FACULTAD
DE
CIENCIAS

CATEGORIZACIÓN DE TEXTOS: Asignación de temas

(Text categorization (assigning subject categories, topics)

Informe trabajo de Semántica

MÁSTER DATA SCIENCE

Autores: Jose Ney Gandica Cardenas
Julia Ruiz Salmón

Abril-2019

Índice general

1. INTRODUCCIÓN

En la actualidad, el uso de tecnologías de la información como la televisión, el teléfono, el equipamiento informático o acceso a Internet, ha aumentado considerablemente.

En el año 2018 en España, el 86,1 % de la población entre 16 y 74 años ha utilizado Internet en los últimos tres meses. Este porcentaje ha aumentado en un 1,5 con respecto al del año 2017 según fuentes del Instituto Nacional de Estadística (INE) [1]. En ese mismo año en Cantabria, el 84,5 % de población había utilizado Internet en los últimos tres meses, frente a un 84,3 % en el año 2017, según el Instituto Cántabro de Estadística [2].

El uso de la tecnología, concretamente en el ámbito de Internet, es muy común. Sin embargo, la búsqueda, recopilación de información relevante y la forma no estructurada de artículos, documentos o noticias; propicia a invertir mucho tiempo de lectura e investigación debido a los múltiples resultados que devuelve una consulta realizada en Internet.

La categorización de textos (o clasificación de textos) , perteneciente a la rama de Minería de textos, es una tarea que pretende clasificar los documentos en temáticas predefinidas. La exploración de la temática ofrece facilidad y mayores posibilidades de encontrar lo que uno realmente está buscando.

Además, es una de las tareas en el procesamiento del lenguaje natural (PNL: comunicación, desarrollo personal y psicotearapia) que tiene sus aplicaciones en ámbitos como el análisis de sentimientos, la detección de spam o el etiquetado de temas. En este último es en el que nos centraremos.

Para ello, estudiaremos un *modelo probabilístico de temas* que se explicará en las siguientes secciones de manera más profunda.

Además, se abordará más profundamente los conceptos más relevantes de la minería de textos, así como la sección de categorización. Concretamente la clasificación en temas de los mismos.

2. MARCO TEÓRICO

2.1 Minería de textos

Actualmente, los avances tecnológicos han propiciado la aparición de nuevas fuentes de datos textuales: mensajes de texto, actividades en medios sociales, búsquedas en la web, etc.

Sin embargo, esta variedad de datos textuales no siguen una estructura estándar, son lo que se conoce como datos no estructurados. Por esta razón, se requiere de técnicas novedosas que ayuden a la interpretación y uso de este tipo de fuentes.

Por otro lado, el número elevado de publicaciones y textos, las tecnologías sofisticadas y el creciente interés en las organizaciones para extraer información del texto, ha propiciado el reemplazo del esfuerzo humano por el de máquinas automáticas. Estos sistemas automáticos se basan en un proceso de extracción de índices numéricos que son significativos del texto.

La Minería de textos se encarga de interpretar este conjunto de datos no estructurados y de hacer su uso lo más sencillo posible a través de métodos informáticos desarrollados para el análisis de textos. De esta manera, se podría destacar dos aspectos importantes del análisis de datos no estructurados:

1. La información que proporcionan sería tan útil como la de los datos estándar y se podrían procesar de la misma manera una vez aplicado el método correspondiente.
2. La información que proporcionan puede ser clasificada o agrupada. De esta manera, a diferencia de los datos estructurados, se conseguirían resultados tales como distribución de frecuencias de palabras, patrones de reconocimiento o, incluso, análisis predictivos.

En el tiempo actual, la Minería de textos se entiende como una combinación de diferentes ramas de conocimiento como son la Minería de datos, Inteligencia Artificial, Estadística, Gestión de Bases de Datos, Biblioteconomía y lingüistica.

