Universidad Politécnica
de Madrid

Escuela Técnica Superior de
Ingenieros Informáticos

Grado en Ingeniería Informática

Trabajo Fin de Grado

Extracción de Información a partir de Descripciones de Cursos

Autora: Ruth Verónica Ocampo Prado

Tutora: María Navas Lora

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: Extracción de Información a partir de Descripciones de Cursos
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 objetivo de este Trabajo de Fin de Grado es identificar las habilidades y tareas similares en diferentes textos, utilizando técnicas de Procesamiento del Lenguaje Natural.

De esta forma, partiremos de conjuntos de datos que contienen las habilidades y tareas descritas en inglés de los cursos pertenecientes a diferentes plataformas MOOC como son Coursera y EdX.

Primero, comenzaremos estudiando los modelos de lenguaje basados en Transformers para poder seleccionar el modelo que mejor se adapte a nuestras necesidades. A continuación, entrenaremos a nuestro modelo con bases de datos de habilidades ESCO y O*NET, y realizaremos una comparación de estas habilidades con las habilidades extraídas de los textos, utilizando el coseno como unidad de medición de similitud semántica.

Palabras Clave: similitud semántica, modelos de lenguaje, Transformers, inglés, plataformas MOOC, base de datos de habilidades, ajuste fino.

Abstract

The objective of this Final Degree Project is to identify similar skills and tasks in different texts, using techniques of Natural Language Processing.

In this way, we will start from datasets that contain all the skills and tasks with English descriptions of courses belonging to different MOOC's platforms, such as Coursera and EdX.

We will study the models based on Transformers to select the model that best suits our needs. Once this is done, we will train our model with ESCO and O*NET skills databases, and then we will make a comparison of these, with the skills extracted from the texts, using the cosine as a unit of measurement of semantic similarity.

Keywords: semantic similarity, language models, Transformers, english, Mooc platforms, skills databases, fine tuning.

Tabla de contenidos

1 Planificación

La planificación actual se podrá visualizar desde el siguiente enlace: Diargrama_Gantt_TFG.xlsx

1.1 Cambios en planificación

Respecto a los cambios realizados en relación con la planificación inicial, solo se han visto afectadas las tareas de "Análisis del estado del arte" y "Documentación y escritura del TFG 1º parte".

La primera tarea comenzó como estaba planificada, en la 3ª semana. El cambio que se realizó en esta tarea fue que nos vimos obligadas a extenderla hasta la 5ª semana, puesto que en la planificación inicial se dedicaban 16 horas en la 4ª semana, consideramos que al tener que realizar el borrador del documento era mejor organizarlo e ir investigando con más tiempo y calma.

Respecto a la segunda tarea, se planificó en un principio que se comenzaría a redactar la memoria en la 7ª semana, pero para lograr que la información sea tan fiable como cuando se realizó la investigación se empezó a realizar un borrador de dicha memoria desde la 2ª semana.

2 Introducción

En un mundo impulsado por datos, la extracción de la información relevante de grandes textos ha cobrado una gran importancia en términos de ahorro de tiempo. Es aquí donde podemos percibir que el área del Procesamiento del Lenguaje Natural [1] ha destacado debido a su combinación de la Inteligencia Artificial con la Lingüística Aplicada, pretendiendo 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.), en el análisis de sentimientos de las redes sociales, entre otros.

La similitud semántica juega un papel fundamental en muchas aplicaciones que utilizan el Procesamiento del Lenguaje Natural, dado que permite que las máquinas comparen textos no solo 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 las plataformas MOOC. Por otro lado, también abordaremos la problemática de la extracción de dichas habilidades en textos como descripciones de cursos o guías de aprendizajes.

Para ser más exactos, nos centraremos sólo en los cursos impartidos en inglés ofrecidos en las plataformas de Coursera y EdX.

Para poder realizar este estudio utilizaremos un enfoque híbrido combinando un algoritmo supervisado con uno no supervisado, con el objetivo de lograr una identificación más precisa de las habilidades.

Usaremos modelos grandes basados en la arquitectura Transformer, la cual ha tomado una gran importancia en los últimos años gracias a que están tienen como base mecanismos de atención, lo que les permite capturar las relaciones semánticas con una gran eficacia. Estos modelos se pueden pre-entrenar en grandes corpus y, además, permiten ser ajustados para realizar tareas específicas de PLN, como es en nuestro caso, la medición de la similitud semántica y la extracción de información de textos.

Podemos dividir este proyecto en dos fases, una fase por cada tarea que debemos realizar:

- Diseño, implementación y evaluación de un algoritmo que identifique las habilidades y tareas similares de distintas bases de datos.
- Diseño, implementación y evaluación de un algoritmo para la extracción de dichas tareas y habilidades a partir de descripciones textuales de cursos.

