Minería de texto para estudiar algunas terminologías sobre la ciencia de datos en Colombia.

Text mining to study some terminologies about data science in Colombia

Yubar Daniel Marín Benjumea[1],[1], Heber Esteban Bermúdez Gonzalez[2],[2],

Karen Andrea Amaya[3],[3], René Iral Palomino[4],[4],

[1]Escuela de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Medellín, Colombia
[2]Escuela de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Medellín, Colombia
[3]Escuela de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Medellín, Colombia
[4]Escuela de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Medellín, Colombia

Resumen

El objetivo de este trabajo es estudiar algunas de las terminologías orientadas a la ciencia de datos (aprendizaje automático, big-data, inteligencia artificial, analitica, etc), y cómo estas influyen en el desarrollo de la comunidad Estadística.

Este estudio se lleva a cabo mediante análisis de minería de texto, análisis de sentimientos y análisis sobre redes sociales y usando técnicas de análisis multivariado y estadística descriptiva.

Palabras clave: Aprendizaje automático, big-data, inteligencia artificial, Ciencia de datos, Minería de texto, Análisis de sentimientos.

Abstract

The objective of this paper is to study some of the terminologies oriented to data science (machine learning, big-data, artificial intelligence, analytics, etc.), and how these influence the development of the statistical community.

This study is carried out through analysis of text mining, sentiment analysis and analysis on social networks and using multivariate analysis techniques and descriptive statistics.

Keywords: Machine learning, big-data, artificial intelligence, data science, text mining, sentiment analysis.

[1]Estudiante E-mail: ydmarinb@unal.edu.co
[2]Estudiante E-mail: hebermudezg@unal.edu.co
[3]Estudiante E-mail: kaamayam@unal.edu.co
[4]Profesor asociado. E-mail: riral@unal.edu.co

1. Introducción

Debido a las grandes cantidades de datos que se generan hoy en día, se han desarrollado técnicas y softwares que nos ayudan a procesar toda esta información dando paso así al apogeo de nuevas terminologías relacionadas con la ciencia de datos y estadística, términos como: machine learning, big-data, inteligencia artificial, analitica, entre otras se escuchan frecuentemente y son metodologías que pocas personas saben su concreta utilidad, de cómo se diferencian entre sí o en qué medida están relacionados unas con otras. Es por eso que el presente trabajo se pretende clarificar estos interrogantes mediante análisis de texto, pues aunque los analistas solemos estar entrenados para manejar datos que son en su mayoría numéricos y en arreglos rectangulares, actualmente gran parte de los datos que encontramos no están estructurados y se encuentran en forma de textos. Para el desarrollo del trabajo es preciso sacar provecho de la abundancia de información en la web, es por esto que se adoptan las metodologías de web scraping o raspado web para obtener la materia prima y posteriormente limpiar y analizar con técnicas de análisis de texto.

2. Definiciones

Corpus

Un corpus lingüístico es un conjunto de muestras reales de uso de la lengua, designa la recopilación de material lingüístico hecha con un propósito de investigación concreto, ya sean muestras de oraciones, de enunciados o de textos.

Token

Un token es una unidad significativa de texto, por lo general palabras.

3. Metodología

3.1. Recolección de la información.

El enfoque de este trabajo radica en extraer la información de la web con la metodología web scraping o raspado web haciendo uso de la librería rvest del software estadístico R. Para la obtención de estos datos se buscaron los términos: ciencia de datos, machine learning, big data, inteligencia artificial, ofertas laborales ciencia de datos y ofertas laborales estadística, y se tomó los resultados de la primera página de Google, la cual se llevan la mayor parte del tráfico de los resultados, también se tomó información la red social linkedIn y Twitter. Es importante precisar que para hacer esta búsqueda en la web se eliminaron los cookies del navegador con el fin de evitar sesgos en la recolección de la información.

