Extracción de conocimiento a partir de textos obtenidos de Twitter

Extraction of knowledge from texts obtained from Twitter

Ronny Adalberto Cortez-Reyes[1]
ronny.cortez@utec.edu.sv

Recibido: 22/02/18- Aceptado: 10/04/18

URI: http://hdl.handle.net/11298/451
DOI: http://dx.doi.org/10.5377/entorno.v0i65.6048

Resumen

El trabajo "Extracción de conocimiento a partir de textos obtenidos de Twitter" tiene como objetivo aplicar técnicas de *data mining* para, a partir de un conjunto de tuits, extraer información que permita conocer acerca de lo que se está hablando y así generar conceptos o ideas por medio del uso de diferentes tipos de representaciones gráficas.

Para el análisis se ha utilizado un conjunto de tuits, transmitidos del 1 de enero al 21 de febrero de 2018, relacionados con el tema de la inteligencia artificial. El proceso se dividió en tres fases principales, que incluyen: recolección de tuits, procesamiento de texto y visualización de resultados.

Utilizando diferentes tipos de gráficos, fue posible extraer información comprensible para los lectores, permitiendo tener una idea de los conceptos que se expresan en los textos y seleccionar las ideas principales.

Palabras clave

Lenguajes de procesamiento de texto. Archivos de texto. Visualización de la información. Recuperación de información. Sistemas de almacenamiento y recuperación de información. Tecnología de las comunicaciones. Telecomunicaciones – Innovaciones.

Abstract

The objective of "*Extraction of knowledge from texts obtained from Twitter*" is to apply data mining techniques to a set of tweets, in order to extract information to be able to find out what people are talking about and thus generate ideas or concepts, with the use of a variety of graphic representations.

A set of tweets generated between January 1 and February 21, 2018 has been used to conduct this analysis, and the topic is related to artificial intelligence. The process was divided in three main phases: tweets collection, text processing and display of results.

The use of a variety of graphs facilitated the presentation of comprehensible data for the reader; this allowed them to have an idea on the concepts being expressed in the texts and the selection of the main ideas.

Keywords

Text processing languages. Text files. Information display. Information retrieval. Information storage and retrieval systems. Communications technology. Telecommunications—Innovations.

[1] Investigador Utec

Introducción

En los últimos años se ha generado una gran cantidad de textos en formato digital en diferentes plataformas, como por ejemplo, redes sociales, correos, publicaciones científicas, foros, comentarios, periódicos, entre otros.

Es probable que continúe el aumento de la información en internet y el crecimiento continuo del acceso en todo el mundo y mediante diferentes tecnologías, y cambie la comunicación y la forma en que accedemos a los contenidos (Murphy & Roser, 2018).

Según el estudio elaborado por Open Broadcaster Software, el volumen de datos generados en 2014 se ha multiplicado. En un minuto, en internet se generan 4.1 millones de búsquedas en Google, se escriben 347 mil tuits, se comparten 3.3 millones de actualizaciones en Facebook, se suben 38 mil fotos a Instagram, se visualizan 10 millones de anuncios, se suben más de 100 horas de video a YouTube, se escuchan 32 mil horas de música en *streaming*, se envían 34.7 millones de mensajes instantáneos y se descargan 194 mil aplicaciones. En total, en un minuto se transfieren más de 1.570 terabytes de información.

El texto digital se ha convertido en una forma de intercambio de información y ha crecido a tal punto que cada vez es más difícil poder procesarla; no solamente localizarla rápida y eficientemente, sino también la extracción de conocimiento para ser utilizado en la toma de decisiones. Para solventar dicho problema, podemos usar la minería de textos, que consiste en el proceso de analizar colecciones de materiales textuales con el objeto de capturar los temas y conceptos clave, y descubrir las relaciones ocultas y las tendencias existentes sin necesidad de conocer las palabras o los términos exactos que los autores han utilizado para expresar dichos conceptos (IBM, s. f.).

Entre las redes sociales, los datos de Twitter constituyen una fuente rica que puede usarse para capturar información sobre cualquier tema imaginable. Estos datos se pueden utilizar en diferentes casos como encontrar tendencias relacionadas con una palabra clave específica, medir el sentimiento de la marca y recopilar comentarios sobre nuevos productos y servicios (Moujahid, 2014).

Minería de textos (MT)

En la actualidad, existe una gran cantidad de información digital en textos, por ejemplo, en periódicos, foros, redes sociales, correos electrónicos, revistas, libros y otros.

La proliferación del uso de dispositivos computacionales y de comunicación virtual para la producción de información digital, y en particular en la producción de documentos textuales, ha generado la necesidad de desarrollar métodos, algoritmos y sistemas capaces de realizar el procesamiento automatizado de datos textuales estructurados, semiestructurados y no estructurados, para su organización y consulta, y con ello el surgimiento de áreas de estudio de la información como la MT (Contreras Barrera, 2014).

La MT surge como un enfoque particular del proceso de descubrimiento de conocimiento, específicamente, orientado al descubrimiento en fuentes textuales y no estructuradas (Rodríguez Blanco, Cuevas, & J, 2013).

La MT comprende las siguientes tres actividades fundamentales: (Rochina, 2017)

- Recuperación de la información: consiste en seleccionar los textos pertinentes.
- Extracción de la información incluida en esos textos mediante el procesamiento del lenguaje natural: hechos, acontecimientos, datos clave, relaciones entre ellos, etc.
- Minería de datos para encontrar asociaciones entre los datos clave previamente extraídos de entre los textos.