Para la primera fase, utilizaremos un modelo no supervisado, ya que partimos de bases de datos con datos no etiquetados. Este algoritmo será entrenado y evaluado con el fin de que se ajuste lo máximo posible a las necesidades del proyecto, sabiendo que el objetivo es que mida la similitud de las palabras con la mayor eficiencia posible.

Para la segunda fase, utilizaremos un modelo supervisado para poder identificar y extraer las habilidades, a partir de la base de datos de entidades creada anteriormente.

3 Estado de la cuestión

3.1 Marco Tecnológico

3.1.1 Plataformas MOOC

Las plataformas MOOC (Massive Open Online Course) se caracterizan por ofrecer de forma masiva una gran variedad de cursos formativos online. Los cursos están disponibles para cualquier persona con ganas de aprender y acceso a Internet y aunque por lo general, la mayoría son gratuitos, también existen opciones de pago.

Los cursos que ofrecen estas plataformas tienen 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.

Muchos de los cursos de estas plataformas están acreditados por universidades o instituciones conocidas mundialmente y los cursos pueden abarcar diferentes áreas como el de Ciencia de Datos, Tecnología de la Computación, aprendizaje de un Idioma, Salud, Artes y humanidades, entre otros.

En este trabajo se estudiarán las habilidades que ofrecen los diferentes cursos en inglés de las plataformas Coursera y EdX.

3.1.1.1 Coursera

La plataforma Coursera [2], una de las plataformas más populares, fue fundada por Daphne Koller y Andrew Ng en 2012 con el objetivo de proporcionar experiencias de aprendizaje a estudiantes alrededor de todo el mundo. Trabaja con diferentes universidades reconocidas como Universidad de Yale o Michigan, además de importantes instituciones como IBM, Google, entre otros.

3.1.1.2 EdX

La plataforma Edx [3] es una plataforma de aprendizaje en línea fundada por Harvard y el MIT. Lo que en un principio fue un experimento para poder ofrecer a las personas la posibilidad de tener una educación buena en cualquier parte del mundo, ha sido una de las plataformas más revolucionarias que conecta a más de 78 millones de personas. Colaboran con ella universidades importantes

como la Universidad de Columbia, Universidad de Brown, entre otras muchas más instituciones.

3.1.2 Redes Neuronales

En la actualidad, las redes neuronales [4] se han convertido en una herramienta muy útil para abordar los problemas del PLN, como lo es encontrar la similitud semántica entre palabras (esta problemática se desarrollará en el apartado 2.1.3). En este apartado explicaremos qué son las redes neuronales, cómo funcionan y las aplicaciones que nos proporcionan en el análisis de habilidades.

Las redes neuronales son redes neuronales artificiales que imitan el sistema nervioso del ser humano, están compuesta por un conjunto de nodos, los cuales representan a las neuronas de nuestro sistema nervioso, dando lugar a capas de nodos. En estas redes nos podemos encontrar normalmente con una capa de entrada, una o varias capas ocultas y una capa de salida.

Estas redes neuronales tienen el siguiente funcionamiento cada nodo de la red neuronal contiene un peso, un umbral y está conectado a otro nodo. A modo que, si la salida del primer nodo supera el valor del umbral que se ha especificado, el nodo se activará enviando datos a la siguiente capa de red.

A través de entrenamientos con 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, la clasificación de texto, el resumen texto, entre otras tareas.

Antes del 2017, el tipo más destacado de redes neuronales eran las redes neuronales recurrentes, las cuales utilizaban bucles de retroalimentación con el fin de que la información dure varias etapas de entrenamiento. A partir de dicho año, la arquitectura Transformer fue desarrollada y en poco tiempo logró reemplazar a las redes neuronales recurrentes como veremos en el siguiente punto.

3.1.3 Arquitectura Transformer

Los modelos de lenguaje basados en redes neuronales recurrentes [5] como Elmo, 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 [6], desarrollada en 2017 por Google, 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.

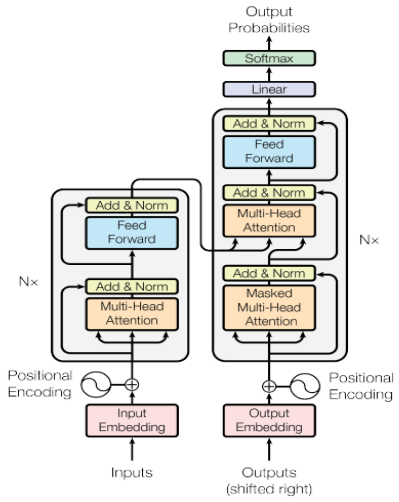

Figure 1: The Transformer - model architecture.

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. A pesar de que en un principio se entrenó especialmente para la traducción automática (del inglés al alemán y del inglés al francés), en la actualidad se puede utilizar para una gran variedad de tareas de PLN como la clasificación de texto, el análisis de sentimientos, etc...

