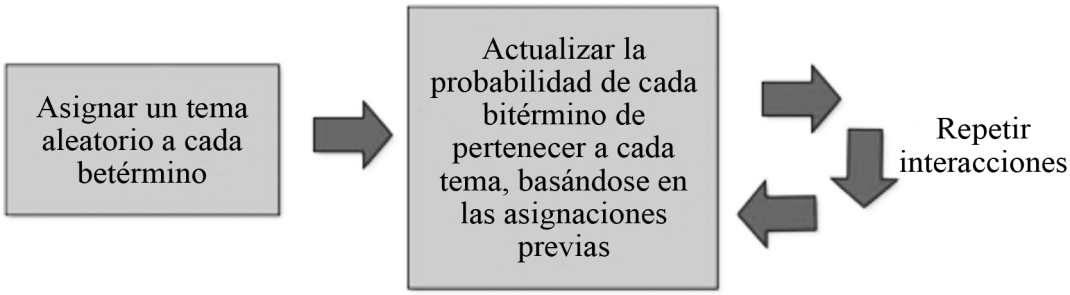


FIGURA 4

El muestreo de Gibbs permite aproximar la distribución de bitérminos por tema, a partir de una serie de muestras aleatorias.

Finalmente, cada tema es presentado como un conjunto de palabras clave. Estas palabras corresponden con las palabras que obtuvieron mayor probabilidad de pertenecer a cada tema. Por ejemplo, para el tema de los premios Oscar, un posible grupo de palabras clave sería *oscar, network speech, social, King, nominations.*

3.1. Identificación de bitérminos emergentes

BBTM debe calcular la probabilidad de que un bitérmino pertenezca a un tema emergente. Esta información es utilizada, luego, para aproximar la distribución de los temas emergentes. El cálculo de la probabilidad de que un bitérmino pertenezca a un tema emergente es el siguiente:

Sea X la frecuencia del bitérmino dentro un periodo de tiempo determinado (un día, por ejemplo). Sea Y el promedio de las frecuencias de ese bitérmino en periodos anteriores (en los 10 días anteriores, por ejemplo). La probabilidad N de ser emergente es:

$$(1) N = \frac{X-Y}{X}$$

En la Figura 5 se muestra un ejemplo de cálculo de probabilidad repentina.

Si la frecuencia del bitérmino en el periodo actual es repentinamente más alto, en comparación con el promedio de los periodos anteriores, significa que a sobrepasado su uso normal y se ha convertido en emergente.

BBTM entonces, descarta aquellos términos con probabilidad casi nula de ser emergentes. Los bitérminos con probabilidad alta de ser emergentes son los que terminan siendo asignados a un tema. Por consiguiente, si se desea aplicar alguna técnica de discriminación de términos para reducir el procesamiento necesario, esta debe asegurar que la mayor parte de los bitérminos emergentes sean conservados. A continuación, se presenta el método propuesto para realizar tal discriminación.