Estas actividades se dividen en tres etapas fundamentales siguientes:

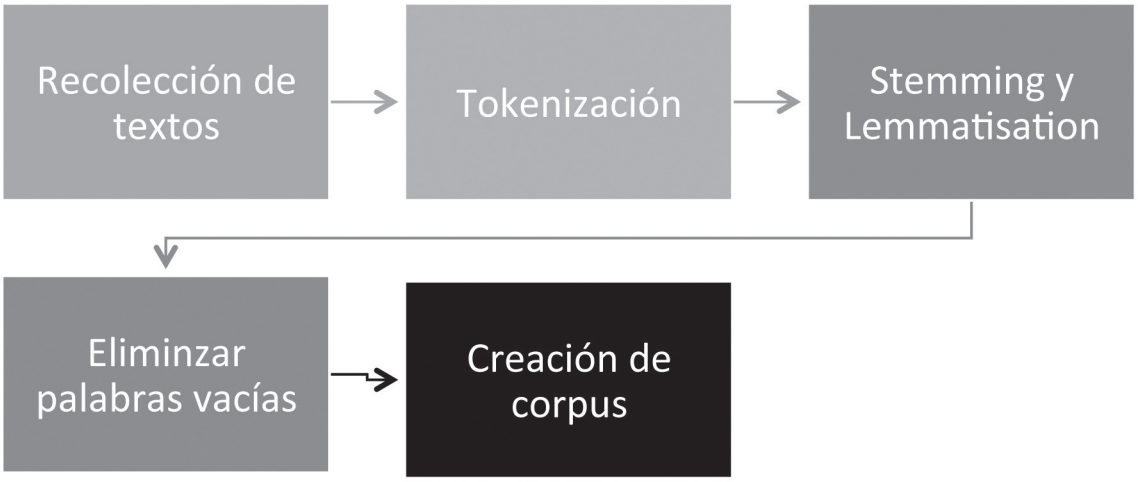
- Etapa de preprocesamiento: en esta etapa, los textos se transforman en algún tipo de representación estructurada o semiestructurada que facilite su posterior análisis. Es decir, el primer paso dentro de la MT sería definir el conjunto (*corpus*) de documentos. Estos documentos deben ser representativos y

seleccionarse aleatoriamente o mediante algún método de muestreo probabilístico. Se debe evitar, en esta etapa, la duplicación de documentos dentro del *corpus*. Una vez seleccionado y estructurado este, debemos reconocer los *tokens* (unidades gramaticales más pequeñas), lo que implica mostrar el texto como una lista de palabras mediante una representación vectorial.

- Etapa de descubrimiento: en esta etapa, las representaciones internas se analizan con el objetivo de descubrir en ellas algunos patrones interesantes o nueva información.
- Etapa de visualización: es la etapa en la que los usuarios pueden observar y explorar los resultados.

En la etapa de preprocesamiento de texto se debe realizar una serie de tareas, teniendo en cuenta que el orden en que se apliquen puede variar de acuerdo con nuestras necesidades; y no siempre se usarán todas. Una posible opción es la mostrada en la figura 1.

Figura 1. Fases del preprocesamiento. Fuente: elaboración propia.



Tokenización

Es el proceso de separar una cadena de texto en palabras, frases, símbolos u otros elementos significativos llamados *tokens*. El objetivo de la tokenización es la exploración de las palabras en una oración. La lista de *tokens* se convierte en entrada para el procesamiento posterior, como el análisis sintáctico o la MT. La *tokenización* es útil tanto en la lingüística (donde es una forma de segmentación de texto) como en la informática, donde forma parte del análisis léxico (Kannan & Gurusamy, 2014).

Stemming y lemmatization

Ambos son los métodos básicos de procesamiento de texto. Su objetivo es reducir las formas inflexionales, y, a veces derivadas, las formas relacionadas de una palabra con una forma de base común (TextMiner, 2014).

Palabras vacías (*stop words*)

Muchas palabras, en los documentos, se repiten frecuentemente, pero son esencialmente sin sentido, ya que se utilizan para unir palabras en una oración. Se entiende comúnmente que las palabras vacías no contribuyen al contexto o al contenido de los documentos textuales. Debido a su alta frecuencia de ocurrencia, su presencia en la minería de textos presenta un obstáculo para entender el contenido de los documentos (Kannan & Gurusamy, 2014).

Las palabras vacías son las más encontradas en cualquier lenguaje natural, que llevan muy poco o ningún contexto

semántico significativo en una oración; solo tienen importancia sintáctica en la medida que ayuda en la formación de la oración (K Raulji & R Saini, 2016). No existe una lista de palabras vacías única, pues puede variar de acuerdo con el idioma, área de trabajo o las palabras propias del momento en que se está trabajando con texto.