En definitiva, la minería de textos proporciona información desde la evidencia y puede sustentar muchas tomas de decisiones en diferentes ámbitos como la empresa o la formulación de políticas [3].

2.1.1 Ámbitos de la Minería de textos

El conocimiento que ofrece esta disciplina es aplicado en múltiples ámbitos [4]. A continuación , se destacan diferentes tareas:

1. Análisis sentimental (sentiment analysis): detectar actitudes o ideologías a través de tweets.
2. Resumen de documentos (document summarization): intentar crear un resumen representativo o resumen de todo el documento, encontrando las frases más informativas.
3. Agrupación de documentos (text clustering): aplicar análisis cluster a documentos de textos.
4. Extracción de nombres de entidades (entity extraction): extraer información mediante un proceso de identificación y clasificación de elementos del texto en categorías predefinidas.

Transforma datos no estructurados en datos estructurados.

5. Modelo de relación entre entidades (entity-relationship model): intenta describir elementos que se definen como tipos de entidades y encontrar relaciones entre estas entidades en un campo específico.
6. Categorización de textos (text categorization): trata de clasificar los textos en grupos organizados a través de modelos.

En este ámbito es en el que se centrará el estudio y se hablará en la siguiente sección.

2.2 Categorización de textos

La categorización o clasificación de textos pretende aplicar una o más clases a un documento dependiendo de su contenido. Esta asignación de clases se realiza a través de la aplicación de un modelo.
Los modelos forman parte del aprendizaje automático y la tarea que ejecutan es establecer etiquetas a través de unos clasificadores predefinidos de forma automática.
Estos clasificadores se entrenan para realizar predicciones particulares para los textos y se basan en las siguientes características:

1. Definir un conjunto de etiquetas con las que el modelo trabajará.
2. Establecer asociaciones entre los trozos de texto y la etiqueta o etiquetas correspondientes.

Una vez que el proceso de etiquetado finaliza en suficientes textos, el clasificador puede aprender de esas asociaciones y comenzar a realizar predicciones con nuevos textos.

Algunos de los ejemplos en los que se utilizan los clasificadores y sus correspondiente conjunto de etiquetas son el análisis de sentimientos (mencionado anteriormente), detección del idioma , clasificación de productos (mediante su descripción) o asignación de temas y categorías.

En este último ejemplo es en el que se centrará este estudio: asignación de temas y categorías.

Sin embargo, existen otros muchos ejemplos donde se aplica la clasificación de textos como son la detección de spam, la identificación de autoría o la identificación de edad/género [5] [6].

2.3 Asignación de temas y modelo de temas

Anteriormente, se ha mencionado la gran cantidad de documentos, publicaciones y artículos disponibles en medios tecnológicos que se pueden analizar. El procesamiento del lenguaje natural (PNL) es un término basado en una serie de algoritmos que pueden llegar a procesar grandes

volúmenes de texto de manera automática.

Este tipo de métodos pueden ser supervisados y no supervisados. En los primeros, se requiere de una clasificación manual previa de una muestra de documentos. Posteriormente, este aprendizaje de clasificación manual serviría para conseguir asociaciones de palabras y se aplicaría el mismo sistema a una mayor cantidad de documentos. En los segundos, no supervisados, no se requiere de un entrenamiento previo que clasifique manualmente los documentos. Sin embargo, este tipo de algoritmos "aprenden" del uso de las palabras en los documentos que se examinan. De esta manera, se recogen pautas y, a su vez, una estimación del contenido de los documentos sin realizar una lectura directa del mismo. Entre los métodos no supervisados que se utilizan para el análisis de textos se encuentra el modelado de temas.

Estos métodos se encargan de examinar un conjunto de documentos, denominado *corpus*, a través de las palabras y conjunto de palabras (frases) que se encuentran en esos documentos. A partir del examen de estos elementos, el algoritmo aprende automáticamente los grupos de palabras que mejor describen o explican el contenido de los documentos, es decir, el conjunto de palabras que mejor caracterizan a esos documentos. Este grupo de palabras resultante es lo que representaría un tema del documento [7].

