

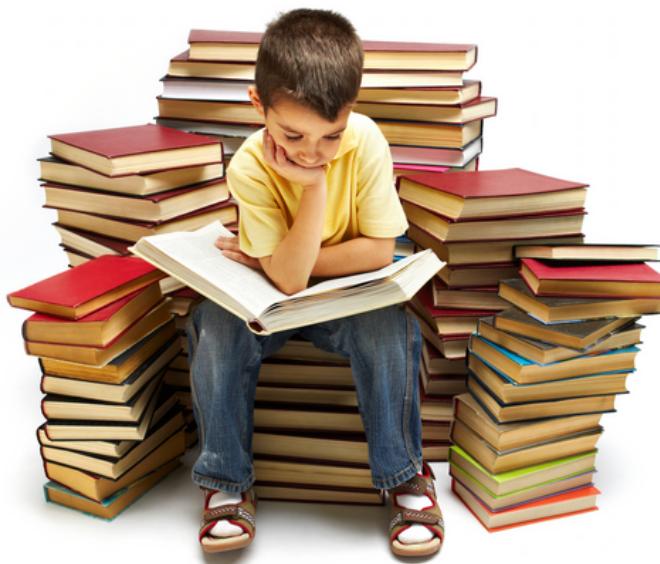
Integración de Técnicas de Aprendizaje de Maquina Hacia la Mejora de calidad en Síntesis de Noticias en Castellano

VÍCTOR MARIANO VILLACORTA PLASENCIA
CURSO: TESIS II



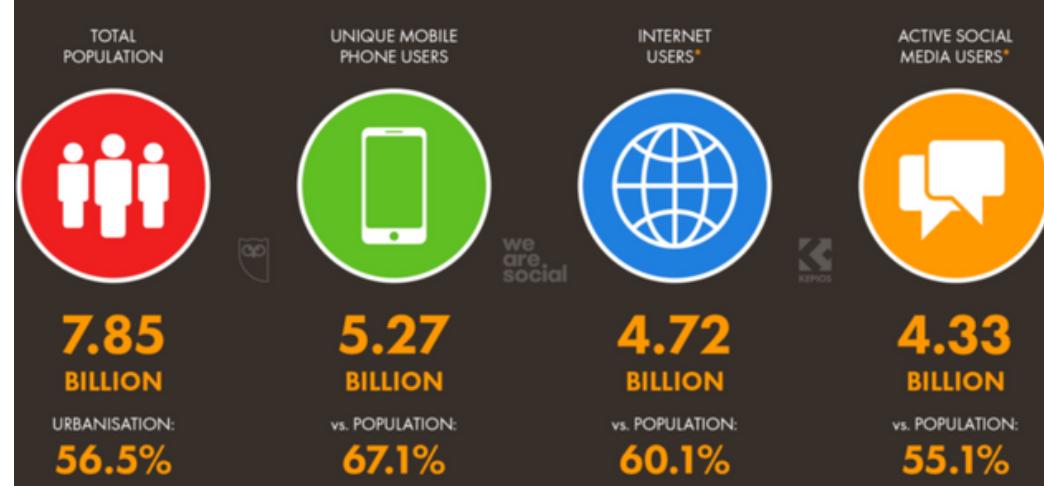
INTRODUCCION

El valor de obtener una síntesis de calidad sobre cualquier tema recae en el aumento en la factibilidad tanto en el entendimiento como en la transmisión del conocimiento



Cada vez somos mas digitales, cada vez delegamos mas tareas rutinarias a las maquinas, cada vez tenemos mucha mas información a la mano, cada vez surgen cosas nuevas que aprender, cada vez aumenta la necesidad de siempre estar enterados de nuestro entorno de forma rápida y concisa.





Digital 2021 April Global Statshot Report

DEFINICIÓN DEL PROBLEMA



Tendencia SMART MIRROR



"Las Nuevas Generaciones preffieren información despiezada en vez textos completos"

"Los usuarios de internet son inexpertos en sus hábitos de búsqueda "

"Generación del copia y pega "

El gran apogeo del Procesamiento Lenguaje Natural hacia los idiomas de mayor habla en el mundo a extendido la brecha existente frente al idioma Castellano en investigaciones recientes como lo es el Resumen de Automático de textos.



Coronavirus | ¿Qué dicen los expertos sobre el contagio a través de los asintomáticos?

Especialista de la OMS declaró que la propagación del coronavirus a través de pacientes sin síntomas sería inusual. ¿Cuán cierto es esto?