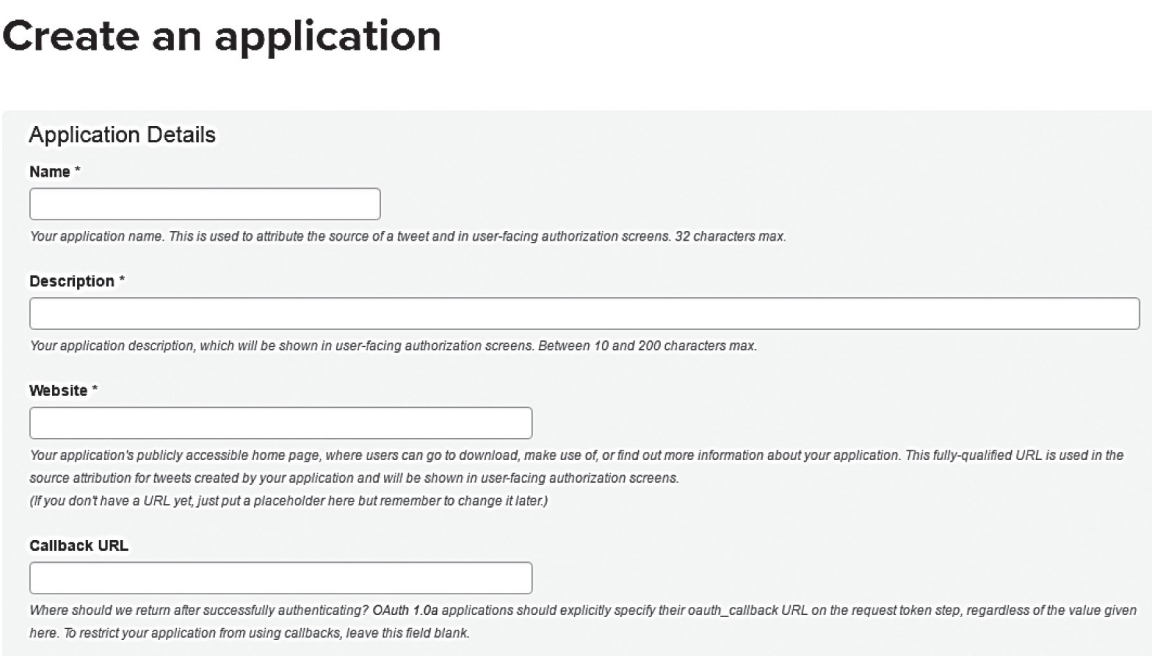
Método

Para la extracción de conocimiento, a partir de tuits, se dividió el proceso en tres fases principales: recolección de tweets, procesamiento de texto y visualización de resultados, todas ellas utilizando R, el cual es un lenguaje y un entorno para la informática estadística y los gráficos. Es un proyecto GNU que es similar al lenguaje S y el entorno que fue desarrollado en Bell Laboratories[3] y RStudio, el cual es un entorno de desarrollo integrado (IDE) para R; incluye una consola, un editor de resaltado de sintaxis que admite la ejecución directa de código, así como herramientas para trazar la historia, la depuración y la gestión del espacio de trabajo.[4]

Recolección de tuits

R cuenta con una serie de librerías que nos permiten recolectar tuits por medio de una API creada desde el sitio web para desarrolladores de tuits. Para crear la cuenta debemos seguir los pasos listados a continuación.

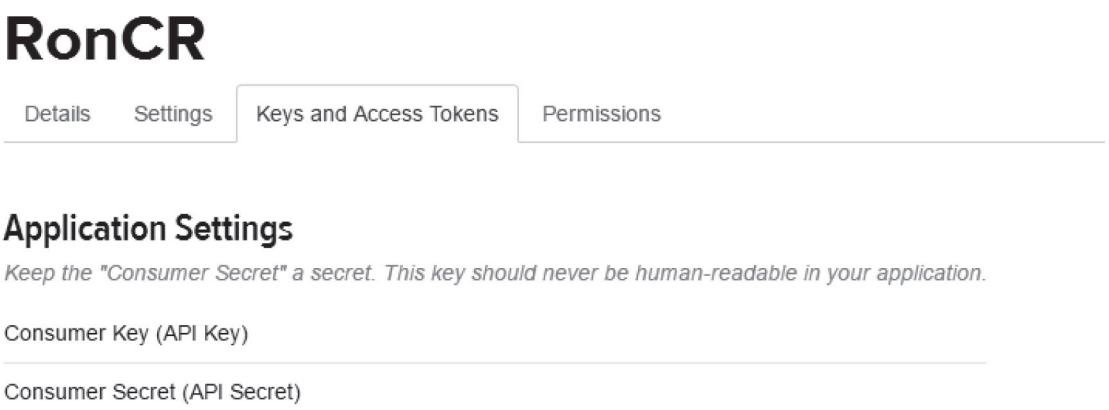
Figura 2. Datos necesarios para la creación de la API en el sitio web de desarrolladores de Twitter



1. Contar con una cuenta activa en Twitter a la cual previamente se haya agregado un número telefónico, pues es requisito para la creación de la API.
2. Ir a https://dev.twitter.com/ en la sección *My apps*.
3. Creamos una nueva aplicación llenando todos los campos obligatorios.
4. Obtenemos los datos que serán utilizados como conexión:
   a. *Consumer Key (API Key)*
   b. *Consumer Secret (API Secret)*
   c. *Access Token*
   d. *Access Token Secret*

[2] https://www.r-project.org/about.html
[3] https://www.rstudio.com/products/rstudio/

Figura 3. Acceso a las credenciales para ser utilizadas en R



Cuando ya tenemos los datos, podemos conectarnos y hacer una búsqueda de tuits con los siguientes parámetros:

1. *Search Terms*: el término de búsqueda que puede incluir varias palabras clave unidas con operadores booleanos; también se puede hacer una búsqueda por cuentas. Para nuestro análisis, se ha decidido utilizar la búsqueda por términos.
2. *n*: el número de tuits que queremos recuperar.
3. *lang*: define el idioma de búsqueda.
4. *geocode* = "*lat*, *lng*": define las coordenadas para buscar en un área específica.
5. *since* y *until*: para poner rangos en las fechas, tienen el formato yyyy-mm-dd.

