Publicación Electrónica **ISSN 2718- 6636**

UNIVERSIDAD NACIONAL DE ROSARIO
FACULTAD DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA
SECRETARIA DE CIENCIA Y TECNOLOGIA E INSTITUTOS DE INVESTIGACIONES

Resumen Ampliado

Jornadas Anuales
"Investigaciones en la Facultad"
Ciencias Económicas y Estadística

Collado, Facundo José
Vitelleschi, María Susana
Quaglino, Marta Beatriz
Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística, Escuela de Estadística

Algoritmos automatizados para el análisis textual de búsquedas laborales en la *web*[1]

Resumen

Las responsabilidades específicas para los puestos de profesionales especializados en el área de análisis de datos se encuentran en constante evolución, generando nuevas barreras de acceso a puestos de analistas, ingenieros y científicos de datos. Tanto profesionales del área y quienes están aún en proceso de capacitación para adquirir las habilidades necesarias de estos puestos laborales, no sólo deben disponer con cierta capacidad de adaptación a las demandas actuales, sino que deberán estar al tanto de las tendencias del mercado laboral; como las que se pueden encontrar reflejadas en búsquedas laborales dentro de la *web*. El objetivo de este estudio consiste en analizar, mediante la aplicación de técnicas de minería de texto, propuestas laborales en formato textual no estructurado orientadas a profesionales de datos, con el objeto de encontrar las responsabilidades laborales específicas de mayor demanda que están ocultas en la redacción de las solicitudes publicadas, mediante la aplicación de técnicas de minería de texto. Los resultados de este trabajo permiten descubrir aquellos patrones, relaciones, tendencias y otros conocimientos útiles que pueden servir a los aspirantes a este tipo de ofertas para adaptar su currículum o postular a tareas más específicas a su capacitación. Asimismo, las organizaciones pueden nutrirse de este conocimiento para instruir al personal de recursos humanos a los efectos de generar propuestas laborales que permitan atraer talentos idóneos a sus necesidades o bien adaptar sus procesos para aprovechar aquellas habilidades de mayor importancia actual. Se aplican dos técnicas de análisis de texto las cuales son evaluadas y contrastadas según diferentes indicadores de su eficacia.

Palabras clave: Búsquedas laborales, Análisis de texto, *Big data*.

Abstract

Specific responsibilities for data professional positions are constantly evolving and creating new barriers for data analysts, engineers and scientists who wish to access these positions. Not only must professionals in the field and those who are still in the training process to acquire the necessary skills for these job positions, have a certain ability to adapt to current demands, but also they should be aware of labor market trends. This could be reflected in job searches on the

[1] Trabajo elaborado en el marco del Proyecto 80020180300041UR, titulado: “Métodos de clasificación y predicción en contextos de Big-Data y tres modos”, dirigido por la Dra. Marta Quaglino.

web. The objective of this study is to analyze job proposals in unstructured textual format. It is aimed at data professionals, in order to find specific job responsibilities in great demand hidden in the wording of the demands, through the application of text mining techniques. The results of this study would allow us to discover patterns, relationships, trends and other useful knowledge that could help applicants for this type of offer to adapt their resume or apply for positions more akin to their training. Likewise, organizations could draw on this knowledge in order to instruct human resources personnel to generate job proposals. This would allow them to attract suitable talents or adapt their processes to take advantage of skills that are currently in great demand. Two text analysis techniques are applied, which are evaluated and contrasted according to their effectiveness.

Keywords: Job search, Text analysis, Big data.

Introducción

La ciencia de datos es un área formada por múltiples campos del conocimiento, los cuales son ciencia de la computación, matemática y estadística. Las matemáticas, por su lado, proveen los conceptos y fundamentos que acompañan el desarrollo de los métodos estadísticos, los cuales incluyen obtención, evaluación, análisis, visualización e interpretación de datos, así como el planteo y estimación de modelos que permiten descubrir las interrelaciones entre las múltiples variables que describen los fenómenos estudiados. Por otra parte, la informática provee las herramientas para procesar grandes volúmenes de datos y ejecutar cálculos complejos y procesos iterativos en tiempos muy reducidos.

La aplicación de los modelos para el análisis de datos textuales se logra a través de algoritmos informáticos que permiten extraer grupos de palabras que conforman las frases clave de interés, las cuales son interpretadas considerando el contexto del estudio. Los resultados se generan con la intención de identificar aquellas actividades que son de mayor valor contemporáneo para las empresas que trabajan con datos y formar, en base a esta información, una generalización apropiada que pueda ser utilizada por diferentes profesionales como guía de encaje al tipo de búsquedas analizadas.

