

Desarrollo de una herramienta de aprendizaje automático (*machine learning*) para establecer relaciones entre ocupaciones y programas de capacitación en el Uruguay

Miguel Omar Velardez
Germán César Dima



NACIONES UNIDAS

CEPAL

Gracias por su interés en esta publicación de la CEPAL



Si desea recibir información oportuna sobre nuestros productos editoriales y actividades, le invitamos a registrarse. Podrá definir sus áreas de interés y acceder a nuestros productos en otros formatos.

 www.cepal.org/es/publications

 www.cepal.org/apps

Desarrollo de una herramienta de aprendizaje automático (*machine learning*) para establecer relaciones entre ocupaciones y programas de capacitación en el Uruguay

Miguel Omar Velardez
Germán César Dima



NACIONES UNIDAS



Este documento fue preparado por Miguel Omar Velardez, Consultor de la Unidad de Estudios del Empleo de la División de Desarrollo Económico de la Comisión Económica para América Latina y el Caribe (CEPAL), y Germán César Dima, en el marco de las actividades del proyecto de la Cuenta de las Naciones Unidas para el Desarrollo “Transformaciones tecnológicas en América Latina: promover empleos productivos y enfrentar el desafío de las nuevas formas de empleo informal”.

Los autores agradecen a María José González, Leticia Lado y Gabriel Vázquez por los aportes conceptuales, el acceso a los datos, las discusiones y la revisión del manuscrito; a Pablo Melgar, de la Universidad del Trabajo del Uruguay (UTU), por proveer los datos de los programas de capacitación, y a Sonia Gontero, Oficial de Asuntos Económicos de la CEPAL, por las correcciones y aportes al documento.

Las opiniones expresadas en este documento, cuyos elementos gráficos no han sido sometidos a revisión editorial, son de exclusiva responsabilidad los autores y pueden no coincidir con las de la Organización o las de los países que representa.

Publicación de las Naciones Unidas
LC/TS.2022/2
Distribución: L
Copyright © Naciones Unidas, 2022
Todos los derechos reservados
Impreso en Naciones Unidas, Santiago
S.21-00597

Esta publicación debe citarse como: M. O. Velardez y G. C. Dima, “Desarrollo de una herramienta de aprendizaje automático (*machine learning*) para establecer relaciones entre ocupaciones y programas de capacitación en el Uruguay”, *Documentos de Proyectos* (LC/TS.2022/2), Santiago, Comisión Económica para América Latina y el Caribe (CEPAL), 2022.

La autorización para reproducir total o parcialmente esta obra debe solicitarse a la Comisión Económica para América Latina y el Caribe (CEPAL), División de Documentos y Publicaciones, publicaciones.cepal@un.org. Los Estados Miembros de las Naciones Unidas y sus instituciones gubernamentales pueden reproducir esta obra sin autorización previa. Solo se les solicita que mencionen la fuente e informen a la CEPAL de tal reproducción.

Índice

Introducción	5
I. Metodología	9
II. Resultados	13
III. Validación	23
IV. Conclusiones	27
Bibliografía	29
Anexo 1.....	31
Cuadros	
Cuadro 1	Ranking de cercanía de programas de capacitación para cada una de las 22 ocupaciones de la primera ola de relevamiento de ONET UY 14
Cuadro 2	Familias de programas de capacitación: ejemplo de familias de programas de capacitación altamente relacionados.....17
Cuadro 3	Familias de programas de capacitación: ejemplo de familias de programas de capacitación escasamente relacionados 18
Cuadro 4	Listado de programas de capacitación recomendados para cada ocupación de la primera ola de relevamiento de ONET UY 18
Cuadro 5	Resumen de resultados de estrategia mixta TF-iDF-Doc2Vec 22
Cuadro 6	Resumen de los resultados obtenidos con la estrategia mixta TF-iDF-Doc2Vec en función de los resultados de la validación manual llevada a cabo por el equipo de expertos 26
Diagramas	
Diagrama 1	Procesamiento de textos de ocupaciones y programas de capacitación mediante NLP 10
Diagrama 2	Árbol de decisión para validar la herramienta NLP 24

Gráficos

Gráfico 1	Proporción de programas de capacitación recomendados por la herramienta TF-IDF que cumplen con el criterio de pertinencia según número de recomendaciones seleccionadas	16
Gráfico 2	Número medio de programas pertinentes según la clasificación manual realizada por el equipo de validación.....	25
Gráfico A1	Frecuencia de los programas de capacitación considerados pertinentes según la clasificación manual realizada por el equipo de validadores.....	32
Gráfico A2	Frecuencia de los programas de capacitación considerados no pertinentes según la clasificación manual realizada por el equipo de validadores.....	35

Introducción

El mundo del trabajo está atravesando un proceso de cambios cada vez más acelerados en medio de la globalización, el envejecimiento de la población, el cambio climático y los cambios tecnológicos, sobre todo aquellos que se relacionan con la adopción de la inteligencia artificial y los procesos globales de transformación digital (Banco Mundial, 2016; McKinsey Global Institute, 2017; OIT, 2019). En un trabajo reciente de la Organización de Cooperación y Desarrollo Económicos (OCDE) se muestra que estas tendencias también afectan a los países latinoamericanos, aunque con alguna diferencia respecto de otras regiones (OCDE, 2020). Debido a que su población es más joven y la velocidad para adoptar tecnologías avanzadas es menor, los países latinoamericanos hasta ahora han estado levemente protegidos de los efectos del envejecimiento poblacional y el cambio tecnológico. Por otro lado, en un informe reciente de la Organización Internacional del Trabajo (OIT) se indica que, si no se enfrentan los desafíos que surgen como resultado de estos cambios, se corre el riesgo de incrementar las brechas interregionales y generacionales, así como las de género e ingresos (OIT, 2019). Un factor clave que puede aumentar las desigualdades es la naturaleza cambiante del trabajo y la baja velocidad a la que las personas se adaptan a estos cambios, lo cual disminuye la empleabilidad y limita las oportunidades para encontrar un trabajo decente que asegure una mejor calidad de vida.

En los últimos años el mercado de trabajo se ha fragmentado y polarizado de manera creciente, y eso ha dado como resultado que uno de cada cuatro trabajadores no posea las habilidades necesarias para desempeñar las tareas de su ocupación (Autor, 2010; McGowan y Andrews, 2015).

A la luz de estos cambios, el desarrollo de carrera para adultos representa una herramienta fundamental destinada a contribuir a que estas personas naveguen con éxito las aguas de un mercado laboral en constante evolución. En una revisión reciente, la OCDE definió el desarrollo de carrera como un conjunto de servicios que ayudan a las personas a tomar decisiones educativas, de formación y ocupacionales informadas (OCDE, 2021a). Existe evidencia en la literatura sobre el impacto positivo que el desarrollo de carrera tiene en los resultados de aprendizaje, la participación en la formación y, en menor medida, los resultados relacionados con el empleo (OCDE, 2021b).

El estudio más reciente de caracterización de los servicios de desarrollo de carrera en América Latina fue realizado por la OCDE en 2020 (OCDE, 2021b). El documento en que se informa sobre dicho estudio contiene los resultados de la Encuesta sobre Desarrollo de Carrera para Adultos (Survey of Career Guidance for Adults, SCGA), y esos resultados indican que alrededor del 57% de los adultos de América Latina no

se capacita y no desea hacerlo. También muestran que, a la hora de analizar y decidir sobre las mejores opciones de capacitación para el desarrollo profesional, solo el 9% acude a un servicio prestado por las oficinas de empleo, mientras que el 34% contrata un proveedor privado (OCDE, 2021b). El uso limitado del servicio que ofrecen las oficinas de empleo podría estar relacionado con el hecho de que el objetivo de estas es que las personas encuentren trabajo rápidamente en lugar de hacer foco e invertir en el desarrollo de carrera para que las personas obtengan mejoras profesionales a largo plazo

Lo ideal sería que en los servicios de desarrollo de carrera se ayudara a las personas a analizar tanto las oportunidades laborales como las de capacitación a lo largo de la vida, para alinearlas con sus objetivos personales y con la demanda del mercado laboral. Sin embargo, las oficinas de empleo no suelen tener suficiente personal capacitado como para brindar un servicio de desarrollo de carrera, y tampoco suelen contar con herramientas que permitan ofrecer un servicio personalizado de evaluación de las habilidades, la educación formal, las experiencias, la historia laboral, el perfil de personalidad, los intereses personales y los objetivos de largo plazo de la persona que demanda el servicio. La mayoría de las recomendaciones de capacitación se limitan a cursos o programas relacionados con la ocupación actual de la persona, con la rama de actividad en la que se desempeña o con la oferta de cursos disponibles en la región para su nivel educativo.

Por otro lado, en el momento de recomendar programas de capacitación a una persona, las opciones se pueden ver limitadas o afectadas por el sesgo propio del proveedor del servicio. Es muy difícil que una persona, por más experiencia y capacitación que tenga, conozca toda la información sobre el contenido de la amplia gama de programas de capacitación existentes en un país. El sesgo personal se refiere a aspectos como las experiencias de vida y los intereses del proveedor del servicio, que podrían hacer que este se desviara inconscientemente hacia ciertos programas de capacitación en detrimento de otros que quizás podrían ajustarse mejor a los intereses y objetivos profesionales del trabajador.

En el conjunto de recomendaciones del trabajo de la OCDE (2021b) se resalta la conveniencia de fortalecer los estándares de calidad de la prestación de servicios públicos generales e incluir un componente que considere específicamente el desarrollo de carrera, así como de desarrollar una herramienta de elaboración de perfiles de habilidades que puedan utilizar los proveedores, especialmente cuando se trabaja con adultos que no tienen calificaciones formales.

En concordancia con esta última recomendación, el Ministerio de Trabajo y Seguridad Social (MTSS) del Uruguay viene desarrollando desde 2017 el Sistema de Información Ocupacional - Encuesta "Perfil de Ocupaciones" - ONET Uruguay (ONET UY)¹ que permite elaborar perfiles ocupacionales de forma exhaustiva y estandarizada, replicando la metodología utilizada por la plataforma ONET de los Estados Unidos². En ONET UY hay dos tipos de cuestionarios: los comunes, en que se relevan las mismas variables en relación con todas las ocupaciones, y los específicos, en que se relevan variables de cada ocupación. En los cuestionarios comunes se indaga sobre las destrezas, las habilidades, los conocimientos, los estilos de trabajo, las actividades generales, los contextos de trabajo, el nivel educativo y la formación profesional. En los específicos se indaga sobre tareas específicas de cada ocupación, así como sobre herramientas y tecnologías que se utilizan en dicha ocupación.

Como puede observarse, el volumen de información que ONET UY proporciona sobre cada ocupación, así como el hecho de que sean miles los programas de capacitación que se ofrecen en las instituciones educativas, las de formación profesional y las plataformas virtuales, hacen que para el proveedor de servicios de orientación profesional sea prácticamente imposible analizar, asimilar y manejar esa información de forma completa y no sesgada.

Frente a esta problemática surgen diversas alternativas destinadas a analizar grandes volúmenes de información. En los últimos 30 años, la rama del aprendizaje automático (*machine learning*) llamada procesamiento del lenguaje natural (*Natural Language Processing* (NLP)) se ha coronado como una de las

¹ Véase <https://www.mitrabajofuturo.gub.uy/>.

² Véase <https://www.onetonline.org/>.

proveedoras más relevantes de herramientas de análisis e interpretación de textos. Las aplicaciones de estas herramientas abarcan desde la clasificación de documentos identificando el tópico de los artículos o detectando posibles autores dado el estilo de la escritura, hasta los traductores automáticos que dejan atrás la traducción literal y se centran en el contexto en que las palabras están inmersas, el análisis de sentimientos en los textos, la extracción automática de conceptos y el autocompletado de textos sobre la base de palabras clave. En este marco, la diversidad de instrumentos que ofrece el NLP lo convierte en un candidato plausible para llevar a cabo un análisis objetivo, sistemático y automatizado que permita establecer un vínculo entre textos similares.

En el campo de la caracterización de ocupaciones según el modelo ONET, se ha utilizado el NLP recientemente para identificar tareas específicas emergentes en nuevas ocupaciones (Dahlke y Putka, 2021). En dicho trabajo se describe el desarrollo de una herramienta de NLP que analiza el texto de 136.955 tareas específicas de ocupaciones de los Estados Unidos y clasifica esas tareas como aceptables o no aceptables en función de criterios previamente establecidos, y, por otro lado, identifica el nivel de redundancia entre tareas. En la publicación, los autores son capaces de establecer grados de similitud entre las distintas tareas específicas mediante el uso de la distancia coseno. Estas métricas les permitieron, por un lado, definir si una tarea específica debía ser sumada como nueva tarea a la base de datos de ONET y, en ese caso, ser asociada a una o más ocupaciones, y, por el otro, identificar redundancias de tareas equivalentes pero con redacción levemente diferente en el conjunto de tareas actualmente en uso por ONET Estados Unidos.

El objetivo del presente trabajo es desarrollar una herramienta que permita relacionar las ocupaciones y los programas de capacitación utilizando herramientas de NLP. En primer lugar, se presenta la aproximación metodológica que se aplica para procesar y analizar el contenido tanto de las ocupaciones como de los programas de capacitación. En segundo lugar, se presentan los resultados y su validación y, por último, se discuten las implicancias de estos hallazgos para el servicio de desarrollo de carrera del sector público y del privado, y las futuras investigaciones y desarrollos que podrían llevar a cabo las instituciones que forman parte del sistema de aprendizaje a lo largo de la vida en el Uruguay.

