



Universidad Politécnica
de Madrid

**Escuela Técnica Superior de
Ingenieros Informáticos**



Grado en Ingeniería Informática

Trabajo Fin de Grado

**Similitud semántica para la
armonización de habilidades**

Autora: Ruth Verónica Ocampo Prado

Tutora: María Navas Loro

Cotutora: Patricia Martín Chozas

Madrid, octubre de 2023

Este Trabajo Fin de Grado se ha depositado en la ETSI Informáticos de la Universidad Politécnica de Madrid para su defensa.

Trabajo Fin de Grado

Grado en Ingeniería Informática

Título: Similitud semántica para la armonización de habilidades

Octubre 2023

Autora: Ruth Verónica Ocampo Prado

Tutora:

María Navas Loro

Dpto. Inteligencia Artificial

ETSI Informáticos

Universidad Politécnica de Madrid

Cotutora:

Patricia Martín Chozas

Dpto. Lingüística Aplicada a la Ciencia y a la Tecnología

ETSI Informáticos

Universidad Politécnica de Madrid

Resumen

El propósito central de este Trabajo de Fin de Grado es la identificación de habilidades equivalentes en la clasificación ESCO, extrayendo datos de Coursera mediante técnicas de Procesamiento del Lenguaje Natural (PLN).

Este estudio, con el fin de hallar la similitud entre vectores, se enfocará en la evaluación de cuatro modelos de lenguaje basados en la arquitectura Transformer explorando cómo realizan la representación vectorial. Se busca determinar cuál de estos modelos arroja los resultados más destacados en la tarea de hallar la similitud semántica, utilizando el coseno como unidad de medición.

La metodología adoptada se alinea con un enfoque no supervisado, partiendo de un conjunto de datos no etiquetados que recoge las habilidades en inglés provenientes de cursos ofrecidos en plataformas MOOC, específicamente en Coursera. A través de este estudio, se busca no solo identificar la habilidad equivalente en la clasificación ESCO, sino también establecer cuál de los modelos Transformer seleccionados destaca en la representación vectorial de habilidades.

Palabras Clave: similitud semántica, modelos de lenguaje, Transformers, MOOC, ESCO.

Abstract

The main goal of this Final Degree Project is the identification of equivalent skills in the ESCO classification, extracting skills from Coursera using Natural Language Processing (NLP) techniques.

In order to find the similarity between vectors, this research will focus on the evaluation of four language models based on the Transformer architecture, analysing how it performs vector representation. The aim is to determine which of these models yields the most outstanding results in the task of finding semantic similarity, using the cosine as a measurement unit.

The methodology adopted pursues an unsupervised approach, starting from a non-label dataset that contains English skills from courses offered on MOOC platforms, specifically Coursera. Through this study, we seek not only to identify the equivalent skill in the ESCO classification, but also to establish which of the selected Transformer models stands out in the vector representation of skills.

Keywords: semantic similarity, language models, Transformers, MOOC, ESCO.

Contenido

1	Introducción	1
2	Estado de la cuestión	3
2.1	Marco Tecnológico	3
2.1.1	Redes Neuronales Artificiales	3
2.1.2	Arquitectura Transformer	4
2.1.3	Hugging Face	5
2.1.4	Similitud Semántica	5
2.2	Trabajos relacionados	6
2.2.1	Recuento de habilidades	6
2.2.2	Modelado de tópicos para extraer habilidades	7
2.2.3	Representación vectorial de habilidades	8
2.2.4	Técnicas de aprendizaje automático (ML)	8
3	Desarrollo	10
3.1	Fuente de datos	10
3.1.1	Plataforma MOOC: Coursera	10
3.1.2	Base de datos ESCO	12
3.2	Algoritmo de identificación de habilidades	13
3.2.1	Modelos utilizados	14
3.2.2	Métrica similitud	15
3.2.3	Evaluación de modelos	15
4	Resultados y conclusiones	19
4.1	Resultados	19
4.2	Conclusiones	22
5	Análisis de Impacto	23
6	Bibliografía	24

Índice de tablas

Tabla 1. Ejemplo información curso de Coursera	11
Tabla 2. Ejemplo de información de habilidad ESCO.....	12
Tabla 3. Resultados de T5 en la comparación de habilidades	16
Tabla 4. Resultado de Roberta en la comparación de habilidades.....	16
Tabla 5. Resultados All-MiniLM-L6v2 en la comparación de habilidades	17
Tabla 6. Resultados de Albert en la comparación de habilidades	18

Índice de figuras

Figura 1. Arquitectura original de un Transformer, originalmente en [2]	4
Figura 2. Ejemplo de curso en la página oficial de Coursera	11
Figura 3. Ejemplo de habilidad en página oficial ESCO	13
Figura 4. Algoritmo final identificación de habilidades	14
Figura 5. Ejemplo búsqueda habilidad: “VR”	20
Figura 6. Ejemplo de búsqueda habilidad: “JAVA”	20
Figura 7. Ejemplo de búsqueda de habilidad “DATABASE MANEGEMENT” ..	21
Figura 8. Ejemplo de búsqueda habilidad “A general understanding of traditional Chinese medicine and Indian Ayurveda”	21

1 Introducción

En la actualidad, la identificación precisa de habilidades laborales similares es fundamental para optimizar procesos de desarrollo profesional y de selección. Gracias a los notables avances en Inteligencia Artificial y a la Lingüística Aplicada, se percibe un progreso significativo en el campo del Procesamiento del Lenguaje Natural. Esta área tiene como objetivo lograr que las máquinas tengan la misma capacidad de comprensión que una persona real. Este campo ha obtenido muy buenos resultados puesto que lo podemos emplear en nuestro día a día a través de diversas aplicaciones y tecnologías como, por ejemplo, en los traductores de texto, en asistentes de voz (Siri, Alexa, etc.), o en el análisis de sentimientos de las redes sociales, entre otros.

La similitud semántica, cuya problemática abordaremos en el apartado 2.1.4, juega un papel fundamental en muchas aplicaciones que utilizan el Procesamiento del Lenguaje Natural, PLN en adelante. El PLN permite que las máquinas comparen textos no sólo teniendo en cuenta la aparición de palabras idénticas, si no también teniendo en cuenta la relación semántica que poseen.

En el contexto de este Trabajo de Fin de Grado, abordaremos el problema de la similitud semántica, llevada al ámbito profesional, es decir, realizaremos comparaciones de habilidades profesionales con habilidades requeridas para poder realizar un curso en plataformas Massive Open Online Courses [1], que a partir de ahora será nombrada plataforma MOOC. Para ser más exactos, nos centraremos sólo en la fuente de datos de los cursos impartidos en inglés de la plataforma de Coursera¹, la cual trataremos más a fondo en el apartado 3.1.1.

Para poder realizar este estudio utilizaremos la fuente de datos perteneciente a la Clasificación multilingüe de habilidades, competencias y ocupaciones europeas, denominada ESCO², la cual también contiene las habilidades de datos en inglés y que veremos más en detalle en el apartado 3.1.2.

Para esta investigación usaremos modelos grandes basados en la arquitectura Transformer [2], detallada en el apartado 2.1.2, la cual ha tomado una gran importancia en los últimos años gracias a su base formada por mecanismos de atención, los cuales le permiten capturar las relaciones semánticas con una mayor eficacia en comparación a otras arquitecturas. Esto ha llevado a que los modelos basados en esta arquitectura alcancen resultados de última generación en diversas tareas del PLN y del aprendizaje profundo.

El objetivo principal de este trabajo es evaluar qué modelo basado en la arquitectura Transformer es adecuado para poder realizar la comparación entre

¹ <https://about.coursera.org/>

² https://esco.ec.europa.eu/en/use-esco/download/privacy-statement?packages=v110_classification_en_csv/

habilidades con mayor exactitud y fiabilidad. Para llevarlo a cabo se realizará un estudio, el cual se encuentra detallado en el apartado 3.2 y se divide en 3 fases:

- Estudio de los modelos a utilizar, desarrollado en el apartado 3.2.1.
- Explicación de la métrica utilizada para poder realizar las comparaciones, la cuál es tratada en el apartado 3.2.2.
- Evaluación de los modelos respecto a los resultados obtenidos que será explicada en el apartado 3.2.3.

En cuanto a los resultados del estudio, tratado en el apartado 4, se realizará una demo basada en el modelo que presente un mejor desempeño a la hora de realizar las comparaciones entre las habilidades.

Por último, abordaremos en el apartado 5 las bases y planteamientos para futuros Trabajos de Fin de grado o estudios que estén relacionados con la similitud semántica o el análisis de habilidades profesionales.