Para hallar la similitud semántica, evaluaremos los modelos pre-entrenados basados en los tres modelos de lenguajes más destacados: BERT, GPT y T5.

3.1.4 Hugging Face

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

Hugging Face

Destaca principalmente por permitir el acceso fácil a los modelos basados en Transformer a través de la biblioteca que recibe el nombre de "SentenceTransformer". Esto, 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 de un ahorro en tiempo, puesto que los modelos a los que se accede en dicha biblioteca ya están entrenados.

3.1.5 Similitud Semántica

La similitud semántica ha sido un tema ampliamente investigado en el área del Procesamiento del lenguaje natural. 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 ha jugado un papel muy importante.

En un principio, los expertos e investigadores del PLN hacían referencia a la similitud cuando dos oraciones contenían las mismas palabras. Esta idea, aunque servía para comparar cierto tipo de oraciones, seguía teniendo una taza 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, ni los casos en que las oraciones contenían una misma palabra, pero según el contexto tenían diferentes significados.

Para resolver esta problemática, durante las últimas décadas los expertos han llevado a cabo grandes investigaciones y desarrollado diferentes técnicas de similitud semántica, las cuales pueden medir la clasificación o el porcentaje de similitud entre los textos u oraciones examinados.

Según el estudio realizado sobre la evolución de la similitud semántica [7], 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 una o más fuentes de conocimientos, como, por ejemplo, bases de datos léxicas, diccionarios, etc.
- 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 del texto en cuestión y para estimar la similitud entre estas representaciones, utilizaban medidas de distancia semántica basadas en la Hipótesis Distributiva.
- Métodos basados en Redes neuronales profundas: Estos métodos aprovechan los desarrollos en redes neuronales para mejorar el rendimiento y estiman la similitud entre vectores de palabras, aprovechando las incrustaciones de palabras.
- 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, para obtener la similitud entre textos.

3.2 Trabajos relacionados

En este apartado hablaremos sobre los diferentes métodos empleados para la identificación de las habilidades en trabajos similares o relacionados con el tema en cuestión.

El estudio “A Survey on Skill Identification From Online Job Ads” [8] 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:

1. Recuento de habilidades.
   Este método suele ser 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 el análisis de la lista de habilidades más frecuentes detectadas. Puede llevarse a cabo de dos maneras: bien contando con una base de habilidades, o bien mediante un grupo de profesionales que lleven a cabo el trabajo de manera manual.

   Esta técnica la podemos ver empleada en el estudio "Identifying Critical 21st-Century Skills for Workplace Success: A Content Analysis of Job Advertisements" [9] realizado por Rios J. A., Ling, G., Pugh, R., Becker, D., y Bacall, A., en el cual investigaban las habilidades fundamentales del siglo XXI para tener éxito en un trabajo después de graduarse en la universidad. En él, podemos notar como los investigadores recopilaron 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 Psycinfo y Merriam-Webster. 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.

2. Modelado de temas para extraer las habilidades.
   Es un método no supervisado, es decir, es un método 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 de 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 "Big Data Software Engineering: Analysis of Knowledge Domains and Skill Sets Using LDA-Based Topic Modeling" [10], realizado por F. Gurcan y N. E. Cagiltay, 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 Lantent Dirichlet Allocation (LDA), gracias a la cual pueden agrupar semánticamente las habilidades o conocimientos recogidos en temas o áreas específicos. Por último, realizan una interpretación en función de los datos obtenidos en los anteriores pasos.

   Aunque aplicar estas técnicas 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.

3. Incorporación de habilidades.
   Este método, se propusó 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 "Skill Embedding" en la

que se crea una representación vectorial por cada habilidad, a 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 trabajos como "Learning skills adjacency representations for optimized reskilling recommendations" [11], en el que utilizan el modelo Word2vec para crear las representaciones de las habilidades que querían estudiar y una vez hecho esto, medían el grado de similitud entre pares de habilidades.

El tener cada palabra una representación única permitió reducir los falsos positivos. Pero, por el contrario, también aumentó el número de falsos negativos, y al ser las palabras representadas de esta forma, no permite tener en cuenta el contexto de las frases y manejar bien las relaciones semánticas entre las diferentes palabras que componen el texto.