I. Metodología

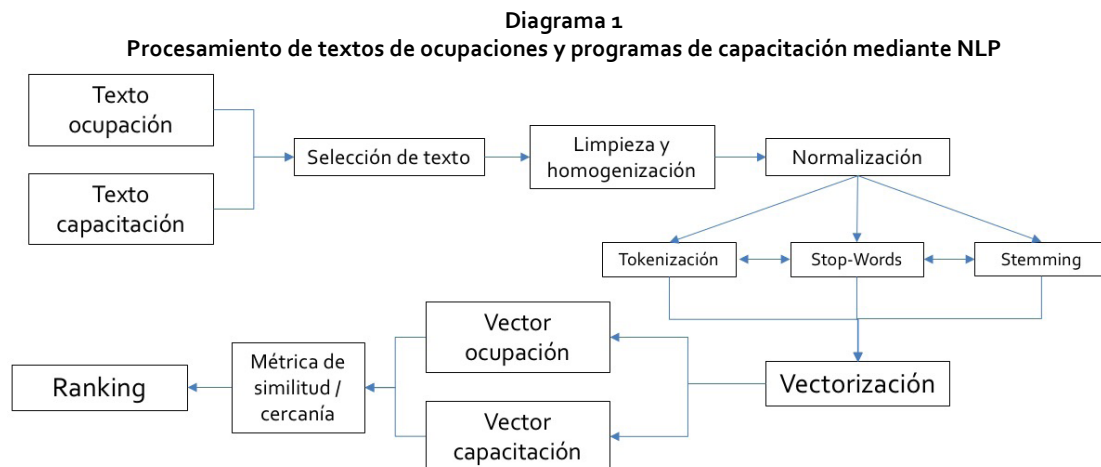
Los códigos y requisitos de Python utilizados en el presente trabajo se pueden consultar solicitándolos al MTSS del Uruguay.

Con el fin de establecer un vínculo entre las ocupaciones y los programas de capacitación, fue necesario generar un corpus descriptivo que contuviera la mayor cantidad de información pertinente que describiera el objeto de estudio. Por lo tanto, en un primer paso se identificaron los textos que mejor explicarían cada una de las ocupaciones y los programas, y que más información aportasen a su descripción.

El corpus de las ocupaciones se construyó a partir de los microdatos de las 22 ocupaciones de la primera ola del relevamiento de ONET UY (Velardez, 2021), que fueron proporcionados por el MTSS en un archivo .csv. A partir de esa base de microdatos se extrajeron los valores de los seis bloques de preguntas de los cuestionarios comunes, el contenido de las tareas específicas, y el listado de herramientas y tecnologías. Por otro lado, se utilizaron las descripciones traducidas y adaptadas de las variables de los cuestionarios comunes de ONET de los Estados Unidos. Por último, se utilizó la definición de cada ocupación según la Clasificación Internacional Uniforme de Ocupaciones de 2008 (CIUO-08).

Los datos de los programas de capacitación fueron proporcionados por la Universidad del Trabajo del Uruguay (UTU) en tres documentos en formato de texto que contenían una descripción parcial de los programas. Se generó de forma manual un archivo que contenía la identificación y el nombre de los programas, así como el texto correspondiente al campo "perfil_egresado_programa". Se incluyeron solamente los 203 programas respecto de los cuales había información disponible en dicho campo.

Una vez generado el corpus bruto que contenía el texto de las ocupaciones y de los programas de capacitación, se procedió a preprocesarlo para poder aplicar correctamente las herramientas de NLP. Con el fin de determinar vínculos entre las ocupaciones y los programas de capacitación, se optó por utilizar NLP para vectorizar los textos, es decir, para transferir las palabras a un espacio matemático de múltiples variables. Dado que no existe una única manera de llevar a cabo esa vectorización, se exploraron diversos métodos, en cada uno de los cuales se resaltaban distintos aspectos de los textos. Una vez transferidos a un espacio vectorial, se estableció una métrica de similitud. En este marco fue posible determinar un *ranking* de similitudes sobre la base de la cercanía. En el diagrama 1 se brinda un esquema resumido de este proceso.



Fuente: Elaboración propia.

Nota: *Stop-words* se refiere a palabras sin un significado (artículos, pronombres). *Stemming* se refiere a la reducción de una palabra a su raíz. "Tokenización" se refiere al proceso de agrupar una secuencia de caracteres que serán considerados como una unidad semántica.

El texto de entrada de las ocupaciones contenía los siguientes seis datos: i) código CIUO-08; ii) nombre de la ocupación; iii) descripción de la ocupación según CIUO-08; iv) nombre de la pregunta del cuestionario común; v) descripción de la pregunta del cuestionario común, y vi) descripción de tareas específicas. Los campos "nombre_pregunta" y "descripcion_pregunta" contenían la totalidad del nombre y la descripción de las preguntas de los bloques de cuestionarios comunes seleccionados cuyas respuestas promedio en cada ocupación sobrepasaban el umbral de 3,5. Eso significaba que, en promedio, los trabajadores de la ocupación habían valorado dicha variable como importante, muy importante o extremadamente importante.

El texto de entrada de los programas de capacitación contenía los siguientes tres datos: i) identificación del programa; ii) nombre del programa, y iii) perfil del egresado.

A partir de estas dos fuentes se realizó un preprocesamiento. Respecto de cada ocupación se concatenaron los datos correspondientes al nombre, la descripción y los campos "nombre_pregunta", "descripcion_pregunta" y "tareas_especificas". Respecto de los programas de capacitación se concatenaron los campos correspondientes al nombre del programa y al perfil del egresado. Se procedió a eliminar los caracteres especiales y ortográficos de estos dos conjuntos de texto, a convertir la totalidad del texto en minúscula, a eliminar los caracteres HTML (que estaban presentes dado el origen de los datos) y a quitar los caracteres numéricos. Esto se hizo con el fin de homogeneizar los corpus.

En un segundo paso se tokenizaron los corpus y cada palabra se convirtió en una unidad semántica mínima (un *token*). Se eliminaron las palabras cuya función en el texto no aportaba a la descripción del mismo. Esas palabras, que en su mayoría son proposiciones, pronombres y conjunciones, suelen denominarse *stop-words*. También se eliminaron específicamente las palabras "capacidad" y "conocimiento", pues se observó que aparecían con más frecuencia y no agregaban información adicional alguna. Finalmente, se optó por realizar un *stemming* (radicalización) de los *tokens* (palabras). Este procedimiento automático consiste en llevar las formas derivadas de una palabra a una forma básica común: palabras como electricidad, electricistas y eléctricamente son mapeadas en el elemento "electr".

Una vez obtenidos los textos semánticamente ilegibles pero conteniendo la información pertinente para su descripción, se procedió a vectorizarlos. Se unieron ambos corpus y se aplicó la herramienta TF-IDF (Chowdhury, 2010), que consiste en generar un espacio vectorial de dimensión equivalente a la cantidad de *tokens* distintos que hay en el corpus. Una vez establecido el espacio de trabajo, cada texto es llevado hasta allí teniendo en cuenta dos aspectos: la frecuencia con que los términos aparecen en el texto y en

el corpus. Es decir, se consideró la cantidad de veces que cada *token* se repetía en cada descripción de las ocupaciones y los programas de capacitación (*term frequency* (TF)) dentro de su propia definición y se la pesó por la cantidad de veces que el término se repetía a lo largo del corpus, en escala logarítmica (*inverse document frequency* (iDF)). El resultado de esta transformación del texto es un conjunto de vectores separados espacialmente pero que, por su construcción, se encuentran dispersos por regiones cuyos textos originarios mantienen determinada cercanía y similitud entre *tokens*.

Definido ya un marco vectorial, se procedió a establecer una relación matemática entre las ocupaciones y los programas de capacitación. Con ese fin se calculó la matriz de distancias (euclidianas) entre todos los vectores generados por los textos limpios de entrada de las ocupaciones y los textos limpios de entrada de los programas de capacitación. Para cada vector de ocupación se seleccionaron los diez vectores correspondientes a los programas de capacitación más próximos como posibles candidatos, siempre y cuando esos vectores estuvieran dentro de la hipersfera de radio 1,45. Este valor se estableció con el fin de no obtener recomendaciones de programas de capacitación que estuvieran demasiado alejados de la definición de la ocupación.

Si bien con esta metodología se logró establecer un vínculo entre las ocupaciones y los programas de capacitación sobre la base de la repetibilidad de los conceptos clave (véase la sección II), se exploró una segunda manera de vectorizar los textos. Con el fin de no limitarse a la frecuencia de los términos, se optó por emplear la herramienta Doc2Vec para hacer una representación numérica del corpus (Le y Mikolov, 2014). En esta herramienta, creada sobre la base de su antecesor Word2Vec, se utilizan redes neuronales que generan *embeddings* (asignaciones), determinando la posición de los textos (en su totalidad) en un espacio vectorial. A diferencia de lo que ocurre con la herramienta TF-iDF, en donde la dimensión del espacio de salida queda unívocamente determinada por la cantidad de términos distintos del corpus, en el caso de Doc2Vec la dimensión es un hiperparámetro a fijar y no guarda ningún correlato directo con las características del texto. Más aún, los pesos de las redes necesarias para convertir los textos en vectores se descarga previamente de otras bases ya entrenadas que pueden volver a entrenarse con corpus nuevos o específicos. Una ventaja destacable de Doc2Vec frente a TF-iDF en lo que respecta a la comparación de textos es que en el primero se tiene en cuenta la totalidad del texto y no cada elemento de manera individual. Por ejemplo, en el espacio creado con la herramienta TF-iDF, las oraciones “me gusta el auto” y “el vehículo es de mi agrado” se encontrarán separadas porque no contienen términos en común, salvo “el”; por otro lado, si la conversión se hace mediante Doc2Vec, esas oraciones se encontrarán próximas, ya que la herramienta reconoce que el par “gusta” y “agrado” y el par “auto” y “vehículo” son similares.

La metodología que se utilizó para llevar a cabo la vectorización vía Doc2Vec fue análoga a la que se empleó en el caso de TF-iDF: limpieza y homogeneización, normalización, tokenización, eliminación de *stop- words* y *stemming*. Se optó por volver a entrenar el modelo preexistente con el corpus completo para garantizar una mayor convergencia. Sin embargo, pese a haber realizado una exploración en el espacio de parámetros, los resultados de la vinculación entre las ocupaciones y los programas de capacitación no fue convincente y dio lugar a múltiples asignaciones sin sentido (dato no mostrado). Esto podría estar muy ligado a la descripción de las ocupaciones. Sin embargo, se observó que este hecho no ocurría si se analizaba la cercanía entre los programas de capacitación. Por consiguiente, se optó por determinar los diez programas de capacitación más próximos (y por ende más similares) a cada programa de capacitación. Como ya se explicó, dada la interpretación del espacio vectorial en que se trabajó, no fue necesario fijar un valor límite de cercanía. Para determinar la dimensión del espacio de trabajo se tomó una muestra aleatoria constituida por el 20% de todos los programas de capacitación y se analizó el efecto del tamaño de la dimensión evaluando el número de programas correctos, es decir, que estuvieran relacionados o fueran cercanos al programa de capacitación de referencia (clasificación realizada por un humano). Se evaluaron espacios de dimensión 10, 20, 50 y 100, y los mejores resultados se obtuvieron con el espacio vectorial de dimensión 10, que se utilizó como hiperparámetro en todo el trabajo.

Habiendo aplicado Doc2Vec y establecido una correspondencia de cercanía entre los programas de capacitación sobre la base de su descripción, se retomó el análisis previo obtenido mediante TF-IDF. Para cada uno de los diez programas de capacitación candidatos del primer método, se expandió el listado con las similitudes determinadas por el segundo método. Posteriormente se estudió la frecuencia de repetición de las recomendaciones y eso se cristalizó en un nuevo *ranking* basado en las recomendaciones de ambos métodos (recomendación mixta).

II. Resultados

En una primera etapa del trabajo se llevó a cabo un análisis de cercanía de textos entre ocupaciones y programas de capacitación mediante TF-IDF, utilizando la distancia euclidiana entre los vectores generados. Se confeccionó un *ranking* de cercanía y se seleccionaron los diez programas de capacitación más cercanos considerando que los mismos deberían ser aquellos programas más relacionados con una ocupación dada en términos de frecuencia de las palabras utilizadas. Se evaluó la pertinencia del listado de diez programas recomendados de cada ocupación y se consideraron pertinentes los programas que cumplían alguna de las siguientes tres condiciones:

- i) El programa de capacitación pertenece a la rama de actividad de la ocupación.
- ii) La persona que se dedica a esa ocupación posee el tipo de habilidades necesarias para cursar el programa (habilidades cognitivas o manuales).
- iii) La persona que se dedica a esa ocupación posee conocimientos relacionados con el programa de capacitación.

Como se observa en el cuadro 1, los resultados obtenidos en el *ranking* de programas de capacitación recomendados mediante TF-IDF son heterogéneos. Por un lado, hay ocupaciones, como peón industrial, diseñador/a gráfico/a, operario/a y peón rural, en que todos los programas recomendados parecen ser pertinentes (véanse las celdas sombreadas con texto en negritas). En contraposición, las ocupaciones empleado/a de inventario, chofer de auto, taxi o camioneta, limpiador/a, peón de carga y reponedor/a, muestran un bajo porcentaje de pertinencia en el listado de diez programas recomendados (10% a 30%).