2.3.1 ¿Cómo funcionan los modelos de temas?

Los siguientes apartados explican el funcionamiento de los modelos de temas:

- Los documentos se basan en un cierto número de temas que se suponen fijos.
  - Los temas preceden a los documentos.
- Se evalúan las palabras dentro de los documentos y se intenta encontrar las agrupaciones de palabras que mejor se adapten o describan el conjunto de documentos (corpus) en base a la restricción anterior.
- Finalmente se obtienen dos resultados:
  1. Una matriz de palabras y temas: desglose de los temas en términos de su composición de palabras.
  2. Una matriz de documentos y temas: descripción de los documentos en términos de sus temas.
- Una palabra puede ser asignada a múltiples temas (en proporción) o asignada a un único tema.

Hay una variedad de algoritmos de modelado de temas de uso común, incluyendo la factorización de matriz no negativa, el *Latent Dirichlet Allocation* (LDA) y los modelos de temas estructura-

les [8].

A su vez, otro tipo de tarea relacionada con la asignación de temas, es la asignación de categorías a los textos. Esta asignación se refiere a la agrupación efectiva de contenidos similares bajo una etiqueta de clase o descriptor. En concordancia con *Harter, 1986*, los descriptores facilitan el trabajo del investigador ya que logran capturar el concepto intrínseco de un texto, lo que hace que búsquedas genéricas de información sean posibles a gran escala.
Las etiquetas, descriptores o metadatos informativos pueden estar agrupados por temas: economia, deportes, ciencia; por géneros: noticias, hogar, moda [9].

En la siguiente sección se estudiará la estructura y proceso del algoritmo LDA de manera más detallada. Del mismo modo, se expondrá un ejemplo de aplicación de forma ilustrativa.

3. APLICACIONES

3.1 Algoritmo LDA

La idea básica del algoritmo LDA es conseguir representar los documentos como una variedad aleatoria de temas latentes. Además, cada tema se caracteriza mediante una distribución de palabras.

En primer lugar, se le pide a un anotador que realice una asignación a cada tema de una o más etiquetas de clase o categorías, en función de sus palabras más probables.
Una vez realizado este primer paso, el algoritmo clasificaría un documento en función de las proporciones posteriores del tema y de las categorías de los temas.

Más generalmente, el algoritmo LDA consiste en:

1. Un modelo de temas no supervisado muy simple.
2. Extrae varios temas de un documento.
3. Cada tema tiene una correspondencia con una distribución sobre un vocabulario fijo. Esta distribución es la distribución de Dirichlet [10] que se basa en el Teorema de Bayes.
4. En primer lugar, se generan los temas. En segundo, los documentos.
    - Cuando se genera o escribe un documento, se comienza por tener una idea o temática de la que va a hablar el documento.
5. Consiste en un proceso generativo que:
    - Realiza muchas iteraciones.

- Prueba diferentes combinaciones y configuraciones.
- Aprende de cada iteración lo que funciona y lo que no funciona.

El proceso generativo del LDA consta de dos etapas:

1. Elige al azar una distribución sobre temas.
2. Para cada palabra en el documento realiza lo siguiente:
   - Aleatoriamente, elige un tema de la distribución sobre temas.
   - Aleatoriamente, escoge una palabra de la correspondiente distribución sobre vocabulario.

