Exploring Topics in Information Technology Open Educational Resources through the LDA Algorithm

ARTICLE HISTORY

Received 18 September 2023
Accepted 27 November 2023

René Ludeña
Universidad Técnica Particular de Loja
Loja, Ecuador
rjludena@utpl.edu.ec

Verónica Segarra-Faggioni
Universidad Técnica Particular de Loja
Loja, Ecuador
Ecole de Technologie Supérieure, Quebec, Canadá
vasegarra@utpl.edu.ec
ORCID: 0000-0002-7275-7411

Audrey Romero-Peláez
Universidad Técnica Particular de Loja
Loja, Ecuador
Universidad Politecnica de Madrid Madrid, España
aeromero2@utpl.edu.ec
ORCID: 0000-0002-3710-1257

Juan Carlos Morocho-Yunga
Universidad Técnica Particular de Loja
Loja, Ecuador
jcmorocho@utpl.edu.ec
ORCID: 0000-0001-6387-8054

Explorando Temas en Recursos Educativos Abiertos de Tecnologías de la Información a través del Algoritmo LDA

Exploring Topics in Information Technology Open Educational Resources through the LDA Algorithm

René Ludeña
Universidad Técnica Particular de Loja
Loja, Ecuador
rjludena@utpl.edu.ec

Verónica Segarra-Faggioni
Universidad Técnica Particular de Loja
Loja, Ecuador
Ecole de Technologie Supérieure,
Quebec, Canadá
vasegarra@utpl.edu.ec
ORCID: 0000-0002-7275-7411

Audrey Romero-Peláez
Universidad Técnica Particular de Loja
Loja, Ecuador
Universidad Politecnica de Madrid
Madrid, España
aeromero2@utpl.edu.ec
ORCID: 0000-0002-3710-1257

Juan Carlos Morocho-Yunga
Universidad Técnica Particular de Loja
Loja, Ecuador
jcmorocho@utpl.edu.ec
ORCID: 0000-0001-6387-8054

***Abstract*—** This paper explores the application of machine learning and text mining techniques to discover OER issues in the context of Engineering Education. Applying the LDA (Latent Dirichlet Allocation) algorithm, themes are extracted from OER, it is possible to consider them as additional metadata. This augmentation serves to enhance the description and categorization of OER. Furthermore, this study introduces a methodology to automatically identify topics in open educational resources. In this research, a dataset of 80 OER was obtained from the Skills Commons repository. The highest coherence value achieved at 0.42, emerged when the number of topics was 9 in the LDA model. These nine topics are closely associated with Information Technology Education.

***Keywords*—** *metadata, OER, LDA, text mining, topic modeling*

***Resumen*—** Este artículo aplica el algoritmo Latent Dirichlet Allocation, LDA, como una técnica de aprendizaje de máquina y minería de texto para descubrir temas en OER en el contexto de la educación en ingeniería. El algoritmo LDA permite extraer temas, en este estudio los temas que se extraen de OER pueden ser considerados como metadatos adicionales que enriquecerán la descripción y clasificación de los mismos. Además, se define una metodología para la identificación automática de temas en los recursos educativos abiertos. En esta investigación, se utiliza un dataset de 80 OER extraído del repositorio Skills Commons. El valor más alto de coherencia es 0.42, cuando el número de temas en el modelo LDA es 9. Estos nueve temas están relacionados con Educación en Tecnologías de la Información.

***Palabras clave*—** *metadata, OER, LDA, minería de texto, modelado de temas*

I. Introducción

En la actualidad, la computación en la educación es un componente fundamental para el desarrollo de habilidades tecnológicas que impulsan el progreso de la sociedad. Con el avance constante de la inteligencia artificial (IA) y el aprendizaje automático (ML), es fundamental que las personas adquieran competencias en el campo de la informática para hacer frente a los desafíos de un mundo cada vez más digitalizado. En este contexto, los Recursos Educativos Abiertos (OER) han surgido como una solución prometedora para democratizar el acceso a materiales educativos de alta calidad.

Los Recursos Educativos Abiertos ofrecen una estrategia y oportunidad para mejorar el acceso a la educación de manera libre y abierta a través de sus materiales [1]. Los recursos educativos abiertos son materiales educativos de acceso gratuito para el público en general, los cuales proporcionan derechos de uso, reutilización y pueden adaptarse de acuerdo a las necesidades específicas. Con la abundancia de OER disponibles, descubrir contenido relevante y útil de manera eficiente se ha convertido en un desafío para los usuarios, debido a que los metadatos no siempre describen el recurso lo más completo posible [2].