Cuadro 1
Ranking de cercanía de programas de capacitación para cada una de las 22 ocupaciones de la primera ola de relevamiento de ONET UY

Nombre de la Ocupación	Recomendación_1	Recomendación_2	Recomendación_3	Recomendación_4	Recomendación_5	Recomendación_6	Recomendación_7	Recomendación_8	Recomendación_9	Recomendación_10
Peón Industria	Carpintería (1.317)	Carpintería de Ribera (1.318)	Técnicas Creativas Énfasis en Joyería y Fibra (1.332)	Mecánica Automotriz (1.332)	Construcción Énfasis Arquitectura (1.34)	Industrias Lácteas (1.348)	Construcción Énfasis Obra Seca (1.349)	Belleza capilar (1.352)	Diseño Aplicación En Madera (1.353)	Automatización - Agrícola /Ganadero/Agrícola /Forestal (1.355)
Diseñador/a Gráfico/a	Diseño (1.257)	Diseño Gráfico en Comunicación Visual (1.294)	Audiovisual (1.313)	Comunicación Social Opción Publicidad (1.313)	Comunicación Social Opción Televisión (1.315)	Comunicación Social Opción Prensa (1.316)	Comunicación Social Opción Radiodifusión (1.318)	Audiovisual con Mención en Producción, Dirección de Arte y Guión (1.319)	Artes Gráficas (1.325)	Audiovisual con Mención en Captura y Procesamiento de Imagen y Sonido (1.327)
Contador/a	Administración (1.193)	Administración (1.252)	Administración (1.286)	Administración (1.304)	Gestión Humana (1.311)	Logística (1.319)	Agencia de Viajes y Alojamiento (1.329)	Ingeniero Tecnológico Aeronáutica (1.332)	Producción Agropecuaria Familiar (1.337)	Química Industrial (1.338)
Representante Comercial	Alojamiento (1.237)	Agencia de Viajes y Alojamiento (1.292)	Gastronomía (1.307)	Gestión de Alojamiento (1.309)	Logística (1.316)	Gestión de Empresas Turísticas (1.318)	Tecnólogo Prevencionista (1.33)	Producción Lechera (1.331)	Ingeniero Tecnológico Prevencionista (1.335)	Gastronomía (1.337)
Oficinista	Alojamiento (1.332)	Comunicación Social Opción Prensa (1.347)	Comunicación Social Opción Radiodifusión (1.351)	Gestión Humana (1.352)	Administración (1.356)	Administración (1.357)	Agencia de Viajes y Alojamiento (1.357)	Secretariado Ejecutivo Bilingüe (1.358)	Química Industrial (1.36)	Gastronomía (1.36)
Empleado/a Contable	Administración (1.228)	Administración (1.295)	Agencia de Viajes y Alojamiento (1.331)	Redes y Software (1.342)	Agrario (1.349)	Administración (1.351)	Ingeniero Tecnológico Aeronáutica (1.353)	Alojamiento (1.36)	Minería (1.361)	Robótica (1.361)
Empleado/a Inventario	Administración (1.335)	Ingeniero Tecnológico Aeronáutica (1.339)	Química Industrial (1.354)	Comunicación Social Opción Televisión (1.359)	Reparación de PC (1.361)	Alojamiento (1.361)	Instalaciones Sanitarias (1.361)	Mecánica Automotriz (1.361)	Químico Plan 2010 (1.362)	Mecánica De Producción (1.363)
Cocinero/a	Gastronomía (1.311)	Alojamiento (1.318)	Gastronomía (1.331)	Químico Plan 2010 (1.332)	Gastronomía (1.339)	Tecnólogo Prevencionista (1.345)	Producción Lechera (1.346)	Mecánico (1.346)	Logística (1.346)	Ingeniero Tecnológico Prevencionista (1.347)
Asistente Venta	Gastronomía (1.285)	Alojamiento (1.294)	Gastronomía (1.315)	Agencia de Viajes y Alojamiento (1.337)	Barman (1.356)	Telecomunicaciones (1.36)	Mecánica Automotriz (1.365)	Logística (1.365)	Comunicación Social Opción Relaciones Públicas (1.366)	Comunicación Social Opción Radiodifusión (1.368)
Cajero/a	Alojamiento (1.344)	Gastronomía (1.345)	Gastronomía (1.365)	Robótica (1.367)	Agencia de Viajes y Alojamiento (1.368)	Barman (1.369)	Administración (1.372)	Organización de Eventos (1.373)	Audiovisual con Mención en Captura y Procesamiento de Imagen y Sonido (1.373)	Logística (1.374)
Cuidador/a	Reacondicionamiento Físico (1.307)	Ingeniero Tecnológico Prevencionista (1.31)	Masaje y Rehabilitación Deportiva (1.313)	Tecnólogo Prevencionista (1.318)	Gestión Humana (1.329)	Belleza capilar (1.331)	Alojamiento (1.339)	Agro - Energético (1.346)	Comunicación Social Opción Radiodifusión (1.351)	Comunicación Social Opción Prensa (1.352)

Cuadro 1 (conclusión)

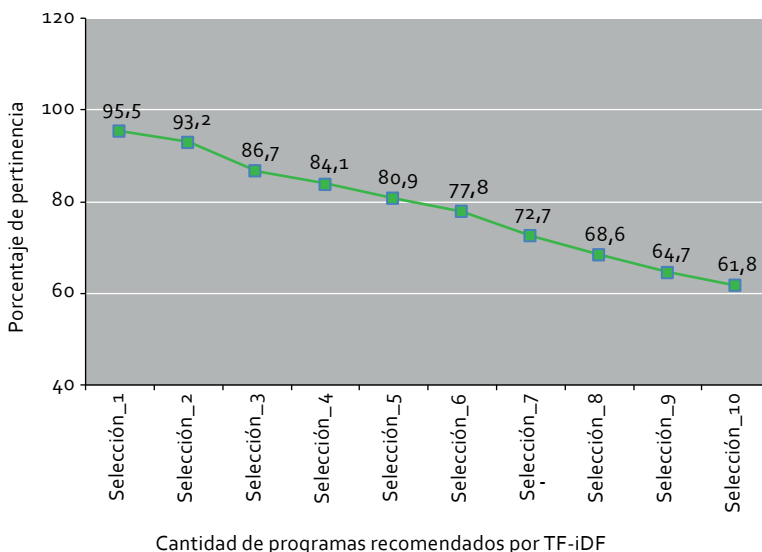
Nombre de la Ocupación	Recomendación_1	Recomendación_2	Recomendación_3	Recomendación_4	Recomendación_5	Recomendación_6	Recomendación_7	Recomendación_8	Recomendación_9	Recomendación_10
Guardia	Náutica Y Pesca (Ciencias Del Mar) (1.328)	Conservación de Recursos Naturales (1.344)	Minería (1.345)	Instalaciones Eléctricas (1.346)	Guardaparque (1.349)	Logística (1.351)	Organización de Eventos (1.353)	Comunicación Social Opción Relaciones Públicas (1.353)	Ingeniero Tecnológico Aeronáutica (1.356)	Alojamiento (1.357)
Albañil	Construcción (1.341)	Soldadura y Calderero Industrial (1.342)	Parques y Jardines (1.35)	Construcción (1.355)	Construcción (1.36)	Construcción de Viviendas y Aberturas (1.363)	Construcción (1.365)	Infografía y Tecnología de la Arquitectura (1.37)	Comunicación Social Opción Publicidad (1.373)	Comunicación Social Opción Radiodifusión (1.374)
Carnicero/a	Producción Animal de carnes y ovinos (1.277)	Ganadería en Régimen de Alternancia (1.326)	Gastronomía (1.326)	Cárnico (1.327)	Náutica Y Pesca (1.336)	Sistema de Producción Arroz Pastura (1.346)	Náutica Y Pesca (Ciencias Del Mar) (1.346)	Patrón de Pesca de Altura (1.347)	Producción Agrícola Ganadera (1.35)	Vitivinicultura (1.35)
Operario/a	Minería (1.291)	Carpintería (1.325)	Carpintería de Ribera (1.326)	Producción Agropecuaria Familiar (1.333)	Mecánica De Producción (1.334)	Maquinista Naval (1.335)	Técnicas Creativas Énfasis en Joyería y Fibra (1.336)	Construcción Énfasis Obra Seca (1.337)	Mecánico (1.337)	Mantenimiento Electro-Mecánico Industrial (1.339)
Chofer Auto, Taxi, Camioneta	Especialista en Logística del Transporte Carretero y Ferroviario (1.269)	Mecánica Automotriz (1.339)	Electro - Electrónica Automotriz (1.343)	Telecomunicaciones (1.356)	Alojamiento (1.36)	Automatización - Instrumentación y Control Plan 2020 (1.363)	Minería (1.363)	Comunicación Social Opción Radiodifusión (1.364)	Carpintería (1.369)	Carpintería de Ribera (1.37)
Chofer Camión	Especialista en Logística del Transporte Carretero y Ferroviario (1.292)	Mecánica Automotriz (1.303)	Maquinista Naval (1.347)	Mecánica De Producción (1.355)	Minería (1.356)	Automatización - Instrumentación y Control Plan 2020 (1.358)	Ingeniero Tecnológico Aeronáutica (1.36)	Soporte y Mantenimiento Informático (1.361)	Mecánica General (1.361)	Carpintería (1.364)
Operario/a Agrícola	Mecanización Agrícola (1.252)	Automatización - Instrumentación y Control Plan 2020 (1.296)	Automatización - Agrícola /Ganadero/Agrícola /Forestal (1.318)	Mecánica Naval (1.319)	Producción Hortícola (1.32)	Mecánica General (1.328)	Mantenimiento Industrial Electromecánico (1.331)	Mecanización Agroindustrial (1.333)	Forestal (1.335)	Maquinista Naval (1.335)
Limpiador/a	Ingeniero Tecnológico Aeronáutica (1.341)	Tecnólogo Aeronáutica (1.351)	Bioteología Industrial (1.356)	Aeronáutica Énfasis Sistemas de Aeronaves (1.357)	Belleza capilar (1.358)	Mecánica Automotriz (1.366)	Aeronáutica Énfasis Motopropulsores (1.371)	Alojamiento (1.372)	Gestión de Alojamiento (1.38)	Técnico en Instalaciones Eléctricas Industriales (1.381)
Peón Rural	Producción Animal de carnes y ovinos (1.227)	Producción Agrícola Ganadera (1.244)	Agrícola Ganadero (1.269)	Ganadería en Régimen de Alternancia (1.291)	Agrícola Ganadero (1.308)	Producción Lechera (1.34)	Producción Lechera (1.342)	Agrario (1.346)	Producción Agropecuaria Familiar (1.347)	Mecanización Agrícola (1.347)
Peón Carga	Especialista en Logística del Transporte Carretero y Ferroviario (1.225)	Maquinista Naval (1.3)	Patrón de Pesca de Ultramar (1.341)	Electro-Eléctrico Naval OMI (1.344)	Patrón de Hidrovía (1.357)	Construcción de Muebles por Diseño (1.366)	Termodinámica (1.366)	Patrón de Pesca de Altura (1.367)	Maquinista Naval (1.368)	Náutica Y Pesca (1.37)
Reponedor/a	Alojamiento (1.349)	Bioteología Industrial (1.374)	Metalúrgica (1.376)	Ingeniero Tecnológico Aeronáutica (1.38)	Agencia de Viajes y Alojamiento (1.381)	Textil (1.381)	Mantenimiento Electro-Mecánico Industrial (1.382)	Hortifruticultura (1.383)	Comunicación Social Opción Radiodifusión (1.384)	Comunicación Social Opción Publicidad (1.385)

Fuente: Elaboración propia, sobre la base de información proporcionada por la Dirección de Estadísticas Laborales, Ministerio de Trabajo y Seguridad Social.

Nota: Las cifras entre paréntesis indican la distancia relativa. Las celdas sombreadas en rosa con texto en negritas contienen los programas que se consideraron pertinentes.

Dado que los programas recomendados se encuentran ordenados por cercanía, es esperable que un programa recomendado como primera opción tenga más probabilidades de ser pertinente que un programa recomendado en décimo lugar. Esto efectivamente sucede y se puede observar que 21 de los 22 programas que están listados como primera recomendación fueron clasificados como pertinentes (celdas sombreadas en columna "Recomendación_1" del cuadro 1), mientras que solo 8 de los 22 programas recomendados en posición 10 fueron clasificados como tales.

Gráfico 1
Proporción de programas de capacitación recomendados por la herramienta TF-IDF que cumplen con el criterio de pertinencia según número de recomendaciones seleccionadas
(En porcentajes)



Fuente: Elaboración propia, sobre la base de información proporcionada por la Dirección de Estadísticas Laborales, Ministerio de Trabajo y Seguridad Social.

Con estos resultados se podría crear una herramienta que fuera capaz de recomendar, para cada ocupación, un solo programa de capacitación cuya probabilidad de ser pertinente fuera superior al 95%. Sin embargo, la finalidad de la herramienta es ofrecer un listado de posibles programas para que el usuario evalúe y tome una decisión informada sobre las trayectorias formativas posibles en función de sus objetivos profesionales. Si se consideran los dos primeros programas recomendados, el porcentaje cae al 93,2%, valor que sigue siendo alto en términos comparativos. Sin embargo, esta probabilidad decrece a medida que se incorporan más programas al listado de recomendaciones y, si se recomiendan diez programas de capacitación para las 22 ocupaciones de la primera ola de relevamiento, el valor cae al 61,8%.

Una alternativa para recomendar más de un programa de capacitación y mantener un alto grado de pertinencia sería tomar el primer programa recomendado por el método TF-IDF y buscar los programas que estuvieran más relacionados con él. Para ello sería necesario determinar no solo si es posible formar familias de programas relacionados, sino también si esas familias tienen sentido desde el punto de vista conceptual.

Con el fin de evaluar la cercanía entre el contenido de los programas de capacitación, se analizó el texto de los 203 programas y se evaluó la cercanía mediante la distancia coseno entre vectores generados con Doc2Vec. En el cuadro 2 se muestran a modo de ejemplo los diez programas más cercanos a dos programas de capacitación de referencia: "Electricidad y Cerámica". Como se puede observar, en ambos casos nueve de los diez programas sugeridos por Doc2Vec tienen un alto grado de relación. En el caso del

primero, se observa que los programas relacionados tienen en común aspectos energéticos, ópticos o de la rama de actividad de la construcción. En color se resalta el programa “Sala-Bar”, que *a priori* parece guardar menos relación con el programa “Electricidad” que los otros nueve programas sugeridos para integrar la misma familia. Con el programa de referencia “Cerámica” se observan resultados similares: nueve de los programas sugeridos comparten con él aspectos vinculados con el diseño, la creatividad y el arte. El único que *a priori* estaría menos relacionado sería el de “Producción Agropecuaria Sustentable”.