Se han descargado un total de 10 mil tuits con el parámetro de búsqueda *Artificial intelligence* OR *#ArtificialIntelligence*, solamente los de idioma inglés; y a partir del 1 de enero de 2018, de este conjunto se eliminan los repetidos y se seleccionan los 300 más importantes ordenados por el número de retuit.

Procesamiento de texto

Debido a que los tuits vienen con una gran cantidad de caracteres que pueden no ser relevantes para su análisis, es necesario un proceso de limpieza. En este punto se hicieron varias pruebas hasta lograr que el texto quedase completamente libre de elementos innecesarios.

Entre los problemas más comunes, se incluye la presencia de URL, ya que algunas de ellas se encuentran incompletas, por lo que no todas eran detectadas, por lo que se hizo una combinación de ellas. Convertir todas las palabras a minúsculas también presentaba algunos errores, por lo que se utilizó una función que detectara los errores y los eliminara.

Ya que las palabras utilizadas por las personas al momento de redactar los tuits son muy variadas, no todas se incluyen en las listas disponibles en R, haciendo necesario agregar otras listas para completarlas y obtener mejores resultados.

Además, se hizo una limpieza propia del contenido de los tuits, eliminando

1. las entidades de retuits,
2. las @ y el texto irrelevante,
3. todos los símbolos no numéricos o que no estén en el idioma inglés, y
4. los *hashtags*.

Finalmente se eliminaron los números y los signos de puntuación y se sustituyeron por espacios en blanco. En caso de hacer búsquedas en español, además de tener una nueva lista de palabras vacías, es necesario eliminar tildes y caracteres especiales para que no haya problemas en la visualización de los resultados.

Figura 4. Librería utilizada para eliminar tildes y caracteres especiales del idioma español

```
#Just for Spanish text
library(stringi)
tweets.corpus <- stri_trans_general(tweets.corpus, "Latin-ASCII")
tweets.corpus <- Corpus(VectorSource(tweets.corpus))
```

No se implementó el *stemming* para la reducción de sus palabras a sus raíces, ya que en este proceso se puede perder información. En caso de que se quiera hacer una aplicación dedicada a la búsqueda de un tema en específico, se puede personalizar el proceso para ciertas palabras.

Resultados

Para la extracción de conocimiento a partir de texto, es necesario hacer uso de diferentes técnicas de representación gráfica. Por lo que, luego de que los tuits han sido procesados, se pueden presentar los resultados mediante gráficos de frecuencia de palabras, clúster o agrupamiento de palabras que comparten un conjunto de características, relaciones y creaciones de temas mencionados.

Frecuencia

El texto cuenta con una frecuencia variada de las palabras; algunas aparecen solamente una vez, y otras, más de cien veces, debido a esto vamos a representar solamente las 20 que más se repiten. No se ha definido un número mínimo de repeticiones, sino que se seleccionan aquellas que aparecen con mayor frecuencia.

Se han creado dos gráficos. En el primero se incluyen las palabras clave utilizadas para la búsqueda de los tuits; en el segundo no se han tomado en cuenta para tener una mejor percepción de las otras palabras que no se incluyen en la primera búsqeuda. Se debe tener en cuenta que las palabras clave, al ser utilizadas en la búsqueda, aparecen en todos los tuits, por eso su frecuencia es elevada y su importancia irrelevante.

Figura 5. Lista de palabras más frecuentes, incluyendo las palabras clave o de búsqueda