Los metadatos son una colección de atributos que describen e identifican un recurso. En consecuencia, el uso de un único metadato estándar no es factible, ya que podría existir carencias en la cobertura de datos referentes a los recursos [3]. Debido a esta diversidad de información, los metadatos no siempre pueden controlar la calidad de los temas tratados en los recursos. Para abordar este tema, en este artículo exploramos el algoritmo LDA (Latent Dirichlet Allocation) como una técnica de aprendizaje de máquina y minería de texto para descubrir patrones y estructuras en los OER

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2024
DOI: 10.5281/zenodo.10402399

seleccionados. El algoritmo LDA es una herramienta eficaz para analizar un conjunto de datos de texto y descubrir temas en estos documentos, por esta razón, se emplea LDA para identificar temas en recursos educativos abiertos de diferentes escenarios de formación en Ingeniería. Al emplear técnicas de aprendizaje automático no supervisado y minería de datos en los OER, se busca enriquecer la experiencia de aprendizaje al permitir que los usuarios encuentren los recursos apropiados según sus requerimientos de forma eficaz. Además, siguiendo la línea de investigación de varios autores que enfocan sus trabajos en la evaluación de recursos educativos abiertos desde un punto de vista técnico, se considera que con la aplicación de LDA se mejorará la organización y acceso al contenido educativo.

El presente trabajo está organizado de la siguiente manera: Sección 2. Se presentan trabajos relacionados al tema de estudio; Sección 3. Metodología aplicada; Sección 4. Experimentos con el algoritmo LDA, análisis y resultados obtenidos; y, Sección 5. Conclusiones y trabajos futuros.

II. Trabajos Relacionados

El modelado de temas es una herramienta que brinda una visión de conjuntos de documentos individuales y las interconexiones entre ellos [4].Varios autores han publicado y presentando resultados utilizando el algoritmo LDA en diferentes campos. En este trabajo, se realizó una búsqueda de artículos relacionados a la aplicación del algoritmo LDA en el contexto de los recursos educativos abiertos y se seleccionaron aquellos que aportan al objetivo de este trabajo (Ver Tabla I).

Una propuesta específica para analizar la calidad de los metadatos es aplicar el algoritmo Random Forest (RF) [5], el análisis está enfocado en aplicar un modelo de predicción para anticipar la calidad de los OER. Mediante el uso de metadatos como característica de entrada, los hallazgos indican que este modelo alcanza el 94.6% de precisión al identificar de manera correcta los OER de alta calidad.

Por otro lado, el estudio de [6] combina la técnica *ElasticSearch* y el modelo LDA con la finalidad de clasificar artículos académicos de acuerdo a temas relevantes y características extraídas de los resúmenes de los documentos. Se utilizaron como variables de entrada las palabras clave y el resumen de los artículos académicos.

Otro estudio [7] compara los métodos: LDA, BERT y Tf–idf para determinar el más eficaz para la clasificación de texto en base a una etiqueta predefinida. El conjunto de datos en este estudio corresponde a temas deportivos y educativos que fueron recopilados por estudiantes de postgrado. Así también, el estudio de [8] extrae temas de OER, aplicando técnicas de minería de texto, para generar metadatos de alta calidad.

Los trabajos relacionados utilizan Random Forest y BERT para tareas de clasificación y regresión utilizando etiquetas en un documento, mientras que LDA permite extraer temas de documentos. El enfoque de utilizar el algoritmo LDA para investigar el tema de documentos o palabras en diversos campos porque permite analizar grandes cantidades de datos. En este estudio, se aplica LDA para descubrir temas dentro de los recursos educativos abiertos con la finalidad de considerarlos como metadatos adicionales que enriquecerán la descripción y clasificación de los OER.