Cuadro 2
Familias de programas de capacitación: ejemplo de familias de programas de capacitación altamente relacionados

Ranking	Capacitación de referencia	Capacitación relacionada	Distancia
1	Electricidad	Instalaciones Eléctricas	0.89798892
2	Electricidad	Optica	0.88876820
3	Electricidad	Instalaciones Sanitarias	0.88767183
4	Electricidad	Instalaciones Eléctricas	0.84422404
5	Electricidad	Energías Renovable - Fotovoltaica - Eolica	0.84020722
6	Electricidad	Construcción	0.83706355
7	Electricidad	Redes y Comunicaciones Ópticas	0.81667823
8	Electricidad	Mantenimiento Industrial Electromecánico	0.80248296
9	Electricidad	Sala - Bar	0.79803336
10	Electricidad	Seguridad Electrónica para la vivienda	0.79646420
1	Cerámica	Dibujo y Pintura	0.99950242
2	Cerámica	Joyería	0.99843043
3	Cerámica	Escultura	0.99738884
4	Cerámica	Diseño	0.94229805
5	Cerámica	Artes Y Artesanias Figari	0.94159222
6	Cerámica	Audiovisual	0.86896104
7	Cerámica	Diseño Aplicación En Madera	0.85882789
8	Cerámica	Diseño Gráfico en Comunicación Visual	0.84556431
9	Cerámica	Técnicas Creativas Enfasis en Joyería y Fibra	0.79743159
10	Cerámica	Produccion Agropecuaria Sustentable	0.76646674

Fuente: Elaboración propia, sobre la base de información proporcionada por la Dirección de Estadísticas Laborales, Ministerio de Trabajo y Seguridad Social.

Nota: Las filas sombreadas en rosa indican programas que guardan escasa relación con el programa de referencia.

En el otro extremo se observan ejemplos en que Doc2Vec parece sugerir programas de capacitación que están escasamente relacionados con el programa de referencia. Este es el caso de los programas “Biotecnología industrial” o “Químico Plan 2010”: siete de los diez programas sugeridos para el primero no guardan relación con él, y ninguno de los diez programas sugeridos para el segundo están relacionados con él (véase el cuadro 3).

Una vez obtenidos estos resultados se procedió a definir el listado de programas de capacitación recomendados para cada ocupación mediante un proceso en el que se combina el *ranking* de programas recomendados a partir de los resultados obtenidos con TF-IDF y el conjunto de programas más cercanos obtenidos a partir de los resultados de Doc2Vec.

Los resultados de la recomendación de programas de capacitación para cada ocupación se muestran en el cuadro 4. Dado que la condición para integrar el listado de programas de capacitación recomendados para una ocupación dada es tener una frecuencia igual o mayor a dos y se utilizaron los primeros cuatro programas recomendados por TF-IDF, el número de programas a recomendar puede tomar valores entre 1 y 20. La estrategia utilizada consistió en tomar un número de programas recomendados por TF-IDF cuyo porcentaje de pertinencia fuera alto, y aumentar el número de recomendaciones utilizando las familias de programas relacionados construidas a partir de los resultados de Doc2Vec. La estrategia se considera exitosa si permite aumentar el número de programas recomendados sin disminuir el porcentaje de pertinencia comparado con el resultado de TF-IDF.

Cuadro 3
Familias de programas de capacitación: ejemplo de familias de programas de capacitación escasamente relacionados

Ranking	Capacitación de referencia	Capacitación relacionada	Distancia
1	Biotecnología Industrial	Recreación	0.87639773
2	Biotecnología Industrial	Control Ambiental - Plan 2019	0.85823953
3	Biotecnología Industrial	Conservación de Recursos Naturales	0.84964508
4	Biotecnología Industrial	Ingeniero Tecnológico Prevencionista	0.82960290
5	Biotecnología Industrial	Tecnólogo Prevencionista	0.82592452
6	Biotecnología Industrial	Construcción Énfasis Arquitectura	0.82559878
7	Biotecnología Industrial	Construcción Énfasis Obra Seca	0.81945169
8	Biotecnología Industrial	Minería	0.76312983
9	Biotecnología Industrial	Manejo de Sistemas de Producción Lechera	0.75999641
10	Biotecnología Industrial	Infografía y Tecnología de la Arquitectura	0.73508441
1	Químico Plan 2010	Construcción	0.82607168
2	Químico Plan 2010	Construcción	0.82568949
3	Químico Plan 2010	Construcción	0.81466591
4	Químico Plan 2010	Cárnico	0.80585086
5	Químico Plan 2010	Refrigeración	0.79732752
6	Químico Plan 2010	En Madera	0.79346585
7	Químico Plan 2010	Agro - Energético	0.79247171
8	Químico Plan 2010	Minería	0.75914359
9	Químico Plan 2010	Dibujo de la Construcción	0.75861752
10	Químico Plan 2010	Gastronomía	0.75456423

Fuente: Elaboración propia, sobre la base de información proporcionada por la Dirección de Estadísticas Laborales, Ministerio de Trabajo y Seguridad Social.

Nota: Las filas sombreadas en rosa indican programas que guardan escasa relación con el programa de referencia.

Cuadro 4
Listado de programas de capacitación recomendados para cada ocupación de la primera ola de relevamiento de ONET UY

CIUO-08	Ocupación	Programa	Frecuencia
932	Peón Industria	Mecánica De Producción	3
		Mantenimiento Industrial Electromecánico	2
		Carpintería de Ribera	2
		Construcción	2
		Mecánico	2
		Mecánica General	2
		Refrigeración	2
		Soporte y Mantenimiento Informático	2
		Refrigeración	2
		Carpintería	2
		2166	Diseñador/a Gráfico/a
		Audiovisual	4
		Cerámica	3
		Escultura	3
		Joyería	3
		Dibujo y Pintura	3
		Técnicas Creativas Énfasis en Joyería y Fibra	2
		Producción Agropecuaria Sustentable	2
		Comunicación Social Opción Prensa	2
		Comunicación Social Opción Publicidad	2
		Comunicación Social Opción Radiodifusión	2
		Artes Y Artesanías Figari	2
		Diseño	2
2411	Contador/a	Administración	3
		Robótica Y Telecomunicaciones	2
		Telecomunicaciones	2
		Instalaciones Eléctricas	2
		Administración	2
		Gestión de Alojamiento	2
		Administración	2
Aeronáutica Énfasis Avionica	2		
		IT y Redes	2

Cuadro 4 (continuación)

CIUO-08	Ocupación	Programa	Frecuencia
3322	Representante Comercial	Administración	3
		Agencia de Viajes y Alojamiento	3
		Gestión de Alojamiento	2
		Audiovisual con Mención en Producción, Dirección de Arte y Guión	2
		Comunicación Social Opción Relaciones Públicas	2
		Comunicación Social Opción Televisión	2
		Gestión de Empresas Turísticas	2
		Operador Inmobiliario	2
		Gastronomía	2
		Textil	2
4110	Oficinista	Comunicación Social Opción Prensa	3
		Comunicación Social Opción Relaciones Públicas	3
		Comunicación Social Opción Televisión	3
		Agencia de Viajes y Alojamiento	3
		Diseño Gráfico en Comunicación Visual	2
		Comunicación Social Opción Publicidad	2
		Comunicación Social Opción Radiodifusión	2
		Vestimenta Prendas de Punto	2
		Especialista en Logística del Transporte Carretero y Ferroviario	2
		Turismo	2
4311	Empleado/a Contable	Administración	3
		Administración	3
		Gestión de Alojamiento	3
		Telecomunicaciones	2
		Instalaciones Eléctricas	2
		Aeronáutica Énfasis Avionica	2
		IT y Redes	2
		Audiovisual con Mención en Producción, Dirección de Arte y Guión	2
		Comunicación Social Opción Relaciones Públicas	2
		Comunicación Social Opción Televisión	2
4321	Empleado/a Inventario	Comunicación Social Opción Prensa	2
		Audiovisual con Mención en Producción, Dirección de Arte y Guión	2
		Comunicación Social Opción Relaciones Públicas	2
		Comunicación Social Opción Televisión	2
		Agencia de Viajes y Alojamiento	2
		Agencia de Viajes y Alojamiento	2
5120	Cocinero/a	Refrigeración	3
		Gastronomía	3
		Soporte y Mantenimiento Informático	2
		Agencia de Viajes y Alojamiento	2
		Gastronomía	2
		Alojamiento	2
5223	Asistente Venta	Agencia de Viajes y Alojamiento	3
		Gastronomía	3
		Soporte y Mantenimiento Informático	2
		Refrigeración	2
		Administración	2
		Comunicación Social Opción Relaciones Públicas	2
		Comunicación Social Opción Televisión	2
		Textil	2
Gastronomía	2		
5230	Cajero/a	Alojamiento	2
		Gastronomía	3
		Soporte y Mantenimiento Informático	2
		Refrigeración	2
		Autotrónica	2
		Agencia de Viajes y Alojamiento	2
Gastronomía	2		
Alojamiento	2		

Cuadro 4 (continuación)

CIUO-08	Ocupación	Programa	Frecuencia
5321	Cuidador/a	Reacondicionamiento Físico	2
		Gastronomía	2
		Biología Industrial	2
		Tecnólogo Prevencionista	2
		Ingeniero Tecnológico en Electrotecnia	2
		Ingeniero Tecnológico Prevencionista	2
		Hípica	2
		Deporte Y Recreación	2
		Belleza postizos	2
		Estética Integral	2
		Conservación y Gestión de Áreas Naturales	2
		Masaje y Rehabilitación Deportiva	2
		Estética Integral Masaje	2
		Estética Integral Capilar	2
		Construcción	2
		Conservación de Recursos Naturales	2
		Recreación	2
		Instalaciones Sanitarias	2
		Control Ambiental - Plan 2019	2
		Manejo de Sistemas de Producción Lechera	2
5414	Guardia	Control Ambiental - Plan 2019	2
7112	Albañil	Minería	2
		Construcción	2
		Construcción	2
		Operación y Mantenimiento de Instalaciones de Transmisión	2
7511	Carnicero/a	Construcción	2
		Refrigeración	2
		Forestal	2
		Arrocero	2
		Agroforestal	2
			2
8183	Operario/a	Construcción	3
		Mecánica De Producción	2
		Mantenimiento Industrial Electromecánico	2
		Carpintería de Ribera	2
		Mecánico	2
		Mecánica General	2
		Refrigeración	2
		Soporte y Mantenimiento Informático	2
		Refrigeración	2
Carpintería	2		
8322	Chofer Auto, Taxi, Camioneta	Mecánica Automotriz	2
		Autotrónica	2
		Robótica	2
		Mecánica Automotriz	2
		Termodinámica	2
		Motores de Combustión Interna	2
		Electro - Electrónica Automotriz	2
		Chapa Y Pintura	2
		Electromecánica Automotriz	2
8332	Chofer Camión	Mecánica De Producción	2
		Refrigeración	2
		Mecánica Automotriz	2
		Mecánica Automotriz	2
		Termodinámica	2
		Electro - Electrónica Automotriz	2
		Chapa Y Pintura	2
		Electromecánica Automotriz	2
		Sala - Bar	2
Refrigeración	2		
8341	Operario/a Agrícola	Mecánica General	2
		Refrigeración	2
		Motores de Combustión Interna	2
		Automatización - Instrumentación y Control Plan 2020	2
		Mecánica Automotriz	2
		Mecatrónica Industrial	2
		Automatización - Agrícola /Ganadero/Agrícola/Forestal	2
		Reparación de PC	2
		Mantenimiento Electro-Mecánico Industrial	2

Cuadro 4 (conclusión)

CIUO-08	Ocupación	Programa	Frecuencia
9112	Limpiador/a	Tecnólogo Aeronáutica	3
		Ingeniero Tecnológico Aeronáutica	3
		Automatización - Instrumentación y Control Plan 2020	3
		Aeronáutica Énfasis Sistemas de Aeronaves	3
		Aeronáutica Énfasis Motores	3
		Motores de Combustión Interna	2
		Mecanización Agroindustrial	2
		Mecánica Automotriz	2
		Mecatrónica Industrial	2
9212	Peón Rural	Automatización - Agrónoma /Ganadero/Agrícola/Forestal	2
		Forestal	3
		Agrícola Ganadero	3
		Producción Agrícola Ganadera	3
		Producción Lechera	3
		Agrícola Ganadero	3
		Ganadería en Régimen de Alternancia	3
		Hortifruticultura	3
		Producción Lechera	2
		Hortifruticultura	2
		Arrocero	2
Producción Agropecuaria Familiar	2		
9333	Peón Carga	Patrón de Pesca de Ultramar	3
		Electro-Electrónico Naval OMI	3
		Maquinista Naval	3
		Sala - Bar	2
		Administración	2
		Aeronáutica Énfasis Avionica	2
		Aeronáutica Énfasis Sistemas de Aeronaves	2
		Patrón de Pesca de Altura	2
		Maquinista Naval	2
		Barman	2
		Producción Animal de carnes y ovinos	2
Producción Granjera	2		
9334	Reponedor/a	Minería	2

Fuente: Elaboración propia, sobre la base de información proporcionada por la Dirección de Estadísticas Laborales, Ministerio de Trabajo y Seguridad Social.

Nota: Las celdas sombreadas en rosa con texto en negritas indican que el programa está escasamente relacionado con la ocupación.

Los resultados muestran que la estrategia utilizada efectivamente permitió aumentar el número de programas recomendados por ocupación: dicho número pasó de 4 en los resultados obtenidos con TF-iDF a 8,9 en los resultados obtenidos mediante la recomendación mixta (TF-iDF-Doc2Vec) (véase el cuadro 5). Solamente dos ocupaciones ("guardia" y "reponedor/a") obtuvieron menos de cuatro programas recomendados (véanse las celdas sombreadas en rosa en la columna Total del cuadro 5). Más aún, al aplicar la estrategia de recomendación mixta aumenta el porcentaje promedio de programas que cumplen el criterio de pertinencia, que pasa del 65% en el método TF-iDF (tomando nueve programas recomendados) al 74% en la recomendación mixta, con un número comparable de programas recomendados (8,9). En síntesis, la herramienta desarrollada fue capaz de recomendar nueve programas de capacitación por cada ocupación con una probabilidad de 0,74 de que estos cumplan con el criterio de pertinencia (lo que significa que 1 de cada 4 programas recomendados no cumple dicho criterio).