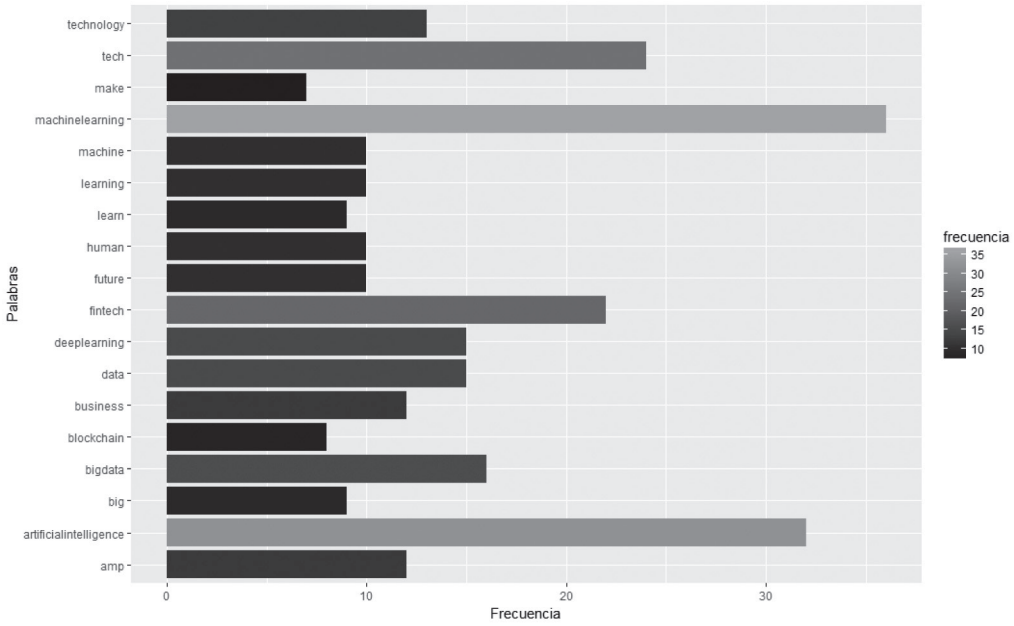
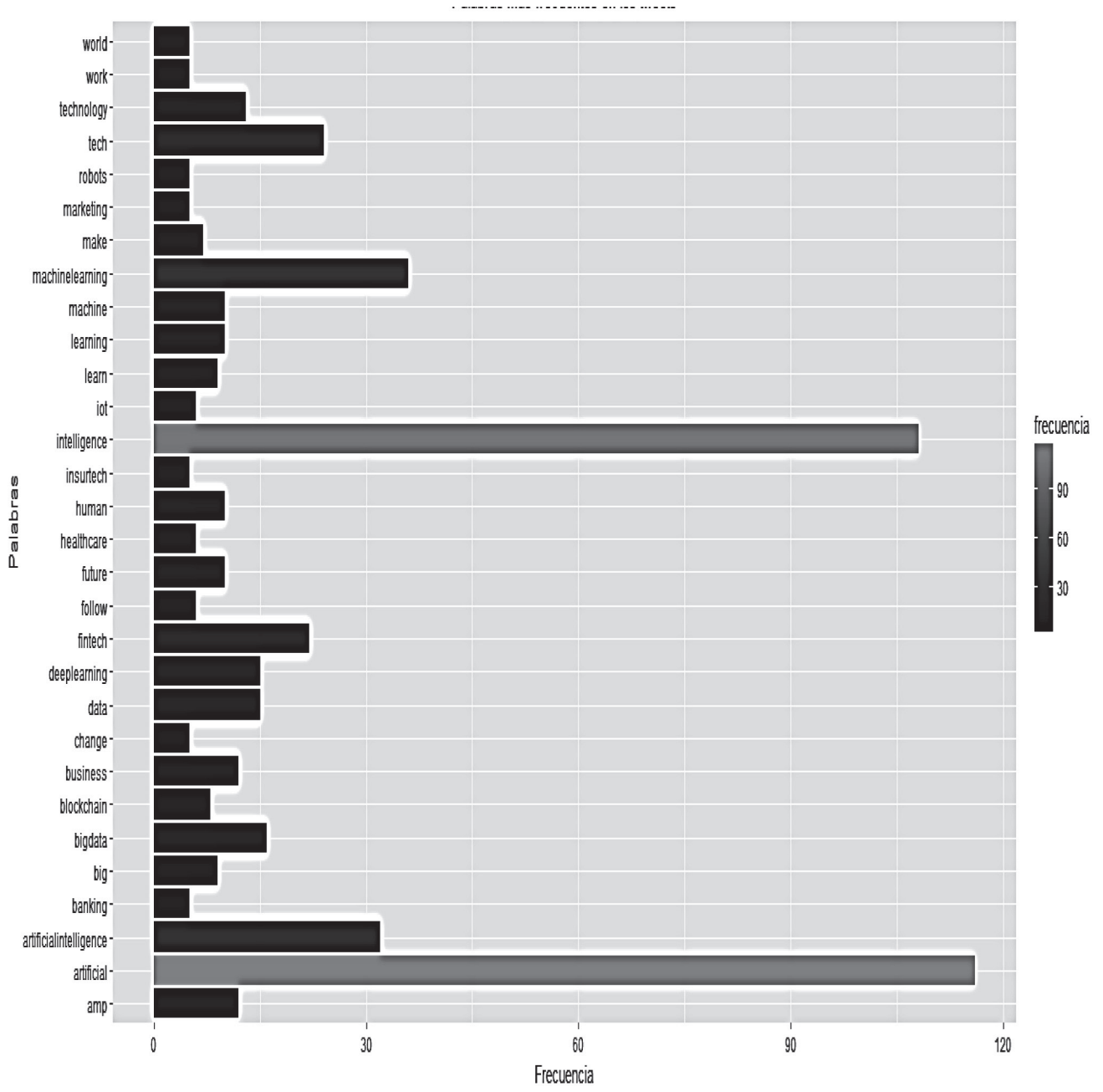


Figura 6. Lista de palabras más frecuentes, sin incluir las palabras clave o de búsqueda



Comparando ambos gráficos (figuras 4 y 5) podemos observar que existe una gran diferencia entre la frecuencia de las palabras clave y las otras. Quitando las palabras clave, podemos formarnos una idea de lo que se está hablando en lo que se refiere a inteligencia artificial, sobresaliendo el aprendizaje de máquina (*machine learning*), que se define como la rama de la inteligencia artificial, que tiene como objetivo desarrollar técnicas que permiten a las computadoras aprender (Sancho Caparrini, 2017).

Otras palabras de importancia que aparecen en el listado son *big data*, *deep learning*, *technology*, *blockchain* y *human*, todas ellas también relacionadas con la inteligencia artificial. Mediante la representación de

la frecuencia, utilizando graficas de barra, podemos hacernos una idea de la temática de la que se habla en los tuits utilizados en las pruebas.

