ANEXO 1

FORMATO PARA LA PRESENTACIÓN DE PROYECTOS DE INVESTIGACIÓN CON EL FINANCIAMIENTO DEL FEDU

1. Título del proyecto

Modelo de aprendizaje automático para el análisis semántico de textos en lengua castellana

2. Área de Investigación

Área de investigación	Línea de Investigación	Disciplina OCDE
INGENIERÍA DE SISTEMAS	SISTEMAS, COMPUTACIÓN E INFORMÁTICA	CIENCIA DE LA COMPUTACIÓN

3. Duración del proyecto (meses)

12

4. Tipo de proyecto

Individual	●
Multidisciplinario	○
Director de tesis pregrado	○

4. Datos de los integrantes del proyecto

Apellidos y Nombres	**SOSA MAYDANA CARLOS BORIS**
Escuela Profesional	**INGENIERIA DE SISTEMAS**
Celular	**986739074**
Correo Electrónico	**cbsosa@unap.edu.pe**

I. Título (El proyecto de tesis debe llevar un título que exprese en forma sintética su contenido, haciendo referencia en lo posible, al resultado final que se pretende lograr. Máx. palabras 25)

Modelo de aprendizaje automático para el análisis semántico de textos en lengua castellana

II. Resumen del Proyecto de Tesis (Debe ser suficientemente informativo, presentando -igual que un trabajo científico- una descripción de los principales puntos que se abordarán, objetivos, metodología y resultados que se esperan**)**

El presente proyecto nace con la idea de hacer uso del procesamiento de lenguaje natural, el mismo que esta dentro del área de inteligencia artificial para efectuar el análisis semántico de textos de diferentes formatos electrónicos (PDF, TXT, DOC, etc.) de forma automática, haciendo uso de herramientas y técnicas actuales para el caso; contextualizando el mismo para usuarios de habla hispana. El procesamiento del lenguaje natural funciona a través del aprendizaje automático (ML o machine learning). Los sistemas de aprendizaje automático

almacenan las palabras y las formas en que se unen como cualquier otra forma de datos. Frases, oraciones y a veces libros enteros se introducen en los motores de ML donde se procesan en base a reglas gramaticales, los hábitos lingüísticos de la vida real de la gente, o ambos.
La traducción automática es una de las mejores aplicaciones de PLN, pero no la más utilizada: lo es la búsqueda. Cada vez que se busca algo en Google o Bing, se está alimentando el sistema con datos. Cuando hacemos clic en un resultado de búsqueda, el sistema lo ve como una confirmación de que los resultados que ha encontrado son correctos y utiliza esta información para mejorar la búsqueda en el futuro; pero todo ello queda en una búsqueda por lo que se hace necesario generar un modelo que permita ir más allá como el que planteamos.

III. Palabras claves (Keywords) (Colocadas en orden de importancia. Máx. palabras: cinco**)**

Aprendizaje automático, procesamiento de lenguaje natural, inteligencia artificial, inteligencia computacional, análisis semántico, procesamiento de textos

IV. Justificación del proyecto (Describa el problema y su relevancia como objeto de investigación. Es importante una clara definición y delimitación del problema que abordará la investigación, ya que temas cuya definición es difusa o amplísima son difíciles de evaluar y desarrollar**)**

La importancia del lenguaje natural en la creación de contenidos y su redacción implica utilizar palabras claves de la misma manera y con el mismo sentimiento que la audiencia o los seres humanos de habla castellana deseamos. Conseguir este objetivo implica de tareas y acciones muy complejas sobre todo cuando se quiere realizar el mismo de forma automática (utilizando computadores). Al conseguir ello podremos optimizar el tiempo de procesamiento para el mismo y también pragmatizar un enfoque imparcial de análisis o interpretación del contexto.

V. Antecedentes del proyecto (Incluya el estado actual del conocimiento en el ámbito nacional e internacional. La revisión bibliográfica debe incluir en lo posible artículos científicos actuales, para evidenciar el conocimiento existente y el aporte de la Tesis propuesta. Esto es importante para el futuro artículo que resultará como producto de este trabajo)

PROCESAMIENTO DE LENGUAJE NATURAL
Según. Gelbukh (2010) menciona que el procesamiento de lenguaje natural o también NLP es la habilidad que una maquina tiene de procesar información comunicada, es decir , no solamente letras sino que también sonidos del lenguaje y también menciona que la ciencia que estudia el procesamiento de lenguaje natural se denomina lingüística computacional en donde lingüistas definen reglas y diccionarios que cada vez son más exactos y detallados para dotar a las computadoras con la capacidad de poder entender el lenguaje humano.
Cortez Vásquez, A. et al. (2009) consideran que el NLP consiste en utilizar el lenguaje natural para poder comunicarnos con la computadora, debiendo ésta entender oraciones o textos que le sean proporcionadas, para poder comprender mecanismos humanos que están relacionados con el lenguaje. También menciona algunas de las aplicaciones del procesamiento de lenguaje natural, como la traducción automática, recuperación de la información, extracción de la información,

resúmenes, entre otros. como por ejemplo el lenguaje sesgado.

LENGUAJE SESGADO

Según Alonso. (2020) haciendo referencia a la American Psychological Association («Formato APA 7.a edición», 2020) menciona que se reconoce la importancia de la utilización de un lenguaje más preciso e inclusivo para que se siga generando prejuicios y sesgos contra diferentes tipos y grupos de personas, también menciona que APA hace énfasis en la necesidad de hablar sobre todas las personas con inclusión y respeto, así usar un lenguaje libre de prejuicios, las normas para un lenguaje libre de perjuicios contienen temas que abarcan características de individuales de edad, género, nivel socioeconómico, discapacidad, entre otros.