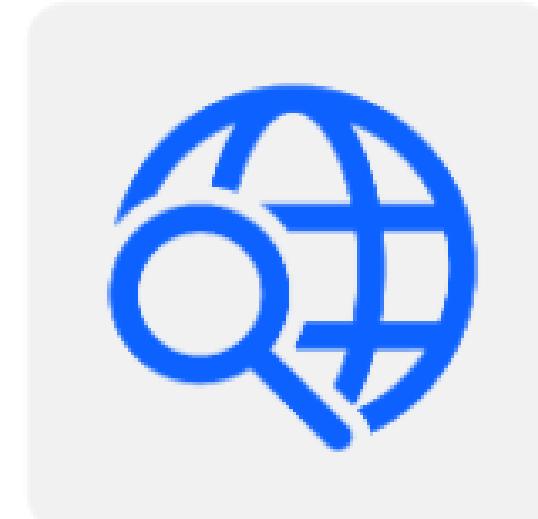


Se delimitará a la selección de noticias con un tipo de redacción informativa

Contexto en común: COVID-19

Publicaciones en un periodo mayor a Enero 2021

DELIMITACIÓN

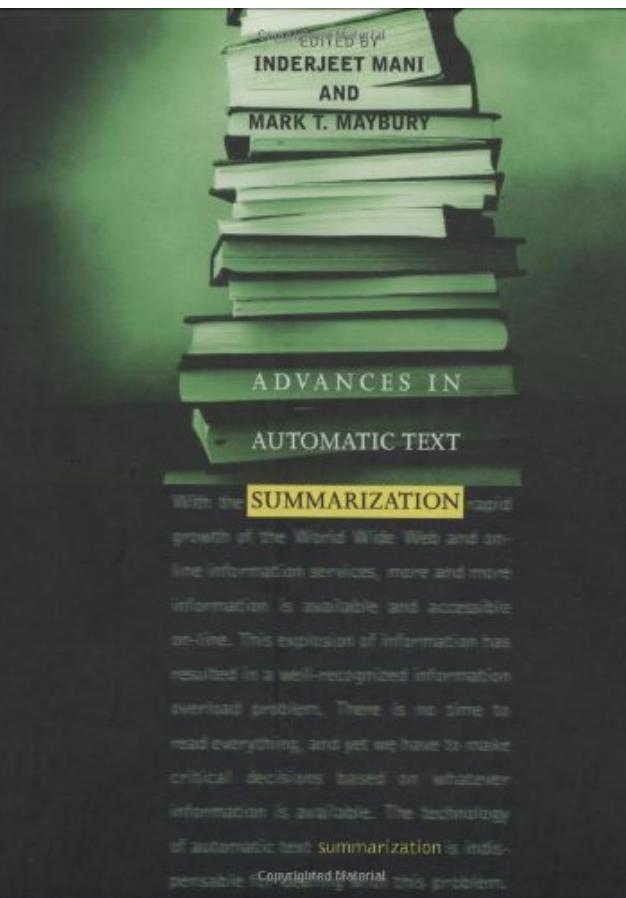


El motivo de porqué enfocarnos en noticias es por ya contener un correcto orden gramatical en su redacción

La obtención de su información es libre mediante técnicas de web scraping y web crawling limitados solamente por la paginación de listado de noticias con el contexto seleccionado

MARCO HISTORICO

1960 Se oficializa la necesidad



1999
Mani y
Maybury

2002
Lal y Ruger



2005
Teufel y
Moens

VECTORES



2020
Transformers

Gregory
Valderrama
Vilca
2017
PUCP

Generación
automática de
resúmenes
abstractivos mono
documento
utilizando análisis
semántico



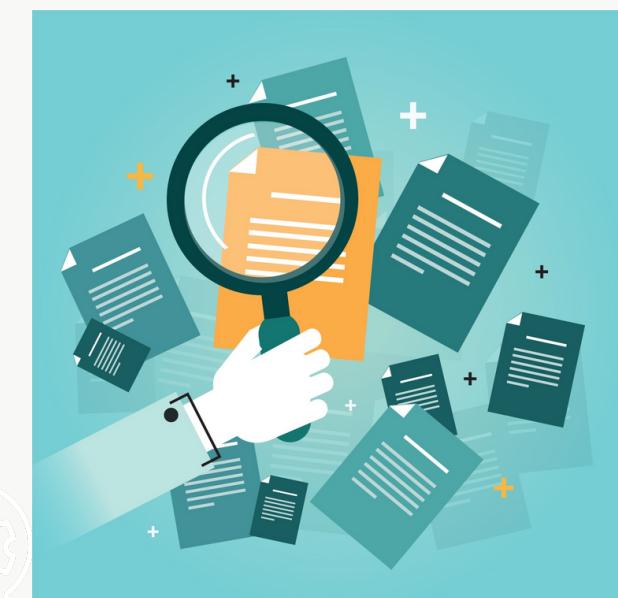
Implementación de
un software de
apoyo a la escritura
de resúmenes de
textos científicos

Irvin Vargas
Campos
2013
PUCP

ANTECEDENTES

Angel Alonso
Hernandez
2017

Deep Learning
aplicado al resumen
de textos



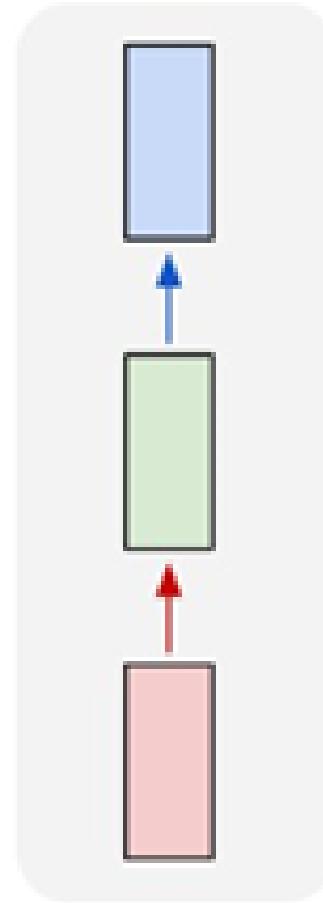
Implementación de
un software de
apoyo a la escritura
de resúmenes de
textos científicos

Selene
Sepúlveda
2020

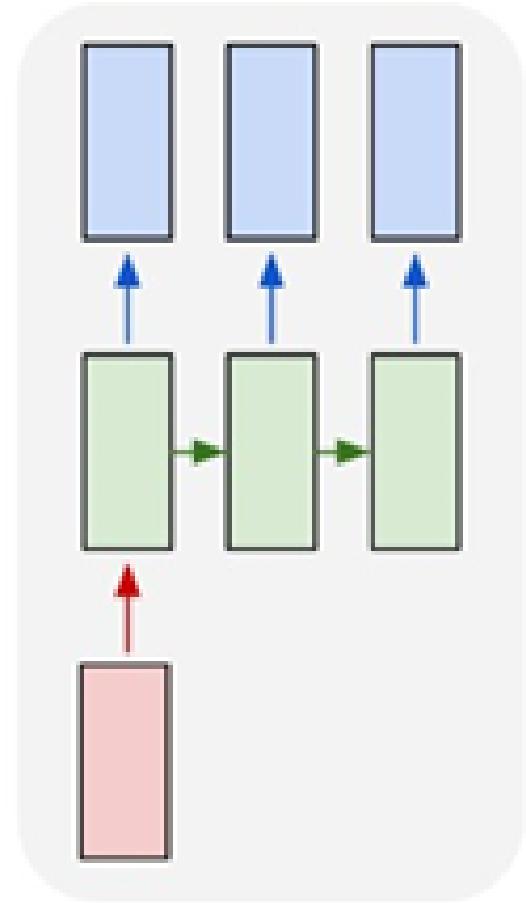
MARCO TEORICO

Diferentes tipos de aplicación de las redes neuronales:

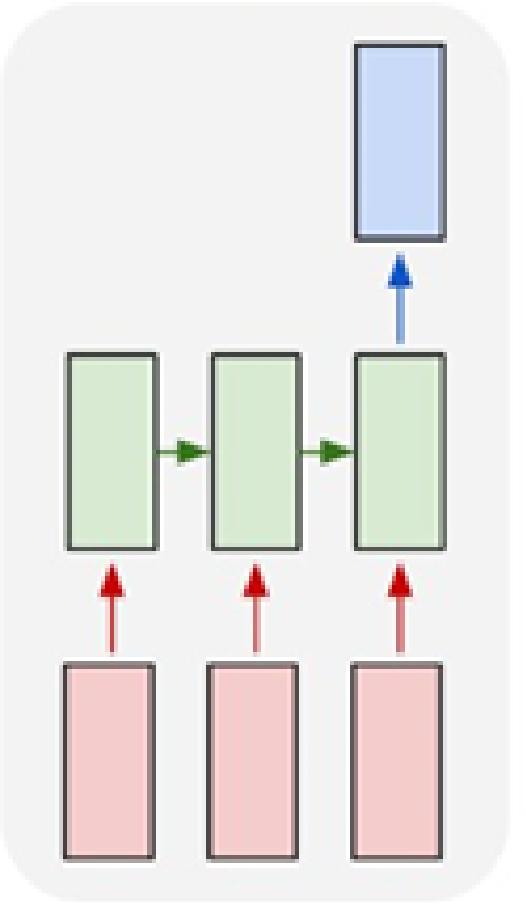
one to one



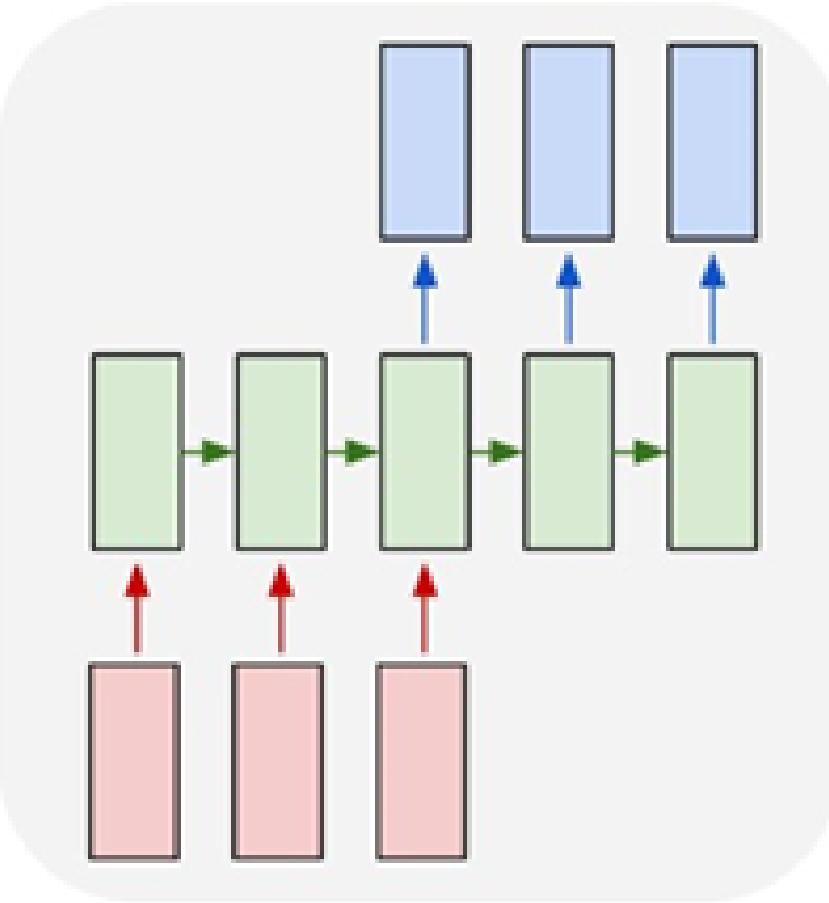
one to many



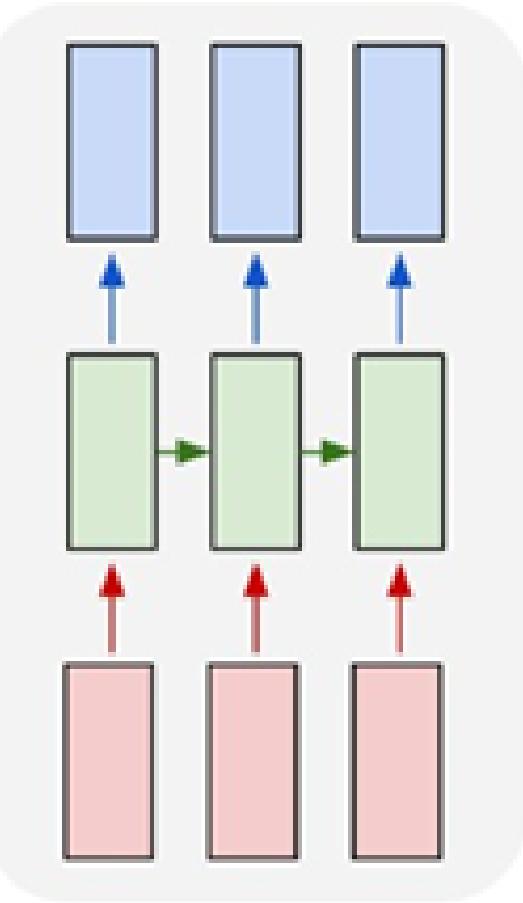
many to one



many to many



many to many



Machine
Learning
Tradicional

Descripción
de imágenes

Analisis de
sentimientos

Traducción
de textos /