Conocer la estructura de una página web es el primer paso para extraer y usar los datos, las páginas en su gran mayoría son construidas con el lenguaje demarcado conocido como HTML, el cual da la estructura y el contenido de la página. Una vez familiarizado con este lenguaje es fácil identificar las etiquetas que contiene la información de interés, las cuales suelen estar en una clase definida dentro de la estructura propia de la página. Para esto, utilizamos la herramienta SelectorGadged de Google

Chrome y el paquete rvest de R para extraer así la información en forma de cadenas de caracteres de manera automática.

3.2. Preparación de los datos.

Una vez obtenida esta información en formato de cadenas de texto se procedió a la limpieza de las mismas, a continuación se muestra una fragmento de cadena de caracteres recuperada mediante web scraping sobre la consulta ofertas laborales ciencia de datos.

```
  [1] "\r\n         Si quieres desarrollar una carrera en el campo más sexy para
trabajar en el siglo 21"
```

Se puede observar que esta cadena de texto contiene caracteres especiales, puntuaciones, espacios en blanco, números, stopwords que son palabras vacías como artículos, preposiciones, conjunciones, pronombres, que no son interesantes para nuestro análisis puesto que carecen de contenido semántico por sí solas. Para limpiar la cadena mostrada es preciso eliminarlos, esto con la ayuda la librería tm del software R para minería de texto. Una vez limpia la cadena se ve de la siguiente manera:

```
 [1] " Si quieres desarrollar carrera campo sexy trabajar siglo"
```

Limpias la cadenas de texto se define un formato de texto ordenado como una tabla, con un token por fila, esto se hace usando la función `tibble()` del paquete dplyr una vez así la cadena de texto puede ser trabajada como n-gramas donde n es el número de palabras por fila a considerar para el análisis.

3.3. Análisis de la información

Una vez preparadas las cadenas de texto y puestas en un marco de datos puede resultar de particular interés observar gráficamente cuales son las palabras con mayor frecuencia dentro de toda la información recopilada, esto con ayuda de la función `unnest_token()` del paquete tidytext para dividir nuestro texto por palabras individuales (token) y se cuentan las frecuencias de la palabras. Luego, se procede a graficar una nube de palabras con ayuda de la función `wordcloud()` del paquete wordcloud de R.

En la Figura 1 podemos observar cuales son la palabras con mayor frecuencia dentro de todo el conjunto de textos o corpus

Figura 1: Nube de palabras de los 50 términos con mayor frecuencia dentro del corpus.

También resulta de particular interés observar la frecuencia de pares de palabras por término o tema consultado tanto para token de una palabra (unigrama) como de dos (bigrama), a continua se muestra el el gráfico de frecuencias para pares de palabras por término consultado (bigrama).