Cuadro 5
Resumen de resultados de estrategia mixta TF-iDF-Doc2Vec

CIUO-08	Ocupación	Total	Pertinentes	Porcentaje Mixta	Porcentaje TF-iDF_9
932	Peón Industria	10	9	90%	100%
2166	Diseñador/a Gráfico/a	13	12	92%	100%
2411	Contador/a	9	7	78%	55%
3322	Representante Comercial	11	11	100%	66%
4110	Oficinista	10	8	80%	88%
4311	Empleado/a Contable	13	11	85%	66%
4321	Empleado/a Inventario	5	0	0%	22%
5120	Cocinero/a	6	4	67%	67%
5223	Asistente Venta	10	8	80%	66%
5230	Cajero/a	7	6	86%	77%
5321	Cuidador/a	20	10	50%	66%
5414	Guardia	1	1	100%	66%
7112	Albañil	5	5	100%	77%
7511	Carnicero/a	4	4	100%	44%
8183	Operario/a	10	10	100%	100%
8322	Chofer Auto, Taxi, Camioneta	9	8	89%	33%
8332	Chofer Camión	10	9	90%	55%
8341	Operario/a Agrícola	9	8	89%	100%
9112	Limpiador/a	10	2	20%	33%
9212	Peón Rural	11	11	100%	100%
9333	Peón Carga	12	4	33%	11%
9334	Reponedor/a	1	0	0%	33%
	PROMEDIO TOTAL			74%	65%

Fuente: Elaboración propia, sobre la base de información proporcionada por la Dirección de Estadísticas Laborales, Ministerio de Trabajo y Seguridad Social.

Nota: Las celdas sombreadas en rosa con texto en negrita indican disminución con respecto a la obtenida con TF-iDF. Las celdas sombreadas en verde con texto en negrita indican aumento con respecto a la obtenida con TF-iDF. Total = número de programas recomendados por la estrategia mixta. Pertinentes = número de programas pertinentes con la estrategia mixta. Porcentaje Mixta = porcentaje de programas pertinentes respecto de los programas totales recomendados por la estrategia mixta. Porcentaje TF-iDF_9 = Porcentaje de programas pertinentes utilizando los primeros nueve programas recomendados utilizando TF-iDF.

III. Validación

A continuación se intentó comprender las razones por las que se habían obtenido algunas asignaciones incorrectas. Hay varias causas posibles, y una de ellas podría atribuirse a la calidad del texto de entrada correspondiente a los programas de capacitación. El texto de las ocupaciones era de alta calidad dado que se encuentra estandarizado y redactado con el propósito de caracterizar e identificar explícitamente cada ocupación; no obstante, esto no ocurría así con los perfiles de egreso de los programas de capacitación, ya que en algunos casos la descripción del perfil era muy genérica.

Otra posible fuente de error radica en el hecho de que no hubiera programas de capacitación pertinentes para una dada ocupación. La metodología obliga a la herramienta a recomendar los programas de capacitación más cercanos independientemente de que estos sean o no correctos desde el punto de vista conceptual. En otras palabras, si entre los 203 programas analizados solo uno cumple el criterio de pertinencia, es muy probable que la herramienta recomiende dicho programa, y además liste los nueve programas siguientes en orden de cercanía, aunque estos no cumplan el criterio de pertinencia.

Con el fin de validar la herramienta destinada a relacionar programas de capacitación y ocupaciones, un equipo de seis expertos analizó la calidad del texto del campo "perfil_egresado_programa" y la pertinencia de los 203 programas para cada una de las 22 ocupaciones. A los efectos de analizar la calidad del texto de los programas de capacitación se intentaron responder las siguientes preguntas:

- ¿El texto da una idea acabada del programa de capacitación al cual hace referencia?
- ¿El contenido es suficiente para decidir si el programa es pertinente o no para una ocupación dada?

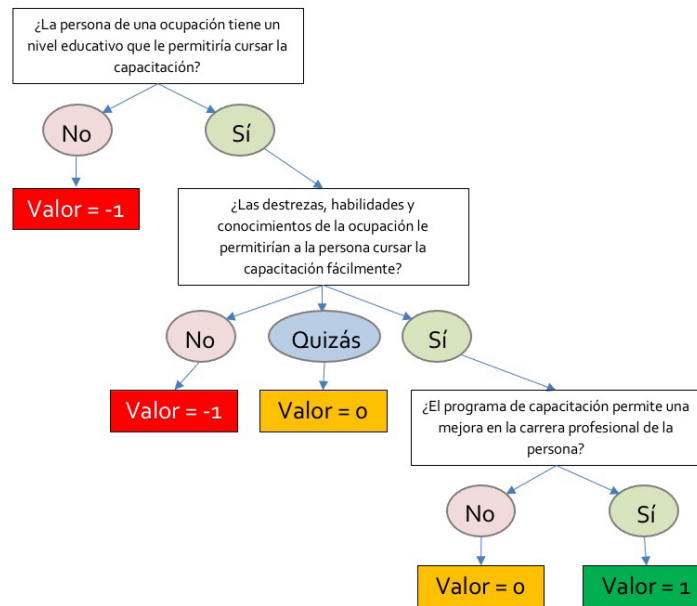
De no cumplirse ninguno de los dos requisitos, el texto era clasificado como de baja calidad. Si cumplía solo con uno de los dos criterios, era clasificado como de calidad media, y si cumplían con ambos, entonces era considerado como de alta calidad. Los programas cuyo texto se consideró de baja calidad no se tomaron en cuenta para el posterior análisis de pertinencia. Solamente el 3% de los programas (6 de ellos) fueron clasificados como de calidad baja, media-baja o media, mientras que un 97% (197 programas) contenían textos de calidad media-alta o alta.

Seguidamente se solicitó al equipo de validación que clasificara los programas en tres categorías:

- i) Los que deberían ser incluidos en el grupo de programas pertinentes para una ocupación.
- ii) Los que deberían ser excluidos del grupo de programas pertinentes para una ocupación.
- iii) Los que podrían o no estar relacionados con una ocupación.

Para llevar a cabo esa tarea se utilizó el árbol de decisión que se muestra en el diagrama 2.

Diagrama 2
Árbol de decisión para validar la herramienta NLP

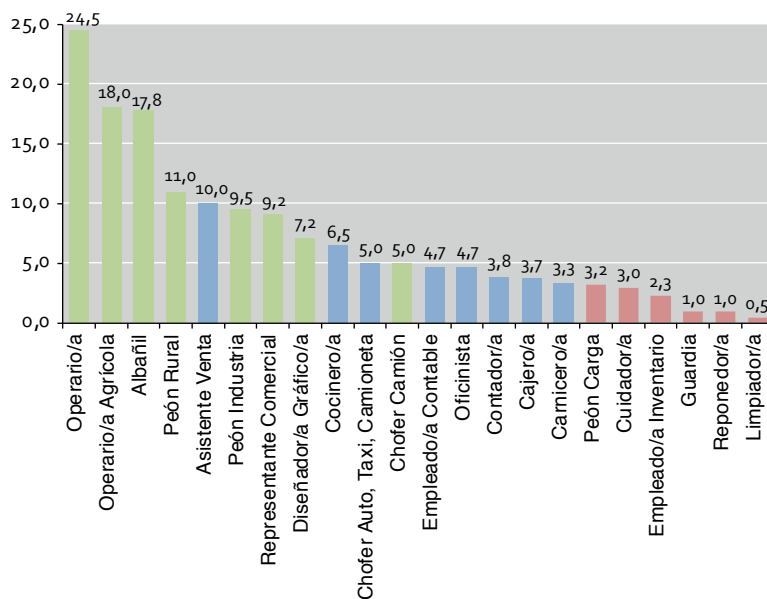


Fuente: Elaboración propia.

Los programas que obtuvieron un valor de 1 se clasificaron como pertinentes y cabría esperar que la herramienta los incluyera en el listado de programas recomendados. Los programas que obtuvieron un valor nulo podrían aparecer o no en el listado de recomendados por la herramienta, y los que obtuvieron un valor negativo no deberían aparecer en él.

En un primer análisis se determinó el número de programas pertinentes para cada ocupación según la clasificación manual llevada a cabo por el equipo de validación. En el gráfico 2 se muestra el promedio del número de programas clasificados como pertinentes (valor = 1) para cada ocupación. Se observa que, según el criterio de validación, entre los 203 programas proporcionados por la Universidad del Trabajo del Uruguay (UTU) hay un número muy bajo que sea pertinente para las ocupaciones limpiador/a, reponedor/a, guardia, empleado/a de inventario, cuidador/a y peón de carga (barras de color rosa en el gráfico 2). Coincidentemente, estas son las ocupaciones en que la herramienta presentada en este trabajo dio como resultado un porcentaje alto de errores de recomendación o un número bajo de programas en el listado de recomendados (véase el cuadro 4).

Gráfico 2
Número medio de programas pertinentes según la clasificación manual realizada por el equipo de validación
(En número de programas)



Fuente: Elaboración propia, sobre la base de información proporcionada por la Dirección de Estadísticas Laborales, Ministerio de Trabajo y Seguridad Social.

Asimismo se verifica que los mejores resultados de recomendación de la herramienta NLP coinciden con las ocupaciones en que los validadores encontraron un mayor número de programas pertinentes (barras de color verde en el gráfico 2). Estos resultados sugieren que los errores de recomendación de la herramienta NLP podrían deberse a la falta de programas pertinentes en el conjunto de 203 programas que figuran en la base de la UTU.

Seguidamente se analizaron los programas que el equipo de validación clasificó como pertinentes. Los resultados muestran que el número de programas clasificados como pertinentes fue bajo y que la dispersión en las respuestas de los validadores fue elevada (véase el gráfico A1 del anexo). De hecho, solo hubo tres programas que el conjunto de validadores clasificó de forma unánime como pertinentes para una ocupación dada: "agroforestal y mecanización agrícola" para la ocupación "operario agrícola", y "gastronomía" para la ocupación "cocinero". Por otro lado, en el caso de la ocupación "operario" hubo un total de 79 programas que se asignaron como pertinentes por al menos un validador. Esto ilustra lo heterogénea que es la clasificación cuando la llevan a cabo los seres humanos, y refuerza la idea de que existen sesgos en el momento de recomendar programas de capacitación a las personas que necesitan un servicio de desarrollo de carrera.

Otra manera de validar la herramienta NLP es analizando cuántos programas de los que el equipo de validación clasificó como no pertinentes aparecen en el listado de la recomendación mixta. En este caso sí hubo un alto grado de concordancia entre los validadores, que se manifiesta en el número elevado de programas que el equipo de validación clasificó de forma unánime como no pertinentes (véanse las barras con frecuencia = 6 en el gráfico A2 del anexo).

Los resultados muestran que, si se considera como error el hecho de que la herramienta recomendara programas que el equipo de validadores hubiera clasificado unánimemente como no pertinentes, el porcentaje promedio de programas correctamente recomendados asciende a $(85 \pm 21)\%$

(véase el cuadro 6). Los valores de la columna Porcentaje Mixta muestran una distribución notoriamente desplazada hacia los porcentajes altos. En este escenario, asignar la media como número identificador de la herramienta puede llevar a que se subestime su capacidad. Un indicador más sensato podría ser la mediana de aciertos en función de la clasificación humana, que en este caso asciende a $(90 \pm 12)\%$.

Cuadro 6
Resumen de los resultados obtenidos con la estrategia mixta TF-iDF-Doc2Vec en función de los resultados de la validación manual llevada a cabo por el equipo de expertos

CIUO-08	Ocupación	Total	Pertinentes	Porcentaje Mixta	Porcentaje TF-iDF_9
932	Peón Industria	10	9	90%	100%
2166	Diseñador/a Gráfico/a	13	12	92%	100%
2411	Contador/a	9	7	78%	55%
3322	Representante Comercial	11	11	100%	66%
4110	Oficinista	10	10	100%	88%
4311	Empleado/a Contable	13	11	85%	66%
4321	Empleado/a Inventario	5	2	40%	22%
5120	Cocinero/a	6	6	100%	67%
5223	Asistente Venta	10	8	80%	66%
5230	Cajero/a	7	6	86%	77%
5321	Cuidador/a	20	13	65%	66%
5414	Guardia	1	1	100%	66%
7112	Albañil	5	5	100%	77%
7511	Carnicero/a	4	4	100%	44%
8183	Operario/a	10	10	100%	100%
8322	Chofer Auto, Taxi, Camioneta	9	8	89%	33%
8332	Chofer Camión	10	9	90%	55%
8341	Operario/a Agrícola	9	8	89%	100%
9112	Limpiador/a	10	2	20%	33%
9212	Peón Rural	11	11	100%	100%
9333	Peón Carga	12	8	67%	11%
9334	Reponedor/a	1	1	100%	33%
			PROMEDIO TOTAL	85%	65%

Fuente: Elaboración propia, sobre la base de información proporcionada por la Dirección de Estadísticas Laborales, Ministerio de Trabajo y Seguridad Social.

Nota: Las celdas sombreadas en rosa con texto en negrita indican disminución con respecto a la obtenida con TF-iDF. Las celdas sombreadas en verde con texto en negrita indican aumento con respecto a la obtenida con TF-iDF. Total = número de programas recomendados por la estrategia mixta. Pertinentes = número de programas pertinentes con la estrategia mixta según equipo de validación. Porcentaje Mixta = porcentaje de programas pertinentes respecto de los programas totales recomendados por la estrategia mixta. Porcentaje TF-iDF_9 = Porcentaje de programas pertinentes utilizando los primeros nueve programas recomendados utilizando TF-iDF.

En síntesis, los resultados de la validación muestran que la herramienta NLP es capaz de recomendar programas de capacitación pertinentes para una ocupación dada con un alto grado de concordancia con los resultados obtenidos mediante la clasificación humana.