2 Estado de la cuestión

Esta sección se estructura en dos apartados esenciales: el análisis del marco tecnológico, presentado en el apartado 2.1 y revisión exhaustiva de trabajos e investigaciones previas que exploran el tema en cuestión, detallado en el apartado 2.2.

2.1 Marco Tecnológico

En el contexto tecnológico de nuestro estudio, exploramos los elementos que tienen un papel importante en el ámbito del estudio. Este análisis recoge la siguiente información: en la sección 2.1.1, se explicará qué son y cómo funcionan las Redes Neuronales Artificiales; en el apartado 2.1.2, se detalla qué es la arquitectura Transformer y por qué marca un antes y un después en el mundo del PLN y el aprendizaje profundo; en el punto 2.1.3; se aborda la información sobre la biblioteca de Transformers, y finalmente, en la sección 2.1.4, se analiza la problemática de la similitud semántica.

2.1.1 Redes Neuronales Artificiales

En la actualidad, las redes neuronales artificiales [3], en adelante RNA, se han convertido en una herramienta muy útil para abordar las diferentes cuestiones del PLN como, por ejemplo, hallar la similitud semántica entre palabras (esta problemática se desarrollará en el apartado 2.1.3).

Las RNA imitan el sistema nervioso del ser humano, además están compuestas por nodos que representan a las neuronas de nuestro propio sistema nervioso. Estos nodos se pueden agrupar en conjuntos, dando lugar a las denominadas capas neuronales. En una red neuronal existen tres tipos de capas según la función que realicen:

- Capa de entrada. Esta capa es la capa inicial y además es única. Tiene como función recibir la información del exterior, ya que cada uno de sus nodos representa una característica única recogida del conjunto de datos recibido.
- Capa oculta. De este tipo de capa pueden existir varias. Se sitúan entre la Capa de entrada y la Capa de salida. Su función es permitir el paso de información y, en caso de existir varias capas, facilita el aprendizaje de patrones completos observados en los datos de entrada.
- Capa de salida. Es la capa final de la red. Su función es transmitir la función que se ha procesado en nuestra red al sistema que lo emplea.

Como hemos mencionado anteriormente, las RNA están compuestas por nodos que poseen un peso y un umbral especificado y funcionan de la siguiente manera: cada nodo que lo conforma está conectado a otro a modo que, si la salida del primer nodo en la capa de entrada supera el valor del umbral que se ha especificado, el nodo se activará enviando datos a la siguiente capa de red y así sucesivamente entre las diferentes capas que contenga nuestra red.

A través de entrenamientos y alimentándose de conjuntos de datos, las redes neuronales han conseguido destacar ya que han logrado resolver tareas cognitivas consiguiendo dar respuestas muy parecidas a las que daría el cerebro humano. Entre estas tareas se encuentran la traducción automática [4], la clasificación de texto [5], el resumen de texto [6], entre otras.

Antes del 2017, el tipo más destacado de las RNA eran las redes neuronales recurrentes [7], las cuales utilizaban bucles de retroalimentación con el fin de que la información dure varias etapas de entrenamiento. En el año 2017 se produjo un cambio significativo en la historia del PLN y del aprendizaje profundo,

se desarrolló una arquitectura basada en mecanismos de atención la cual, en poco tiempo, logró reemplazar a las redes neuronales recurrentes como se explica en el apartado 2.1.

2.1.2 Arquitectura Transformer

Los modelos de lenguaje basados en redes neuronales recurrentes [8] como Elmo [9] han sido utilizados para tratar con problemas como la ambigüedad de las palabras (polisemia), comprobación de gramática, o el etiquetado de roles semánticos, entre otros.

En la actualidad, los modelos de lenguaje que usan mecanismos de atención han superado a los modelos basados en redes neuronales recurrentes, gracias a que centran una especial atención al contexto de las palabras.

La arquitectura Transformer fue desarrollada en 2017 por Google³. Es una arquitectura basada en las redes neuronales multicapas tradicionales que combina los mecanismos de atención con codificadores posicionales, siendo una arquitectura más simple que la de las redes neuronales recurrentes.

Esta arquitectura trata el resultado de los cálculos subyacentes de manera paralela, teniendo como resultado un uso más eficiente de los avances de Hardware como las GPUs, lo que computacionalmente, se traduce en una gran mejora de la exactitud y de la precisión al construir modelos más grandes mediante entrenamientos con datos más robustos. De este modo, esta arquitectura hace frente a la gran debilidad de las redes neuronales recurrentes, la dependencia sobre los resultados de estados anteriores.

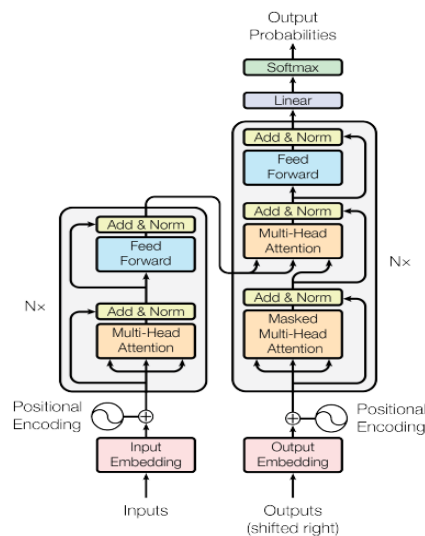


Figura 1. Arquitectura original de un Transformer, originalmente en [2]

El Transformer ha destacado por su flexibilidad y eficiencia, siendo el primer modelo de transducción basado únicamente en el mecanismo de atención “Self-attention”. Este mecanismo le permite capturar las relaciones semánticas entre las palabras de una misma oración. Esta arquitectura a pesar de haber sido

³ https://about.google/?fg=1&utm_source=google-ES&utm_medium=referral&utm_campaign=hp-header

entrenada especialmente para la traducción automática (del inglés al alemán y del inglés al francés), en la actualidad, es posible utilizarla para una gran variedad de tareas de PLN como, por ejemplo, en la clasificación de texto, aplicada en [10], el análisis de sentimientos, desarrollado en [11], o la generación de textos, abordada en [12].

En nuestro estudio, como trataremos en el apartado 3.2.1, con el fin de hallar la similitud semántica se ha realizado una evaluación en base a 4 modelos basados en esta arquitectura: 3 modelos basados en BERT [13] y uno basado en T5 [14].

2.1.3 Hugging Face

Hugging Face⁴ es una plataforma de código abierto que facilita el desarrollo de herramientas y recursos para trabajar con modelos de PLN.

Esta plataforma destaca principalmente por permitir el acceso fácil a los modelos basados en Transformer a través de una biblioteca que recibe el nombre de “SentenceTransformer”⁵. Acceder de esta manera a los modelos se puede traducir como un ahorro de los recursos computacionales puesto que solo se debe descargar dicha biblioteca y no el modelo por completo. Además, nos proporciona un ahorro en cuestión de tiempo puesto que los modelos a los que se accede en dicha biblioteca ya están entrenados.

2.1.4 Similitud Semántica

La similitud semántica ha sido un tema ampliamente investigado en el área del Procesamiento del Lenguaje Natural, siendo una parte esencial para poder desempeñar distintas tareas de PLN, como recuperación de información, la clasificación de texto, la traducción automática, entre otras, medir la similitud semántica entre documentos, oraciones o palabras.

En un principio, los expertos e investigadores del PLN asociaban la similitud entre dos oraciones si ambas contenían las mismas palabras. Aunque esta idea en un principio servía para comparar cierto tipo de oraciones, presentaba una tasa significativa de fallos importante, ya que no se tenía en cuenta ni las propiedades semánticas ni las sintácticas del texto. Por ejemplo, no se tenía en cuenta los casos en que las oraciones a pesar de contener las mismas palabras tenían un significado diferente debido al orden en que iban en la oración. Tampoco abordaba los casos en que las oraciones compartían una misma palabra, pero, según el contexto poseían diferentes significados.

Para resolver esta problemática, durante la última década los expertos han llevado a cabo grandes investigaciones y desarrollado diferentes técnicas respecto a la similitud semántica.

Según el estudio realizado en [15], el cual trata la evolución que ha tenido la similitud semántica durante los últimos años, podemos clasificar las técnicas según la fuente de información:

- Métodos basados en conocimientos. Estos métodos calculan la similitud semántica entre dos términos basándose en la información obtenida en fuentes de conocimientos (bases de datos léxicas, diccionarios, etc.). Una

⁴ <https://huggingface.co/>

⁵ <https://www.sbert.net>

muestra de este tipo de métodos es el Método basado en características [16] o el Método basados en Información contenida [17]. Estas técnicas como se reflejan en sus respectivos nombres se fundamentan en datos provenientes de diferentes fuentes de información.