Resúmenes
de texto

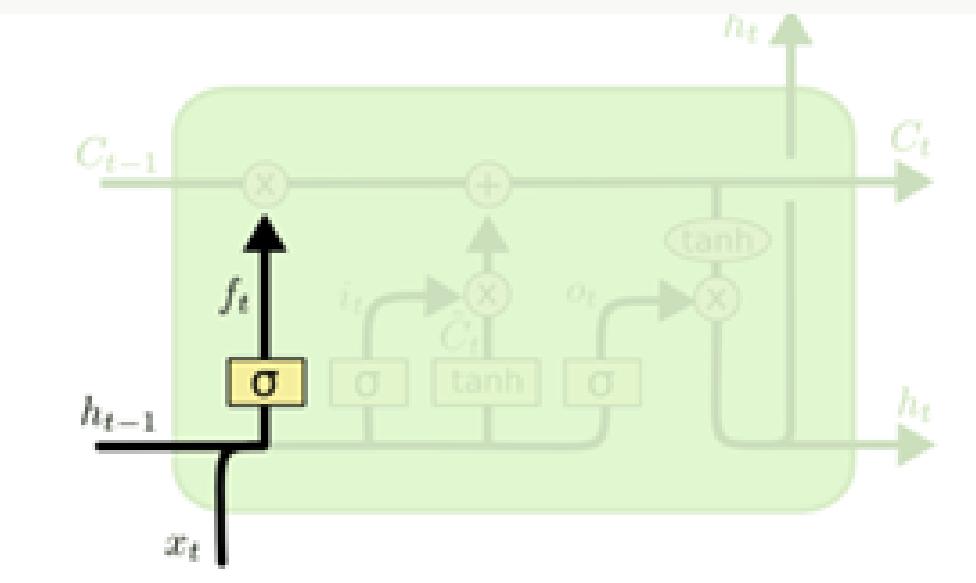
Clasificación
de videos



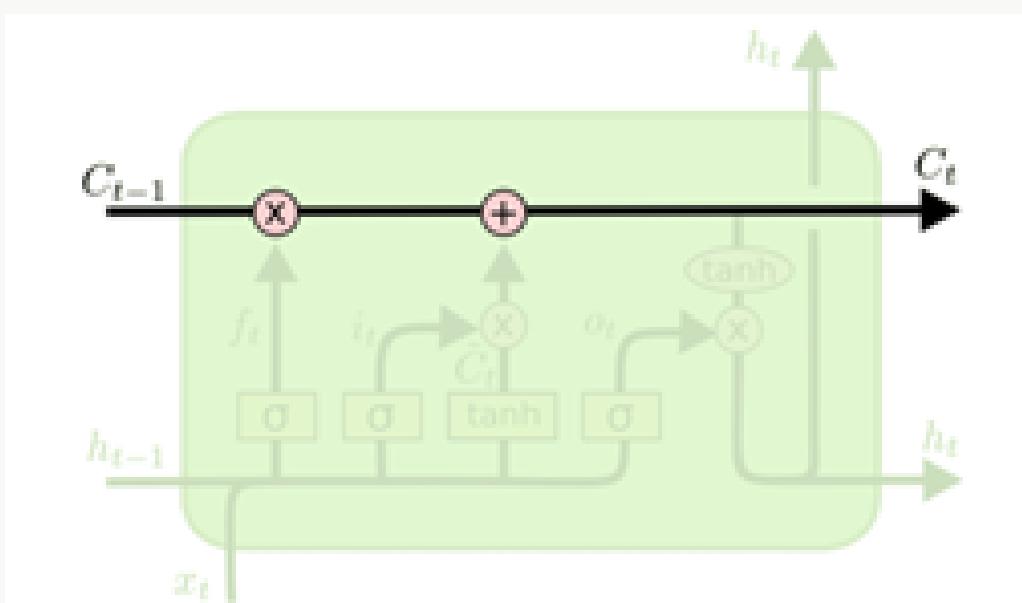
LSTM

Paso a paso

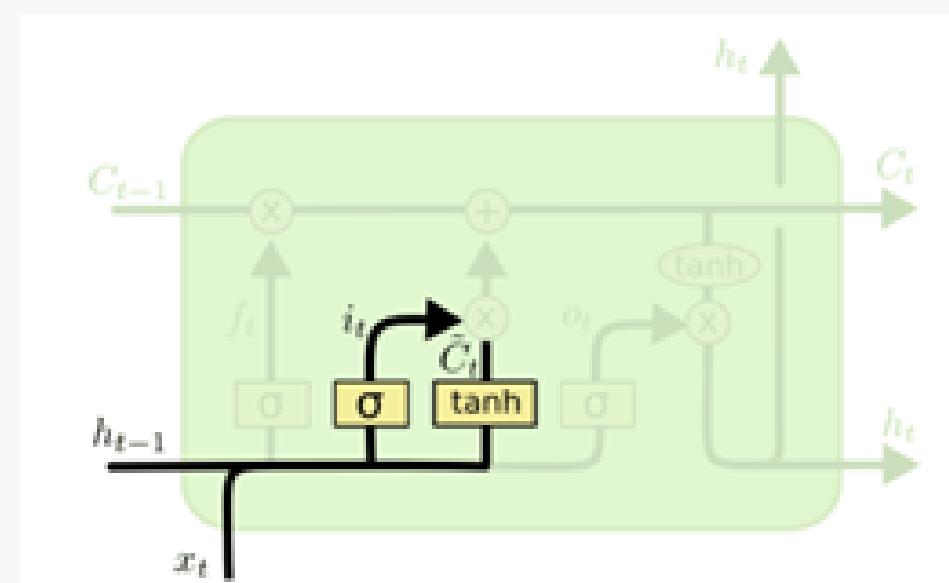
1-
Compuerta
de olvido



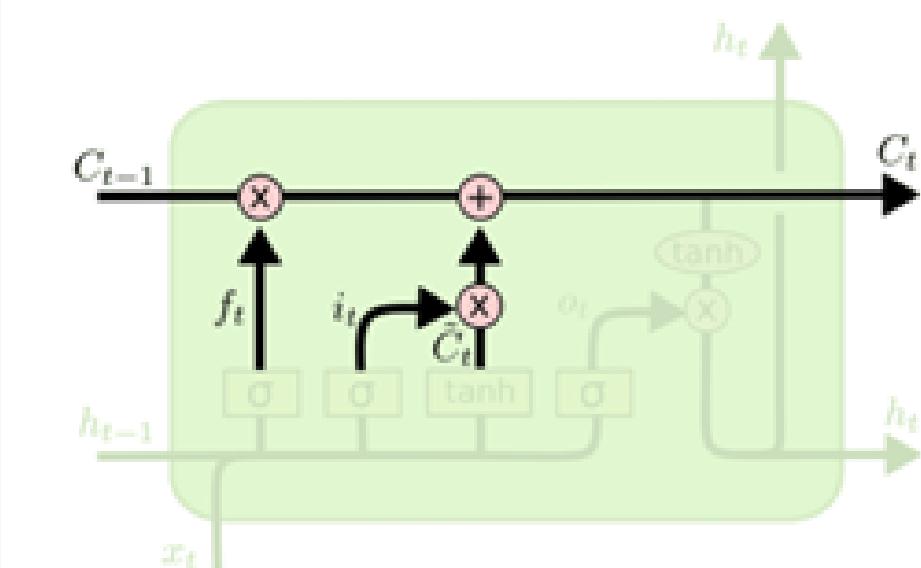
2-
Compuerta
de guardado
de lo mas