PYTHON - SPACY

Spacy(SpaCy · Industrial-Strength Natural Language Processing in Python, s. f.) es un api escrito desde cero en Cython(Behnel et al., 2011) y su aplicación es para procesar vocablos, actualmente SPACY se ha convertido en un estándar de la industria con un enorme ecosistema. Y cuenta con soporte a más de 67 idiomas incluyendo el español. Sus funcionalidades básicas son la tokenización, clasificación de texto, lematización, entre otros.

VI. Hipótesis del trabajo (Es el aporte proyectado de la investigación en la solución del problema)

Un modelo de aprendizaje automático permitirá realizar un óptimo análisis semántico de textos en lengua castellana

VII. Objetivo general

Desarrollar un modelo de aprendizaje automático para el análisis semántico de textos en lengua castellana

VIII. Objetivos específicos

- Identificar los principales componentes para un análisis semántico de textos
- Aplicar técnicas y herramientas actuales para generar un modelo automatizado de análisis de textos
- Probar y Validar el modelo generado

IX. Metodología de investigación (Describir el(los) método(s) científico(s) que se empleará(n) para alcanzar los objetivos específicos, en forma coherente a la hipótesis de la investigación. Sustentar, con base bibliográfica, la pertinencia del(los) método(s) en términos de la representatividad de la muestra y de los resultados que se esperan alcanzar. Incluir los análisis estadísticos a utilizar**)**

Aplicaremos el método de investigación científico con enfoque cualitativo. En la construcción del modelo utilizaremos métodos híbridos que incluyan: Redes neuronales, procesamiento de lenguaje natural y algoritmos de aprendizaje automático.

X. Referencias (Listar las citas bibliográficas con el estilo adecuado a su especialidad)

Behnel, S., Bradshaw, R., Citro, C., Dalcin, L., Seljebotn, D. S., & Smith, K. (2011). Cython: The Best of Both Worlds. *Computing in Science & Engineering*, *13*(2), 31-39. https://doi.org/10.1109/MCSE.2010.118

Cortez Vásquez, A., Pariona Quispe, J., Vega huerta, H., & Huayna, A. M. (2009). *Procesamiento de lenguaje natural*. *6*(2). https://revistasinvestigacion.unmsm.edu.pe/index.php/sistem/article/view/5923

de Marneffe, M.-C., Manning, C. D., & Potts, C. (2012). Did It Happen? The Pragmatic Complexity of Veridicality Assessment. *Computational Linguistics*, *38*(2), 301-333. https://doi.org/10.1162/COLI_a_00097

Formato APA 7.ª edición: Lenguaje libre de sesgos. (2020, septiembre 3). *Psyciencia*. https://www.psyciencia.com/normas-apa-7-edicion-lenguaje-libre-de-sesgos/

IVAN. (2019, diciembre 29). ¿Sesgos en investigación?…¿y eso qué es? *Audeo Dicere*. https://audeodicereblog.wordpress.com/2019/12/29/sesgos-en-investigacion/

Primeros-resultados-encuestra-discriminacion.pdf. (s. f.). Recuperado 12 de enero de 2023, de https://centroderecursos.cultura.pe/sites/default/files/rb/pdf/primeros-resultados-encuestra-discriminacion.pdf

Procesamiento de lenguaje natural y sus aplicaciones.pdf. (s. f.). Recuperado 12 de enero de 2023, de https://www.gelbukh.com/CV/Publications/2010/Procesamiento%20de%20lenguaje%20natural%20y%20sus%20aplicaciones.pdf

SpaCy · Industrial-strength Natural Language Processing in Python. (s. f.). Recuperado 12 de enero de 2023, de https://spacy.io/

Wikipedia:Portada. (2020). En *Wikipedia, the free encyclopedia*. https://es.wikipedia.org/w/index.php?title=Wikipedia:Portada&oldid=123425818

XI. Uso de los resultados y contribuciones del proyecto (Señalar el posible uso de los resultados y la contribución de los mismos)

Los usos serán fundamentalmente de carácter académico

XII. Impactos esperados

i. Impactos en Ciencia y Tecnología

Nuestros resultados contribuirán a la investigación de cómo se comunican las máquinas con las personas mediante el uso de lenguas naturales, como el español y si estos mejoran su nivel de análisis contextual.

ii. Impactos económicos

iii. Impactos sociales

Interacción humano computador

iv. Impactos ambientales

XIII. Recursos necesarios (Infraestructura, equipos y principales tecnologías en uso relacionadas con la temática del proyecto, señale medios y recursos para realizar el proyecto**)**

Computadoras, internet, bibliografía actualizada, licencias para software específico

XIV. Localización del proyecto (indicar donde se llevará a cabo el proyecto)

Universidad Nacional del Altiplano - Puno

XV. Cronograma de actividades

Actividad	Trimestres											
Analizar y seleccionar métodos de PLN	■	■	■									
Seleccionar los casos para ser analizados	■	■	■	■								
Desarrollar el modelo automático de procesamiento automático de imágenes y probarlo				■	■	■	■	■	■			
Validar los resultados								■	■			
Redacción del documento final											■	
Redacción del articulo												■

XVI. Presupuesto

Descripción	Unidad de medida	Costo Unitario (S/.)	Cantidad	Costo total (S/.)
1. Personal	Persona	10000	01	10000
2. Materiales y equipos	Equipo	20000	01	20000
3. Servicios	Servicio	1000	03	3000
4. Imprevistos	Otros	3000	01	3000
TOTAL				36000