- Métodos basados en corpus. Estos métodos miden la similitud utilizando información recuperada de grandes corpus. Hacían uso de diferentes técnicas para obtener la representación vectorial de las palabras (Word Embeddings) del texto en cuestión y para estimar la similitud entre estas representaciones, utilizaban medidas de distancia semántica basadas en la Hipótesis Distributiva la cuál persigue la idea de que las palabras que tienen significados similares tienden a aparecer en contextos parecidos. De estos tenemos como ejemplo los métodos GloVe [18] y LSA [19].
- Métodos basados en redes neuronales profundas. Estos métodos aprovechan los desarrollos en redes neuronales para mejorar su rendimiento y estiman la similitud entre los vectores de palabras aprovechando el uso de la Vectorización de palabras. Ejemplos de métodos pertenecientes a este conjunto son los modelos basados en LSTM [20] o los modelos basados en la arquitectura Transformers.
- Métodos híbridos. Estos métodos aprovechan tanto la eficiencia estructural que ofrecen los métodos basados en el conocimiento, como la versatilidad de los métodos basados en corpus entre otras ventajas de los métodos anteriormente mencionados con el fin de obtener la similitud entre textos. Claros ejemplos pertenecientes a este grupo son la representación vectorial NASARI [21] o el enfoque MSSA [22].

2.2 Trabajos relacionados

En este apartado estudiaremos los diferentes métodos empleados para la identificación de las habilidades en trabajos realizados similares.

El estudio sobre identificación de habilidades realizado en [23] nos permite realizar una clasificación de los métodos utilizados para la identificación de habilidades en anuncios de trabajos en cuatro grandes grupos: métodos de recuento de habilidades, métodos de modelados de temas, métodos de incorporación de habilidades y métodos que utilizan las técnicas de aprendizaje automático. En este caso, al ser bastante similar los requisitos establecidos sobre habilidades en anuncios de empleos con los que se piden para poder realizar una formación en una plataforma MOOC, usaremos esta clasificación para orientarnos en el estudio.

2.2.1 Recuento de habilidades

El Recuento de habilidades es el método más usado para la identificación de habilidades. El objetivo de esta técnica es obtener el análisis del contenido de cada anuncio a estudiar y llevar a cabo un estudio de la lista de habilidades más frecuentes detectadas. Esta técnica puede realizarse de dos maneras: mediante una base de habilidades o mediante un grupo de profesionales que se encarguen de hacer el trabajo de manera manual.

Esta técnica la podemos ver empleada en [24], en este estudio investigaban las habilidades fundamentales del siglo XXI para tener éxito en un trabajo después de graduarse en la universidad. En este estudio los investigadores recopilaban las habilidades más frecuentes en los documentos encontrados en plataformas de

investigación para poder crear una lista de sinónimos de habilidades generadas a partir de los Tesauros Psycinfo⁶ y Merriam-Webster⁷. Partiendo de esta lista de sinónimos como base de habilidades, realizaron una búsqueda de habilidades similares en anuncios de empleos, dando como resultado la lista definitiva de las habilidades más frecuentes en los anuncios de empleo.

Otro ejemplo en el que usan esta metodología es en [25], en el cual los investigadores buscan identificar las habilidades sociales requeridas de ingeniería en un puesto de origen marroquí y analizar el contenido de diferentes anuncios de trabajo con el objetivo de ver si las habilidades sociales requeridas en diferentes campos reflejan las necesidades del mercado laboral marroquí.

Sin embargo, este tipo de estudios tienen un alto costo en cuanto al tiempo ya que primero se debe investigar y realizar una lista con las habilidades más repetitivas en los textos. Además, la fiabilidad en este tipo de estudios recae en la calidad y cantidad del tipo de documentación de la que extraen la información. En otras palabras, cuántos más documentos que contengan información relevante para el estudio sean recopilados, mejores resultados se obtendrán a la hora de realizar el estudio.

2.2.2 Modelado de tópicos para extraer habilidades

El modelado de tópicos es un método no supervisado. Esto quiere decir que no necesita de etiquetas ni de información previa sobre categorías o temas para entrenarse. Estos algoritmos tienen el objetivo de descubrir patrones o relaciones ocultas en los datos estudiados. Por lo general, se centran en un área o tema en específico ya que requieren de la interpretación final de un experto para analizar los resultados.

Podemos ver este método aplicado en el estudio [26], en el cual tratan de identificar las habilidades más solicitadas en las diferentes áreas de la industria del Big Data. Para llevar a cabo el análisis primero realizan una extracción de las habilidades encontradas en anuncios de empleo. A continuación, aplican la técnica Latent Dirichlet Allocation (LDA), gracias a la cual pueden agrupar semánticamente las habilidades o conocimientos recogidos en temas o áreas específicas. Por último, realizan una interpretación en función de los datos obtenidos en los anteriores pasos.

Otro trabajo realizado en el que podemos observar este tipo de metodología es en [27]. En este ejemplo podemos ver cómo se busca hallar las habilidades relacionadas con los temas específicos del área de Big Data y de Business Intelligence a nivel individual con el fin de satisfacer la alta demanda y adaptar los diferentes grados universitarios de las mismas

Aunque aplicar esta técnica ha tenido buenos resultados, dependen mucho de la calidad de los textos que se pasen como entrada para ser analizados y, para proyectos en los que se actualicen constantemente los datos, tendría un costo computacional grande a nivel de tener que reentrenar el modelo o añadirle técnicas complementarias para poder realizar la actualización.

⁶<https://library.bath.ac.uk/psycinfo/apa-thesaurus>

⁷ <https://www.merriam-webster.com/>

2.2.3 Representación vectorial de habilidades

Este tipo de método se propuso con el fin de resolver las limitaciones del recuento de habilidades y de extender la posibilidad de etiquetar habilidades en diferentes áreas. Estos modelos usan la técnica de Vectorización de habilidades (Skill Embedding). Esta tiene como objetivo crear una representación vectorial por cada habilidad a analizar, de modo que las habilidades similares y las más concurrentes tienen un espacio vectorial más corto.

Este tipo de método se puede ver claramente reflejado en [28], en el que emplean el modelo Word2vec [29] para crear las representaciones de vectores de las habilidades requeridas en el estudio, y a continuación, se mide el grado de similitud entre pares de habilidades.

En el experimento realizado en el estudio [30] podemos observar cómo también utilizan la representación vectorial de las habilidades y de los trabajos para lograr un mayor rendimiento a la hora de recomendar trabajos.

Un punto a favor de este tipo de métodos es el poder obtener de cada palabra una representación vectorial única lo cual permite reducir los falsos positivos teniendo en cuenta que las habilidades similares estarán cercanas en el espacio vectorial, facilitando la identificación de relaciones y similitudes semánticas. Por otro lado, también aumentó el número de falsos negativos y al ser representadas las palabras como vectores no permite tener en cuenta el contexto de las frases.

2.2.4 Técnicas de aprendizaje automático (ML)

En el contexto de este estudio, de este tipo de métodos se han analizado las siguientes técnicas relacionadas con el procesamiento de texto:

- Reconocimiento de entidades nombradas (NER) es una tarea del PLN que hace referencia al proceso de extraer información como nombres, lugares, fechas, etc. Esto se consigue mediante el etiquetado de datos para el entrenamiento de los modelos, es decir, se le entrena con un conjunto de ejemplos en los que se le debe indicar a qué entidad pertenece cada palabra. Por lo general, el conjunto debe ser grande para obtener mejores resultados.

Un ejemplo de uso de esta técnica la podemos ver en el estudio [31], en el cuál estudian qué método tiene mejores resultados, aplicando el modelo NER combinando con un diccionario que contiene nombres de habilidades que comúnmente poseen una misma persona y diccionarios obtenidos de los datos provenientes de las plataformas de ESCO, LinkedIn⁸ y Kaggle⁹. Esto lo realizan con el fin de extraer habilidades tanto técnicas como personales de ofertas de trabajo de Google y Amazon.

En el estudio [32] se empleó esta técnica con la finalidad de analizar los requisitos que deben cumplir los aplicantes al puesto de analista de datos. Los datos fueron extraídos de más de 5.00 anuncios y etiquetaron más de 60.000 entidades para realizar dicho experimento.

⁸ <https://about.linkedin.com/es-es?lr=1>

⁹ <https://www.kaggle.com/>

- Clasificación de texto: En esta técnica, como dice su nombre, se asigna etiquetas predefinidas de categorías a fragmentos de texto que contienen habilidades específicas. Esta tarea es fundamental para el PLN y en específico, la tarea de clasificar las habilidades es usada en diversos campos que trabajan con el talento de las personas. Este método se puede emplear tanto de manera supervisada, como mediante un enfoque no supervisado, es decir, sin partir de datos etiquetados.

