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

VÍCTOR MARIANO VILLACORTA PLASENCIA

CURSO: TESIS II

MATRIZ DE CONSISTENCIA

202013427

	Preguntas	Objetvios	Hipótesis	*Variable Independiente*	*Variable Dependiente*
General	¿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
Específicas	¿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 JENSEN SHANON	Grado de coherencia acorde a puntuación experta

FLUJO DE DATOS

FASE 1

Obtención de noticias

Extracción web de noticias disponibilizadas al público

FASE 2

Preprocesamiento

Almacenamiento , limpieza y estructuración del texto

FASE 3

PRE-Experimento

Evaluación de disitntas técncas de Sinteis de textos por separado

FASE 4

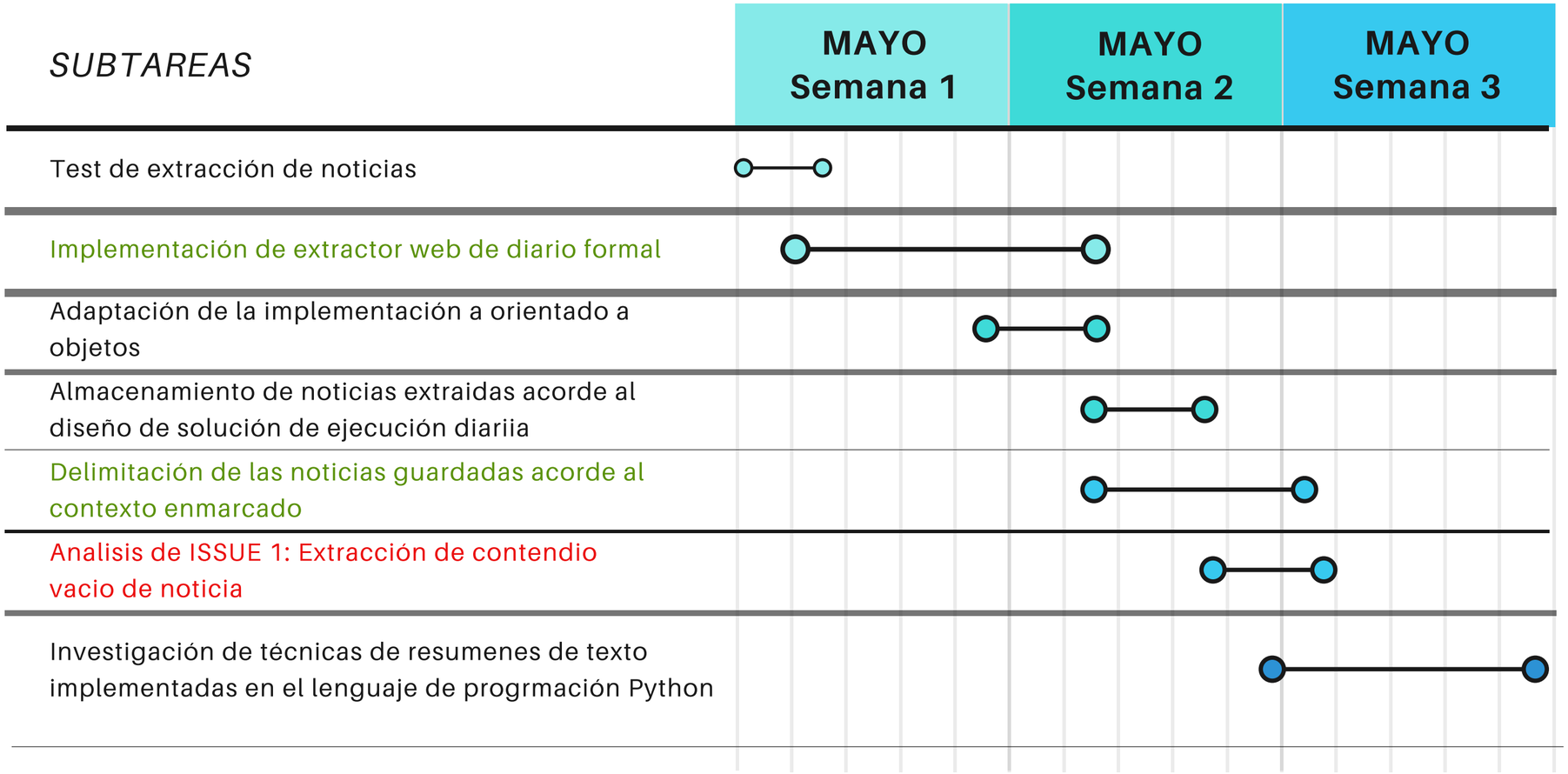
POST-EXPERIMENTO

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



EXTRACCIÓN WEB DE NOTICIAS DISPONIBILIZADAS AL PÚBLICO

El Comercio

https://elcomercio.pe/

Selección de Diario formal

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

1er nivel delimitación a contexto COVID

https://elcomercio.pe/coronavirus/

La extracción se centra a las noticias etiquetadas con el tag "coronavirus"

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

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

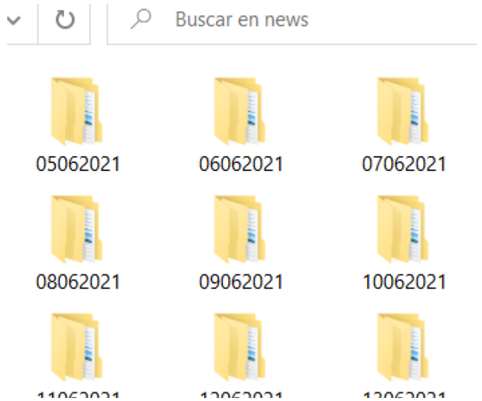
Estrategia de ETL (extracción / transformación y cconsumo de información)