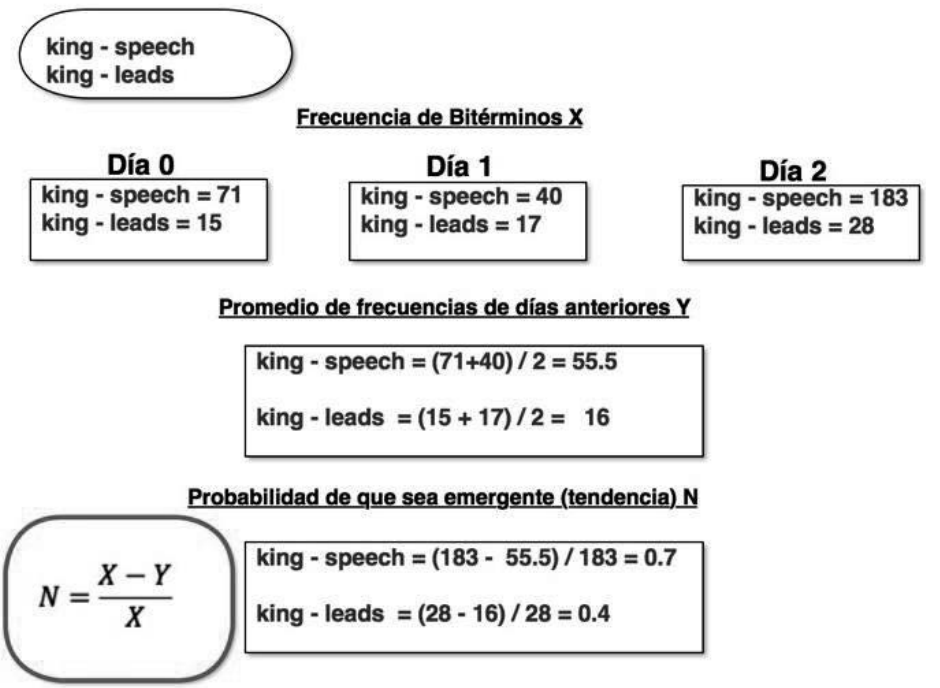


FIGURA 5

Cálculo de probabilidad repentina.

4. Discriminación de términos

Para realizar este proceso es necesario que los bitérminos sean extraídos, junto con su frecuencia y su probabilidad de ser emergentes. De esta manera, se tiene la suficiente información para discriminar los términos y así, procesar con *BBTM* aquellos bitérminos importantes.

4.1. Creación del grafo

Es posible crear un grafo no dirigido a partir de los bitérminos formados por *BBTM*. Cada término es tomado como un nodo con aristas o enlaces hacia los términos con los que forma bitérminos. En la Figura 6 se muestra un ejemplo.

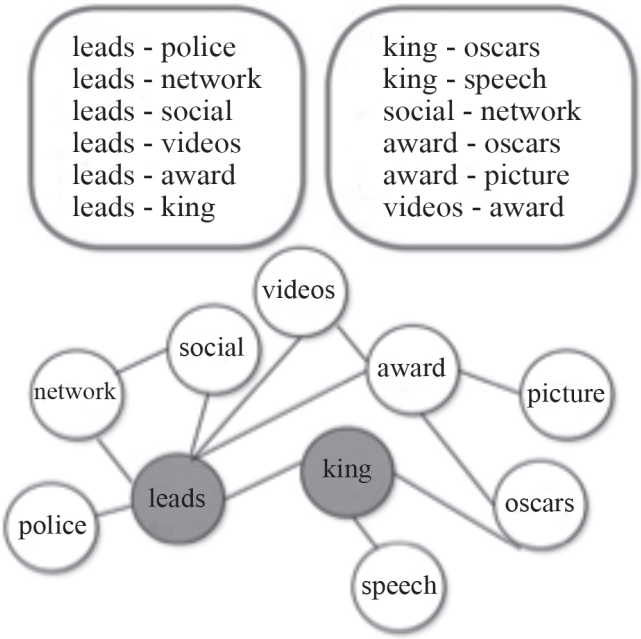


FIGURA 6

Formación de un grafo no dirigido a partir de un conjunto de bitérminos.

Si se creara un grafo a partir de todos los bitérminos encontrados en un conjunto formado por una cantidad considerable de textos, se terminaría con un grafo de tamaño también considerable. En consecuencia, realizar cálculos sobre ese grafo sería computacionalmente costoso.

Por la razón anterior, es preferible construir el grafo a partir de los términos contenidos en bitérminos con alguna probabilidad de ser emergentes, lo cual, reduce el grafo a un tamaño computacionalmente factible.

Una vez construido el grafo, se procede a detectar los nodos que representen términos que probablemente terminen dentro de las palabras claves de los temas emergentes.

4.2. Detección de nodos importantes

La idea es que los nodos más importantes son aquellos que, al ser eliminados del grafo, tengan mayor afectación sobre la entropía del grafo (Shetty y Adibi, 2005). Por ello, lo primero es definir el cálculo de la entropía para el grafo.