Con el fin de entender mejor esta metodología se puede ver su aplicación en [33]. En él podemos observar el desarrollo de un modelo para clasificar las habilidades de múltiples etiquetas. En él, primero extraen las habilidades, clasificándolas a continuación, en grupos de competencias según la similitud de las habilidades, para finalmente, comparar las ofertas de trabajo con el grupo de competencia creado.

También, en el trabajo [34] en el que realizan la clasificación de habilidades con el objetivo de lograr predecir las habilidades que no están presentes en ofertas de trabajo subidas por los reclutadores.

3 Desarrollo

En la sección de desarrollo de este trabajo, trataremos dos aspectos fundamentales para la comprensión y ejecución de este estudio. La sección 3.1 detalla las diferentes fuentes de datos que se han utilizado, ofreciendo una visión más detallada de la información que sustenta el estudio. A continuación, la sección 3.2 se realiza una explicación del Algoritmo de Identificación de habilidades que ha sido empleado en este trabajo, ofreciendo una exposición detallada de la estructura y su funcionamiento.

3.1 Fuente de datos

En esta sección se investiga la procedencia de las fuentes de datos recogidas para la realización del estudio. En el apartado 3.1.1 se explica qué son las plataformas MOOC, cuál es una de las plataformas más importantes y por qué se ha recogido datos de esta.

3.1.1 Plataforma MOOC: Coursera

Las plataformas Massive Open Online Course, se caracterizan por ofrecer de forma masiva una gran variedad de cursos formativos online. La plataforma Coursera, es una de las plataformas más populares. Fue fundada por Daphne Koller y Andrew Ng. en 2012 con el objetivo de extender conocimientos sobre distintas materias a personas en diferentes partes del mundo.

Los cursos que ofrece esta plataforma poseen una descripción en la que se informa de las habilidades que se adquirirán al realizarlos. Estos cursos cuentan también con foros de discusiones para poder resolver dudas o realizar comentarios respecto al tema en cuestión. Los conocimientos aprendidos se evalúan mediante una serie de pruebas o proyectos prácticos y finalmente, para poder superar el curso, se realiza un examen final en el que, en caso de superarlo, se obtiene una insignia o certificado que acredite la realización de este.

Para poder estudiar la información proveniente de la plataforma Coursera, se ha utilizado un dataset que contenía 5.266 cursos en inglés. Este conjunto de datos nos proporciona la siguiente información por cada curso: la referencia a la página oficial del curso, el tema del curso, las habilidades que proporciona dicho curso, el tipo del curso y una descripción de lo que se estudiará en el mismo.

Para entrar más en el contexto de esta fuente se proporciona el siguiente ejemplo de la tabla 1 para visualizar la información que se obtiene del mismo.

URL	https://www.coursera.org/learn/uva-darden-digital-product-management
Nombre del curso	Digital Product Management: Modern Fundamentals Coursera
Tema	Business, Leadership and Management
Habilidades	Product/Market Fit, Product Management, Design Thinking, Innovation Pipeline, Lean Startup.
Tipos	course
Descripción	Not so long ago, the job of product manager was about assessing market data, creating requirements, and managing the hand-off to sales/marketing. Maybe you talk to a customer somewhere in there and they tell you what features they wanted. But companies that manage product that way are dying. Being a product person today is a new game, and product managers are at the center of it. Today,

particularly if your product is mostly digital, you might update it several times a day. Massive troves of data are available for making decisions and, at the same time, deep insights into customer motivation and experience are more important than ever. The job of the modern product manager is to charter a direction and create a successful working environment for all the actors involved in product success. Its not a simple job or an easy job, but it is a meaningful job where you'll be learning all the time.

This course will help you along your learning journey and prepare you with the skills and perspective you need to Create the actionable focus to successfully manage your product (week 1). Focus your work using modern product management methods (week 2). Manage new products and explore new product ideas (week 3). Manage and amplify existing products (week 4).

This course is ideal for current product or general managers interested in today's modern product management methods.

Please note that there are new additions to this course and subtitles for these videos will soon be available.

This course was developed with the generous support of the Batten Institute at UVA Darden School of Business. The Batten Institute mission is to improve the world through entrepreneurship and innovation: www.batteninstitute.org.

Tabla 1. Ejemplo información curso de Coursera

The screenshot shows the Coursera course page for "Data Visualization with OpenAI API: Generate code with GenAI". The page is part of the "coursera project network". The course is taught in English ("Enseñado en Inglés") by instructor Ahmad Varasteh. A blue button labeled "Inicio Proyecto Guiado" is visible. Below this, it says "Incluido con coursera PLUS" and provides a link to "Obtener más información". A navigation bar includes "Acerca de", "Resultados", "Detalles del proyecto", and "Testimonios". The "Qué aprenderás" section lists three learning objectives: creating effective prompts, producing Python code for data preparation, and designing data visualizations. The "Habilidades que practicarás" section lists skills: Data Analysis, Python Programming, prompt engineering, ChatGPT, and Data Visualization.

coursera project network

Data Visualization with OpenAI API: Generate code with GenAI

Enseñado en Inglés

Instructor: Ahmad Varasteh

[Inicio Proyecto Guiado](#)

Incluido con **coursera PLUS** • [Obtener más información](#)

[Acerca de](#) [Resultados](#) [Detalles del proyecto](#) [Testimonios](#)

Qué aprenderás

- ✓ Create effective prompts for communicating with the OpenAI API to generate Python code for data analysis and visualization.
- ✓ Produce Python code for data preparation to ensure data is ready for meaningful visualizations using AI-generated code.
- ✓ Design and implement data visualizations using AI-generated code to effectively communicate insights and patterns in the data.

Habilidades que practicarás

Data Analysis Python Programming prompt engineering ChatGPT Data Visualization

Figura 2. Ejemplo de curso en la página oficial de Coursera

3.1.2 Base de datos ESCO

European Skills, Competences, Qualifications and Occupations¹⁰ (ESCO), en español, Habilidades, Competencias, Cualificaciones y Ocupaciones europeas es la clasificación oficial europea de habilidades, competencias y ocupaciones. Tiene el objetivo de ofrecer un 'lenguaje común' y apoyar a la movilidad laboral en toda Europa para obtener un mercado laboral más integrado y eficiente. Además, está traducida en los 24 idiomas oficiales de la Unión Europea más el islandés, el noruego, el ucraniano y el árabe.

La base de datos que nos proporciona ESCO cuenta con 35.121 habilidades en las que por cada una podemos encontrar información tal y como es la descripción de la habilidad, etiquetas alternativas a la actual, el tipo de habilidad y el nivel de reutilización de habilidades, entre otro tipo de información.

El objetivo de usar la base de datos ESCO es poder obtener los nombres estándar referentes a cada habilidad. Para poder realizar el estudio se ha utilizado la siguiente información de la base de datos en cuestión: los nombres de las habilidades de la base de datos ESCO, las etiquetas alternativas referentes a estas habilidades y los enlaces que redirigen a la habilidad estudiada en la página oficial en la cual se muestra más información sobre la misma.

ConceptType	KnowledgeSkillCompetence
ConceptURI	http://data.europa.eu/esco/skill/0005c151-5b5a-4a66-8aac-60e734beb1ab
SkillType	skill/competence
reuseLevel	Sector-specific
preferredLabel	manage musical staff
AltLabel	manage staff of music coordinate duties of musical staff manage music staff direct musical staff manage musical staff
HiddenLabel	
Status	Released
modifiedDate	20/12/2016 18:43
ScopeNote	
Definition	
InScheme	http://data.europa.eu/esco/concept-scheme/skills , http://data.europa.eu/esco/concept-scheme/member-skills
Description	Assign and manage staff tasks in areas such as scoring, arranging, copying music and vocal coaching.

Tabla 2. Ejemplo de información de habilidad ESCO

¹⁰ <https://esco.ec.europa.eu/en/about-esco/what-esco>

Desarrollo de cascada

[conocimiento](#) > [negocios, administracion y derecho](#) > [negocios y Administración](#) > [gestión y administración](#) > [Metodologías de gestión de proyectos TIC](#) > [Desarrollo de cascada](#)

Descripción

Descripción

El modelo de desarrollo en cascada es una metodología para diseñar sistemas y aplicaciones de software.