Tabla I. Resumen de trabajos relacionados

Propósito	Metodología	Resultados
Explorar la relación entre la calidad de los metadatos y la calidad de OER [5]	1. Análisis de datos exploratorio sobre los metadatos de OER de Youtube, aplicando RF. 2. Predicción basada en metadatos para anticipar la calidad de los OER.	La clasificación de OER de alta calidad alcanza una precisión del 94,6%.
Clasificar automáticamente artículos académicos [6]	1. Aplica el método de clasificación ElasticSearch y el modelo LDA basado en temas para extraer las características de los artículos académicos. La similitud semántica utiliza palabras clave como las variables de entrada.	El valor de $k = 50$ utilizado para aplicar LDA.
Explorar la aplicación práctica de tres métodos de modelización (LDA, BERT y TF-IDF) y determinar el método más eficaz para la clasificación de documentos. [7]	1.Preprocesamiento de los datos. 2.1. Calcular Tf – idf (Frecuencia de término – Documento inverso Frecuencia) 2.2. Aplicar algoritmo LDA: se crea un diccionario y una bolsa de palabras. 2.3. Aplicación del método BERT: se crea vectores de incrustación de oraciones.	El método BERT alcanzó el 92,6% de éxito, en la clasificación de documentos.
Realizar la extracción de temas de OER mediante técnicas de minería de texto para generar metadatos de alta calidad, lo que puede ayudar a los alumnos a construir itinerarios de aprendizaje eficaces hacia sus objetivos de aprendizaje individuales [8].	1. Recolección de datos 2. Preprocesamiento de datos 3. Aplicación del algoritmo Latent Dirichlet Allocation (LDA) 3.1.Calcular la coherencia para seleccionar el valor de k 4. Evaluación del modelo utilizando la métrica *F1-score*	El modelo extrajo temas de OER con un 79% de puntuación F1.

III. Metodología

En el presente trabajo, se aplica el algoritmo LDA (Latent Dirichlet Allocation) para la identificación automática de temas que se aborden en los recursos educativos abiertos utilizados en la formación en Ingeniería.

Con este objetivo se aplicaron las fases: entender el proceso, entender los datos, preparar los datos, construir el modelo y validar el modelo. La Figura 1 presenta el esquema de los pasos propuestos.

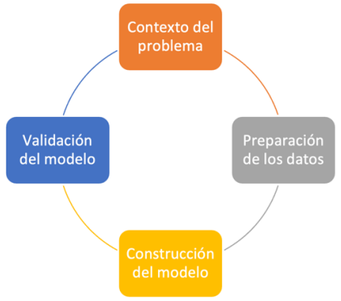


Fig. 1. Metodología aplicada

A. Contexto del problema

Se realiza una definición técnica del problema y, además, precisa los aspectos del dominio de trabajo, en este caso los metadatos de los repositorios OER. Por lo tanto, en esta fase inicial se busca comprender el problema de la búsqueda y recomendación eficiente de los recursos educativos abiertos.

En los repositorios de OER, los metadatos de estos recursos no brindan la eficiencia necesaria para los servicios de búsqueda [11]. Por esta razón, se plantea aplicar el algoritmo LDA, como técnica de aprendizaje de máquina y minería de datos, que permita la extracción de los temas que son tratados en un OER y de esta manera proveer una alternativa que mejore la eficiencia de los usuarios al encontrar y seleccionar materiales según sus requerimientos.

El algoritmo LDA permite identificar la distribución del documento sobre los temas y la distribución de un tema en función de las palabras observadas [12]. De esta manera, se identificarán automáticamente temas de los OER, los mismos que pueden ser considerados como metadatos adicionales que enriquecerán la descripción y clasificación de los recursos educativos abiertos seleccionados.

Además, se realiza un análisis exploratorio con el fin de obtener un panorama de lo que se puede conseguir a través de los datos existentes. En este punto, el conocimiento del dominio de trabajo permite guiar este análisis. Así también, en esta fase se identifican posibles problemas de calidad en los datos y se detectan subconjuntos en los datos que podrían ser interesantes [12].

B. Preparación de los datos

La fase de preparación de datos lleva a cabo todas las actividades para construir un conjunto de datos adecuado y listo para utilizar en modelos de aprendizaje automático. En esta fase el propósito es obtener una comprensión integral del conjunto de documentos que se obtengan del conjunto de datos (dataset). Los metadatos seleccionados del conjunto de datos son los siguientes: *id*, *título*, *url* y *type* (*tipo de material principal*). Con base en estos metadatos, se lleva a cabo la recopilación, exploración y evaluación de los datos para asegurar que se cuente con una colección representativa de recursos educativos abiertos y se abarque una amplia variedad de temáticas utilizadas en la formación en ingeniería.

Durante esta fase, los documentos recopilados pasan por un proceso de preprocesamiento para asegurar que sean adecuados para el análisis [13]. Esta etapa implica diversas tareas de procesamiento de lenguaje natural (en inglés NLP) como: limpieza de texto para eliminar símbolos irrelevantes, puntuación y caracteres especiales y la eliminación de palabras vacías que no aportan información relevante para el análisis. Además, se realiza la *tokenización* para identificar las palabras significativas y la lematización para reducir las palabras a su forma raíz y evitar redundancias con el fin de mejorar la eficacia del análisis.