Relación entre palabras

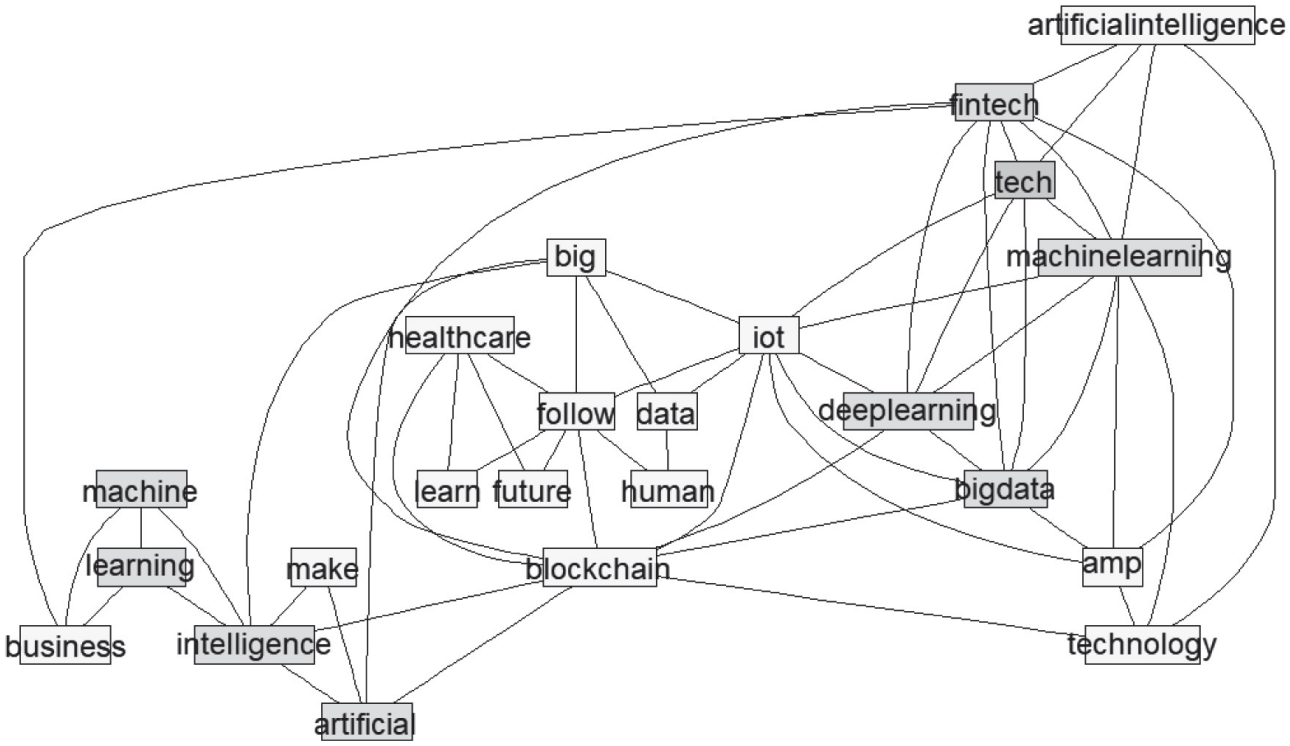
Para la relación de palabras y las siguientes gráficas, es necesario obtener una *term-document matrix* (TDM), o matriz de espacio vectorial, que se genera a partir de un corpus creado con el texto procesado de los tuits. El TDM es un objeto muy importante en extremo en el análisis de texto; es la matriz de documento de términos, porque nos permite almacenar una biblioteca completa de texto en una matriz única. Esto puede usarse para el análisis y para buscar documentos. Forma la base de los motores de búsqueda, análisis de temas y clasificación [filtrado de *spam*] (Das, 2017).

Figura 7. Ejemplo de matriz de términos generada a partir del corpus que contiene los textos de los tuits preprocesados

```
<<TermDocumentMatrix (terms: 11, documents: 8)>>
Non-/sparse entries: 14/74
Sparsity           : 84%
Maximal term length: 12
Weighting          : term frequency (tf)
Sample             :
             Docs
Terms         11 12 13 14 15 16 17 18
  artificial   0  1  1  1  1  1  1  1
  center       0  0  0  0  0  0  0  0
  charles      0  0  0  0  0  0  0  0
  company      0  0  0  0  0  0  0  0
  congrats     0  0  0  0  0  0  0  0
  intelligence 0  1  1  1  1  1  1  1
  invents      0  0  0  0  0  0  0  0
  nigeria      0  0  0  0  0  0  0  0
  onu          0  0  0  0  0  0  0  0
  syst         0  0  0  0  0  0  0  0
```

Teniendo en cuenta lo anterior, podemos generar un gráfico que represente la relación entre las palabras más frecuentes. Para las pruebas, se seleccionaron las palabras con una frecuencia igual o mayor que 6, es decir, todas aquellas que aparecen repetidas más de 5 veces. Este valor puede ser modificado dependiendo de la estructura del texto, ya que puede darse el caso en que se utilice una gran cantidad de palabras con poca frecuencia o en el que se usen pocas palabras con mucha frecuencia.

Figura 8. Representación de la relación entre las palabras basado en su aparición en los tuits



La figura 7 permite apreciar la relación o conexión entre palabras, indicando que estas aparecen en conjunto en los textos. Además de mostrar las palabras que aparecen en la figura 6 muestra otras más, lo que nos permiten expandir la idea del contenido de los tuits. Podemos ver que, al hablar de inteligencia artificial, también se habla de internet de las cosas, fintech (tecnologías financieras) y del cuidado de la salud, temas de mucho auge en la actualidad.

