# Integración de Técnicas de Aprendizaje de Maquina Hacia la Síntesis Coherente de Noticias en Castellano en un Contexto Específico



CURSO: TESIS II



# MATRIZ DE CONSISTENCIA

202013427

Preguntas	Objetvios	Hipótesis	Variable Independiente	Variable Dependiente	
¿Afecta de manera positivia la integración de técnicas de aprendizaje de maquina al grado de coherencia de la síntesis de noticias en castellano de contexto específico?	Desarrollar un prototipo con base en integración de técnicas de aprendizaje de máquina capaz de sintetizar de manera coherente noticias en castellano de contexto específico.	La integración de técnicas de Aprendizaje de Máquina afecta positivamente al grado de coherencia obtenido en la síntesis de noticias en castellano de contexto específico.	Integración de Técnicas de Aprendizaje de Maquina	Síntesis Coherente de Noticias en Castellano de contexto específico	
¿No es factible el desarrollo de un prototipo de resumen automático en el idioma español en la actualidad?	Evidenciar el grado de factibilidad que acarrea el desarrolo de un prototipo de resumen automático en el idioma español	En la actualidad no es factible el desarrollo de un prototipo de resumen automático en el idioma español	DIMENSION		
			INoticias en español en contexto con el COVID (2021 - 2022)	Puntuación recolectada de voluntarios	
¿ La integración de técnias de resumen automático reflejará una mayor coherencia a criterio humano ?	Contrastar con criterio humano el efecto de integración de técnicas sobre el resumen de textos en español.	La integración de técnicas de resumen automático tiene el efecto de reflejar mayor coherencia a criterio humano.	MÉTRICA		
			ROUGE-C	Grado de coherencia acorde a puntuación	



Extracción web de noticias disponibilizadas al público

FASE 1 Obtención de noticias

Almacenamiento, limpieza y estructuración del texto

FASE 2

Preprocesam iento



# **FLUJO DE DATOS**

FASE 3 PRE-Experimento Evaluación de disitntas técncas de Sinteis de textos por separado

FASE 4

POST-**EXPERIMENTO** 

Muestra de Resultados tras la integración de distintas técnicas de síntesis de texto.

Respuesta a la hipótesis general acorde a los resultados obtenidos

# FASE 1 - OBTENCIÓN DE NOTICIAS (STATUS 85%)

Extracción web de noticias disponibilizadas al público

SUBTAREAS	MAYO Semana 1	MAYO Semana 2	MAYO Semana 3	
Test de extracción de noticias	00			
Implementación de extractor web de diario formal	0			
Adaptación de la implementación a orientado a objetos	0			
Almacenamiento de noticias extraidas acorde al diseño de solución de ejecución diariia		0		
Delimitación de las noticias guardadas acorde al contexto enmarcado		0	-0	
Analisis de ISSUE 1: Extracción de contendio vacio de noticia		<u> </u>	-	
Investigación de técnicas de resumenes de texto implementadas en el lenguaje de progrmación Python				



https://elcomercio.pe/

Selección de Diario formal

1er nivel delimitación a contexto COVID

https://elcomercio.pe/coronavirus/

#### **CORONAVIRUS**

2021-6-12

Coronavirus Perú: cifras actualizadas v última hora de este 12 de junio del 2021

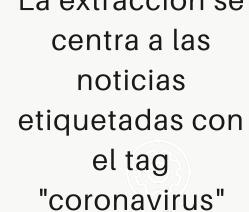
Últimas noticias del 12 de junio, cifras oficiales del Minsa y datos sobre el avance de la pandemia en el país

REDACCIÓN EC

2021-6-12

Uruguay registra 2.999 casos nuevos y 50 muertes por coronavirus en un día





La extracción se

TI CUMETTIO Para acceder a este contenido exclusivo, adquiere tu Plan Digital El Comercio Acceso sin límites a información exclusiva: reportajes, informes y la mejor selección de historias elaboradas por El Comercio. Navegación ilimitada a elcomercio.pe desde todos tus dispositivos. Descuentos especiales del Club de beneficios

El proceso no vulnera la extracción de contenido exclusivo respetando las normativas del sitio web del diario

Detección y habilitación de extracción de contenido público

## **EXTRACCIÓN WEB DE NOTICIAS DISPONIBILIZADAS AL PÚBLICO**

2do nivel delimitación a contexto COVID NACIONAL desccriptivo

Se descartan las noticias con tags: "mundo", "tvmas" y "videos", por carecer de contenido relevante al acontecer nacional

Bolivia recibe una nueva dotación de 100.000 vacunas rusas Sputnik V



Diego Llorente agradece el apoyo en su peor momento





Bolsonaro multado con US\$100 por incumplir medidas contra el Covid-19

Periodo de extracción diaria limitado al 1er nivel del paginado del listado de noticias por tag

Anterior 1 2 3 ... 693 Siguiente >

Límite de busqueda



Aprox. se tiene acceso a 50 noticias por día, de las cuales aprox. 15 cumplen con la delimitación deseada.