C. Construcción del modelo

Durante el proceso de construcción del modelo se eligen y aplican diversas técnicas de modelado, mientras se ajustan los parámetros para alcanzar valores óptimos. En muchos casos, se dispone de múltiples técnicas para abordar el mismo tipo de problema. Algunas técnicas demandan formatos de datos particulares, resaltando así la relación entre la preparación de los datos y el proceso de modelado [9].

En esta fase, se incorpora el algoritmo LDA para extraer temas de los documentos que han sido previamente preprocesados. LDA identifica estos temas basándose en la distribución probabilística de palabras en los documentos y la distribución de temas dentro de los mismos [14]. Es importante determinar el número adecuado de temas, lo que dependerá del conocimiento del dominio y el nivel de granularidad deseado. Una vez definido el modelo basado en el algoritmo de aprendizaje LDA se ejecuta el entrenamiento y validación del modelo.

D. Validación del modelo

En esta etapa final, luego de contar con uno o más modelos que pueden tener alta calidad, desde una perspectiva de análisis de datos. Antes de avanzar con la implementación final del modelo, es necesario realizar una evaluación exhaustiva y revisar los pasos ejecutados para la construcción del modelo, con la finalidad de asegurarse que logre los objetivos propuestos para solucionar el problema planteado al inicio. Al concluir esta fase, se debe llevar a cabo un análisis detallado de los resultados [9], [15]. Los resultados de esta etapa son fundamentales para ajustar y mejorar el modelo LDA y, en última instancia, para lograr un análisis más preciso y útil.

La fase de validación permite garantizar que los temas extraídos sean de alta calidad y puedan proporcionar una comprensión significativa de los contenidos en los OER. La evaluación de la calidad y coherencia de los temas extraídos por el modelo LDA se realiza mediante métricas específicas que miden la relación semántica entre las palabras dentro de cada tema.

IV. Análisis y Discusión de Resultados

A. Conjunto de datos

El conjunto de datos se recopiló del repositorio Skills Commons, una plataforma que cuenta con una amplia variedad de Recursos Educativos Abiertos (OER). Este conjunto de datos incluye metadatos como: *título, url, type* (tipo de material principal), entre otros. Los OER utilizados en este trabajo corresponden al tema "*Tecnologías de la Información*" y los metadatos y recursos se encuentran en idioma inglés.

Para la construcción del corpus, se seleccionó el metadato *type* que significa "tipo de material principal" ya que permite identificar el tipo de recurso educativo abierto. En este estudio, se seleccionaron aquellos recursos que están etiquetados como "*Final Program Report*" en el metadato *type* (Ver Figura 2.). Este criterio proporcionó un total de 80 OER que son de interés para el propósito de este trabajo de investigación. Cada uno de estos recursos contiene una colección de 103 documentos relacionados con diversos aspectos de las *Tecnologías de la Información*.

```
dataFt= data.loc[data['type'] == 'Final Program Report']
dataOpc= dataFt.loc[:, ['title','url', 'type']]
dataOpc
```

id	title	url	type
13172	RITA Consortium Final Evaluation Report Septem...	https://www.skillscommons.org//handle/taaccct/...	Final Program Report
13748	Third party evaluation	https://www.skillscommons.org//handle/taaccct/...	Final Program Report
18125	Get IT Project Evaluation Final Report	https://www.skillscommons.org//handle/taaccct/...	Final Program Report
18467	Third Party Evaluations of the IT Programs	https://www.skillscommons.org//handle/taaccct/...	Final Program Report
18557	Final Evaluation Report: Health Information Te...	https://www.skillscommons.org//handle/taaccct/...	Final Program Report
...	...	...	...
15621	New River Community and Technical College's Fr...	https://www.skillscommons.org//handle/taaccct/...	Final Program Report
15677	Final Evaluation Report Developing Pathways fo...	https://www.skillscommons.org//handle/taaccct/...	Final Program Report
18509	New Jersey Health Professions Pathways to Regi...	https://www.skillscommons.org//handle/taaccct/...	Final Program Report
15562	FINAL EXTERNAL EVALUATION REPORT - Clovis Comm...	https://www.skillscommons.org//handle/taaccct/...	Final Program Report
9295	RCC TAACCCT Final Evaluation Report	https://www.skillscommons.org//handle/taaccct/...	Final Program Report