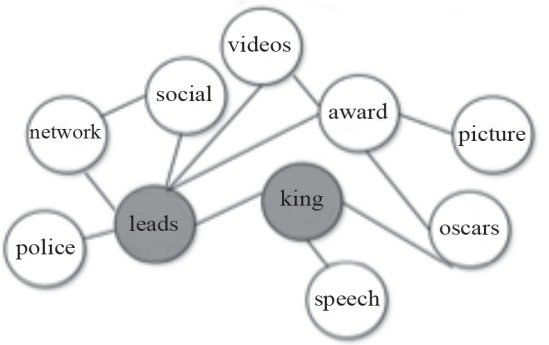
En términos de grafos, una alta entropía indica que muchas aristas son igualmente importantes, mientras que una baja entropía indica que solo unas pocas aristas son relevantes (Navigli y Lapata 2007). Entonces, es posible calcular la entropía del grafo de la siguiente manera (Navigli y Lapata 2007):

$$H(G) = -\sum_{v \in V} p(v) log(p(v)) \qquad (2)$$

Donde es el conjunto de los nodos del grafo. La probabilidad puede ser calculada con:

$$p(v) = \frac{|A|}{2|E|} \quad (3)$$

Donde es la cantidad de aristas del nodo y es la cantidad total de aristas del grafo. En la Figura 7 y Figura 8 se muestran ejemplos de estos cálculos:



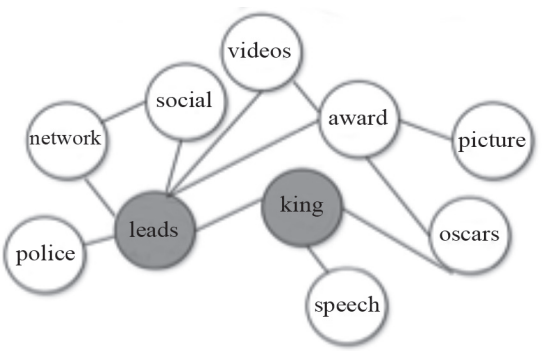
Entropía $H(G) = -\sum_{v \in V} p(v)log(p(v))$

$p(v) = \frac{|A|}{2|E|}$ $\quad p(leads) = \frac{6}{24} = 0.25$

$p(king) = \frac{3}{24} = 0.125$

FIGURA 7

Cálculo de la entropía de un grafo no dirigido. La probabilidad de ocurrencia de un nodo es calculada con base en cantidad de enlaces que posee.



$$H(G) = -\sum_{v \in V} p(v)log(p(v))$$

$$H(G) = 0.925$$

FIGURA 8

Resultado de calcular la entropía del grafo que se presenta como ejemplo.

Con la entropía definida, ahora se debe detectar cuáles nodos provocan mayor afectación sobre esta. Para ello, se puede realizar el siguiente algoritmo propuesto en (Shetty y Adibi 2005):

Para cada nodo N(i) hacer:

1) Computar la entropía del nodo N(i), calculando la entropía del nodo junto con sus vecinos inmediatos, como *E(i).* En la Figura 9 se muestra un ejemplo de este paso.
2) Eliminar temporalmente el nodo N(i) del grafo principal y calcular la entropía del grafo restante como *EN(i).* En la Figura 9 se muestra un ejemplo de este paso.
3) Calcular la entropía cruzada de la siguiente forma:

$$Importancia(i) = \frac{EN(i)}{\log(\frac{EN(i)}{E(i)})} \quad (4)$$

En la Figura 10 se muestra un ejemplo de este paso.

Ordenar los nodos según su *(i)* obtenida.