Límite de extracción

Estrategia de ETL

(extracción /

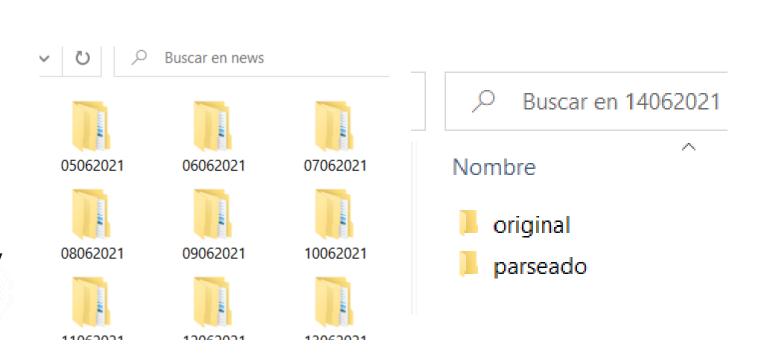
transformación y

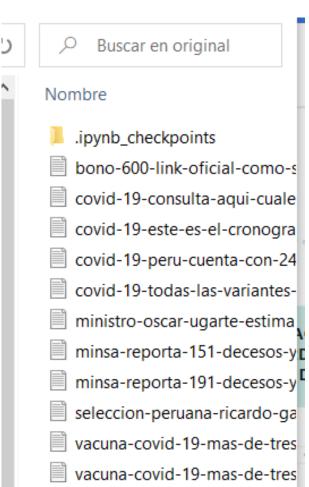
cconsumo de información )

## ALMACENAMIENTO DE NOTICIAS ACORDE AL DISEÑO DE SOLUCIÓN DE EJECUCIÓN DIARIA



- 1.- Ejecución diaria matutina
- 2.- Almacenamiento diferenciado por día y por tratamiento de texto
- 3.- Primer nivel de guardado y consulta sobre un archivo de texto por noticia acorde su tratameinto





# FASE 2: PRE-PROCESAMIENTO (STATUS 35%)

#### Almacenamiento

SUBTAREAS	MAYO Semana 4	JUNIO Semana 1	JUNIO Semana 2	JUNIO Semana 3
Análisis de ISSUE 2: Error en guardado de noticias nuevas	00			
Implementación de limpieza de textos	0	•		
Implementación de Traducción de español a ingles		<b>O</b>		
Test de flujo con implememtaciones de resumen de texto automátco disponibilizadas		0-	•	
Busqueda de implementaciones de métricas rouge-c y jensen-shanon			0-0	
Analisis de ISSUE 3: Incompatibilidad de verisón de librerias de implemnetaciones existentes			<b>O</b>	
Analisis de ISSUE 4: Perdida de mensaje por parseado excesivo de caracteres				
Test Final del flujo de extracción y resumen en modo MBP				

El 90% de publicaciones encontradas referentes a resumen automático se encontraban adaptadas a una solución solo para textos en ingles

Busqueda de implementacio nes en el idioma español

1.- Traducción de texto español a ingles sólo con el fin de test de flujo de datos.

2.- Testeo con implementaciones con autocomplementado de conectores gramaticales en ingles

Alternativas planteadas



Se opta por alternativa 2

Ejecución de TEST

# TEST DE FLUJO DE DATOS CON IMPLEMENTACIONES DE RESUMEN AUTOMÁTICO DISPONIBILIZADAS

OJO: En el test aún no se opta por la versión parseada debido al último issue reportado

El Ministerio de Salud (Minsa) informó este sábado 12 de junio que se elevó a 188.443 la cifra de decesos por coronavirus (COVID-19) en el país. Se trata de 191 nuevos fallecidos frente al reporte de la víspera. Además, informó que se incrementó en 3.003 los contagios (1.772 en las últimas 24 horas), por lo que el número total de personas infectadas llega a 2.001.059. El Minsa reportó también que hay un total de 10.711 pacientes hospitalizados, de los cuales 2.548 están con ventilación mecánica. Para mitigar los efectos de la pandemia, Perú suscribió acuerdos de adquisición de vacunas contra el COVID-19 con Sinopharm (por 2 millones de dosis), Pfizer (32 millones), AstraZeneca (14.04 millones) y con el mecanismo Covax Facility (13.2 millones). La inmunización en el Perú comenzó en febrero, tras el arribo del primer lote de vacunas procedentes de China (Sinopharm). El Gobierno ha anunciado que espera inmunizar a todos los adultos mayores de 60 años antes de que termine la actual gestión. El lunes 31 de mayo, el Gobierno informó que adoptarán los criterios y recomendaciones hechas por el Grupo de Trabajo Técnico (GTT) conformado para establecer la forma en la que se debe llevar el registro de las defunciones por esta enfermedad. Esta comisión determinó que, el número de personas fallecidas por COVID-19 en el Perú del 1 de marzo del 2020 al 22 de mayo de 2021, era de 180.764, mientras que el reporte oficial del Minsa no llegaba a las 70 mil. Desde entonces, el Minsa actualiza sus cifras en función a las recomendaciones del GTT



El Ministerio de Salud (Minsa) informó a 188.443 la cifra de decesos por coronavirus (COVID-19) Se trata de 191 nuevos fallecidos frente al reporte de la víspera. El Minsa reportó también that hay un total of 10.711 pacientes hospitalizados, of los cuales 2.548 están with ventilación mecánica

# **SIGUIENTES PASOS**

### implementaciones

Seguir indagando en la documentación de soluciones disponibilizadas

### Investigar

Ampliar la investigación teórica sobre resumen de textos



#### **ISSUES**

Corregir los errores que se fueron suscitando

# Hipótesis específica

Acorde a lo investigado responder la 1er hipótesis específica

# **Experimentar**

Dar inicio a los experimentos acordes a la respeusat de la 2da hipótesis

# **REPOSITORIO**

https://github.com/kendalvictor/tesis\_maestria\_ciencia\_datos

kendalvictor/ tesis\_maestria\_ciencia\_d...





#### kendalvictor/tesis\_maestria\_ciencia\_datos

Contribute to kendalvictor/tesis\_maestria\_ciencia\_datos development by creating an account on GitHub.

GitHub

# GRACIAS