Modelo de temas

Otra forma interesante de representar la información procesada es la creación de modelos de temas en los cuales podemos definir el número de temas y el conjunto de palabras que los conforman, incluyendo su variación en el tiempo a partir de la fecha en que fueron creados los tuits.

Modificando el número de temas y la cantidad de palabras, podemos jugar para tener una mejor idea de lo que se está hablando en el contenido con base en las palabras que aparecen por cada término, sin embargo, en algunos casos, dependiendo del texto obtenido, a pesar de que definamos un número de temas, puede ser que el resultado sea menor por la fuerte relación entre las palabras o la frecuencia en que ocurren.

Figura 9. Modelo creado definiendo el número de temas y los términos o palabras que lo conforman, con los tuits que van desde el 1 de enero hasta el 21 de febrero de 2018

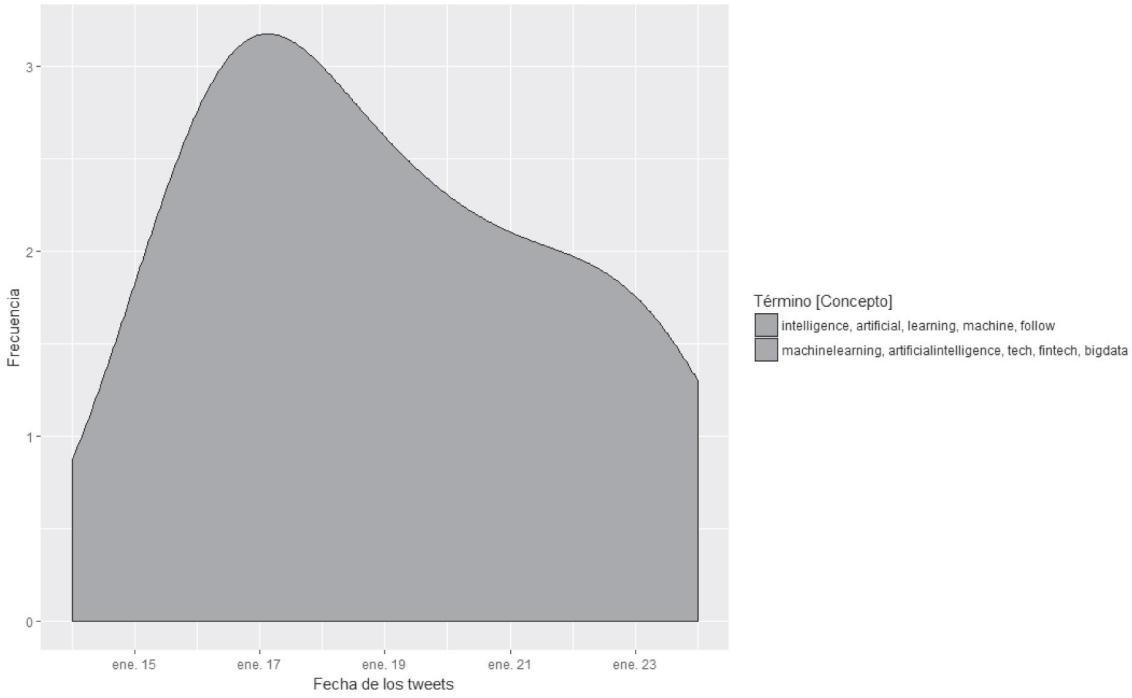
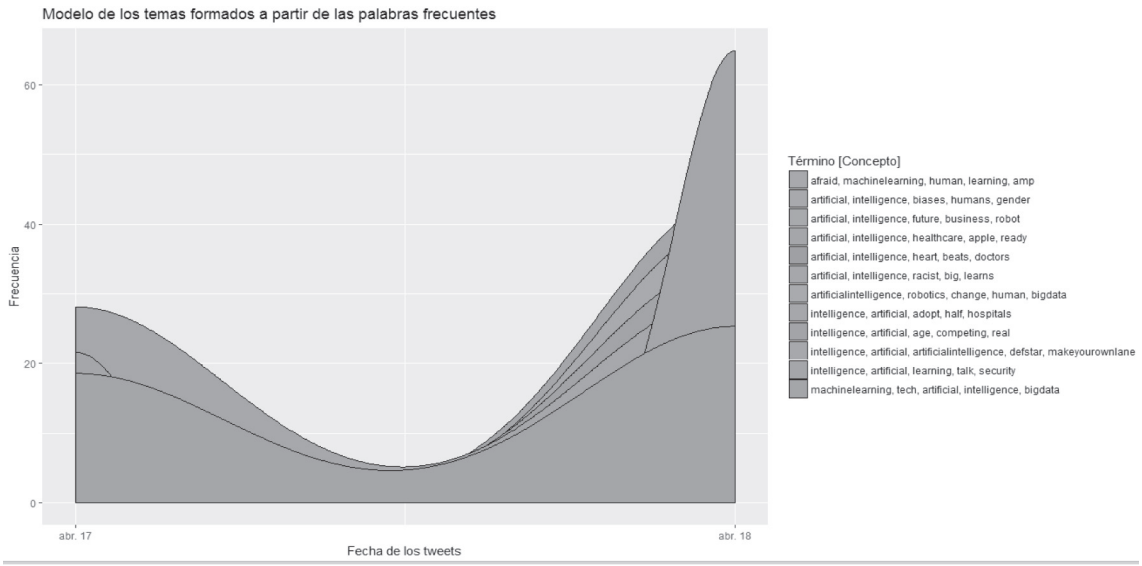