IV. Conclusiones

El presente trabajo tenía por objeto desarrollar una herramienta que facilitara el proceso de recomendación de programas de capacitación que llevan a cabo los proveedores de servicios de desarrollo de carrera. Se trata de un proceso complejo que requiere integrar y analizar un gran número de aspectos que conciernen a las personas que intentan construir un plan de desarrollo profesional, sobre todo en el contexto cambiante del mercado laboral. Este servicio implica que el proveedor evalúe no solo los intereses, el nivel educativo y la experiencia laboral de la persona, sino también habilidades, destrezas y conocimientos que no siempre pueden extraerse fácilmente a partir de esos datos. En este sentido, el proyecto ONET UY, mediante una caracterización estandarizada y exhaustiva de las habilidades, las destrezas y los conocimientos asociados a las ocupaciones, se posiciona como una herramienta fundamental que brinda información útil a los proveedores de servicios de desarrollo de carrera y a los trabajadores. Como se indicó en párrafos anteriores, la cantidad de información que debe manejar un proveedor de servicios de desarrollo de carrera es muy grande, y eso es una limitante en el momento de brindar información no sesgada a las personas que requieren el servicio.

Una de las recomendaciones que surgen a partir del trabajo de la OCDE (2021b) es que se debe desarrollar una herramienta que permita elaborar perfiles de habilidades que los proveedores puedan utilizar, sobre todo cuando trabajan con adultos que no tienen calificaciones formales. En este sentido, este trabajo presenta el primer desarrollo de una herramienta de aprendizaje automático (*machine learning*) que permite vincular ocupaciones con programas de capacitación en América Latina a partir de datos provenientes del perfil de las ocupaciones locales. Esto es de suma importancia, dado que recientemente se demostró que las habilidades, destrezas y conocimientos que caracterizan a una ocupación en el Uruguay difieren significativamente de las que figuran en el perfil de la misma ocupación en los Estados Unidos (Velardez, 2021).

En la herramienta desarrollada mediante NLP se toma en cuenta el texto descriptivo de las ocupaciones y los programas de capacitación, texto cuyo volumen es extraordinariamente grande como para que una única persona lo maneje de manera eficiente y sin sesgo. En los últimos años, el NLP se ha convertido en una herramienta fundamental a la hora de procesar grandes volúmenes de texto (Bird, Klein y Loper, 2009). Como ya se mencionara, en los Estados Unidos recientemente se ha utilizado el NLP para detectar tareas específicas que surgen en nuevas ocupaciones de ONET (Dahlke y Putka, 2021). En la publicación referida a ese trabajo, los autores son capaces de establecer grados de similitud entre las

distintas tareas específicas mediante el uso de la distancia coseno. En la herramienta automática que se describe en el presente trabajo se utiliza tanto la distancia coseno como la euclidiana y se considera la importancia tanto de la frecuencia de las palabras como de la similitud entre conceptos (TF-IDF y Doc2Vec). El resultado final es un listado de programas de capacitación recomendados para cada una de las 22 ocupaciones de la primera ola del relevamiento ONET UY.

En esta prueba de concepto, los resultados se contrastaron con la clasificación manual realizada por un equipo de validadores, que evaluaron tanto la calidad del texto de los 203 programas de capacitación como la pertinencia de estos para cada una de las 22 ocupaciones. Los resultados mostraron que la herramienta NLP es capaz de recomendar hasta nueve programas de capacitación en promedio para cada ocupación, y que el porcentaje medio de éxito es del 85%, mientras que la mediana es del 90%. Cuando se analizan los errores que comete la herramienta al recomendar programas pertinentes para cada ocupación, se observa que las tasas elevadas de error se verifican en las ocupaciones para las que no hay o hay pocos programas pertinentes, según el criterio utilizado por el equipo de validación (véase el cuadro 6). Es de esperar que la tasa de error disminuya a medida que el modelo sea alimentado con un mayor número de programas de capacitación.

Durante el proceso de validación, al analizar el conjunto de programas que se clasificaron manualmente como pertinentes se observó una gran dispersión de datos. En el caso extremo de la ocupación "operarios", se encontró que 79 programas fueron clasificados como pertinentes por al menos un validador. Esto da cuenta de que el resultado de la recomendación depende del validador, lo que refuerza la idea de que en los servicios de desarrollo de carrera se ofrecen recomendaciones que muy probablemente estén sesgadas por la experiencia y los intereses del proveedor. Esto también refuerza la necesidad de contar con una herramienta que asista a los proveedores cuando analizan el contenido de los programas de capacitación. En este sentido, la herramienta que se desarrolló en este trabajo muestra potencial para ser utilizada como un primer filtro al manejar grandes volúmenes de información, filtro que permitiría minimizar el sesgo y tomar en cuenta las habilidades, las destrezas y los conocimientos asociados a las ocupaciones. De esta manera, la herramienta NLP podría brindarle al proveedor del servicio de desarrollo de carrera un listado corto de programas de capacitación a fin de que tenga un número manejable de opciones para luego continuar con un análisis en profundidad.

En resumen, la herramienta presentada en este trabajo ofrece las siguientes cinco ventajas:

- i) Procesa grandes volúmenes de datos.
- i) Recomienda minimizando sesgos.
- ii) Considera habilidades, destrezas, conocimientos y tareas específicas de cada ocupación.
- iii) Disminuye el tiempo de análisis por parte del proveedor de servicios de desarrollo de carrera.
- iv) Es modular, escalable y puede actualizarse de manera continua.

Se observó que la herramienta NLP es muy sensible al contenido de los textos. En el futuro cabe esperar que se pueda mejorar su eficacia mediante la incorporación de datos de nuevas ocupaciones e información más detallada de los programas de capacitación, por ejemplo, el contenido de los temas que se tratan en cada clase. La presente prueba de concepto contribuye al objetivo del MTSS, que pretende diseñar una herramienta que en el futuro permita a una persona introducir, por un lado, sus intereses, nivel educativo, formación en el trabajo y experiencia laboral, y, por el otro, la ocupación a la cual aspira como objetivo profesional, a fin de que obtenga como resultado posibles trayectorias formativas pertinentes que luego pueda analizar con el consejero de desarrollo de carrera.

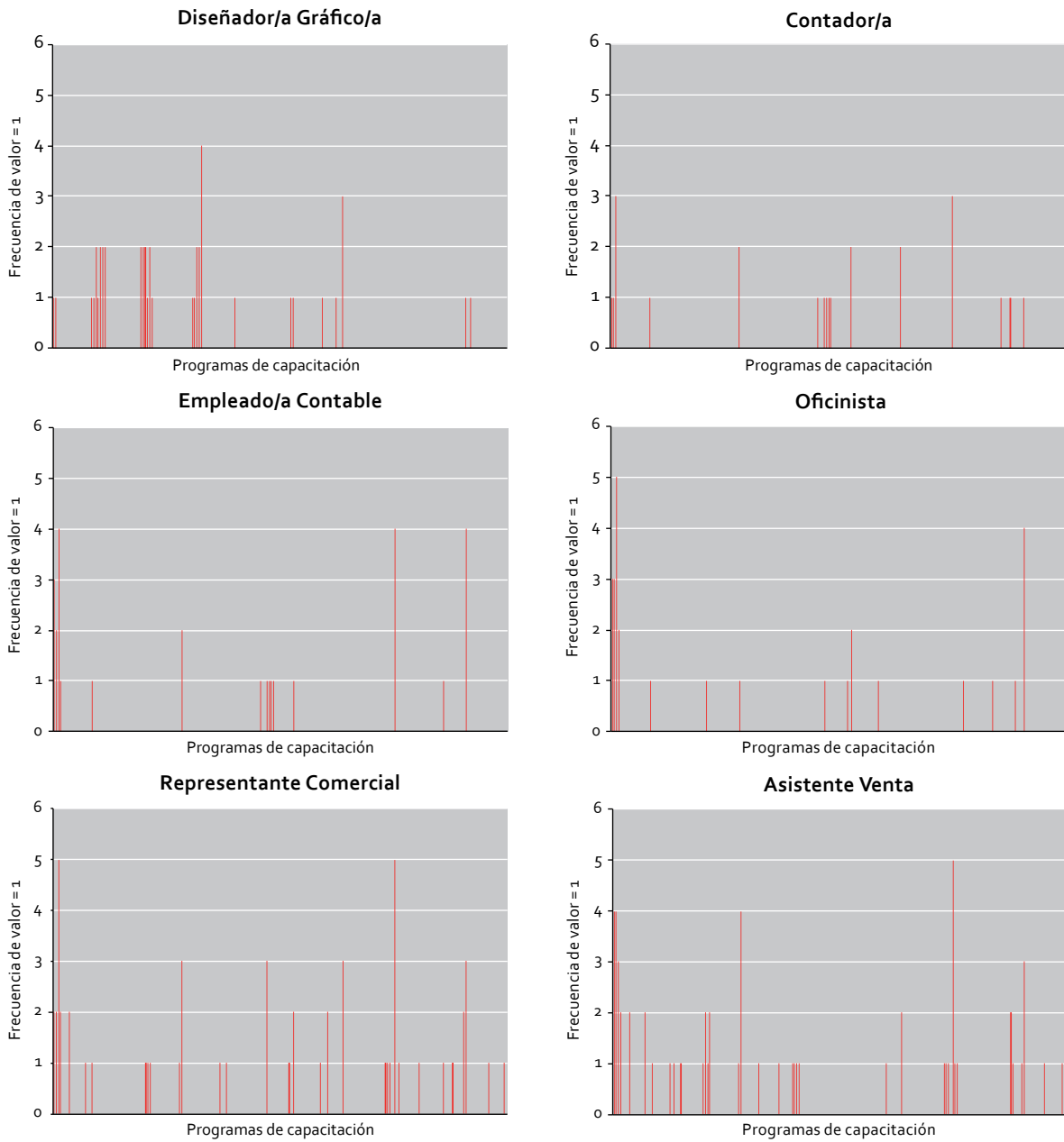
Por último, se espera que los resultados del presente trabajo ayuden a las personas a tomar decisiones informadas al elaborar un plan de carrera profesional, y que contribuyan a mejorar los servicios de desarrollo de carrera que se brindan en los sectores público y privado del Uruguay y de toda América Latina.

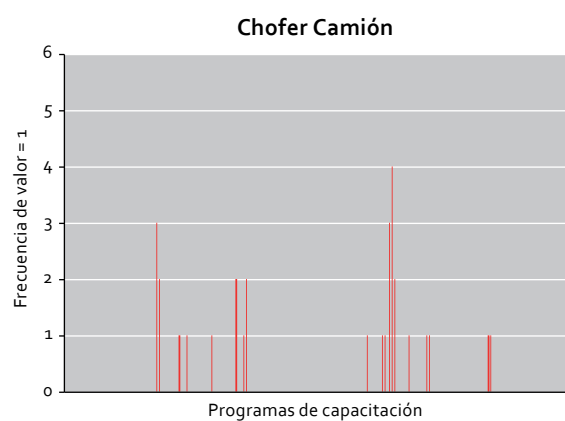
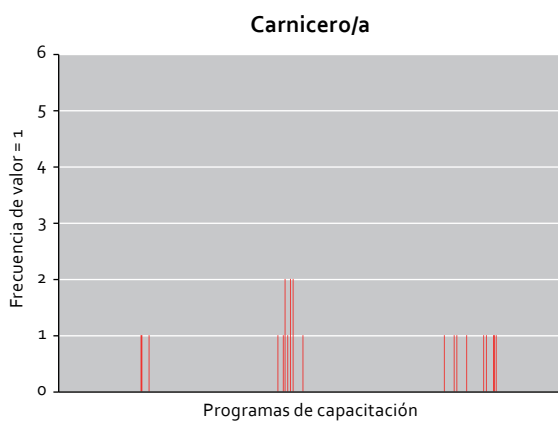
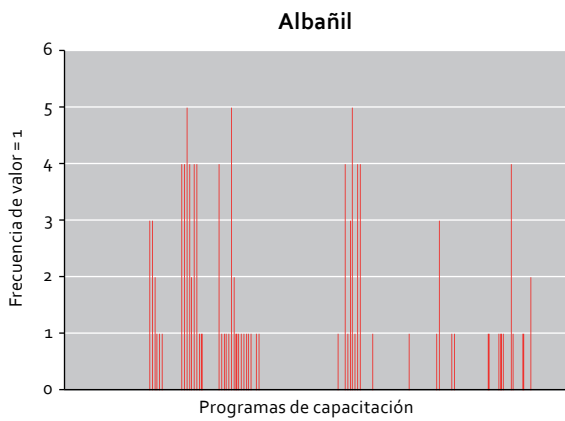
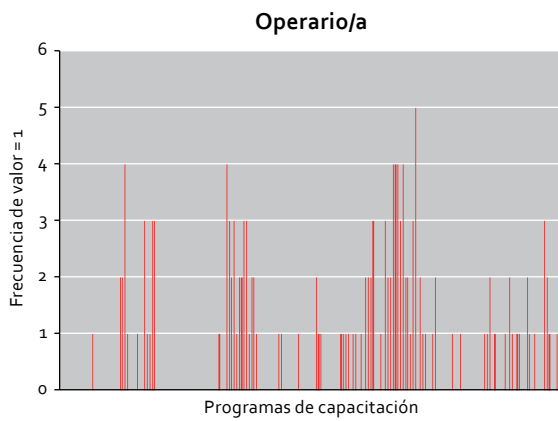
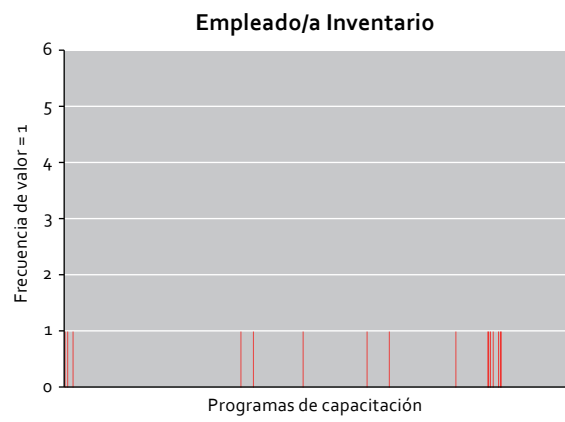
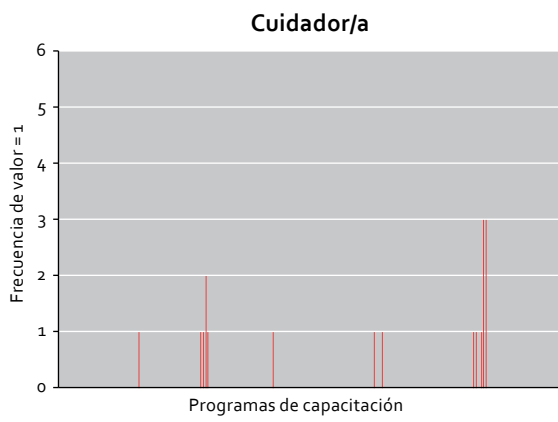
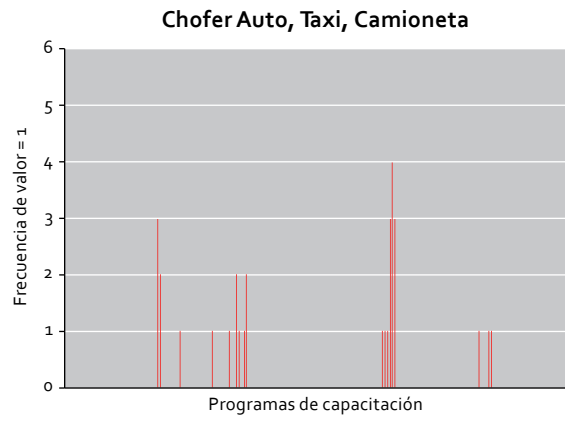
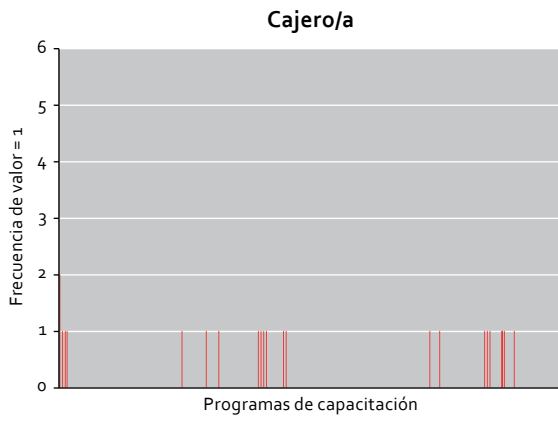
Bibliografía