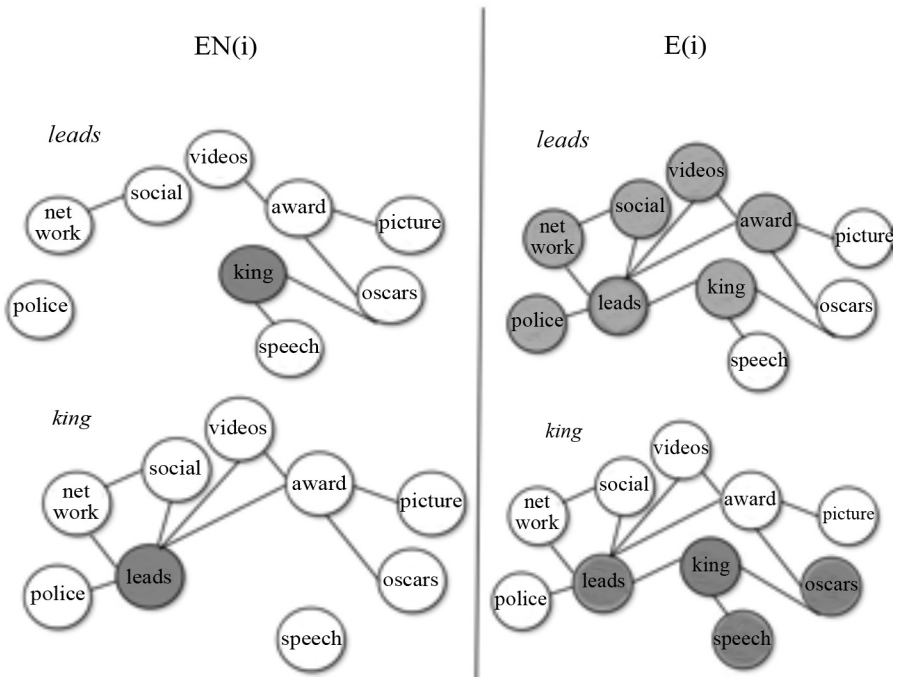


FIGURA 9

Para calcular la importancia de un nodo, es necesario calcular la entropía del grafo sin él y la entropía de él junto con sus vecinos inmediatos.

$$importancia = -\frac{EN(i)}{\frac{\log(EN(i))}{E(i)}}$$

$$EN(leads) = 0.58 \quad E(leads) = 0.72$$
$$EN(king) = 0.71 \quad EN(king) = 0.41$$

$$importancia(leads) = 3.40$$
$$importancia(king) = 11.87$$

FIGURA 10

Ejemplo de cálculo de importancia de dos término.

4.3. Discriminación de términos

La discriminación de los términos se lleva a cabo basándose en el grado de importancia obtenido en el proceso anteriormente descrito. Para esto se define un umbral , de tal forma que cualquier término que haya obtenido una importancia mayor o igual a se clasifica como importante. Aquellos que, por el contrario, tengan un valor menor a , se toman como términos que probablemente no aparezcan como palabra clave de ningún tema.

Utilizando los resultados obtenidos con (4) se realiza el proceso de reducción de bitérminos como sigue:

a) Para cada término cuya importancia sea menor que , se deben eliminar todos los bitérminos en los que esté contenido, excepto aquel que tenga la mayor probabilidad de correlación.
b) Para cada término cuya afectación sea mayor que , se deben mantener todos los bitérminos formados por términos que también hayan superado .

Finalmente, los bitérminos resultantes del proceso anterior son los únicos procesados por *BBTM*. En consecuencia, cada iteración de cálculos realizada por *BBTM* requiere menos procesamiento, en comparación a si se utilizara el conjunto completo de bitérminos.

5. Experimentos

Los experimentos preliminares se llevaron acabo sobre 3 días tomados del corpus *Tweets2011*, colección publicada en *TREC 2011 microblog track*[1]. No se aplicó lematización, se eliminaron los 100 términos más frecuentes y se eliminaron aquellos que aparecían en menos de 10 textos.

Para medir qué tan similares fueron los resultados de la versión original de *BBTM*, en comparación con la versión con discriminación de temas, se utilizará la siguiente medida de distancia:

$$Distancia = \frac{|W_1| - |W_1 \bigcap W_2|}{T \times K} \quad (5)$$

Donde W_1 es el conjunto de términos claves encontrados con la versión de *BBTM* original, W_2 es el conjunto de términos claves encontrados con la versión de *BBTM* con discriminación de términos. T es la Cantidad de términos por tema y K es la cantidad de temas encontrados por el algoritmo. En la figura 11 se muestra un ejemplo de este cálculo de distancia.