Etiquetas alternativas

Desarrollo de cascada

Tipo de habilidad

conocimiento

Nivel de reutilización de habilidades

habilidades y competencias específicas del sector

Figura 3. Ejemplo de habilidad en página oficial ESCO

3.2 Algoritmo de identificación de habilidades

La meta de este algoritmo es realizar una comparación entre cada habilidad extraída de la plataforma Coursera con cada una de las habilidades procedentes del dataset de ESCO con el fin de obtener la habilidad ESCO con mayor similitud a la habilidad estudiada proveniente de la plataforma MOOC. Una de las principales dificultades y de la que partimos como base para poder llevar a cabo el estudio de similitud es que el conjunto son datos no etiquetados.

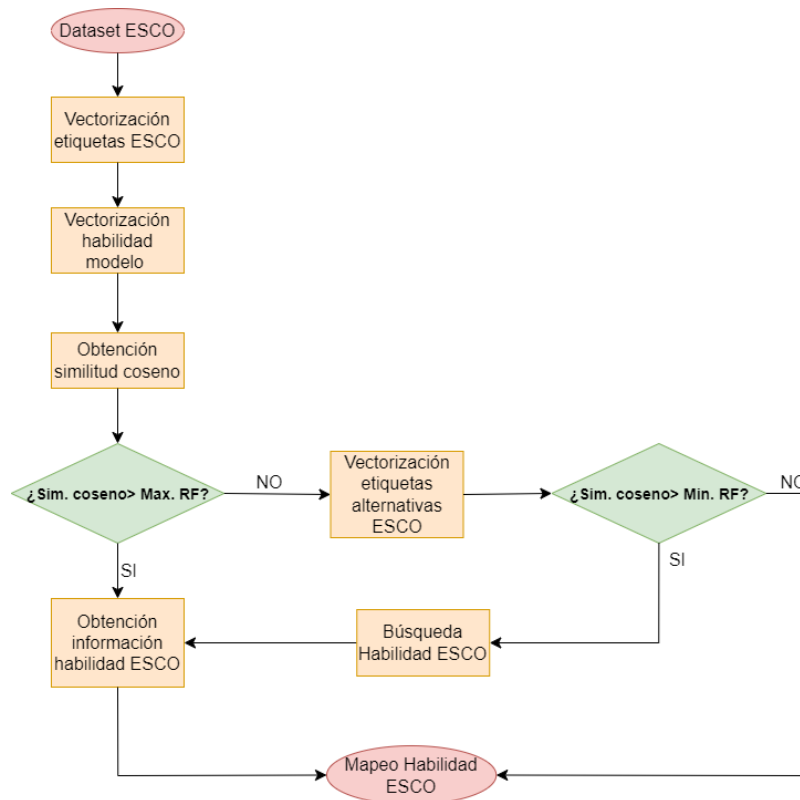


Figura 4. Algoritmo final identificación de habilidades

Con el objetivo de mostrar el procedimiento utilizado por nuestro algoritmo se proporciona la Figura 4.

El algoritmo parte de la base de datos ESCO, la cual contiene datos sin procesar, es decir, sin ninguna modificación. Además, se tiene como entrada la habilidad para la cual se requiere buscar su equivalente en ESCO.

En este proceso, comenzamos vectorizando tanto las habilidades presentes en la base de datos como la habilidad que se desea encontrar.

Posteriormente, se calcula la similitud máxima mediante el uso del coseno. El resultado obtenido se utiliza para llevar a cabo la evaluación. Si la similitud máxima supera la cota máxima establecida en el rango de fiabilidad, se procede a buscar la información correspondiente en el conjunto de datos ESCO relacionada con la habilidad resultante.

En caso contrario, es decir, si la similitud máxima es inferior a la cota mencionada, se realiza la vectorización de las etiquetas alternativas asociadas a las habilidades. A continuación, se compara esta información con la habilidad proporcionada como entrada. En este escenario, evaluamos si el resultado del cálculo del coseno supera el valor mínimo establecido en el rango de fiabilidad. Si esta condición se cumple, se obtiene la etiqueta general a partir de la etiqueta alternativa, con el objetivo de recuperar la información restante y así devolver la habilidad equivalente en ESCO. En caso de no satisfacerse esta condición, no se obtiene ningún resultado.

3.2.1 Modelos utilizados

En esta investigación utilizamos los codificadores de cuatro modelos sobresalientes: T5 [14], Roberta [35], All-MiniLM-L6-v2 [36] y Albert [37]. Estos modelos desempeñan un papel fundamental al proporcionar representaciones vectoriales de las habilidades extraídas de los cursos en Coursera, así como de aquellas pertenecientes a la base de habilidades de ESCO. Nuestro análisis se

centra en evaluar y comparar los resultados obtenidos por cada uno de estos modelos, todos basados en la arquitectura Transformers. El objetivo es determinar cuál de estos modelos muestra un rendimiento superior en relación con la tarea específica que abordamos en el estudio.

3.2.2 Métrica similitud

Como mencionamos previamente, una vez representadas las habilidades a estudiar y teniendo en cuenta que el coseno es una medida ampliamente utilizada y efectiva para evaluar la similitud entre vectores, ésta será utilizada para evaluar la similitud entre los pares de habilidades del estudio.

El resultado de esta métrica varía entre 0 y 1, donde un valor de 1 indica que los vectores tienen la misma dirección, sugiriendo una igualdad entre las habilidades. En el contexto de nuestra comparación de habilidades, un valor cercano a 1 señala un alto grado de similitud entre dos habilidades.

Por lo contrario, un resultado de 0 indica que los vectores son ortogonales. En términos de comparación de habilidades, esto nos sugiere que las dos habilidades no comparten similitud alguna. Por lo tanto, todos los pares de habilidades que resulten en valores cercanos a 0 tendrán un grado muy bajo de similitud.

En resumen, empleamos la medida de similitud del coseno para evaluar la relación entre los vectores que representan las habilidades pertenecientes al dataset ESCO y las pertenecientes a la plataforma Coursera. Los valores próximos a 1 indican una alta coincidencia y, por tanto, una similitud significativa, mientras que valores cercanos a 0 muestran una baja similitud entre las habilidades comparadas.

3.2.3 Evaluación de modelos

Respecto a la evaluación del algoritmo de identificación de habilidades, se ha tomado una muestra de 514 habilidades y sus respectivos resultados en cada modelo. En cuanto a la explicación de los resultados obtenidos, se entenderá como “alta coincidencia” cuando dos habilidades se hayan evaluado como similares por tener significados iguales. De modo opuesto, “baja coincidencia” será interpretado cuando dos habilidades no sean similares por tener significados diferentes.

Por cada uno de ellos, se ha analizado el número de altas coincidencias que han obtenido en los resultados, por lo que, si la habilidad de ESCO obtenida y la habilidad de Coursera eran realmente similares, el resultado era una alta coincidencia (“SI”). Por consecuencia, de manera contraria, si se observaba que el resultado asignado de ESCO no era una habilidad similar a la estudiada se tenía en cuenta como baja coincidencia (“NO”).

Esta evaluación nos permitió conocer el rango de fiabilidad que tenía cada modelo a la hora de detectar la similitud. Llamaremos rango de fiabilidad a el intervalo de números que permiten obtener de manera fiable la habilidad similar a la buscada. El valor máximo de este intervalo se marca en el valor donde los resultados empiezan a ser “altas coincidencias” sin tener ninguna perteneciente al grupo contrario. Para marcar el valor mínimo del intervalo se realizó un estudio más a fondo basado en la observación de habilidades acumuladas por debajo del valor máximo.

Una aclaración respecto a este tema es que hablamos de un intervalo y no de un número en específico, para que en caso de que el resultado pertenezca al rango

de fiabilidad, se pueda realizar una comparación con las etiquetas alternativas de las habilidades para un mejor resultado.

3.2.3.1 Evaluación modelo T5

En el caso del modelo T5, la muestra estudiada obtuvo 205 altas coincidencias, siendo 4 de ellas habilidades escritas de la misma forma tanto en el dataset de Coursera, como en el de ESCO. El rango de fiabilidad para este modelo oscila del [0,93-0,96], siendo a partir de 0,96, el umbral donde se puede encontrar habilidades similares altamente coincidentes.

T5	No	Si	Total general
0,7-0,8	1		1
0,75-0,80	1		1
0,8-0,9	185	61	246
0,80-0,85	36	8	44
0,85-0,9	149	53	202
0,9-1	123	140	263
0,90-0,95	119	74	193
0,95-1	4	66	70
1		4	4
1		4	4
Total general	309	205	514

Tabla 3. Resultados de T5 en la comparación de habilidades

3.2.3.2 Evaluación modelo Roberta

El modelo Roberta demostró la capacidad de identificar 189 habilidades altamente coincidentes, incluyendo 4 casos en los que las habilidades eran idénticas en su escritura. Además, el modelo estableció un rango de confiabilidad declarado de [0,90-0,95], destacando que un valor de 0,95 representa el umbral a partir del cual se pueden obtener habilidades notablemente similares.