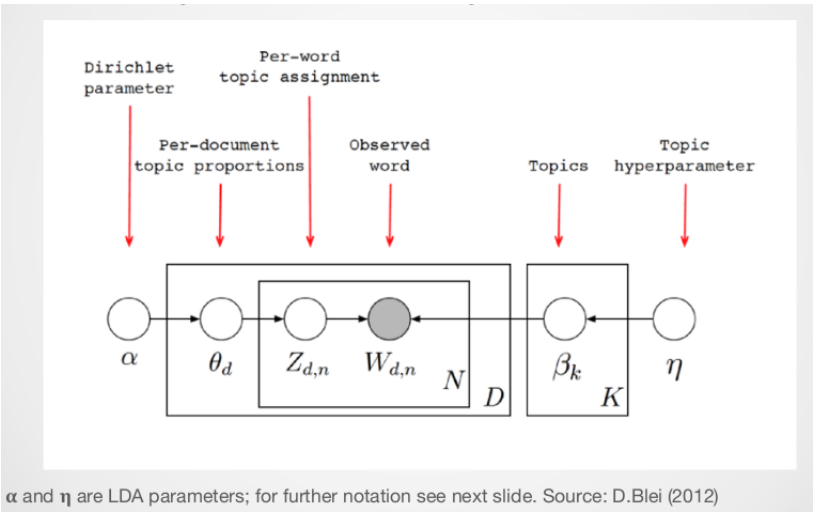

Figura 1: Esquema de diseño del documento a través de palabras y temas. Fuente: *D.Blei (2012)*

En la imagen, se puede observar que el algoritmo LDA se basa principalmente en la elección de dos parámetros: α y η.
Estos dos parámetros se encargan de regularizar los resultados del algoritmo de tal manera que,

- α establece el prior en la distribución de temas por documento. Un α elevado hace que cada documento esté representado por una mezcla mayor de temas, por el contrario un α pequeño hace que cada documento sea representado por una mezcla de pocos temas.
- η establece el prior en la distribución de palabras por tema. Un η elevado hace que cada tema pueda ser representado por una mezcla mayor de palabras, por el contrario un η pequeño hace que cada tema sea representado por una mezcla menor de palabras.

De este modo, existe una correlación entre ambos parámetros: cuando α y η son altos, los documentos y los temas son más similares entre sí. Y viceversa, cuando ambos parámetros son bajos, la similitud entre documentos y temas decrece.

Como se ha dicho anteriormente, la distribución de temas y de palabras está relacionado con la regla de Bayes. A continuación, se menciona *grosso modo* en qué consiste la regla de Bayes y cómo se relaciona con LDA.

3.1.1 Teorema de Bayes

La distribución de temas y de palabras en la que se apoya el algoritmo LDA es una distribución conjunta de probabilidades que depende de unos parámetros asociados a los temas y a las palabras del documento (ver anterior Figura 1). Esta distribución conjunta se puede expresar matemáticamente de la siguiente forma:

$$p(\beta_{1:K}, \delta_{1:K}, z_{1:K}, \omega_{1:K}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\delta_d) \Big(\prod_{n=1}^{N} p(z_{d,n}|\delta_d) p(\omega_{d,n}|\beta_{1:K}, z_{d,n}) \Big) \tag{3.1}$$

- $\beta_{1:K}$: temas
- $\delta_{1:D}$: proporción de temas por documentos
- $z_{1:D}$: tema asignado por palabra
- $\omega_{1:D}$: palabras observadas

El cálculo de esta distribución conjunta se basa en la regla bayesiana que dice que si A y B son dos eventos, entonces,

$$p(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3.2}$$

donde cada elemento de la expresión se refiere a un tipo de información sobre el ejemplo, es decir,

- $P(A)$: basado en información a priori.
- $P(A|B)$: basado en información a posteriori.
- $P(B|A)$: basado en información sobre el modelo estadístico también denominada verosimilitud.
- $P(B)$: basado en una probabilidad marginal.