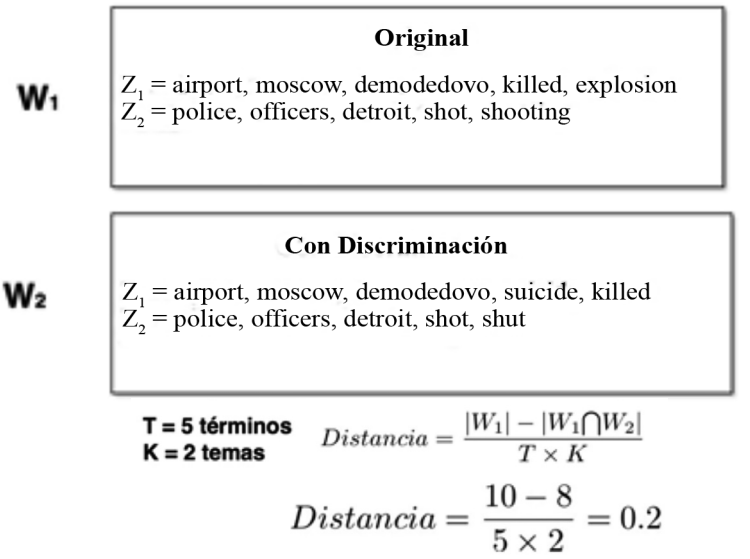


FIGURA 11

Ejemplo del cálculo de distancia entre el conjunto resultado de *BBTM* original y el conjunto resultado de *BBTM* con discriminación de términos.

5.1. Resultados preliminares

Ambos algoritmos fueron ejecutados durante 11 repeticiones. La cantidad de temas solicitados fue de 20 y con 10 términos clave cada uno. La cantidad de bitérminos utilizados por cada algoritmo también fue cuantificado.

En la Figura 12 y la Figura 13, se puede observar cómo la cantidad de bitérminos utilizados por *BBTM* y por *BBTM* con discriminación de términos tiene una diferencia de más del doble. Esto representa también una disminución sustancial del tiempo de procesamiento necesario.

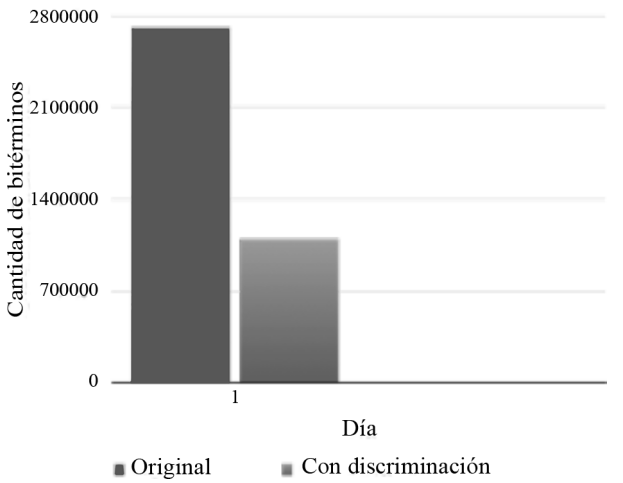


FIGURA 12

Cantidad de bitérminos utilizados por BBTM y BBTM con discriminación de términos durante el día 1.