Roberta	No	Si	Total general
0,7-0,8	1		1
0,75-0,8	1		1
0,8-0,9	231	15	246
0,8-0,85	44		44
0,85-0,9	187	15	202
0,9-1	93	170	263
0,9-0,95	86	107	193
0,95-1	7	63	70
1		4	4
1		4	4
Total general	325	189	514

Tabla 4. Resultado de Roberta en la comparación de habilidades

3.2.3.3 Evaluación modelo All-MiniLM-L6v2

En cuanto al modelo All-MiniLM-L6-v2, los resultados obtenidos estuvieron muy dispersos ya que en la muestra se encontraron tanto grados de similitud que pertenecían al intervalo [0,3-0,4], como grados que pertenecían al intervalo [0,9-1]. Respecto al resultado, este modelo obtuvo 221 altas coincidencias de las cuales 31 habilidades eran habilidades similares escritas del mismo modo en ambos datasets. Además, el rango de fiabilidad que se obtuvo en este modelo oscila entre [0,75-0,83]. Siendo el 0,83 el valor fiable para poder obtener coincidencias fiables.

All-Mini	No	Si	Total general
0,3-0,4	14		14
0,30-0,35	5		5
0,35-0,40	9		9
0,4-0,5	47		47
0,40-0,45	16		16
0,45-0,50	31		31
0,5-0,6	83	1	84
0,50-0,55	33		33
0,55-0,60	50	1	51
0,6-0,7	108	19	127
0,60-0,65	57	7	64
0,65-0,70	51	12	63
0,7-0,8	36	73	109
0,70-0,75	25	31	56
0,75-0,80	11	42	53
0,8-0,9	5	70	75
0,80-0,85	5	32	37
0,85-0,90		38	38
0,9-1		27	27
0,90-0,95		17	17
0,95-1		10	10
1		31	31
1		31	31
Total general	293	221	514

Tabla 5. Resultados All-MiniLM-L6v2 en la comparación de habilidades

3.2.3.4 Evaluación modelo Albert

Analizando los resultados del modelo Albert [37], se obtuvieron un total de 148 altas coincidencias que, al igual que el anterior modelo, estuvieron bastante dispersos entre varios intervalos. Además, este modelo logró detectar 25 pares de habilidades descritas completamente igual y el rango de fiabilidad que se obtuvo fue del [0,85-0,90], siendo el 0,90 el umbral a partir del cual se pueden obtener resultados altamente fiables.

Albert	No	Si	Total general
0,3-0,4	3		3
0,35-0,4	3		3
0,4-0,5	42		42
0,4-0,45	14		14
0,45-0,5	28		28
0,5-0,6	90	4	94
0,5-0,55	31	2	33
0,55-0,6	59	2	61
0,6-0,7	108	5	113
0,6-0,65	55	3	58
0,65-0,7	53	2	55
0,7-0,8	99	23	122
0,7-0,75	61	7	68
0,75-0,8	38	16	54
0,8-0,9	24	58	82
0,8-0,85	18	28	46
0,85-0,9	6	30	36
0,9-1		33	33
0,9-0,95		16	16
0,95-1		17	17
1		25	25
1		25	25
Total general	366	148	514

Tabla 6. Resultados de Albert en la comparación de habilidades

4 Resultados y conclusiones

En esta sección, como indica su nombre, se divide en dos apartados: la explicación de los resultados obtenidos recogidos en la sección 4.1 y la realización de una breve conclusión del trabajo en la sección 4.2.

4.1 Resultados

En el centro de nuestro estudio sobre la similitud semántica entre habilidades, en este apartado trataremos los resultados recogidos para la realización del Trabajo de Fin de Grado presente y las conclusiones obtenidas del mismo. Cabe destacar que el código y la aplicación desarrollados para esta investigación se encuentran en el repositorio Github denominado “tfg_project”¹¹.

En el marco del estudio sobre el análisis de la similitud semántica entre habilidades, es importante destacar que se ha orientado hacia un enfoque no supervisado debido a que la fuente no poseía datos etiquetados.

En términos de resultados, al analizar las evaluaciones de los modelos mencionados en la sección 3.2.3, se destaca que el modelo All-MiniLM-L6v2 ha mostrado una ventaja considerable en la tarea de representación vectorial de las habilidades, frente al resto de los modelos. Dada esta destacada actuación y para su mejor comprensión, se ha desarrollado una pequeña aplicación que tiene como propósito principal de asignar la habilidad equivalente en la clasificación de ESCO. Esta aplicación utiliza como referencia los intervalos de fiabilidad mencionados previamente, específicamente [0,75-0,83], con el objetivo de proporcionar una interpretación más clara y efectiva de los resultados.

Con el propósito de facilitar una comprensión más completa de los resultados de nuestro estudio, presentamos ejemplos detallados en la Figura 4 que ilustran el proceso de identificación de habilidades similares en la clasificación ESCO, tomando como ejemplo la habilidad "VR" (Virtual Reality).

Se proporciona un desglose de la información resultante de la Figura 4. En primer lugar, se presenta la habilidad en cuestión, identificada como "SKILL". A continuación, se revela la etiqueta correspondiente en la clasificación ESCO, designada como "LABEL ESCO", que en este caso es "Virtual Reality". Además, se ofrece una descripción detallada de la habilidad de ESCO bajo el encabezado "DESCRIPTION ESCO", que muestra el significado, según ESCO, de "Virtual Reality": "The process of simulating real-life experiences in a completely immersive digital environment. The user interacts with the virtual reality system via devices such as specifically designed headsets.". Finalmente, se proporciona la URL de la página web asociada a esta habilidad particular en la clasificación ESCO, bajo la etiqueta "URL ESCO": "http://data.europa.eu/esco/skill/5da42cfd-1da8-4e4f-b68e-4f821d005fc5", facilitando la obtención de detalles adicionales. Este mismo formato explicativo se aplica de la misma manera a los diversos resultados presentados en las Figuras 5, 6 y 7. Nótese que se proporcionan diferentes ejemplos de habilidades a identificar: iniciales de una habilidad (Figura 4), una habilidad de una palabra (Figura 5), una habilidad descrita de dos palabras (Figura 6) y una habilidad resumida en una frase (Figura 7).

¹¹https://github.com/rvocampo26/tfg_project/tree/60e75f3349501a60756d082b6345da51b1a1cbd4/AI4LABOUR_course-ESCO_GIT



The screenshot shows a web browser window titled "Similar SKILL". At the top center is a logo with the text "SIMILAR SKILL" in a stylized, orange, italicized font. Below the logo, there is a text input field with the placeholder text "Introduce la habilidad de la que deseas obtener información ESCO:". Below the input field is a button labeled "Buscar". Below the button, there is a text area containing the following information:

```
SKILL
VR
LABEL ESCO
virtual reality
DESCRIPTION ESCO
The process of simulating real-life experiences in a completely immersive digital environm
ent. The user interacts with the virtual reality system via devices such as specifically d
esigned headsets.
URL ESCO
http://data.europa.eu/esco/skill/5da42cfd-1da8-4e4f-b68e-4f821d005fc5
```