Figura 10. Modelo creado definiendo el número de temas y los términos o palabras que lo conforman, con los tuits del mes de abril de 2017



Comparando las figuras 8 y 9, se puede observar una evolución en los temas relacionados con inteligencia artificial. A pesar de utilizar el mismo término de búsqueda, los resultados son diferentes, lo que nos indica una evolución con el paso del tiempo. El tema central, durante los primeros meses del año 2018, es la inteligencia artificial relacionada con el aprendizaje de máquinas, *big data* y *fintech*, mientras que el año pasado se hablaba de cómo la inteligencia artificial se relaciona con el racismo, esto debido a que muchos tuits hablan sobre cómo dicha tecnología puede "aprender" términos racistas si son alimentados con dicho contenido, como sucedió con el chatbot de Microsoft.[4]

Nube de palabras

Las nubes de palabras permiten visualizar la información de los textos en forma clara y sencilla con un mayor número de palabras, cuyo tamaño nos indica la frecuencia con que aparecen en los textos. Mientras mayor sea el tamaño mayor será la frecuencia. En la nube de la figura 10, podemos observar que entre todas las palabras sobresale *machine learning*, y le siguen *big data*, *tech*, *fintech* y *deep learning*.

Figura 11. Nube de palabras con las 100 más frecuentes quitando *artificial intelligence*, mientras mayor es la frecuencia de las palabras mayor será el tamaño observado

[4] https://qz.com/646825/microsofts-ai-millennial-chatbot-became-a-racist-jerk-after-less-than-a-day-on-twitter/

Conclusiones

El análisis de texto es una tarea que requiere de diferentes pruebas para obtener buenos resultados y la extracción de información de utilidad, estos varían de acuerdo con el contexto, el idioma y la cantidad de información con la que contamos. De todas las etapas, la de mayor importancia es la de preprocesamiento; y es a la que más tiempo se le debe dedicar porque se debe depurar todo aquello que no sea de utilidad.

El análisis de texto de tuits y otras redes sociales se vuelve aún más complicado que el texto proveniente de otras fuentes más formales, o que pasan por un filtro antes de ser publicadas, debido a la libertad con que las personas los escriben, como el uso de abreviaturas, escritura incorrecta, uso de términos propios de cada país y la inclusión de caracteres especiales, como por ejemplo, símbolos, URL, entidades de retuit, entre otros.

Algunas técnicas de procesamiento, como el stemming, puede generar pérdida de información, sin embargo, en caso de que se trabaje con un tema específico, es posible crear un diccionario de palabras, conociendo aquellas que más se repiten y sus variaciones.

R es una herramienta poderosa que cuenta con una gran variedad de librerías para el análisis de texto desde diferentes fuentes y la visualización de resultados, esto último es muy importante, ya que sin representación gráfica el texto procesado, en algunas ocasiones, carece de sentido o relevancia, por lo que es altamente recomendable seleccionar adecuadamente los gráficos de acuerdo con las necesidades y con la idea de qué se quiera mostrar con la información, como se hizo para conocer la frecuencia de las palabras, relaciones, temas y nubes de palabras.

Referencias

Contreras Barrera, M. (2014). *Minería de texto: una visión actual*. Recuperado a partir de http://www.redalyc.org/pdf/285/28540279005.pdf

Das, S. R. (2017). *Data science: theories, models, algorithms, and analytics*. Recuperado de http://srdas.github.io/MLBook/TextAnalytics.html#term-document-matrix-tdm

IBM. (s. f.). *About Text Mining.* Recuperado de https://www.ibm.com/support/ knowledgecenter/es/SS3RA7_16.0.0/com.ibm.spss.ta.help/textmining/shared_entities/tm_intro_tm_defined.htm

Raulji, J., & Saini, J. (2016). Stop-Word Removal Algorithm and its Implementation for Sanskrit Language. *International Journal of Computer Applications*. *150(2)*, 15-17. Recuperado de http://www.ijcaonline.org/archives/volume 150/number2/raulji-2016-ijca-911462.pdf

Gurusamy, V., & Kannan, S. (2014). *Preprocessing Techniques for Text Mining*. Recuperado de https://www.researchgate.net/publication/273127322 _ Preprocessing_ Techniques_for_ Text_Mining

Moujahid, A. (2014). *An introduction to text mining using twitter streaming api and python*. Recuperado de http://adilmoujahid.com/posts/2014/07/twitter-analytics/

Murphy, J., & Roser, M. (2018). *Internet*. Recuperado de https://ourworldindata.org/ internet

Rochina, P. (2017). ¿Qué es el Text Mining?. *Revistadigital inesem*. Recuperado de https://revistadigital.inesem.es/informatica-y-tics/text-mining/

Rodríguez Blanco, A., Cuevas, S., & J, A. (2013). Método para la extracción de información estructurada desde textos. *Revista Cubana de Ciencias Informáticas*. 7(1), 55-67.

Sancho Caparrini, F. (2017). *Introducción al aprendizaje automático*. Recuperado de http://www.cs.us.es/~fsancho/?e=75

Text Mining Online. (2014). *Dive into nltk, part IV: stemming and lemmatization* Recuperado de http://textminingonline.com/dive-into-nltk-part-iv-stemming-and-lemmatization