3.2 Ejemplo

A continuación, se muestra de forma ilustrada y sencilla un ejemplo de asignación de temas a partir de un único documento y de la recopilación de temas en base a una distribución de palabras. En esta imagen, se puede ver como a través de una serie de palabras extraídas del

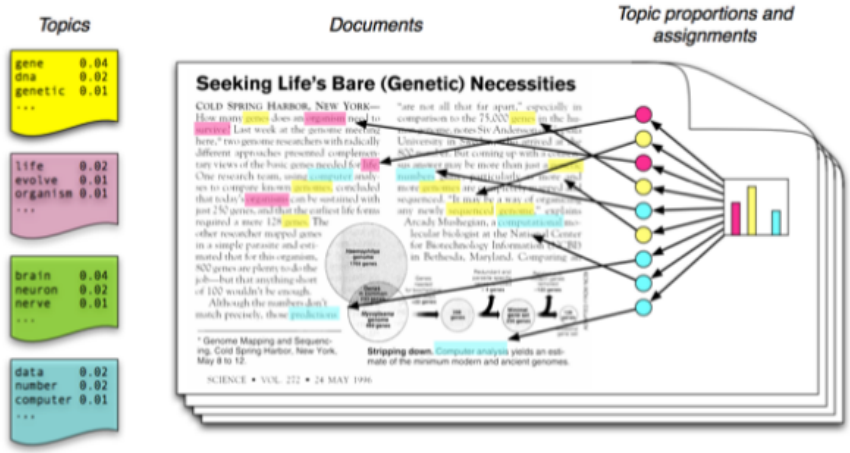

Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.

Figura 2: Ejemplo ilustrativo de un modelo de temas y basado en la asignación de temas. Fuente: *D.M .Blei (2012), 'Probabilistic topic models. Communications of the ACM', 55(4), 77-84.*

documento (y en base a una proporción de cada palabra dada por el algoritmo) se puede representar diversos temas que explican el contenido del mismo.
Los diferentes colores que se muestran se refieren a diferentes temas extraídos del documento: *genética* (amarillo), *vida* (rosa), *neurología* (verde) y *tecnología/informática* (azul).
Si se comparan los resultados con el tema o título del documento *Buscar las necesidades (genéticas) básicas de la vida* (recordar que la temática precede al documento; el documento se basa en una temática para lograr escribirse, etc..), se puede ver una correlación sencilla entre los cuatro temas que el algoritmo ha mostrado y el tema real del documento.

Este tipo de algoritmos, como ya se ha dicho anteriormente, permite al ser humano trabajar con una gran cantidad de documentos sin necesidad de realizar una lectura profunda de los mismos y establecer criterios de clasificación de textos muy útiles en diferentes ámbitos [8] [11].

Estas tareas de asignación de temas están a la orden del día y, a continuación, se mostrará su estado del arte.

4. ESTADO DEL ARTE

En este apartado se presenta de manera ilustrativa el estado de la técnica en un desarrollo reciente relacionado a la asignación de temas a partir de aprendizaje no supervisado.

En una publicación del 2016 de manera conjunta la universidad de *York, Reino Unido, Singapore Management University, Singapore y la Humboldt Universität de Berlin, Alemania* presentan **"Dynamic Topic Modelling for Cryptocurrency Community Forums"** [12]. Allí

mencionan el dinamismo actual de las cripto-monedas y el aumento de su uso y presencia para el canje de dinero en efectivo o bienes. Así como también es mencionada la tecnología de 'block chain' que sustenta el manejo descentralizado de este producto financiero.

Dada la alta volatilidad de las cripto-monedas y el hecho que no están reguladas o respaldadas por ningún ente financiero, es necesario recurrir a medios de información no estructurada como los foros de la comunidad de usuarios para extraer conocimiento e información de las tendencias y opiniones del mercado. A diferencia de las bolsas de valores donde los inversores suelen ser organizaciones financieras en el mundo de las cripto-monedas cualquiera puede participar. Esto da lugar a esquemas fraudulentos de inversión que una vez son descubiertos originan cambios significativos en el valor bajo el cual están cotizadas las cripto-monedas en el mercado.