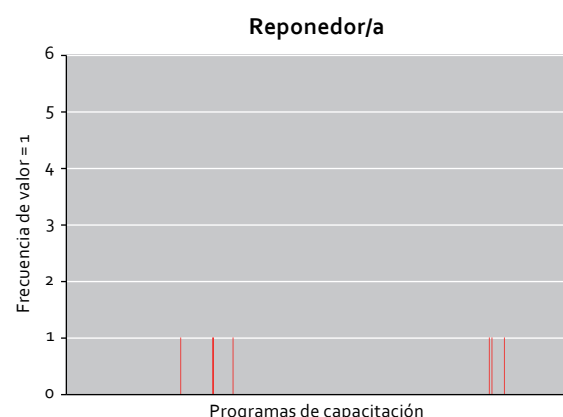
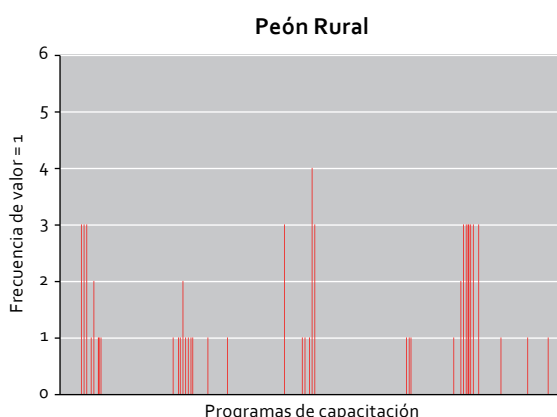
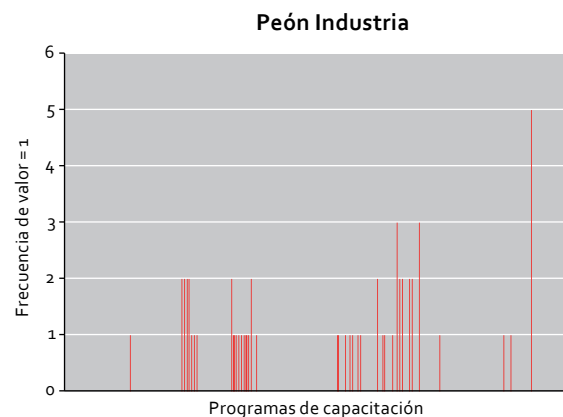
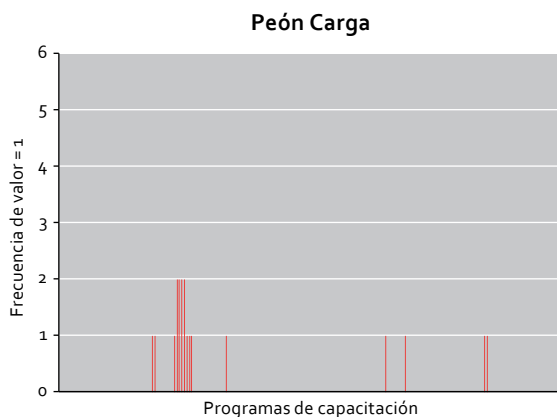
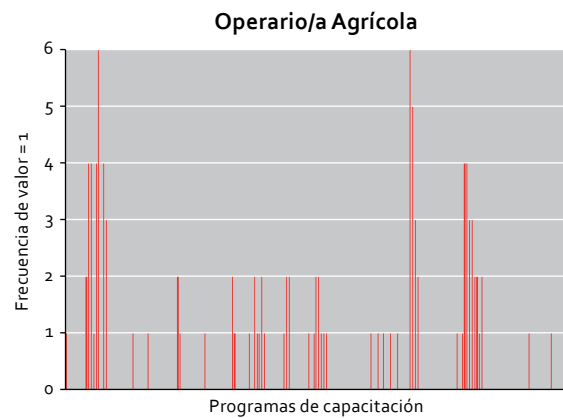
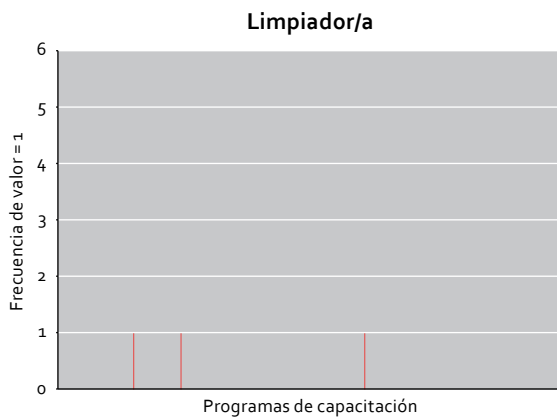
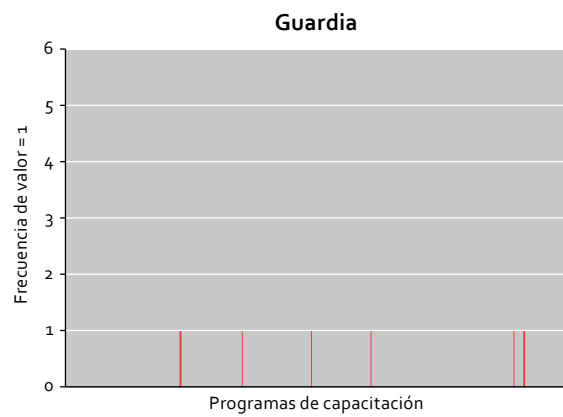
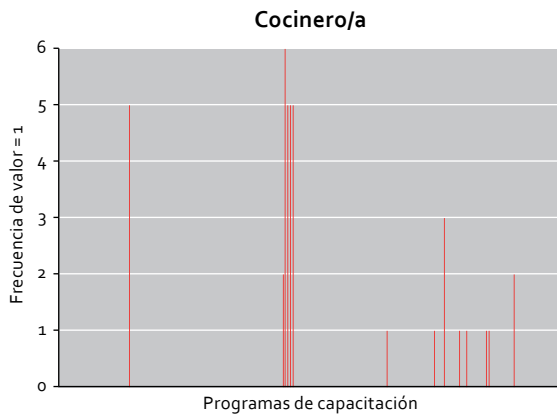
- Autor, D. (2010), *The Polarization of Job Opportunities in the U.S. Labor Market: Implications for Employment and Earnings*, Washington, D.C., Center for American Progress/The Brookings Institution.
- Banco Mundial (2016), *World Development Report 2016: Digital Dividends*, Washington, D.C.
- Bird, S., E. Klein y E. Loper (2009), *Natural Language Processing with Python*, O'Reilly Media, Inc.
- Chowdhury, G. G. (2010), *Introduction to Modern Information Retrieval*, Facet Publishing.
- Dahlke, J. A. y D. J. Putka (2021), *Streamlining the Identification of Emerging Tasks in the O*NET System Using Natural Language Processing (NLP): Technical Summary*, Human Resources Research Organization (HumRRO).
- Le, Q. y T. Mikolov (2014), "Distributed representations of sentences and documents", *Journal of Machine Learning Research*, vol. 32, N° 2, International Machine Learning Society (IMLS).
- McGowan, M. A. y D. Andrews (2015), "Skill mismatch and public policy in OECD countries", *Economics Department Working Paper*, N° 1210, Organización de Cooperación y Desarrollo Económicos (OCDE).
- McKinsey Global Institute (2017), *A Future that Works: Automation, Employment, and Productivity*.
- OCDE (Organización de Cooperación y Desarrollo Económicos) (2021a), *Career Guidance for Adults in a Changing World of Work*, París, OECD Publishing.
- _____ (2021b), *Career Guidance for Adults in Latin America*, París, OECD Publishing.
- _____ (2020), *Effective Adult Learning Policies: Challenges and Solutions for Latin American Countries*, París, OECD Publishing.
- OIT (Organización Internacional del Trabajo) (2019), *Trabajar para un futuro más prometedor - Comisión Mundial Sobre el Futuro del Trabajo*, Ginebra.
- Velardez, M. O. (2021), "Análisis de distancias ocupacionales y familias de ocupaciones en el Uruguay", *Documentos de Proyectos (LC/TS.2021/36)*, Santiago, Comisión Económica para América Latina y el Caribe (CEPAL).

Anexo 1

Gráfico A1
Frecuencia de los programas de capacitación considerados pertinentes según la clasificación manual realizada por el equipo de validadores