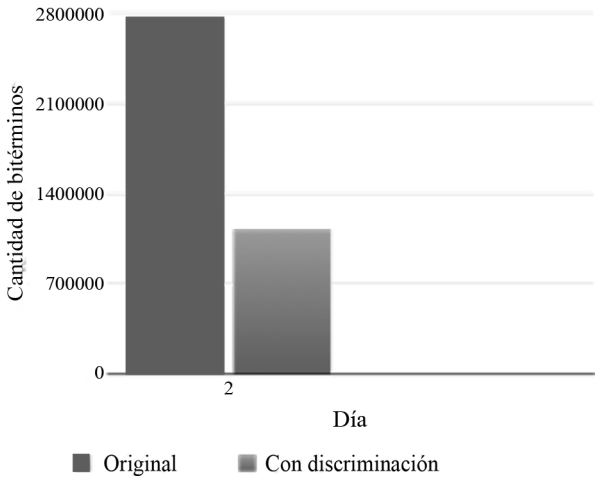


FIGURA 13

Cantidad de bitérminos utilizados por BBTM y BBTM con discriminación de términos durante el día 2.

En la Figura 14 y la Figura 15, se puede apreciar cómo pese a la diferencia de cantidad de bitérminos mostrada anteriormente, la distancia entre los resultados de ambos algoritmos es relativamente baja. Esto significa que ambos resultados comparten la mayoría de términos clave para cada tema.

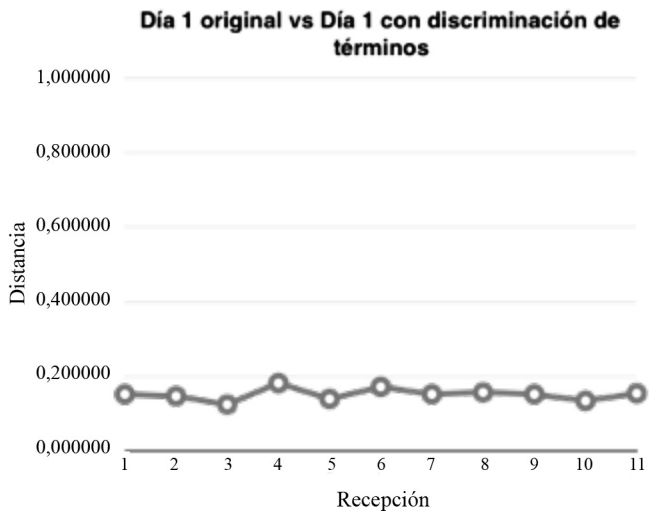


FIGURA 14

A lo largo de todas las repeticiones los resultados entre ambos algoritmos se mantuvieron similares.

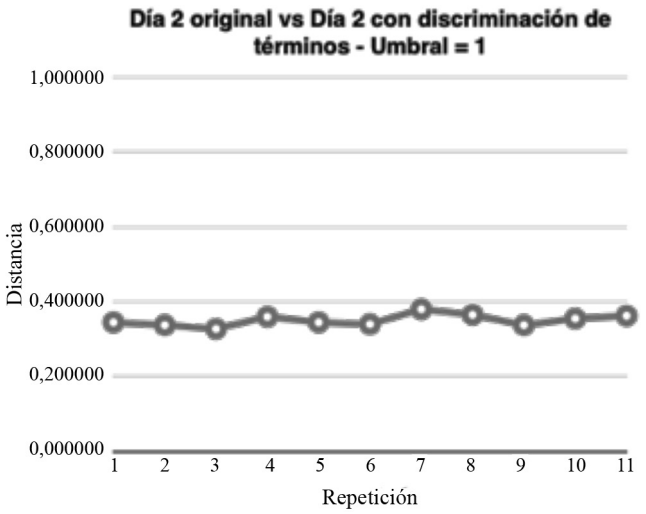


FIGURA 15

Aunque la similitud disminuyó en comparación con día 1, la distancia se mantuvo por debajo del 0.4. Por lo tanto, también se puede afirmar que ambos resultados comparten gran parte de los mismos términos claves para cada tema.

6. Conclusiones

Los resultados preliminares muestran que la adición de un proceso de discriminación de términos a *BBTM*, permitió reducir la cantidad de bitérminos manteniendo resultados similares a los obtenidos por la versión original de *BBTM*. También muestran que el grado de reducción de bitérminos y de similitud de los resultados, varía con la escogencia del umbral que separa los términos importantes de los que no.