Material y Métodos

Se seleccionaron 5.331 propuestas laborales provenientes de la *web* (*Indeed.com*), durante el periodo octubre de 2022 a noviembre de 2022, disponibles en formato .csv. Las mismas poseen la descripción del puesto de trabajo e información adicional como volumen, ubicación, y sitio *web* de la organización. Cada descripción tiene, aproximadamente, 243 palabras y está redactada en lenguaje humano, de estructura cambiante, afectada por factores psicológicos, sociales y conductuales. Utilizando este corpus, se realiza como primera etapa, una depuración y limpieza de contenido, para la posterior aplicación de los métodos de análisis de texto. Luego, se utilizan dos algoritmos de procesamiento de texto con bases estadísticas, conocidos como N-Gramas (Jurafsky, *et al*., 2021) y BERTopic (Devlin, J. *et al*., 2018), a partir de los que pueden identificarse frases clave, palabras más frecuentes y conceptos de capacitación requerida, entre otros. Como aspectos adicionales se comparan los procedimientos respecto del tiempo de ejecución requerido y la capacidad de memoria utilizada en el proceso.

Resultados

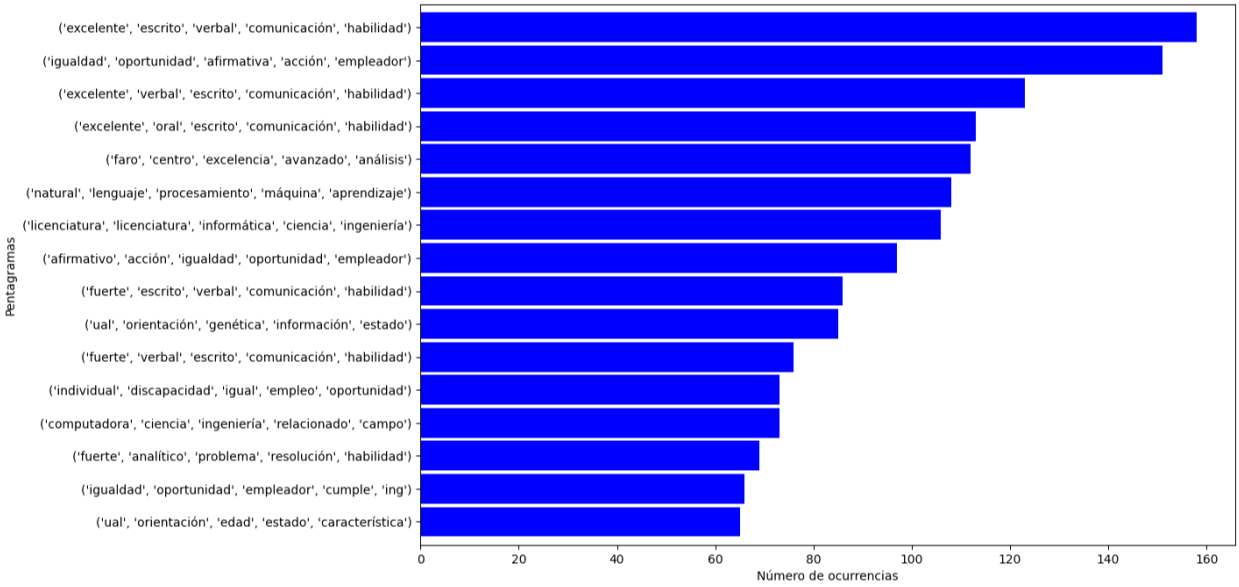
La primera etapa para la obtención de los resultados, fue de depuración de la información original. Cada documento fue recopilado de la *web*, por lo que en su estado más puro, poseía información no relevante, desde caracteres y símbolos *web* especiales necesarios para el formato, hasta referencias específicas de organizaciones o personas. El proceso de limpieza por iteraciones aseguró la reducción de los datos de estudio a un nivel aceptable, que garantice mantener, en su mayoría, aquellas palabras y frases que referencian las actividades y responsabilidades laborales que están referidas en el conjunto total de documentos.

Dentro del amplio volumen de resultados generados por cada método, se presentan las salidas de mayor relevancia para su interpretación. Todos los resultados fueron procesados utilizando las librerías compatibles con el lenguaje de programación “Python”. Los resultados se muestran para cada algoritmo utilizado. Se acompaña la interpretación de algunos de los conjuntos de palabras resultantes, valorando su cualidad de describir contextos específicos, los tiempos de ejecución captados y la cantidad de memoria RAM requerida.