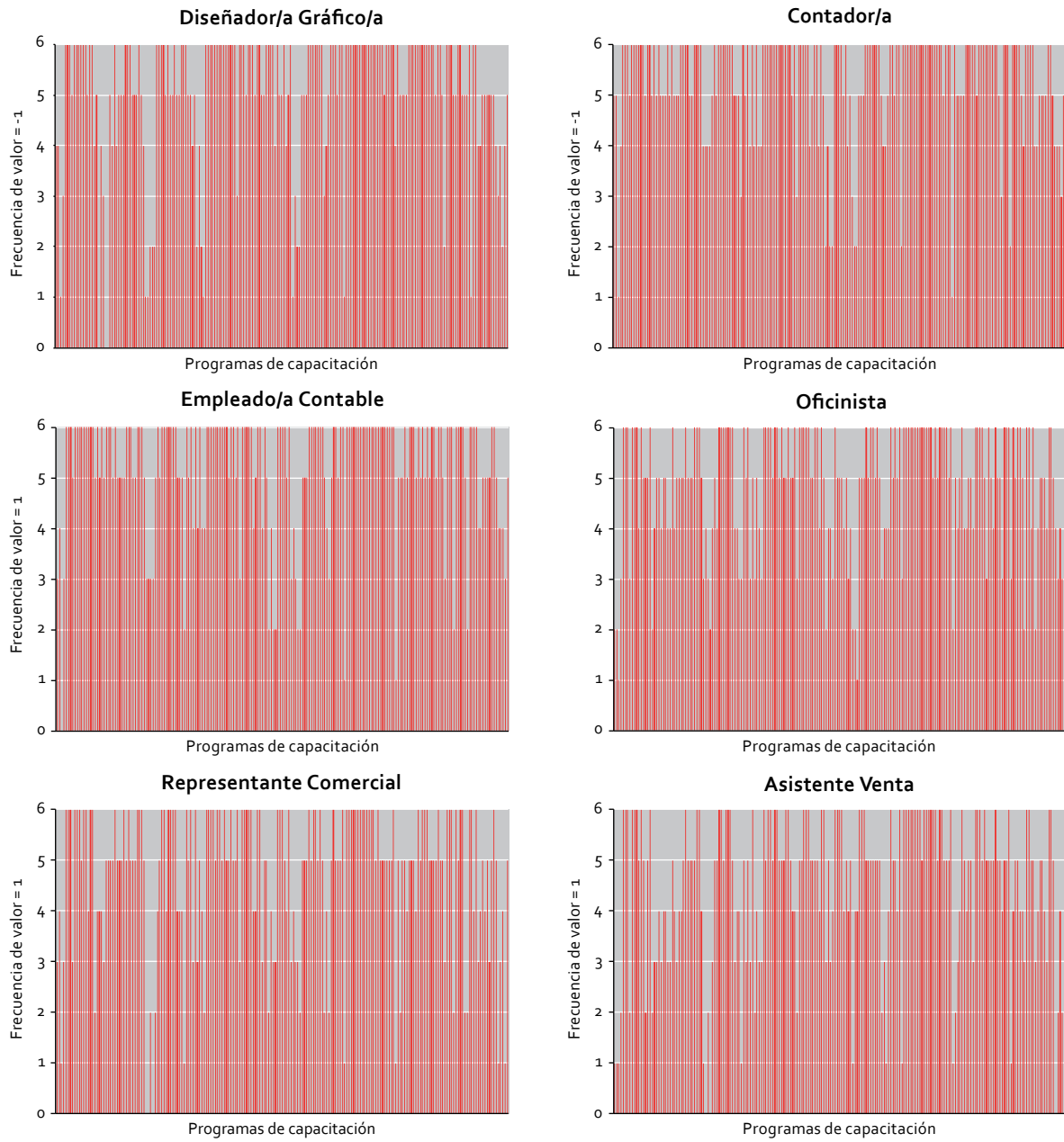


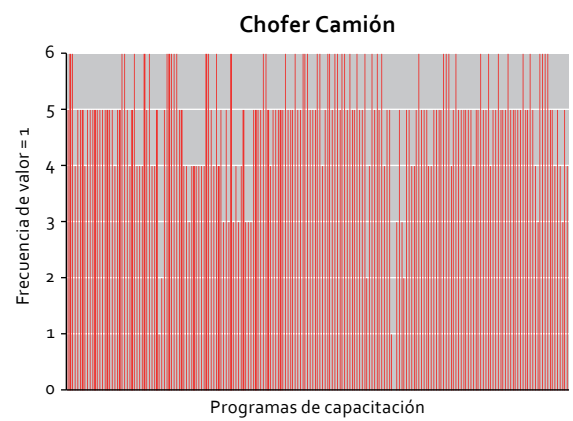
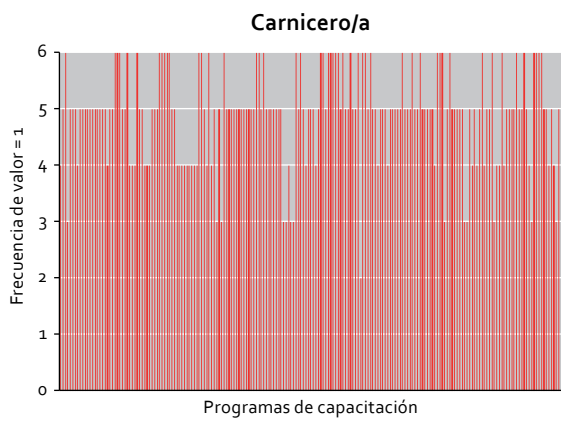
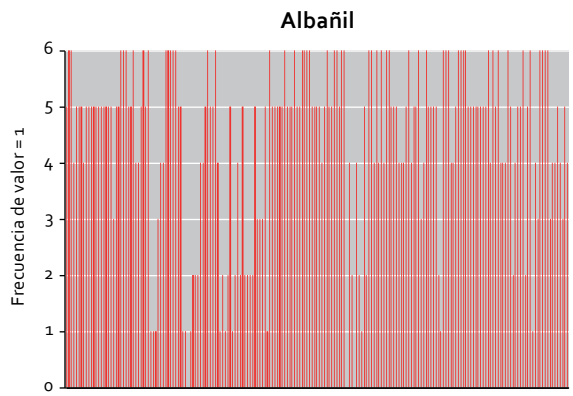
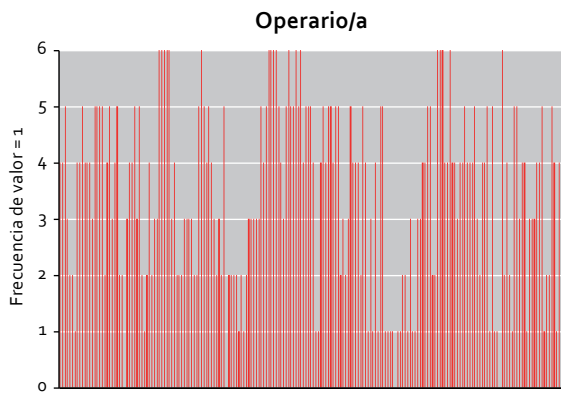
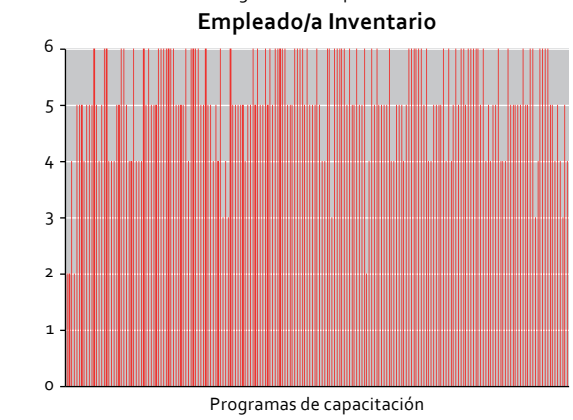
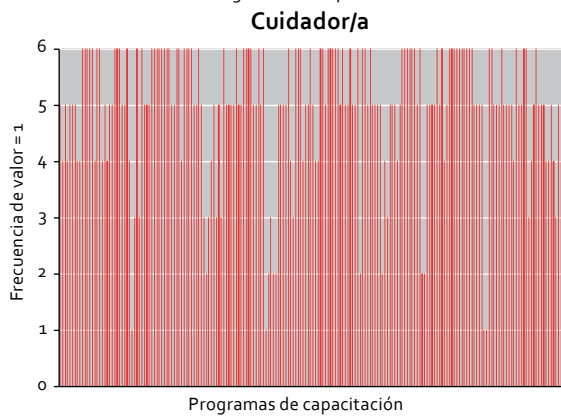
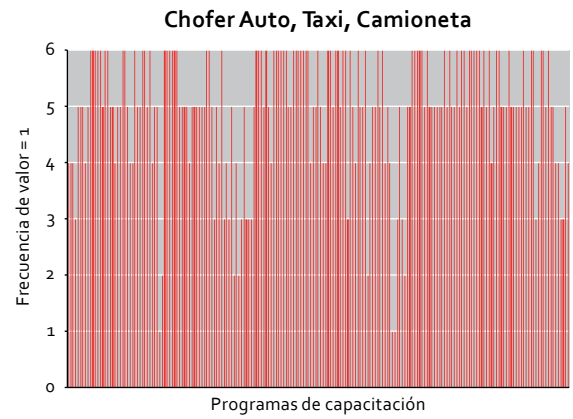
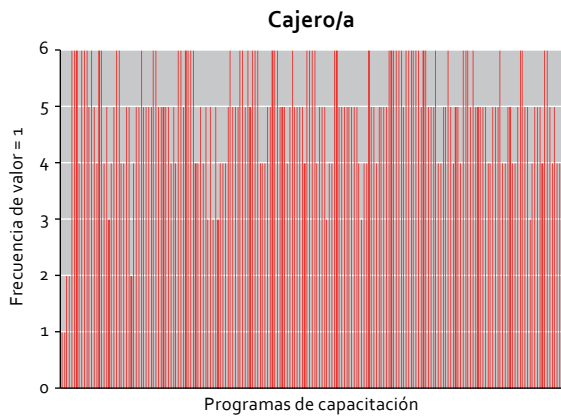


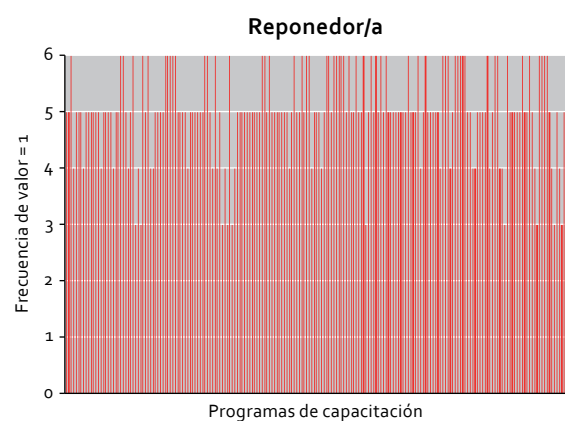
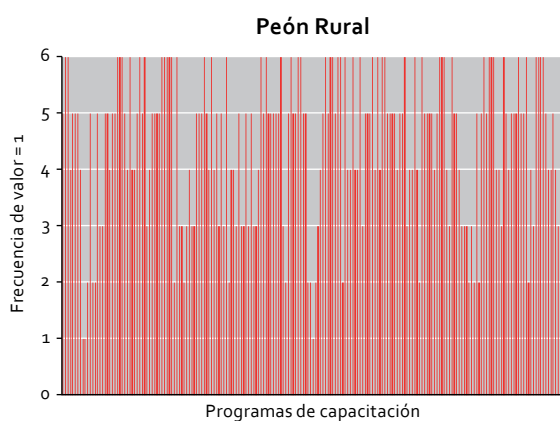
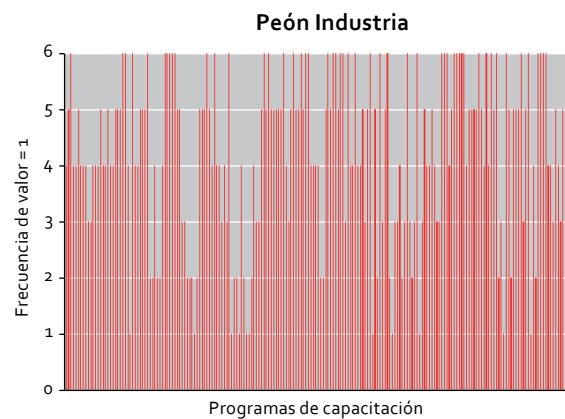
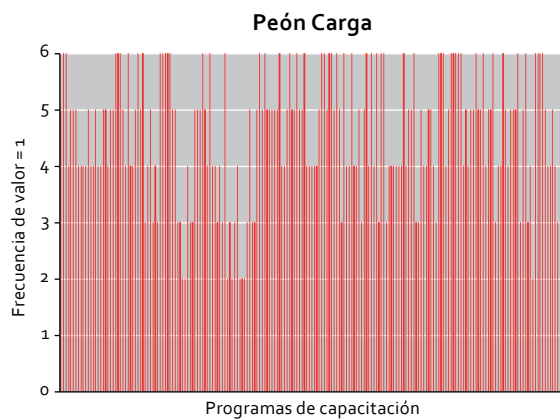
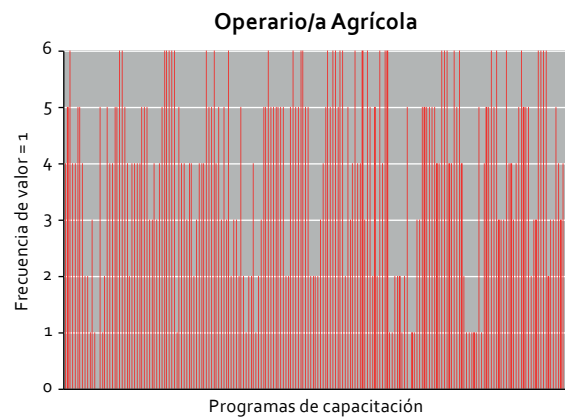
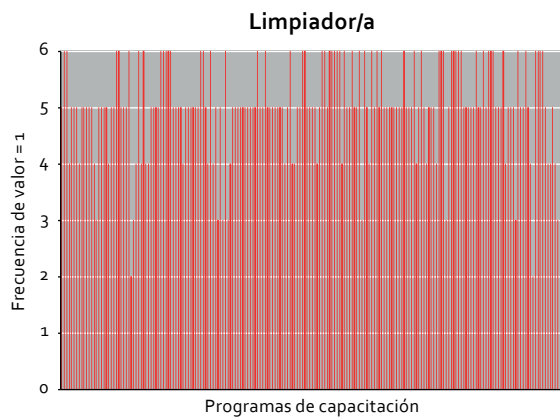
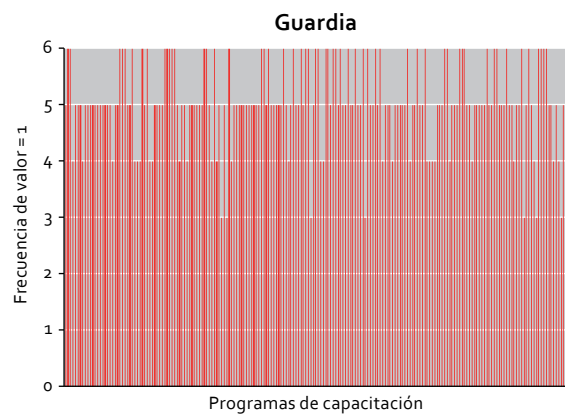
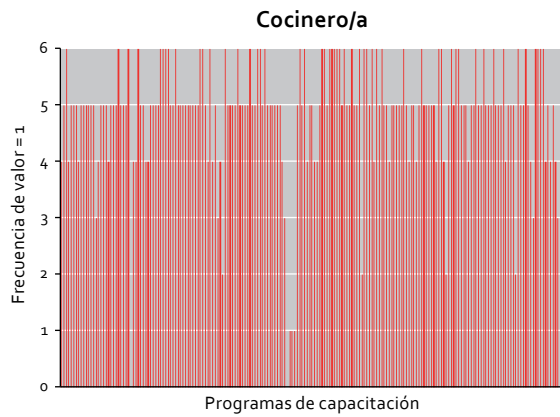


Fuente: Elaboración propia.


Gráfico A2
Frecuencia de los programas de capacitación considerados no pertinentes según la clasificación manual realizada por el equipo de validadores







Fuente: Elaboración propia.



En este trabajo se desarrolló una herramienta automática y no supervisada que tiene por objeto recomendar programas de capacitación para una serie de ocupaciones sobre la base de similitudes entre el perfil de egreso de un conjunto de programas de la Universidad del Trabajo del Uruguay (UTU) y la descripción de las tareas correspondientes a 22 ocupaciones obtenidas a partir del relevamiento de ONET de Uruguay. En la herramienta se utilizan instrumentos del procesamiento de lenguaje natural (Natural Language Processing o NLP) cuya atención se centra en la repetibilidad de conceptos claves y en las similitudes del texto como un todo. Con el fin de evaluar este método, se contrastaron las recomendaciones obtenidas a partir de la herramienta con las que brindó un grupo de personas expertas. Los resultados muestran que la herramienta desarrollada permite recomendar un promedio de hasta nueve programas de capacitación para cada ocupación con un porcentaje de éxito medio del 85%. El potencial de esta metodología radica en que permite manejar de forma eficiente grandes volúmenes de datos que pueden contribuir a brindar información no sesgada en los servicios de desarrollo de carrera.