Fig. 2. Ejemplo de recursos educativos abiertos etiquetados como "Final Program Report"

B. Preparación de los datos

Se recopilaron los documentos relacionados con diversos aspectos de las *Tecnologías de la Información* para el análisis y construcción del modelo. Para la preparación y el preprocesamiento de datos se utilizó el lenguaje de programación Python, reconocido por su amplia gama de librerías especializadas en el procesamiento de lenguaje natural, mediante la biblioteca *nltk* [16]. A continuación se realizan: la implementación del algoritmo LDA, la visualización de datos y la evaluación del modelo [17].

Los pasos realizados en esta fase de preparación y preprocesamiento de datos se detallan en la Tabla II.

C. Representación de documentos (Tf-idf)

Una vez procesados los datos, se realizó el análisis de frecuencia de palabras considerando que es común encontrar palabras que aparecen en múltiples documentos, sin importar la colección de datos a la que pertenecen. La técnica Tf–idf es ampliamente utilizada para reducir el peso de las palabras que aparecen con frecuencia en los vectores de características [14]. Esta técnica funciona mediante la asignación de una ponderación a los términos según la frecuencia de aparición en los documentos, lo que ayuda a destacar las palabras clave que realmente aportan significado a los documentos.

Para aplicar el algoritmo LDA a un conjunto de documentos, primero LDA asume un número fijo de temas (temas), y las relaciones *tema-palabra* y *tema-documento* (*bag of words*) se modelan con matrices de probabilidades *palabra-en-tema* y *tema-en-documento* (Tf-idf).

D. Aplicación del algoritmo LDA

Se utilizó la librería *gensim* en Python con el propósito de construir y entrenar el modelo LDA. En este proceso, los parámetros "*alpha*" y "*eta*" se configuraron con valores automáticos. Estos parámetros afectan la densidad de distribución de temas de los documentos y la palabra perteneciente a un tema.

TABLA II. PREPARACIÓN Y PREPROCESAMIENTO DE DATOS

	Tarea	Descripción
Preparación de datos	Descarga de documentos por OER	Se descargaron los documentos que forman parte del OER y se almacenó en carpetas identificadas por el "*id*".
	Formateo de documentos	Se unificó el formato de los documentos, aquellos que estaban en formato PDF se convirtieron en formato DOCx.
	Transformación de documentos	Los metadatos de los 80 OER seleccionados y el contenido de los documentos de cada OER se almacenó en un archivo CSV para facilitar el manejo y tratamiento de los datos.
Preprocesamiento	Normalización	Para reducir el ruido se transforman las palabras a minúscula, se eliminan siglas, signos de puntuación, números y caracteres especiales.
	Tokenización	Se divide el texto en unidades más pequeñas llamadas tokens. Para tokenizar el texto se utiliza la función *nltk.word_tokenize( ).*
	Eliminación palabras vacías	Se eliminan palabras muy comunes, por ejemplo, "el", "la", "y", "de", que aparecen con frecuencia en el texto pero que no aportan un significado importante.
	Identificación de bigramas y trigramas	Se identifican y agrupan las palabras adyacentes en dos y tres elementos consecutivos, respectivamente, para facilitar la representación adecuada del texto.
	Lematización	Se reducen las palabras a su forma base o raíz, conocida como lema. Para la lematización se utilizó la librería WordNetLemmatizer del paquete *nltk* [16]. En este estudio, la lematización se realiza considerando que los tokens son verbos.

E. Validación y resultados

En esta sección, para evaluar el modelo generado por el algoritmo LDA, se aplica la medida de coherencia de temas (*c_v*) para analizar el desempeño de un conjunto de modelos de temas de diferentes tamaños y tipos. La puntuación de coherencia LDA es una métrica utilizada para evaluar la calidad de los grupos de palabras generados por el modelo LDA. La puntuación de coherencia más alta proporciona una medida cuantitativa que muestra el grado de relación entre las palabras y un tema.

En la figura 3, se presentan los resultados de los valores de coherencia obtenidos para diferentes números de temas (*k*) al

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2024
DOI: 10.5281/zenodo.10402399

aplicar el modelo de matriz Tf–idf con bigramas y trigramas en un análisis de temas. Los valores de *k* van desde 2 hasta 12 y se selecciona el valor más alto. Se puede observar que el valor más alto de coherencia (0.42) se obtiene cuando el valor óptimo de *k* = 9. Esto indica que el modelo de 9 temas tiene una coherencia de temas relativamente alta, lo que sugiere que las palabras dentro de los temas están bien relacionadas y representan patrones claros en el corpus.