importante



3- Calculo
de
información
requerida de
la capa
oculta

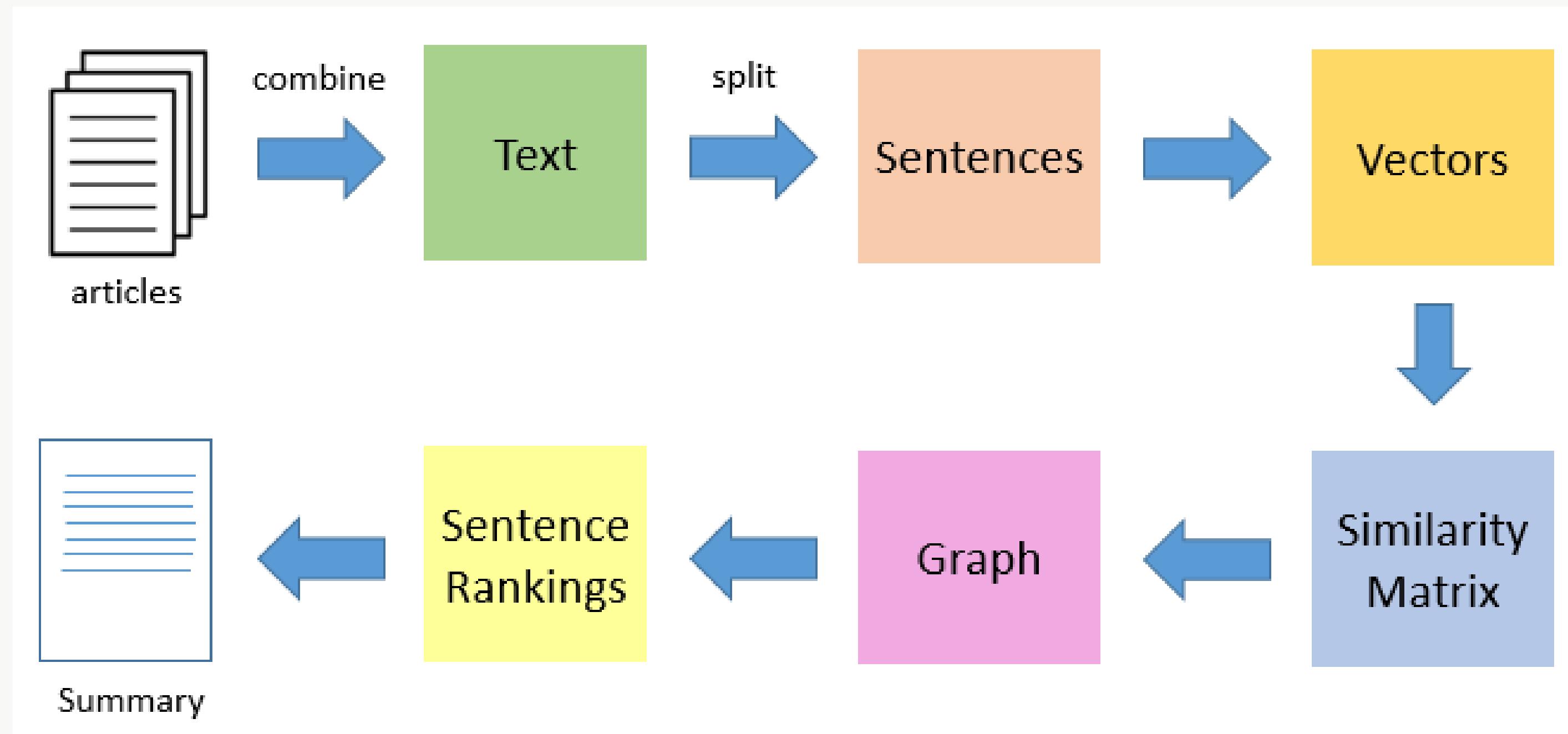


4- Interacción
de compuertas
y actualización
de pesos



TEXT-RANK

Versión adaptada hacia textos del famoso algoritmo de GOOGLE "PageRank", usado para dar relevancia a sus páginas web