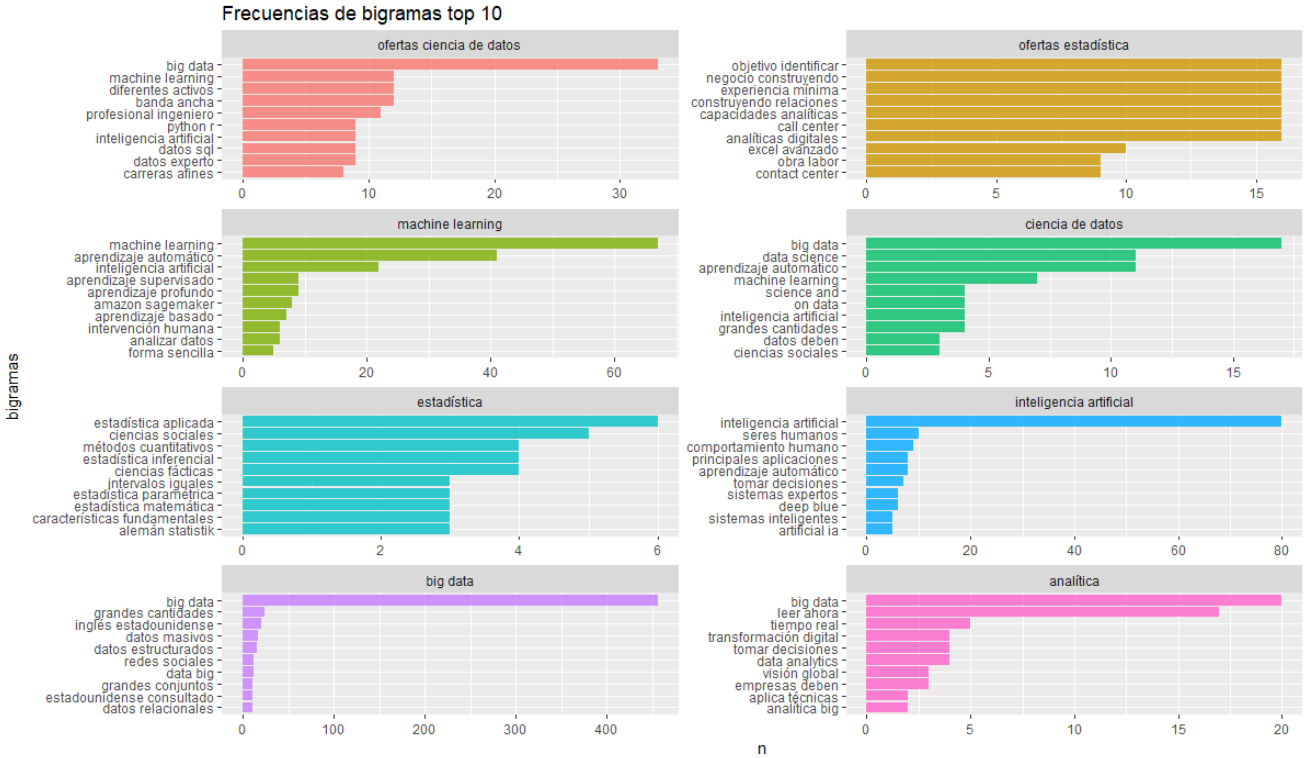

Figura 2: Frecuencias de pares de palabras por término o tema consultado

Se puede observar que el común denominador en todos los términos o temas consultados en Google gira en torno a los datos, puesto que observamos con mayor frecuencia palabras datos, data.

Para inferir cómo se relacionan los términos o temas consultados entre sí es preciso construir un diagrama de correlación entre las frecuencias de las palabras como se puede observar en la Figura 3.

Figura 3: correlaciones de frecuencias de palabras por término o tema

Las palabras que están cerca de la línea roja en las gráficas tienen frecuencias similares en todas los términos o temas consultados, por ejemplo las palabras como datos y data son frecuentes en todos los términos consultados, mientras que palabras como aprendizaje es poco frecuente en ofertas estadística (ofertas laborales estadística) pero es relativamente es común en el resto del conjunto de datos.

4. Conclusiones

Mediante los distintos gráficos descriptivos se puede observar que todos coinciden en el manejo de datos, así entonces, las nuevas terminologías como machine learning, big data, inteligencia artificial, etc. aportan al desarrollo de la comunidad Estadística en la medida en que son técnicas que generan conocimiento a partir de los datos por medio de la programación, permitiendo analizar mayores cantidades de datos, de manera más óptima y automática.

Por otro lado, la palabra estadística y las ofertas laborales estadísticas en esta muestra de datos, se encuentran poco correlacionadas con nuevas terminologías basadas en técnicas y análisis estadístico. Esto permite concluir que el mercado laboral y la documentación analizada no asocia el quehacer estadístico con estas nuevas herramientas.

La abundancia de datos en la web es algo muy positivo y dentro de las habilidades propias de analista radica en la capacidad de recuperar esta información de distintas fuentes para hacer análisis, responder interrogantes y generar conocimiento, el análisis de texto representa una metodología muy útil para lograr este objetivo hoy existen muchas herramientas, librerías en distintos lenguajes de programación y en distintos idiomas que nos facilitan la tarea.

5. Referencias

[1] Text Mining with R, tomado de
https://www.tidytextmining.com/tfidf.html

[2] procesamiento del lenguaje natural:
http://www.iic.uam.es/inteligencia/que-es-procesamiento-del-lenguaje-natural/

[3] Minería de texto frente a analítica de texto, tomado de:
https://www.educba.com/text-mining-vs-text-analytics/