Cabe recalcar que la medida de validación aplicada es coherencia *c_v,* la que es utilizada en el contexto de modelado de temas que permite medir la coherencia semántica entre palabras en un modelo LDA. Mientras que la métrica F1-score utilizada en [8] se aplica para la evaluación de modelos de clasificación.

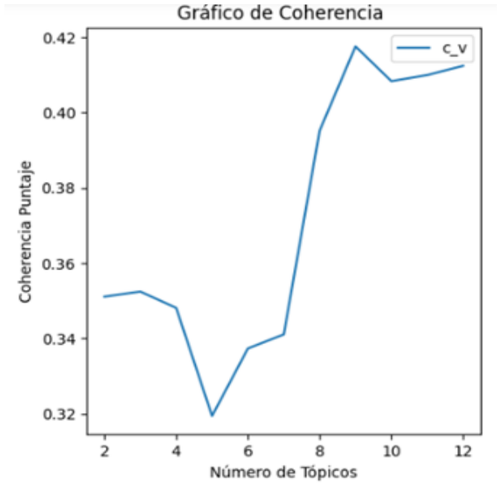


Fig. 3. Métrica de coherencia

Finalmente, para tener una visualización interactiva de los temas y palabras clave se utilizó la librería *pyLDAvis*. Esta librería crea el mapa de distancias entre temas y su distribución de temas para cada OER sobre *Tecnologías de la Información*, con el fin de ayudar en la interpretación de los temas en el modelado (Ver Figura 4).

En la figura 4, se presenta el panel derecho del mapa de distancia, se visualizan las 30 palabras más frecuentes de cada tema calculado con el modelo LDA. Cada tema se representa con una lista de palabras ponderadas por sus respectivos valores TF-IDF. Mientras que, en la interpretación de temas, se analizan las palabras más probables asociadas con cada tema para entender su significado y relevancia.

En la figura 5, se presenta el panel izquierdo del mapa de distancias donde se visualiza el tamaño de cada tema que representa la prevalencia sobre el conjunto de datos. Las burbujas más grandes indican que el tema es más prevalente en comparación con las burbujas más pequeñas.

Los temas 4 y 7 tienen una naturaleza independiente de otros temas. El tema 9 se integra con los temas 1 y 8, es decir, cuentan o comparten términos similares. Entre los términos más frecuentes están *"program", "student", "college", "participant" y "community"* que están relacionados con Educación en Tecnologías de la Información.

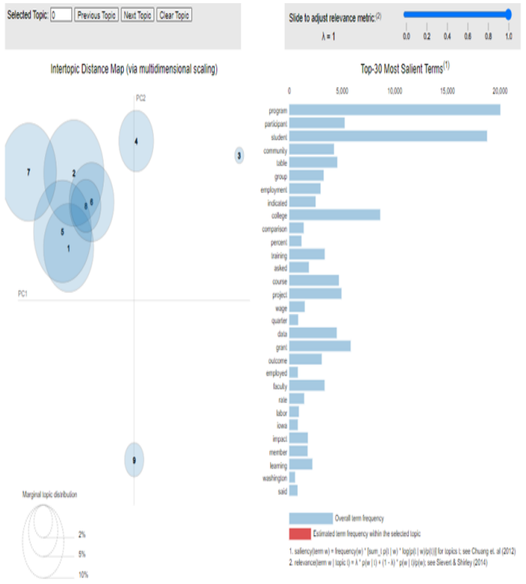


Fig. 4. Visualización de las 30 palabras más frecuentes

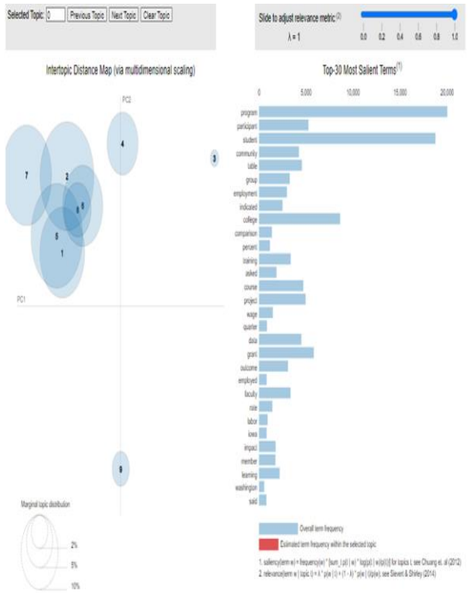


Fig. 5. Mapa de distancia de temas en OER sobre "Tecnologías de la información"