METODOLOGIA

Enfoque
CUANTITATIVO



Indicador
Rouge-l



Pre
experimento

Indicador
Rouge-l de
técnicas por
separado

tratamiento

Integración

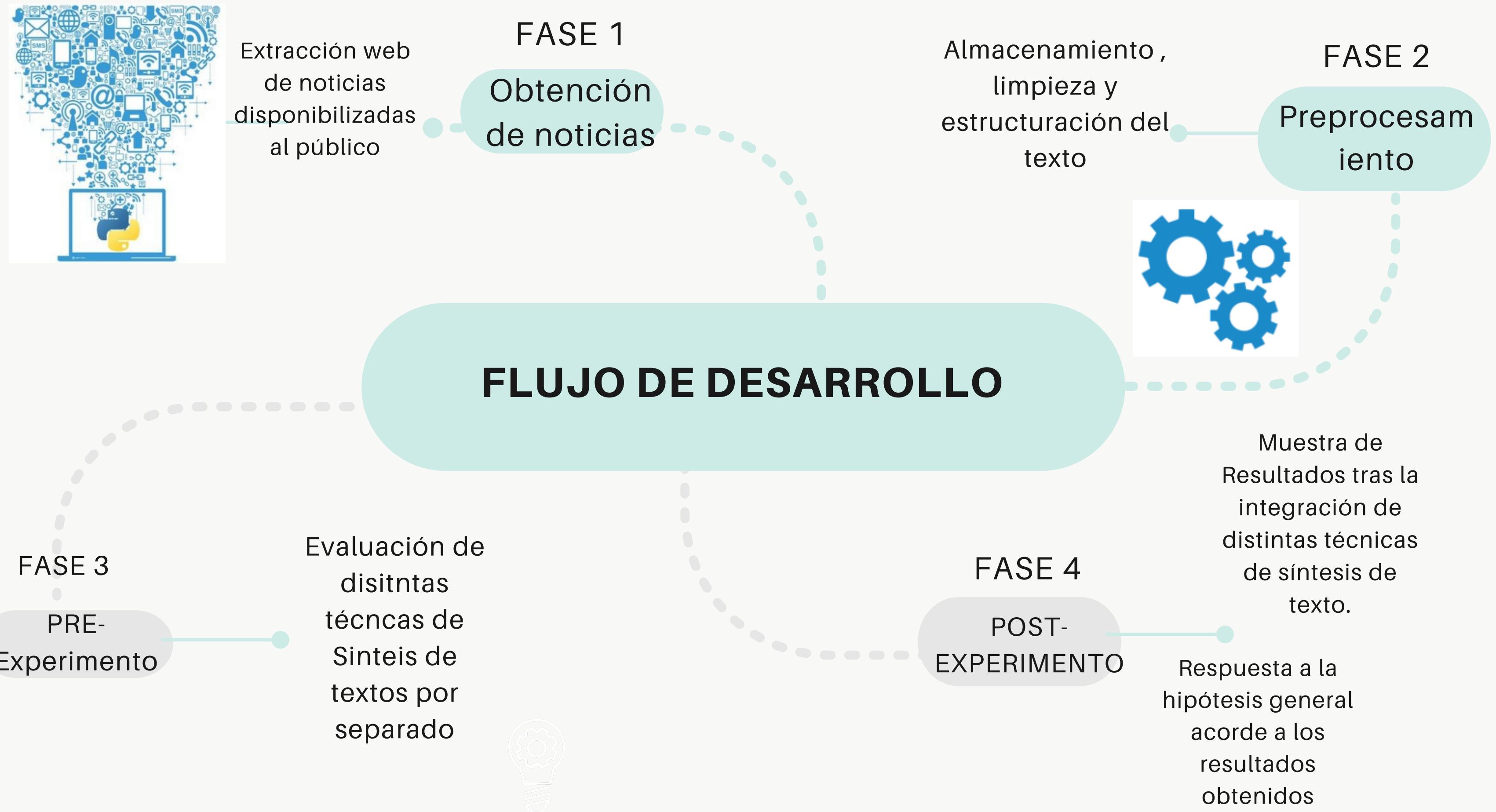
pseudo -
experimento

Diseño
EXPERIMENTAL

Pos
experimento

Indicador
Rouge-l a
la versión
integrada





El Comercio

Selección de
Diario formal

<https://elcomercio.pe/>



PYTHON

1er nivel
delimitación a
contexto
COVID

NAVIRUS

Coronavirus Perú: cifras actualizadas y 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