Por tanto, este estudio recopila mensajes de los principales foros de cripto-monedas y le asigna a cada uno de ellos una estampa de tiempo. A partir del uso de modelos dinámicos de temas, minería de texto y aprendizaje no supervisado se proporciona un indicador de los esquemas fraudulentos, la relación de las opiniones y la evolución de los temas en los foros con grandes acontecimientos del mercado de cripto-monedas. Por su parte este estudio demuestra el potencial predictivo de las técnicas empleadas y la verificación de hipótesis.

Una de las fuentes de información fue *bitcointalk.org* para lo cual fue requerido el uso de 'web scraping'. El estudio estuvo centrado en múltiples sub-temas categorizados como "Bitcoin", "Economy", "Monedas alternativas", "Scam Accusations" entre otros. En total se recopiló información de aproximadamente 200 foros con casi 15 millones de publicaciones.

Entre las técnicas aplicadas se tuvo una variante de LDA propuesta por Blei and Lafferty (2006) en la cual para mantener el carácter temporal de los contenidos analizados se realizaban una serie de cortes discretos que asumían que los temas evolucionaban suavemente con un ruido gaussiano.

Los mayores retos relacionados a este estudio estuvieron relacionados con el pre-procesamiento de los datos dada la naturaleza informal de los foros de la comunidad de usuarios y la cantidad de contenido generada día a día. Se utilizó técnicas de *tokenization, stop-words*, se eliminaron palabras funcionales como verbos adjetivos y adverbios conservando solo los sustantivos y palabras extrañas. Así mismo se eliminaron palabras que aparecieron en menos de 10 documentos y aquellas que aparecieron en más del 10 % del total de documentos. Finalmente se creó un diccionario con casi 500k palabras de donde se extrajo un total de 10k palabras significativas. Una vez realizadas estas tareas se creó una 'sparce matrix' donde cada linea representaba un documento. Por su parte para conservar el carácter de temporalidad de los documentos se creó un archivo que contenía información de los cortes discretos.

Al aplicar el modelo se utilizó "50/kheuristic" de *Griffiths and Steyvers (2004)* [13] para definir el parámetro α con lo que se generaron los temas. Todos los modelos se ejecutaron en cortes discretos de una semana comenzando desde el 22 de noviembre de 2009 hasta el 06 de agosto de 2016. Cada tema es representado a través de una distribución de palabras para facilitar la compresión humana. A continuación se presenta un ejemplo de la tabla de temas/palabras generada:

Topic Number	Most Probable Words
1	value, gold, bar, dollar, rate, demand, interest, asset
2	business, casino, house, trust, gambling, run, strategy, player
5	government, control, criminal, law, study, regulation, state, rule
7	use, service, option, cash, good, spend, fiat, convert
12	account, payment, fund, card, paypal, party, merchant, credit
18	score, online, pay, shop, bill, product, purchase, phone
20	wallet, key, paper, computer, storage, code, data, secure
23	price, trade, market, trader, drop, volume, sell, stock
24	trading, term, hold, buy, pump, dump, earn, gamble
30	exchange, bitfinex, lesson, cryptocurrency, crash, platform, altcoins, popularity
32	investment, risk, invest, aim, impact, salary, making, way
33	year, altcoins, end, today, adoption, prediction, happen, trend
35	transaction, block, fee, chain, confirmation, hour, minute, hardfork
38	altcoin, company, loss, hack, scam, hacker, scammer, road
42	bank, system, security, fiat, banking, role, function, institution
45	ethereum, split, advantage, issue, side, change, fork, core
48	forum, post, topic, member, bitcointalk, thread, index, php
50	mining, miner, network, power, pool, cost, reward, electricity

Figura 3: Temas destacados del modelo de 50 temas en el subforo de discusión de Bitcoin desde 20160731 hasta 2016/08/06. Fuente: *Dynamic Topic Modelling for Cryptocurrency Community Forums*

