UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO

ESCUELA DE POSGRADO

MAESTRÍA EN CIENCIAS MENCIÓN INFORMATICA

TESIS

MODELO DE EXTRACCIÓN AUTOMÁTICA DE GLOSARIO DE TÉRMINOS UTILIZANDO TÉCNICAS DE PROCESAMIENTO DE LENGUAJE NATURAL Y CLUSTERING

PARA OPTAR AL GRADO ACADÉMICO DE MAESTRO EN CIENCIAS MENCION INFORMATICA

AUTOR:

Br. GABRIELA ZUÑIGA ROJAS

ASESOR:

Mgt. HARLEY VERA OLIVERA

CODIGO ORCID:

0000-0003-2011-8797

CUSCO-PERÚ

2024

INFORME DE ORIGINALIDAD

(Aprobado por Resolución Nro.CU-303-2020-UNSAAC)

El que suscribe, asesor del trabajo de investigación/tesis titulado: Modelo de extracción automática de glosario de término utilizando técnicas de procesamiento de lenguaje natural y Clustering

presentado por: Gabriela Zuñiga Rojas

con Nro. de DNI: 45073823, para optar el título profesional/grado académico de Maestro en Ciencias Mención Informática

Informo que el trabajo de investigación ha sido sometido a revisión por 1 veces, mediante el Software Antiplagio, conforme al Art. 6° del *Reglamento para Uso de Sistema Antiplagio de la UNSAAC* y de la evaluación de originalidad se tiene un porcentaje de 9%

Evaluación y acciones del reporte de coincidencia para trabajos de investigación conducentes a grado académico o título profesional, tesis

Porcentaje	Evaluación y Acciones	Marque con una (X)
Del 1 al 10%	No se considera plagio.	X
Del 11 al 30 %	Devolver al usuario para las correcciones.	
Mayor a 31%	El responsable de la revisión del documento emite un informe al inmediato jerárquico, quien a su vez eleva el informe a la autoridad académica para que tome las acciones correspondientes. Sin perjuicio de las sanciones administrativas que correspondan de acuerdo a Ley.	

Por tanto, en mi condición de asesor, firmo el presente informe en señal de conformidad y **adjunto** la primera hoja del reporte del Sistema Antiplagio.

Cusco, 05 de diciembre de 2024

Firma

Post firma Harley Vera Olivera

Nro. de DNI 42542815

ORCID del Asesor 0000-0003-2011-8797

Se adjunta:

1. Reporte generado por el Sistema Antiplagio.
2. Enlace del Reporte Generado por el Sistema Antiplagio: oid::27259:412193963

Gabriela Zuñiga Rojas

MODELO DE EXTRACCIÓN AUTOMÁTICA DE GLOSARIO DE TÉRMINOS UTILIZANDO TECNICAS DE PROCESAMIENTO DE ...

Universidad Nacional San Antonio Abad del Cusco

Detalles del documento

Identificador de la entrega
trn:oid:::27259:412193963

Fecha de entrega
3 dic 2024, 10:44 a.m. GMT-5

Fecha de descarga
3 dic 2024, 11:00 a.m. GMT-5

Nombre de archivo
MODELO DE EXTRACCIÓN AUTOMÁTICA DE GLOSARIO DE TÉRMINOS UTILIZANDO TECNICAS DE P....pdf

Tamaño de archivo
2.4 MB

81 Páginas

20,150 Palabras

112,383 Caracteres

9% Similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para ca...

Filtrado desde el informe

- Bibliografía
- Coincidencias menores (menos de 8 palabras)

Exclusiones

- N.º de coincidencias excluidas

Fuentes principales

7% Fuentes de Internet
1% Publicaciones
6% Trabajos entregados (trabajos del estudiante)

Marcas de integridad

N.º de alerta de integridad para revisión

Texto oculto
284 caracteres sospechosos en N.º de páginas
El texto es alterado para mezclarse con el fondo blanco del documento.

Los algoritmos de nuestro sistema analizan un documento en profundidad para buscar inconsistencias que permitirían distinguirlo de una entrega normal. Si advertimos algo extraño, lo marcamos como una alerta para que pueda revisarlo.

Una marca de alerta no es necesariamente un indicador de problemas. Sin embargo, recomendamos que preste atención y la revise.

Resumen

Para abordar la complejidad y esfuerzo manual que representa la extracción de términos para glosarios a partir de requisitos funcionales en proyectos de desarrollo de software a gran escala, proponemos un enfoque automatizado para la extracción y agrupamiento de términos de glosario. El método combina técnicas de pre-procesamiento y heurísticas para la identificación de términos, junto con embeddings generados con FastText para medir similitudes semánticas. Para el agrupamiento se emplearon los algoritmos *K-means, Expectation Maximization (EM) y Clusterización Jerárquica*. La técnica fue aplicada a un conjunto de 2966 requisitos obteniéndose 318 grupos semánticos, y su eficacia fue evaluada mediante la distancia de Wasserstein (Word Mover's Distance) de 0.0113, el cual comparando los resultados automáticos con agrupamientos manuales es menor. Los experimentos mostraron que el uso de FastText y EM logra una agrupación semántica efectiva y consistente, validando la aplicabilidad del enfoque en entornos reales de desarrollo de software.

Palabras clave: Glosario, Procesamiento de Lenguaje Natural, Documentación de software, Clustering, Extracción Automática de términos

Abstract

To address the complexity and manual effort involved in extracting glossary terms from user requirements in large-scale projects, we propose an automated approach for glossary term extraction and clustering. The method combines preprocessing techniques and heuristics for term identification, along with embeddings generated using FastText to measure semantic similarities. For clustering, the algorithms *K*-means, Expectation Maximization (EM), and Hierarchical Clustering were employed. The technique was applied to a dataset of 2,966 requirements, and its effectiveness was evaluated using the Wasserstein Distance (Word Mover's Distance), comparing the automated results with manual clusterings. The experiments showed that using FastText and EM achieves effective and consistent semantic clustering, validating the applicability of the approach in real-world software development environments.

Keywords: Glossary, Natural Language Processing, Software Documentation, Clustering, Term Extraction