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



<https://elcomercio.pe/coronavirus/>

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



EXTRACCIÓN WEB DE NOTICIAS DISPONIBILIZADAS AL PÚBLICO

2DO nivel descarte de noticias fuera de contexto

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



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

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

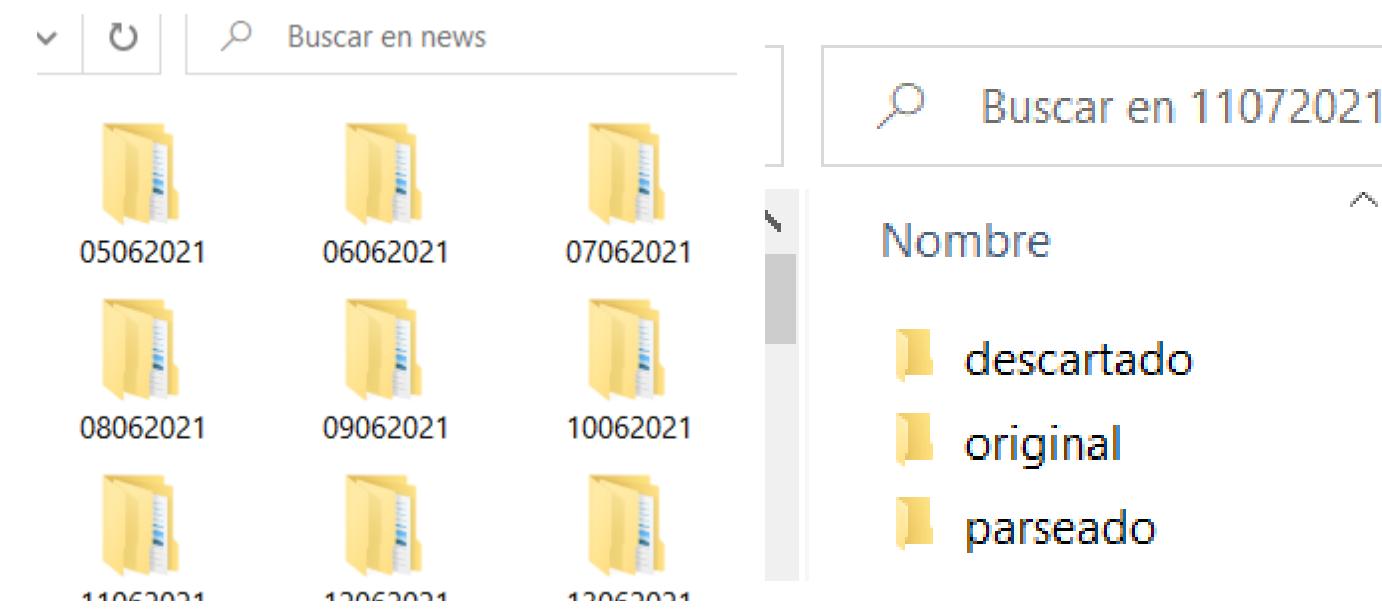
Estrategia de ETL
(extracción /
transformación y
consumo de información)



- 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 tratamiento

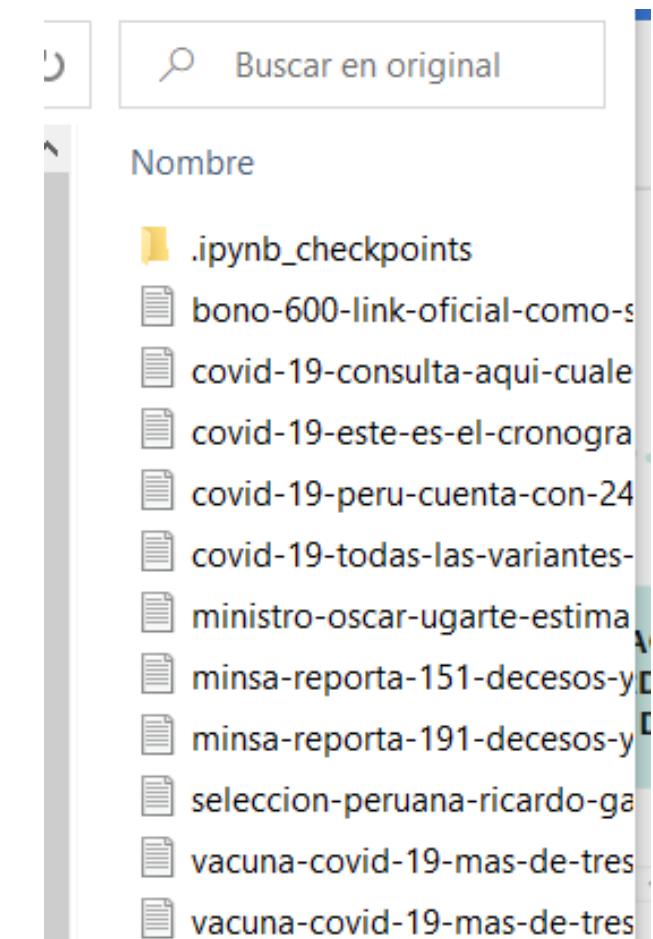


ALMACENAMIENTO DE NOTICIAS ACORDE AL DISEÑO DE SOLUCIÓN DE EJECUCIÓN DIARIA



Aprox. se tiene acceso a 50
noticias por día, de las cuales
aprox. 20 cumplen con la
delimitación deseada.

Límite de
extracción



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

Busqueda de implementaciones en el idioma español

Se opta por alternativa 2

Ejecución de TEST

TEST DE FLUJO DE DATOS CON IMPLEMENTACIONES DE RESUMEN AUTOMÁTICO DISPONIBILIZADAS

- 1.- Traducción de texto español a inglés sólo con el fin de test de flujo de datos.
- 2.- Testeo con implementaciones con autocomplementado de conectores gramaticales en inglés

Alternativas planteadas



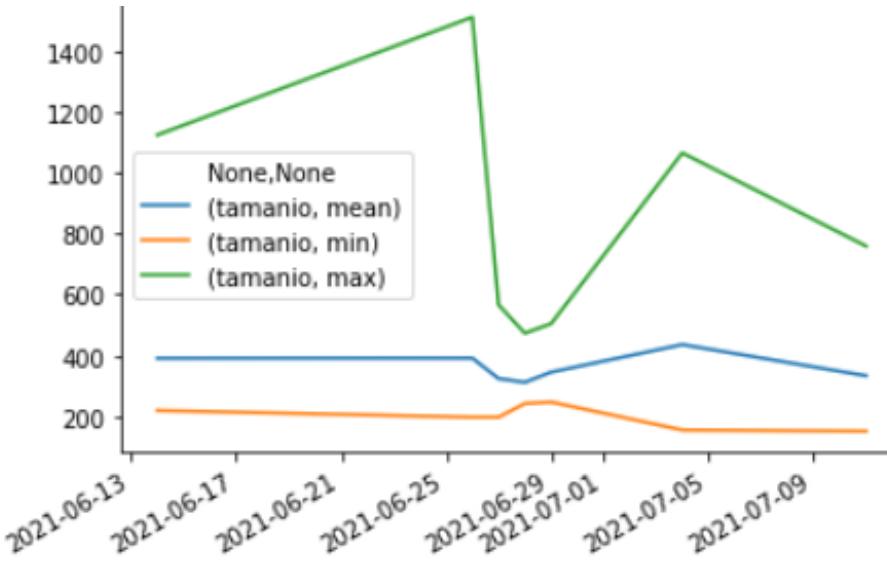
OJO: En el test aún no se opta por la versión parseada debido al último issue reportado