A partir de las series temporales de temas se pudo observar la evolución de algunos términos como CPU a GPU a partir del 2010, donde se hizo extensivo el uso de estos equipos de hardware dada su potencia de calculo matricial. También se analizaron otros temas conocidos como el escandalo de *"Insolvency of the MtGox Bitcoin exchange in 2014"* [14], donde se perdieron miles de cripto-monedas y más de 60k cuentas de usuarios de bitcoin de su plataforma se vieron expuestas. En la siguiente imagen se puede observar el crecimiento en la tendencia del tema 38 relacionado con esta compañía:

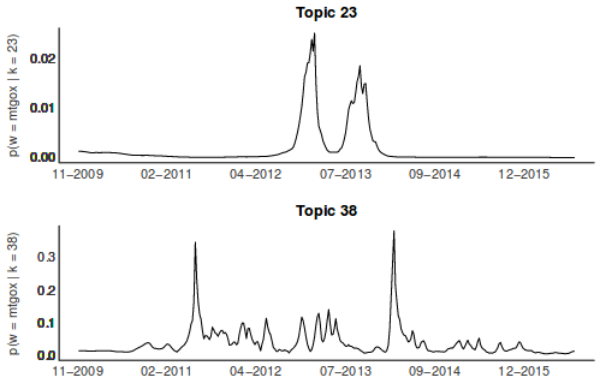

Figure 5: MtGox word evolution 22/11/2009 - 06/08/2016

Figura 4: Evolución de la palabra MtGox 22/11/2009-06/08/2016. Fuente: *Dynamic Topic Modelling for Cryptocurrency Community Forums*

Este estudio también llevó a cabo el análisis del valor óptimo de N (número de palabras por tema) que es un parámetro prior que el usuario debe seleccionar para ejecutar el modelo LDA. Para ello utilizaron *"Umass coherence metric" by Mimno et al. (2011)* [15]. Este método toma el top de N palabras para cada tema y lo hace en consideración a su ocurrecia y co-ocurrencia en el corpus. Queda definido como:

$$\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{D(w_i, w_j) + \epsilon}{D(w_i)} \tag{4.1}$$

Donde w_i y w_j son las palabras *i-th* y *j-th* del rank de un tema dado y $D(w)$ es el número de documentos en el cual esa palabra w occurre. Con esto descubrieron que N = 30 es un número adecuado de palabras que facilita la compresión humana y la coherencia de los temas en las ejecuciones de los cortes discretos.

Finalmente, el párametro usado para encontar temas de relevancia fue la volatilidad del precio del Bitcoin. Por tanto, para cortes discretos donde la pérdida de valor era inferior a 1000 Bitcoins fueron borrados. De los temas encontrados se utilizó la probabilidad de distribución de las palabras de cada temas y se estudiaron aquellos donde la relación de palabras no marcaba un tema de fácil de compresión para el humano y fueron considerados como ruido y eliminados.

4.1 Caso práctico: asignación de temas y categorías

Por último, se ha realizado un ejemplo de clasificación y asignación de temas y categorías para poner en práctica lo aprendido anteriormente. A través del algoritmo LDA se ha realizado una clasificación de diferentes textos según su temática en diferenes ámbitos como la economía, la ciencia y el deporte. Este ejercicio se encuentra en el repositorio de *Github*: https://github.com/rsj9999/semantica [16].

5. CONCLUSIONES

En este informe se hace una breve aproximación a la minería de textos y sus ámbitos relacionados, tales como el análisis de sentimientos, resumen de documentos, agrupación de textos, extracción de nombres de entidades, modelo de relación entre entidades y categorización de textos. En el mismo orden de ideas se presentan las principales técnicas de tratamiento de datos no estructurados y los algoritmos más utilizados en machine learning para la categorización de textos. Particularmente se hace hincapié en la clasificación de documentos a partir de aprendizaje no supervisado utilizando el algoritmo LDA (Latent Dirichlet Allocation).