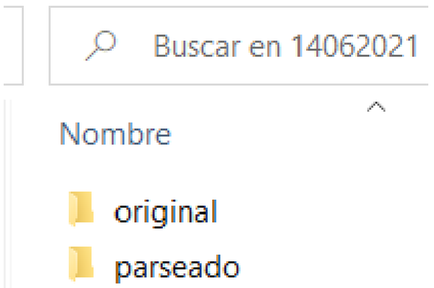
ALMACENAMIENTO DE NOTICIAS ACORDE AL DISEÑO DE SOLUCIÓN DE EJECUCIÓN DIARIA

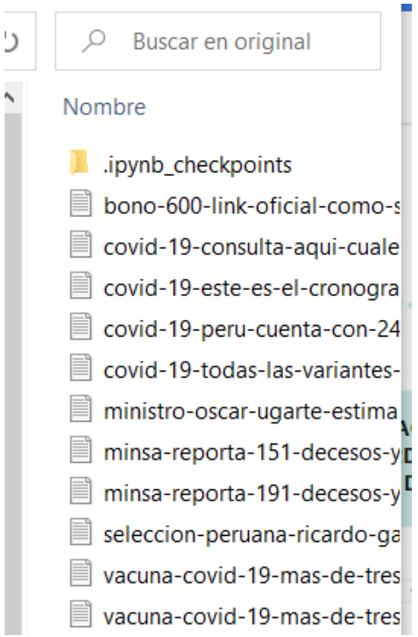
1.- Ejecución diaria matutina

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

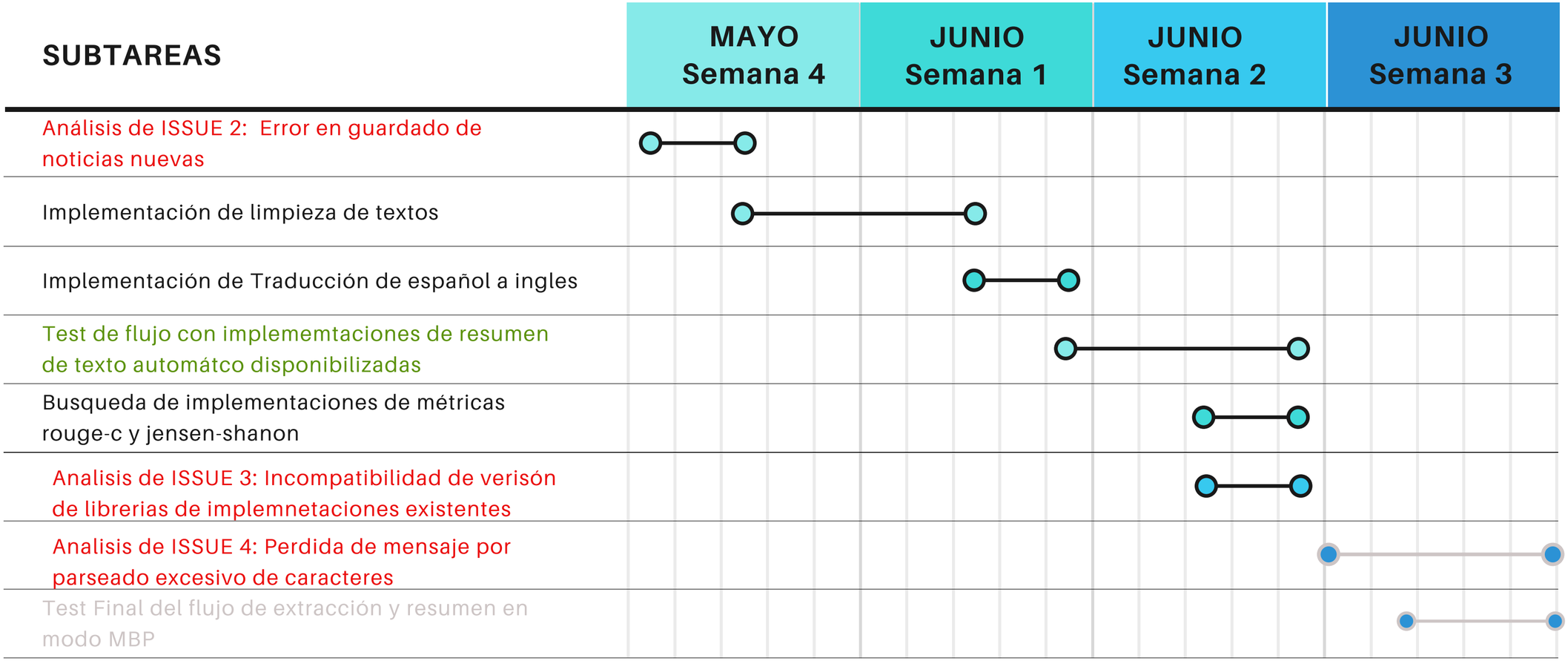






FASE 2: PRE-PROCESAMIENTO (STATUS 35%)

Almacenamiento



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

Busqueda de implementaciones en el idioma español

Alternativas planteadas

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

TEST DE FLUJO DE DATOS CON IMPLEMENTACIONES DE RESUMEN AUTOMÁTICO DISPONIBILIZADAS

Se opta por alternativa 2

Ejecución de TEST

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

ISSUES

Corregir los errores que se fueron suscitando

implementaciones

Seguir indagando en la documentación de soluciones disponibilizadas

Hipótesis específica

Acorde a lo investigado responder la 1er hipótesis específica

Investigar

Ampliar la investigación teórica sobre resumen de textos

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...

1 Contributor 0 Issues 0 Stars 0 Forks

kendalvictor/tesis_maestria_ciencia_datos

Contribute to kendalvictor/tesis_maestria_ciencia_datos development by creating an account on GitHub.

GitHub

GRACIAS