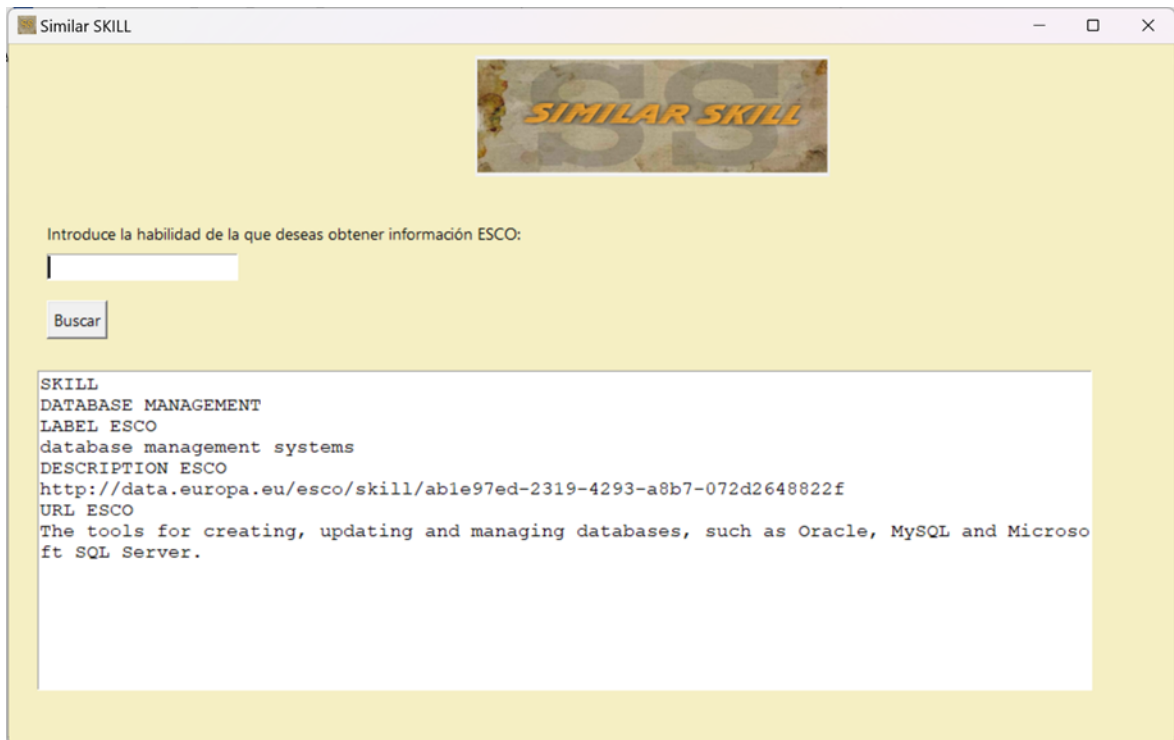
Figura 5. Ejemplo búsqueda habilidad: “VR”



The screenshot shows a web browser window titled "Similar SKILL". At the top center is a logo with the text "SIMILAR SKILL" in a stylized, orange, italicized font. Below the logo, there is a text input field with the placeholder text "Introduce la habilidad de la que deseas obtener información ESCO:". Below the input field is a button labeled "Buscar". Below the button, there is a text area containing the following information:

```
SKILL
JAVA
LABEL ESCO
Java (computer programming)
DESCRIPTION ESCO
http://data.europa.eu/esco/skill/19a8293b-8e95-4de3-983f-77484079c389
URL ESCO
The techniques and principles of software development, such as analysis, algorithms, codin
g, testing and compiling of programming paradigms in Java.
```

Figura 6. Ejemplo de búsqueda habilidad: “JAVA”



Similar SKILL

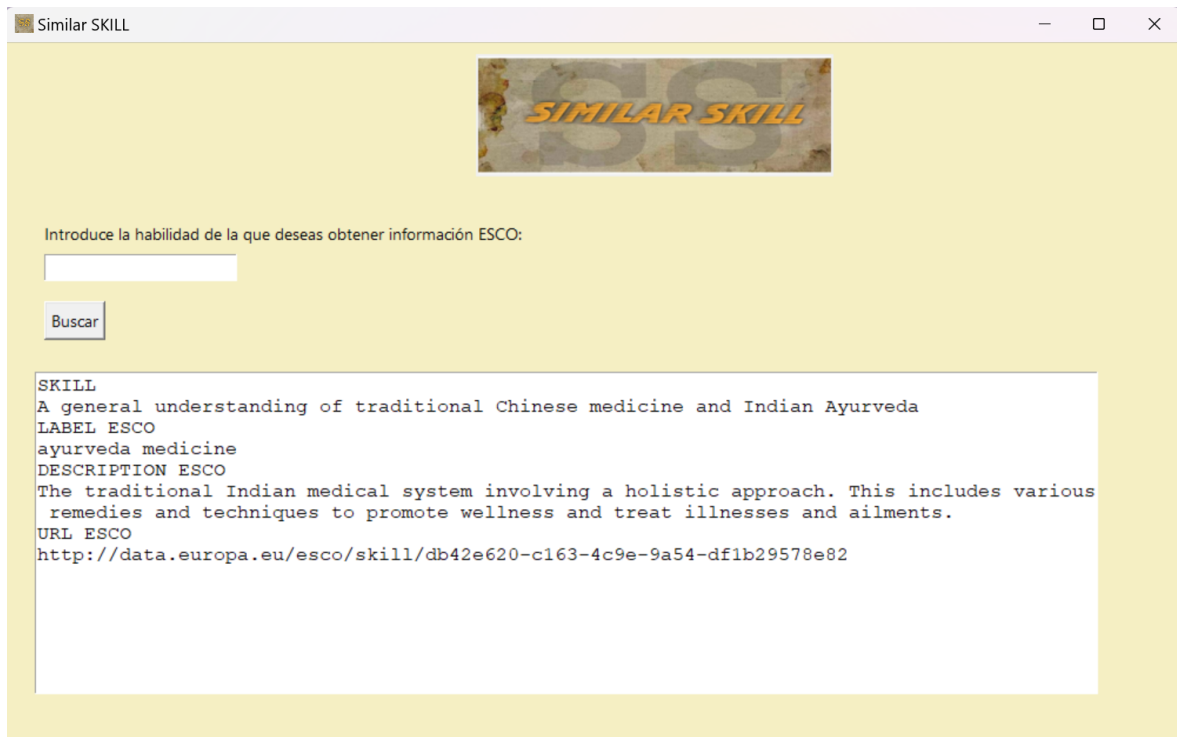
Introduce la habilidad de la que deseas obtener información ESCO:

Buscar

```

SKILL
DATABASE MANAGEMENT
LABEL ESCO
database management systems
DESCRIPTION ESCO
http://data.europa.eu/esco/skill/able97ed-2319-4293-a8b7-072d2648822f
URL ESCO
The tools for creating, updating and managing databases, such as Oracle, MySQL and Microsoft SQL Server.
  
```

Figura 7. Ejemplo de búsqueda de habilidad "DATABASE MANEGEMENT"



Similar SKILL

Introduce la habilidad de la que deseas obtener información ESCO:

Buscar

```

SKILL
A general understanding of traditional Chinese medicine and Indian Ayurveda
LABEL ESCO
ayurveda medicine
DESCRIPTION ESCO
The traditional Indian medical system involving a holistic approach. This includes various remedies and techniques to promote wellness and treat illnesses and ailments.
URL ESCO
http://data.europa.eu/esco/skill/db42e620-c163-4c9e-9a54-df1b29578e82
  
```

Figura 8. Ejemplo de búsqueda habilidad "A general understanding of traditional Chinese medicine and Indian Ayurveda"

4.2 Conclusiones

A continuación, se exponen las 3 principales conclusiones obtenidas al realizar la investigación en cuestión.

En relación con el análisis de los modelos, la conclusión resalta en el interesante apunte de que, a pesar de que los cuatro modelos comparten la misma función para vectorizar las habilidades, cada uno presenta su propia estructura única para dicha función. Por lo que, aunque todos están implementados con el mismo objetivo, los resultados obtenidos difieren considerablemente. El motivo de por qué las mismas habilidades codificadas por diferentes modelos, podría ser un objeto intrigante para realizar investigaciones futuras.

En cuanto al estudio de habilidades, como conclusión personal de la investigación he observado la notable diversidad de trabajos que buscan identificar habilidades en ofertas de empleo y, sin embargo, hay una escasez significativa de enfoques que se centren en la recomendación de cursos para adquirir estas habilidades. Este vacío en la investigación podría ser un terreno bastante motivador para futuros estudios que exploren y propongan soluciones en este sentido.

Por último, y como consecuencia de lo anterior, se muestra una evidente una carencia de recursos de información con datos etiquetados provenientes de descripciones de cursos en plataformas MOOC, como en nuestro caso, Coursera. La posibilidad de realizar etiquetados en estos ejemplos podría facilitar técnicas como el Reconocimiento de Entidades Nombradas (NER) y, al mismo tiempo, simplificar la utilización de modelos de lenguaje ya preentrenados para abordar esta tarea en específico. La creación de conjuntos de datos etiquetados podría abrir nuevas oportunidades para avanzar en la comprensión y aplicación de técnicas de procesamiento del lenguaje natural en el área de la educación a nivel individual online.

5 Análisis de Impacto

El análisis de impacto de este estudio, centrado en la detección de similitud de habilidades, revela perspectivas interesantes y relevantes para el campo de la representación vectorial en el ámbito de las competencias laborales al enfocarse en la gran eficacia que poseen los modelos basados en la arquitectura Transformer. Además, se destaca la capacidad natural de estas representaciones para capturar la relación semántica de las habilidades analizadas.