4. Técnicas de aprendizaje automático (ML):
   De estas técnicas, podemos encontrar dos principales:
   - Reconocimiento de entidades nombradas (NER) es una tarea del PLN que hace referencia al proceso de extraer información de textos estructurados, como pueden ser nombres, lugares, fechas, etc. Un claro ejemplo de uso de esta técnica la podemos ver en el estudio realizado por G. Cenikj, B. Vitanova y T. Eftimov [12], en el cuál estudian qué método tiene mejores resultados, aplicando el modelo NER combinado con diccionarios como ESCO, LinkedIn y Employment para extraer habilidades tanto técnicas como personales de ofertas de trabajo de Google y Amazon.
   - Clasificación de texto: Esta técnica asigna etiquetas de categorías relacionadas con habilidades específicas. Con el fin de entender mejor este método ponemos como ejemplo el estudio "SkillBERT: "Skilling" the BERT to classify skills!" [13], realizado por Amber Nigam, Shikha Tyagi, Kuldeep Tyagi y Arpan Saxena. En este trabajo 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.

Por último, también podemos encontrar trabajos similares en los que se trata la clasificación de habilidades, en diferentes idiomas como es el danés en el estudio "Kompetencer: Fine-grained Skill Classification in Danish Job Postings via Distant Supervision and Transfer Learning" [14] realizado por Mike Zhang, Kristian Norgaard Jensen y Barbara Plank, o también el francés en el estudio "BARThez: a Skilled Pretrained French Sequence-to-Sequence Model", realizado por Moussa Kamal Eddine , Antoine J. -P. Tixier y Michalis Vazirgiannis [15].

4 Desarrollo

En este trabajo, para medir la similitud semántica entre palabras, utilizaremos la medida del coseno.

4.1 Datasets

4.1.1 Base de datos ESCO

4.1.1.1 ¿Qué es ESCO?

4.1.1.2 Utilidades

4.1.2 Base de datos O*NET

4.1.2.1 ¿Qué es O*NET?

4.1.2.2 Utilidades

4.2 Algoritmo de identificación de habilidades: All-MiniLM-L6-v2

4.2.1 Diseño del algoritmo

4.2.2 Evaluación

4.3 Algoritmo de extracción de entidades a partir de cursos: “X”

4.3.1 Diseño del algoritmo

4.3.2 Evaluación

5 Resultados y conclusiones

Resumen de resultados obtenidos en el TFG. Y conclusiones personales del estudiante sobre el trabajo realizado.

6 Análisis de Impacto

7 Bibliografía

[1] Dail Software, «Aplicaciones del Porcesamiento del lenguaje natural,» [En línea]. Available: https://www.dail.es/aplicaciones-del-procesamiento-del-lenguaje-natural/ .

[2] Coursera, «About Coursera,» [En línea]. Available: https://about.coursera.org/ .

[3] Edx, «About us,» [En línea]. Available: https://www.edx.org/about-us.

[4] IBM, «¿Qué son las redes neuronales?,» [En línea]. Available: https://www.ibm.com/es-es/topics/neural-networks .

[5] C. C. Aggarwal, «Recurrent Neural Networks,» de *Machine Learning for Text*, 2 ed., Springer, 2022, pp. 341-363.

[6] N. S. N. P. J. U. L. J. A. N. G. L. K. I. P. Ashish Vaswani, «Attention Is All You Need,» 2017. [En línea]. Available: https://arxiv.org/abs/1706.03762.

[7] V. M. Dhivya Chandrasekaran, «Evolution of Semantic Similarity – A survey,» 2020. [En línea]. Available: arXiv:2004.13820v2 [cs.CL] .

[8] I. K. a. M. G. I. Khaouja, «IEEE Access,» vol. 9, pp. 118134-118153, 2021.

[9] J. A. L. G. P. R. B. D. &. B. A. Rios, «Identifying Critical 21st-Century Skills for Workplace Success: A Content Analysis of Job Advertisements,» Educational Researcher, 2020.

[10] F. G. y. N. E. Cagiltay, «Big Data Software Engineering: Analysis of Knowledge Domains and Skill Sets Using LDA-Based Topic Modeling,» vol. 7.

[11] R. N. y. S. D. S. Gandhi, «Learning skills adjacency representations for optimized reskilling recommendations,» 2022.

[12] B. V. y. T. E. G. Cenikj, «Skills Named-Entity Recognition for Creating a Skill Inventory of Today's Workplace,» 2021. [En línea].

[13] S. T. K. T. A. S. Amber Nigam, «SkillBERT: "Skilling" the BERT to classify skills!,» 2020. [En línea]. Available: https://openreview.net/forum?id=TaUJl6Kt3rW.

[14] K. N. J. B. P. Mike Zhang, «Kompetencer: Fine-grained Skill Classification in Danish Job Postings via Distant Supervision and Transfer Learning,» [En línea]. Available: arXiv:2205.01381.

[15] B. a. S. P. F. S.-t.-S. Model, «Moussa Kamal Eddine, Antoine J.-P. Tixier, Michalis Vazirgiannis,» 2020. [En línea]. Available: arXiv:2010.12321.

[16] ESCO, «What is ESCO?,» 2020. [En línea]. Available: https://esco.ec.europa.eu/en/about-esco/what-esco.