V. Conclusiones y Trabajo Futuro

El modelado de temas cumple un papel importante en la informática para la minería de textos. En este estudio, se aplicó el algoritmo LDA a un conjunto de documentos de Recursos Educativos Abiertos (OER) con el fin de identificar y analizar los temas presentes en toda la colección, así como en cada documento individual y las relaciones entre los documentos. Esta técnica es popular y relativamente simple de

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2024
DOI: 10.5281/zenodo.10402399

implementar, sin embargo, puede ser sensible a la elección de un número de temas específicos.

En este artículo, se presentan los resultados de los experimentos y análisis realizados de OER extraídos del repositorio Skill Commons mediante la aplicación del algoritmo LDA. Los resultados muestran que la mayoría de los temas abordados están relacionados con Educación en Tecnologías de la Información. Estos resultados proporcionan al usuario una forma de identificar los temas principales en los OER sin necesidad de realizar una revisión detallada de cada recurso.

Se propone en futuros experimentos incluir métodos de aprendizaje profundo que permitan incluir más factores para construir un modelo de predicción orientado a la comparación de metadatos extrayendo las características de los OER.

Agradecimiento

El equipo de investigadores agradece a la Universidad Técnica Particular de Loja, especialmente al Grupo de Investigación de Sistemas Basados en Conocimiento (KBS-RG); y, a la SENESCYT de Ecuador.

Referencias

[1] V. Segarra-Faggioni and A. Romero-Pelaez, "Automatic classification of OER for metadata quality assessment," in 2022 International Conference on Advanced Learning Technologies (ICALT), 2022, pp. 16–18. doi: 10.1109/ICALT55010.2022.00011.

[2] X. Ochoa and E. Duval, "Automatic evaluation of metadata quality in digital repositories," Int. J. Digit. Libr., vol. 10, no. 2–3, pp. 67–91, Aug. 2009, doi: 10.1007/s00799-009-0054-4.

[3] J. Chicaiza, N. Piedra, J. Lopez-Vargas, and E. Tovar-Caro, "Recommendation of open educational resources. An approach based on linked open data," null, 2017, doi: 10.1109/educon.2017.7943018.

[4] H. Jelodar et al., "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," Multimed. Tools Appl., vol. 78, no. 11, pp. 15169–15211, Jun. 2019, doi: 10.1007/S11042-018-6894-4/TABLES/11.

[5] M. Tavakoli, M. Elias, G. Kismihók, and S. Auer, "Metadata analysis of open educational resources," in ACM International Conference Proceeding Series, 2021, pp. 626–631. doi: 10.1145/3448139.3448208.

[6] M. Kim and D. Kim, "A Suggestion on the LDA-Based Topic Modeling Technique Based on ElasticSearch for Indexing Academic Research Results," Appl. Sci., vol. 12, no. 6, p. 3118, Mar. 2022, doi: 10.3390/app12063118.

[7] S. Ozdemirci and M. Turan, "Case Study on well-known Topic Modeling Methods for Document Classification," in 2021 6th International Conference on Inventive Computation Technologies (ICICT), Jan. 2021, pp. 1304–1309. doi: 10.1109/ICICT50816.2021.9358473.

[8] M. Molavi, M. Tavakoli, and G. Kismihók, "Extracting Topics from Open Educational Resources," in Addressing Global Challenges and Quality Education, 2020, pp. 455–460.

[9] R. Wirth and J. Hipp, "Crisp-dm: towards a standard process modell for data mining," 2000. [Online]. Available: https://api.semanticscholar.org/CorpusID:1211505

[10] P. Haya, "La metodología CRISP-DM en ciencia de datos," INSTITUTO DE INGENIERÍA DEL CONOCIMIENTO, 2021. https://www.iic.uam.es/innovacion/metodologia-crisp-dm-ciencia-de-datos/

[11] M. Tavakoli, M. Elias, G. Kismihok, and S. Auer, "Quality Prediction of Open Educational Resources A Metadata-based Approach," in 2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT), Jul. 2020, pp. 29–31. doi: 10.1109/ICALT49669.2020.00007.