Por otra parte no se realiza solamente una aproximación teórica de conceptos de la minería de textos, sino que también se presenta el estado de la técnica a través del análisis de una publica-

ción denominada “Dynamic Topic Modelling for Cryptocurrency Community Forums” donde se comenta el motivo de la investigación, los principales retos asociados, las técnicas implementadas y los resultados obtenidos.

Finalmente se realiza una aproximación práctica a la categorización de textos relacionados con el mundo de las noticias, específicamente con temas de ciencia, deporte y economía. Para obtener resultados coherentes, dado el pequeño corpus de noticias empleado, se obtuvieron todos los textos de la misma fuente, con un margen de publicación de escasos días con el objetivo de encontrar noticias de los diferentes temas con una relación temporal. Desde el punto de vista de los investigadores del trabajo, se logró unos resultados congruentes que demuestran las capacidades de la categorización de textos a partir de un modelo LDA.

Bibliografía

[1] INE. Instituto nacional de estadística. https://www.ine.es/ss/Satellite?L=es_ES&c=INESeccion_C&cid=1259925528782&p=1254735110672&pagename=ProductosYServicios%2FPYSLayout.

[2] ICANE. Instituto cántabro de estadística. https://www.icane.es/data/tic-product-use-type-product-from-2014.

[3] Vikas Dhawan Nadir Zanini. Text mining: An introduction to theory and some applications. RESEARCH MATTERS : ISSUE 19 / WINTER 2015 https://www.cambridgeassessment.org.uk/Images/466185-text-mining-an-introduction-to-theory-and-some-applications-.pdf.

[4] Cristina Tirnauca. Semántica, datos conectados y minería de datos textual minería de textos y minería web. Facultad de Ciencias: MCáster Data Science, Dpto Matesco, Universidad de Cantabria.

[5] What is text classification? https://www.meaningcloud.com/developer/text-classification/doc/1.1/what-is-text-classification.

[6] MonkeyLearn Raul Garreta. What is text classification? http://help.monkeylearn.com/text-classification/what-is-text-classification.

[7] Patrick van Kessel. An intro to topic models for text analysis. https://medium.com/pew-research-center-decoded/an-intro-to-topic-models-for-text-analysis-de5aa3e72bdb.

[8] Marco Linton Ernie Gin Swee Teo Elisabeth Bommes Cathy Yi-Hsuan Chen Wolfgang K. Hardle. Dynamic topic modeling for bitcoin message fora. *Ladislaus von Bortkiewicz Chair of Statistics, Sim Kee Boon Institute for Financial Economics, International Research Training Group, Humboldt-Universitat zu Berlin.*

[9] S. P. Harter. Online information retrieval: concepts, principles, and techniques. *Academic Press*, 1986.

[10] WIkipedia. Distribución de dirichlet. https://es.wikipedia.org/wiki/Distribuci\%C3\%B3n_de_Dirichlet.

[11] Michael I. Jordan David M. Blei, Andrew Y. Ng. Latent dirichlet allocation. https://es.wikipedia.org/wiki/Latent_Dirichlet_Allocation:http://jmlr.csail.mit.edu/papers/v3/blei03a.html.

[12] Marco Linton Ernie Gin Swee Teo Elisabeth Bommes Cathy Yi-Hsuan Chen Wolfgang K. Hardle. Dynamic topic modelling for cryptocurrency community forums. *University of York United Kingdom Singapore Management University Singapore Humboldt-Universitat zu Berlin, Germany*, November 24, 2016.

[13] Griffiths and Steyvers. 50/kheuristic. 2004.

[14] Mt. Gox. Insolvency of the mtgox bitcoin exchange in 2014. 2014.

[15] Mimno. Umass coherence metric. 2011.

[16] Susan Li. Modelado de temas y asignación de direccionamiento latente (lda) en python. https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24.