De este modo, cobra sentido investigar más a profundidad los efectos de la discriminación de términos sobre la extracción de temas emergentes realizada por *BBTM*. Asimismo, este proceso de discriminación debe poderse ejecutar en un tiempo suficientemente corto y ser tener un consumo bajo de memoria, para justificar su adición, por tanto, su diseño también merece especial atención y cuidado.

Por consiguiente, es necesario realizar experimentos con diferentes valores del umbral de importancia, y con un corpus mucho más extenso, por ejemplo la colección completa de *Tweets 2011,* con el fin de obtener resultados que ayuden a tender el comportamiento de la combinación con un mayor nivel de detalle. Finalmente, una medición del efecto de la discriminación de términos en *BBTM* con respecto al tiempo de procesamiento, memoria utilizada y calidad de los temas, es parte del trabajo futuro de esta investigación.

Nota

1. URL: http://trec.nist.gov/data/tweets/.

Bibliografía

Blei, D. M. *et al*. 2003. «Latent Dirichlet Allocation». En *Journal of machine Learning research* III (Jan): 993–1022.

Cataldi, M. *et al*. 2010. "Emerging Topic Detection on Twitter Based on Temporal and Social Terms Evaluation". En *Proceedings of the Tenth International Workshop on Multimedia Data Mining :4*. New York, NY, USA: ACM.

Cheng, X. *et al*. 2014. "BTM: Topic Modeling over Short Texts". En *IEEE Transactions on Knowledge and Data Engineering* XXVI (12): 2928-2941.

Jin, O. *et al*. 2011. "Transferring Topical Knowledge from Auxiliary Long Texts for Short Text Clustering". En *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*: 775–784. New York, NY, USA: ACM.

Li, C. *et al*. 2012. "Twevent: Segment-based Event Detection from Tweets". En *Proceedings of the 21st ACM international conference on Information and knowledge management*: 155–164. http://doi.org/10.1145/2396761.2396785

Liu, B. 2012. "Sentiment Analysis and Opinion Mining". En *Synthesis Lectures on Human Language Technologies* V (1): 1–167. http://doi.org/10.2200/S00416ED1V01Y201204HLT016

Liu, B. y Zhang, L. 2012. "A survey of opinion mining and sentiment analysis". En *Mining text data*: 415–463. Springer US. http://doi.org/10.1007/978-1-4614-3223-4

Mathioudakis, M. y Koudas, N. 2010. "Twittermonitor: trend detection over the twitter stream". En *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*: 1155–1158.

Navigli, R., y Lapata, M. 2007. "Graph connectivity measures for unsupervised word sense disambiguation". *IJCAI International Joint Conference on Artificial Intelligence*: 1683–1688.

Phan, X. H. *et al*. 2008. "Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections". En *Proceeding of the 17th international conference on World Wide Web - WWW* :91–100. http://doi.org/10.1145/1367497.1367510

Shetty, Jitesh. y Adibi Jafar. 2005. "Discovering Important Nodes through Graph Entropy The Case of Enron Email Database". En *Proceedings of the 3rd international workshop on Link Discovery* 2005: 74–81.

Wanner, F. *et al*. 2014. "State-of-the-Art Report of Visual Analysis for Event Detection in Text Data Streams". En *Computer Graphics Forum* XXXIII (3) 1–15. http://doi.org/10.2312/eurovisstar.20141176

Xia, Y. *et al*. 2015. "Discriminative Bi-Term Topic Model for Headline-Based Social News Clustering". En *Florida Artificial Intelligence Research Society Conference*:311–316. Recuperado a partir de http://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS15/paper/view/10428

Yan, X. *et al*. 2015. "A Probabilistic Model for Bursty Topic Discovery in Microblogs". En *Twenty-Ninth AAAI Conference on Artificial Intelligence* : 353–359.

Zhu, J. *et al*. 2015. "Coherent Topic Hierarchy: A Strategy for Topic Evolutionary Analysis on Microblog Feeds". En *Web-Age Information Management IXXCVIII* 2015: 70–82. Springer International Publishing. http://doi.org/10.1007/978-3-319-21042-1