Algoritmo N-gramas

Una primera visión global obtenida con N-gramas, está representada por un gráfico de frecuencias que permite identificar conjuntos de palabras mencionadas simultáneamente en los textos, ordenados por su frecuencia de uso en los pedidos laborales, lo cual representa un indicador de su importancia (Figura 1). Para realizar este resumen, pueden utilizarse grupos de distinta cantidad de palabras que se corresponde con la letra “N” del nombre del método. Es usual correr repetidamente el algoritmo fijando distintos valores de este parámetro, porque inicialmente se desconoce su valor óptimo. De las alternativas probadas se eligió N=5 como aquella de resultados más ricos en interpretación (única alternativa mostrada). La elección de este rango se respalda por los hallazgos de Cui H. *et al*. (2006), quienes demuestran cómo frases de tamaño mayor o igual a tres (N=3) tienden a ser menos ambiguas en su interpretación. Estos tamaños ayudan a discriminar la polaridad de cada anuncio en su contexto, mientras que las de menor tamaño fallan en capturar la dependencia de largo alcance entre palabras con cierta lejanía.

Figura 1: Grupos de palabras en avisos laborales en *Indeed.com*, según su frecuencia de uso(1). Periodo octubre - noviembre de 2022



(1) Nota: Solo se muestran las frecuencias mayores a 60 avisos

Para el caso de N-Gramas, las frases son valoradas en relación a su ocurrencia y agrupadas según su cercanía dentro del texto. En los distintos grupos de 5 palabras formados automáticamente por la revisión de todos los textos considerados (5331), aparecen palabras repetidas, acompañadas por grupos diferentes. Si se consideran las palabras comunes a todos los grupos, se puede identificar al conjunto “excelente, escrito, verbal, comunicación, habilidad”, con las frecuencias más altas de uso, entre 60 y 160 citas, lo referencia la importancia de **la capacidad de comunicación** en las demandas laborales. Este agrupamiento se ve repetido en otros 5 resultados, aunque con palabras intercambiadas de lugar o reemplazadas por sinónimos. El mismo caso se observa para el conjunto de palabras “igualdad, oportunidad, afirmativa, acción, empleador”, que indican **la valoración a la igualdad de condiciones por parte de la organización solicitante**, que de forma análoga al grupo anterior, se ve reiterado como grupo, en diferente orden de palabras y con palabras reemplazadas, y un número de frecuencias de aparición dentro del corpus de entre 60 y 155.

Estos resultados permiten obtener las ideas y/o conceptos generales encontrados dentro de las propuestas laborales. No obstante, carecen de expresiones verbales que describen contextos específicos o responsabilidades particulares. Estas particularidades pueden deberse a que este método, se basa en la posibilidad de ocurrencia de una palabra en una frase dada la ocurrencia de sus anteriores, lo cual tiene como consecuencia que los resultados focalizan a las palabras que tengan mayor probabilidad de ocurrencia, ignorando completamente el significado semántico. En consecuencia aparecen resultados duplicados o sintácticamente equivalentes, también explica el por qué en su mayoría solo se identifican frases relacionadas con intereses comunes entre propuestas laborales, como habilidades de comunicación, estudios o beneficios de contratación.

El tiempo de ejecución utilizado por este algoritmo es de aproximadamente 3.41 segundos por variante de N utilizada y se requirió de casi 1.8 gigabytes de memoria RAM.

Algoritmo BERTopic

El algoritmo entrena un modelo de aprendizaje profundo utilizando las descripciones de las propuestas laborales, el cual extrae y agrupa palabras según su relación hacia ciertos temas y la frecuencia con que son utilizadas dentro de cada documento (Grootendorst, M. (2022)). Para proyectar cada conjunto de palabras, se utilizó el método *web* de visualización interactiva de temas llamado “LDAvis”, desarrollado por Sievert, C. *et al*. (2014), el cual es basado en el algoritmo de aprendizaje no supervisado “Asignación Latente de Dirichlet”. Es usado normalmente para detectar temas compartidos por los documentos dentro de un cuerpo de textos. El algoritmo recibe como entrada el modelo entrenado, el corpus de las solicitudes laborales, y es posible asignarle un número específico de temas que se deseen extraer. Con esta información genera una gráfica bidimensional de círculos (Figura 2) que representan cada conjunto de palabras según un número de tema al que representa dentro del corpus. El diámetro de cada círculo se calcula en base a su probabilidad de ocurrencia dentro del corpus. Además, la posición de los círculos respeta una regla de cercanía entre temas con mayor relación semántica.

Figura 2: Temas identificados por BERTopic en solicitudes laborales de *Indeed.com*, según su probabilidad de ocurrencia y cercanía con otros temas. Periodo octubre - noviembre de 2022.

La Figura 2, fue realizada indicando al algoritmo que reduzca el número de temas posibles de forma automática, el mismo proviene de una herramienta *web* interactiva que permite al usuario seleccionar cada círculo y visibilizar las palabras clave que contiene cada tema. Al analizar los diámetros, se puede ver que es alta la frecuencia de los primeros tres círculos de mayor tamaño al corpus, con respecto al resto. Al analizar su distancia, se nota que la mayoría de los círculos menos relevantes se aproximan a los más relevantes, lo que significa que las temáticas que representan son similares a las tres principales.