Este estudio no solo contribuye a entender qué modelos sobresalen en la tarea de detectar similitud entre habilidades en inglés, sino que también asienta las bases para realizar futuros desarrollos. Por consecuente, afirmamos que el presente trabajo puede servir como base para un extractor de habilidades futuro lo que tendrá un gran impacto en la creación de herramientas más especializadas, eficientes y enfocadas en la extracción y comparación de habilidades.

En términos de un impacto más práctico, este estudio podría tener implicaciones directas en el desarrollo de sistemas avanzados de gestión de talentos, facilitando la identificación rápida y precisa de habilidades relevantes, es decir, se podría implementar una aplicación que tengan la función de un buscador para obtener las habilidades más relevantes de textos. A tener en cuenta, que todo lo mencionado anteriormente se basará en la utilidad de la vectorización del modelo All-MiniLM-L6v2.

Por último, este estudio no solo arroja luz sobre el estado actual de la detección de similitud de habilidades, sino que también indica un camino hacia futuras innovaciones en la extracción y comparación de competencias y habilidades laborales presentes en los cursos de diferentes plataformas MOOC.

6 Bibliografía

- [1] M. H. Andreas M. Kaplan, «Higher education and the digital revolution: About MOOCs, SPOCs, social media, and the Cookie Monster,,» 2016. [En línea]. Available: <https://doi.org/10.1016/j.bushor.2016.03.008..>
- [2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser y I. Polosukhin, «Attention Is All You Need,» 2017. [En línea]. Available: [arxiv:1706.03762](https://arxiv.org/abs/1706.03762).
- [3] A. NACELLE y E. MIZRAJI, «Redes neuronales artificiales,» 2009. [En línea].
- [4] A. M. Larriba Flor, «Traducción automática basada en caracteres y redes neuronales,» 2017. [En línea]. Available: <http://hdl.handle.net/10251/89965>.
- [5] M. Campos Mocholí, «Clasificación de textos basada en redes neuronales,» 2021. [En línea]. Available: <http://hdl.handle.net/10251/172276>.
- [6] J. GARCÍA SAN VICENTE, «Resumen abstractivo de textos basado en redes neuronales,» 2020. [En línea]. Available: <http://hdl.handle.net/10251/151561>.
- [7] Z. C. Lipton, J. Berkowitz y C. Elkan, «A Critical Review of Recurrent Neural Networks for Sequence Learning,» 2015. [En línea]. Available: [arXiv:1506.00019](https://arxiv.org/abs/1506.00019).
- [8] C. C. Aggarwal, «Recurrent Neural Networks,» de *Machine Learning for Text*, 2 ed., Springer, 2022, pp. 341-363.
- [9] J. Sarzynska-Wawer, Aleks, e. Wawer, A. Pawlak, J. Szymanowska, I. Stefaniak, M. Jarkiewicz y L. Okruszek, «Detecting formal thought disorder by deep contextualized word representations,» 2018. [En línea]. Available: [arXiv:1802.05365](https://arxiv.org/abs/1802.05365).
- [10] M. R. A. Sofia, Y. V. K. Valentina, M. H. G. Andres y K. S. P. Nimisica, «Modelos de atención aplicados a clasificación de textos narrativos,» 2021. [En línea]. Available: <http://hdl.handle.net/10554/54930>.
- [11] J. ESTÉVEZ ASENSIO, «Análisis emocional en redes sociales basados en modelos de aprendizaje automático transformers BERT.,» 2023. [En línea].
- [12] J. SANCHEZ GONZALO, «Análisis del estado del arte de la generación de texto con redes neuronales mediante modelos de Transformers.,» 2020. [En línea].
- [13] J. Devlin, M.-W. Chang, K. Lee y K. Toutanova, «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,» 2019. [En línea].
- [14] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li y P. J. Liu, «Exploring the limits of transfer learning with a unified,» 2019. [En línea]. Available: [arXiv:1910.10683](https://arxiv.org/abs/1910.10683).
- [15] V. M. Dhivya Chandrasekaran, «Evolution of Semantic Similarity – A survey,» 2020. [En línea]. Available: [arXiv:2004.13820v2 \[cs.CL\]](https://arxiv.org/abs/2004.13820v2) .

- [16] D. Sánchez, M. Batet, D. Isern y A. Valls, «Ontology-based semantic similarity: A new feature-based approach,» 2012. [En línea]. Available: <https://doi.org/10.1016/j.eswa.2012.01.082>.
- [17] D. Sánchez y M. Batet, «A semantic similarity method based on information content exploiting multiple ontologies,» 2013. [En línea]. Available: <https://doi.org/10.1016/j.eswa.2012.08.049>.
- [18] J. Pennington, R. Socher y C. D. Manning, «Glove: Global vectors for word representation.,» *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532-1543, 2014.
- [19] T. Landauer y S. Dumais, «A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge,» 1997. [En línea]. Available: <https://doi.org/10.1037/0033-295X.104.2.211>.
- [20] F. A. Gers, N. N. Schraudolph y J. Schmidhuber., «Learning Precise Timing with LSTM Recurrent Networks,» *Journal of machine learning research*, pp. 115-143, 2002.
- [21] J. Camacho-Collados, M. T. Pilehvar y R. Navigli, «Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities.,» 2016. [En línea]. Available: <https://doi.org/10.1016/j.artint.2016.07.005..>
- [22] A. Aizawa, T. Ruas y W. Grosky, «Multi-sense embeddings through a word sense disambiguation process,» *Expert Systems with Applications*, vol. 136, pp. 288-303, 2019.
- [23] I. Khaouja, I. Kassou y M. Ghogho, «A Survey on Skill Identification From Online Job Ads,» 2021. [En línea]. Available: [10.1109/ACCESS.2021.3106120](https://doi.org/10.1109/ACCESS.2021.3106120).
- [24] B. Joseph, A. Rios, G. Ling, R. Pugh y D. Becker, «Identifying Critical 21st-Century Skills for Workplace Success: A Content Analysis of Job Advertisements,» *Educational Researcher*, 2020.
- [25] H. CHAIBATE, A. HADEK, S. AJANA, S. BAKKALI y K. FARAJ, «Analyzing the engineering soft skills required by Moroccan job market,» 2019. [En línea].
- [26] F. Gurcan y N. E. Cagiltay, «Big Data Software Engineering: Analysis of Knowledge Domains and Skill Sets Using LDA-Based Topic Modeling,» vol. 7.
- [27] S. Debortoli, O. Müller y J. v. Brocke, «Comparing business intelligence and big data skills,» 2014. [En línea]. Available: <https://link.springer.com/article/10.1007/s12599-014-0344-2..>
- [28] S. Gandhi, R. Nagesh y S. Das, «Learning skills adjacency representations for optimized reskilling recommendations,» 2022.
- [29] T. Mikolov, K. Chen, G. Corrado y J. Dean, «Efficient Estimation of Word Representations in Vector Space,» 2013. [En línea]. Available: [arXiv:1301.3781](https://arxiv.org/abs/1301.3781).
- [30] V. S. Dave, B. Zhang, M. A. Hasan, K. AlJadda y M. Korayem, «A Combined Representation Learning Approach for Better Job and Skill Recommendation,» 2018. [En línea]. Available: <https://arxiv.org/abs/1808.08111>.

<https://doi.org/10.1145/3269206.3272023>.

- [31] G. Cenikj, B. Vitanova y T. Eftimov, «Skills Named-Entity Recognition for Creating a Skill Inventory of Today's Workplace,» 2021. [En línea]. Available: 10.1109/BigData52589.2021.9671435.
- [32] L. Cao, «Skill Requirements Analysis for Data Analysts Based on Named Entities Recognition,» *IEEE*, pp. 64-68, 2021.
- [33] A. Nigam, S. Tyagi, K. Tyagi y A. Saxena, «SkillBERT: "Skilling" the BERT to classify skills!,» 2020. [En línea].
- [34] N. Goyal, J. Kalra, C. Sharma, R. Mutharaju, N. Sachdeva y P. Kumaraguru, «JobXMLC: EXtreme Multi-Label Classification of Job Skills with Graph,» 2023. [En línea]. Available: <https://aclanthology.org/2023.findings-eacl.163>.
- [35] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer y V. Stoyanov, «RoBERTa: A Robustly Optimized BERT Pretraining Approach,» 2019. [En línea]. Available: arXiv:1907.11692.
- [36] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang y M. Zhou, «MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers,» 2020. [En línea]. Available: arXiv:2002.10957.
- [37] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. S. y R. Soricut, «ALBERT: A Lite BERT for Self-supervised Learning of Language Representations,» 2019. [En línea]. Available: arXiv:1909.11942.