[12] D. M. Blei, A. Y. Ng, and M. Jordan, "Latent Dirichlet Allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003, Accessed: Mar. 25, 2019. [Online]. Available: http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[13] H. Lane, C. Howard, and H. Max Hapke, Natural Language Processing in Action: Understanding, Analyzing, and Generating Text with Python. Manning Publications., 2019.

[14] V. Mirjalili and S. Raschka, Python Machine Learning. Marcombo, 2019.

[15] S. Weiss, N. Indurkhya, and T. Zhang, Fundamentals of Predictive Text Mining, Second Edi. 2015. doi: 10.1007/978-1-4471-6750-1.

[16] S. Bird and E. Loper, "NLTK: The Natural Language Toolkit," 2006. Accessed: Oct. 14, 2018. [Online]. Available: www.python.org.

[17] S. Raschka, J. Patterson, and C. Nolet, "Machine Learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence," CoRR, vol. abs/2002.0, 2020, [Online]. Available: https://arxiv.org/abs/2002.04803

LATIN-AMERICAN JOURNAL OF COMPUTING (LAJC), Vol XI, Issue 1, January 2024

AUTHORS

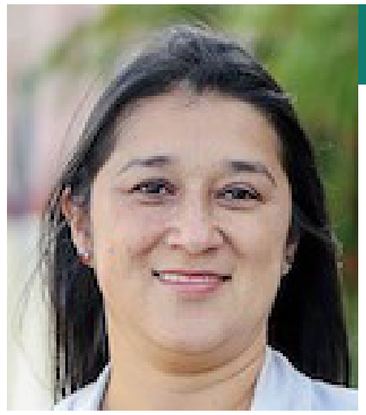
René Ludeña

Estudiante de la carrera de Sistemas informáticos y Computación, Universidad Técnica de Loja.
Participó en el proyecto "Sistema de automatización de riesgos operativos “SARO” del Banco de Loja" donde se implementaron diferentes tecnologías para optimizar procesos, mejorar la eficiencia y prevenir los riesgos, 2021.

Verónica Segarra

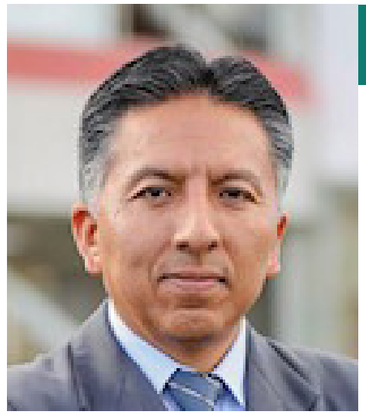
Ph.D en Ingeniería por la École de technologie supérieure, Montreal -Canadá, 2021. La tesis de doctorado fue realizada bajo la dirección de la Prof. Sylvie Ratté, Ph.D. de la ÉTS, Móntreal, Canadá y codirección del Prof. Frank de Jong de la Universidad AERES, Wageningen, Países Bajos.
Participó en un Hackaton en Paises Bajos donde se evaluó los informes escritos de los alumnos de maestría de la Universidad AERES, 2018.
Realizó pasantías de investigación en CRIM (Centre de Recherche Informatique de Montréal) como parte del equipo “Parole et Texte” para análisis de texto utilizando técnicas de procesamiento de lenguaje natural, 2019.
Magíster en Auditoría de Gestión de la Calidad, Universidad Técnica Particular de Loja –Ecuador, 2008.
Diploma en Mejora de Procesos, Tecnológico de Monterrey, 2012.
Ingeniera en Sistemas Informáticos y Computación, Universidad Técnica Particular de Loja - Ecuador.
Licenciada en Ciencias de la Educación Mención Inglés, Universidad Técnica Particular de Loja - Ecuador.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2024



Audrey Romero

Profesor titular a tiempo completo del Departamento de Ciencias de la Computación y Electrónica en la Universidad Técnica de Loja.
Máster en Ciencias y Tecnologías de la Computación por la Universidad Politécnica de Madrid.
Sus campos de investigación son: Educación abierta, Modelos de Calidad, Tecnologías Semánticas, Análisis de Datos y Tecnologías Emergentes.



Juan Carlos Morocho

Profesor titular a tiempo completo del Departamento de Ciencias de la Computación y Electrónica en la Universidad Técnica de Loja.
Graduado de la Maestría de Inteligencia Artificial en la Universidad Politécnica de Madrid.
Sus campos de investigación son: Diseño Ontológico, Tecnologías Semánticas y Análisis de Datos.