Los resultados de este método presentes en este estudio se corresponden con una ejecución en particular y son únicos, por lo que no son exactamente reproducibles, ya que cada iteración produce resultados únicos.

Tabla 1: Temas identificados por BERTopic en solicitudes laborales de *Indeed.com*, ordenados según el número de ocurrencias dentro del corpus. Periodo octubre - noviembre de 2022.

Grupos de palabras	Número de ocurrencias en corpus
Datos, negocios, equipos, analítica, trabajo	1109
Aprendizaje, datos, máquina, análisis, habilidad	707
Cuidado, clínico, paciente, datos, calidad	178
Datos, *spark*, *pipeline*, bases de datos, sistema	125
Beneficios, contrato, contratar, técnico, salario	85
Adoptante temprano, firma, diversos, procesamiento, inteligencia artificial	50
Investigación, ciencia, paciente, datos, cuidados	39
Análisis, entregable, datos, mercancía, reportes	37
Bio informática, computacional, biología, genómica, genómico	35

En la Tabla 1 se observan conjuntos de palabras que conforman las solicitudes, como por ejemplo “Datos, negocios, equipos, analítica, trabajo” (primera fila) que hacen referencia **al análisis de información de negocios** y son las palabras más frecuentemente citadas en los avisos, 1109 veces. En segundo lugar de importancia y con una frecuencia de 707 citas aparecen: “Aprendizaje, datos, máquina, análisis, habilidad”, el cual indica **habilidades analíticas en el aprendizaje y de máquina**. En tercer lugar y con una frecuencia mucho menor, (178) que representa el 6% del grupo más repetido: “Cuidado, clínico, paciente, datos, calidad”, que sugiere **conocimientos en el área de manejo de datos médicos**. Relacionando la Tabla 1 con la Figura 2 se aprecia que en la Figura 2, todos los grupos son cercanos entre sí, debido a que comparten significados semánticos. No obstante, el tercer grupo se ubica en el centro del gráfico, entre el primer y segundo grupo, porque comparten la palabra “datos”. Para estos gráficos, los temas con palabras en común aparecerán a menor distancia. De este modo pueden seguir analizándose los grupos de palabras mencionados en las solicitudes, identificando el área o especialidad solicitada, según su frecuencia de aparición.

El tiempo de ejecución utilizado por este algoritmo es de aproximadamente 22.47 minutos y se requirió de casi 2 gigabytes de memoria RAM. La diferencia en la utilización de recursos computacionales radica en que N-Gramas se basa en un método estadístico que asigna puntuaciones a palabras y frases basadas en la concurrencia y la probabilidad condicional de aparición dado un cierto número de palabras. Tampoco se toma en consideración el contexto en el que aparecen, mientras que BERTopic incluye un modelo de redes neuronales para el aprendizaje profundo, técnicas de reducción de dimensionalidad, agrupamiento jerárquico, y transformaciones vectoriales, lo cual aumenta la complejidad computacional.

CONSIDERACIONES FINALES

Analizando los resultados de ambos algoritmos, se observa que los provistos por BERTopic no solamente son menos redundantes en el conjunto de palabras extraídas, respecto al algoritmo N-Gramas, sino que también referencian a contextos más específicos, lo que los vuelven de mayor valor para descubrir información oculta. La cantidad de memoria RAM requerida es similar entre ambos, aunque BERTopic necesita de un tiempo de ejecución mayor para el conjunto de datos utilizado.

Finalmente, si bien la elección del algoritmo depende de la disponibilidad de información y de recursos del analista, junto a su objetivo de interés, en este estudio se concluye que extraer y generalizar similitudes semánticas a partir de información textual escrita por humanos, es una tarea cuya complejidad va más allá de contabilizar palabras. El método aplicado debe tener un conocimiento previo del lenguaje hablado y comprender el contexto en el que se utilizan las palabras.

REFERENCIAS BIBLIOGRÁFICAS

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional Transformers for language understanding*. arXiv.org. https://arxiv.org/abs/1810.04805

Cui, H., Mittal, V., Datar, M. (2006). *Comparative experiments on sentiment classification for online product reviews*. Guide Proceedings. https://dl.acm.org/doi/10.5555/1597348.1597389

Jurafsky, D., Martin, J. H. (2021). *Speech and Language Processing (3rd ed. draft) Dan Jurafsky and James H. Martin*. Speech and Language Processing. https://web.stanford.edu/~jurafsky/slp3/

Grootendorst, M. (2022). *Bertopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv.org. https://arxiv.org/abs/2203.05794

Sievert, C., Shirley, K. E. (2014). *LDAvis: A method for visualizing and interpreting topics*. The Stanford Natural Language Processing Group. https://nlp.stanford.edu/events/illvi2014/papers/sievert-illvi2014.pdf