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

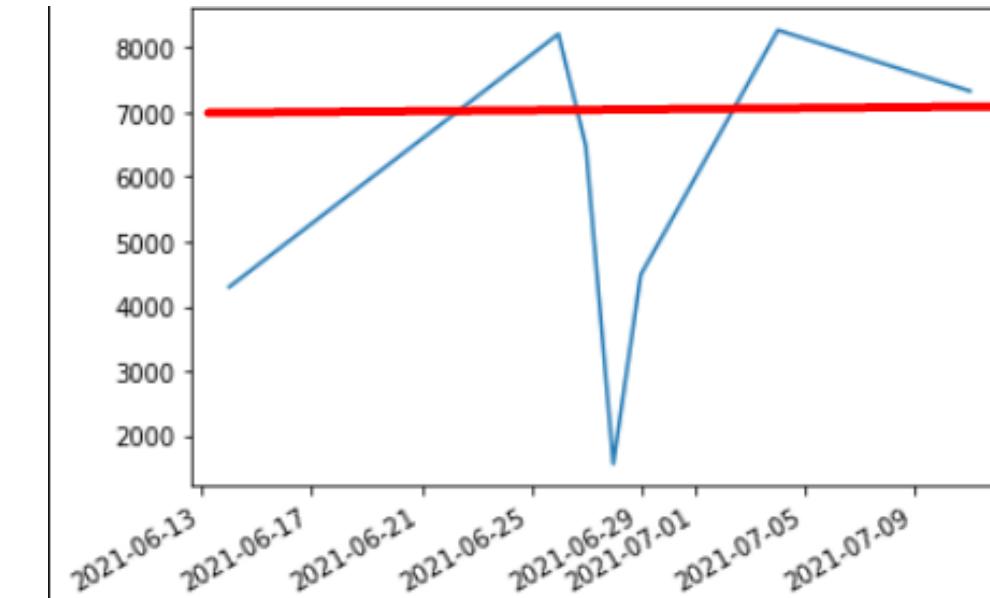


El Ministerio de Salud (Minsa) informó a 188.443 la cifra de decesos por coronavirus (COVID-19) Se trata de 191 nuevos fallecidos frente al reporte de la víspera . El Minsa reportó también that hay un total of 10.711 pacientes hospitalizados, of los cuales 2.548 están with ventilación mecánica

Analisis de taaño de texto



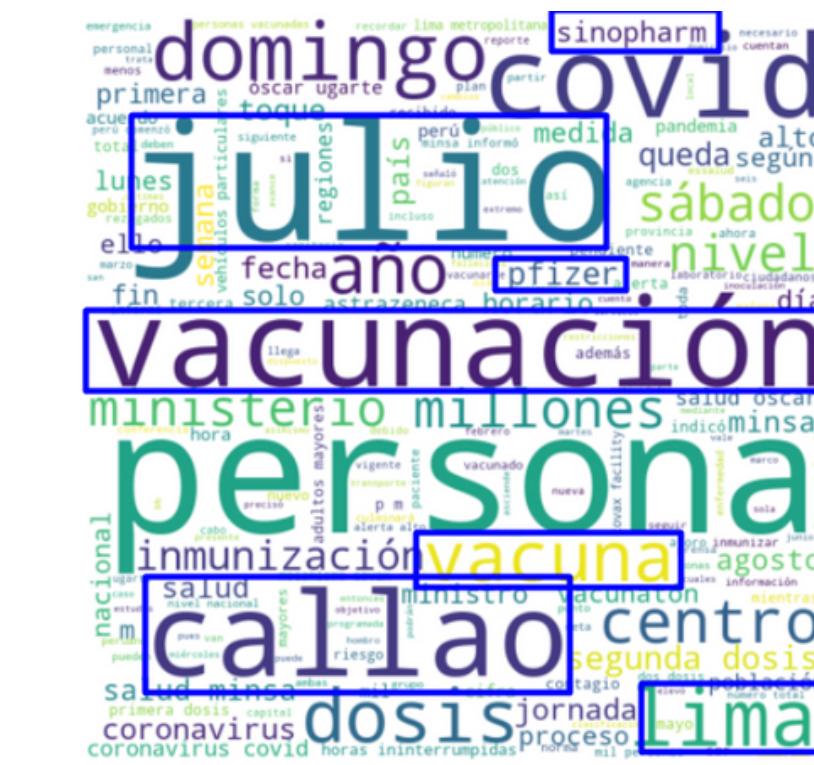
En promedio las 15 a 20 notiicas seleccionadas por d a contiene 400 palabras, lo cual, genera un texto total a resuir de aproximadamente 7000 palabras



ANALISIS DESCRIPTIVO 1

A la izquierda frecuencias de palabras en noticias descartadas y a la derecha de las seleccionadas.

Se ve claramente que el contexto internacional está enfocado en las variantes en distintos países mientras en nuestro país el enfoque es la vacunación y los tipos de vacuna



COMUNIDAD



Como estrategia de inicio a la fase de pre-experimento me he unido a una comunidad con objetivos claros que me ayudarán a responder mi 1er hipótesis específica



https://www.linkedin.com/posts/nlp-en-es_nlp-espaehol-transformers-activity-6805453628304830464-J0vJ/

INICIANDO EL PRE-EXPERIMENTO

Métricas

Resumen de referencia

Texto original

Rouge-n

Rouge-c

Resumen automático

Resumen automático

Implementaciones:

Rouge-1
Rouge-2

Rouge-L

No encontrado

Mencionado en Papers apuntando a automatización completa

Criterios de
contraste

Técnicas

Limpieza de texto

Resumen de
referencia
post-resumen

Resumen
automático
pre-resumen

RESULTADOS PRELIMINARES

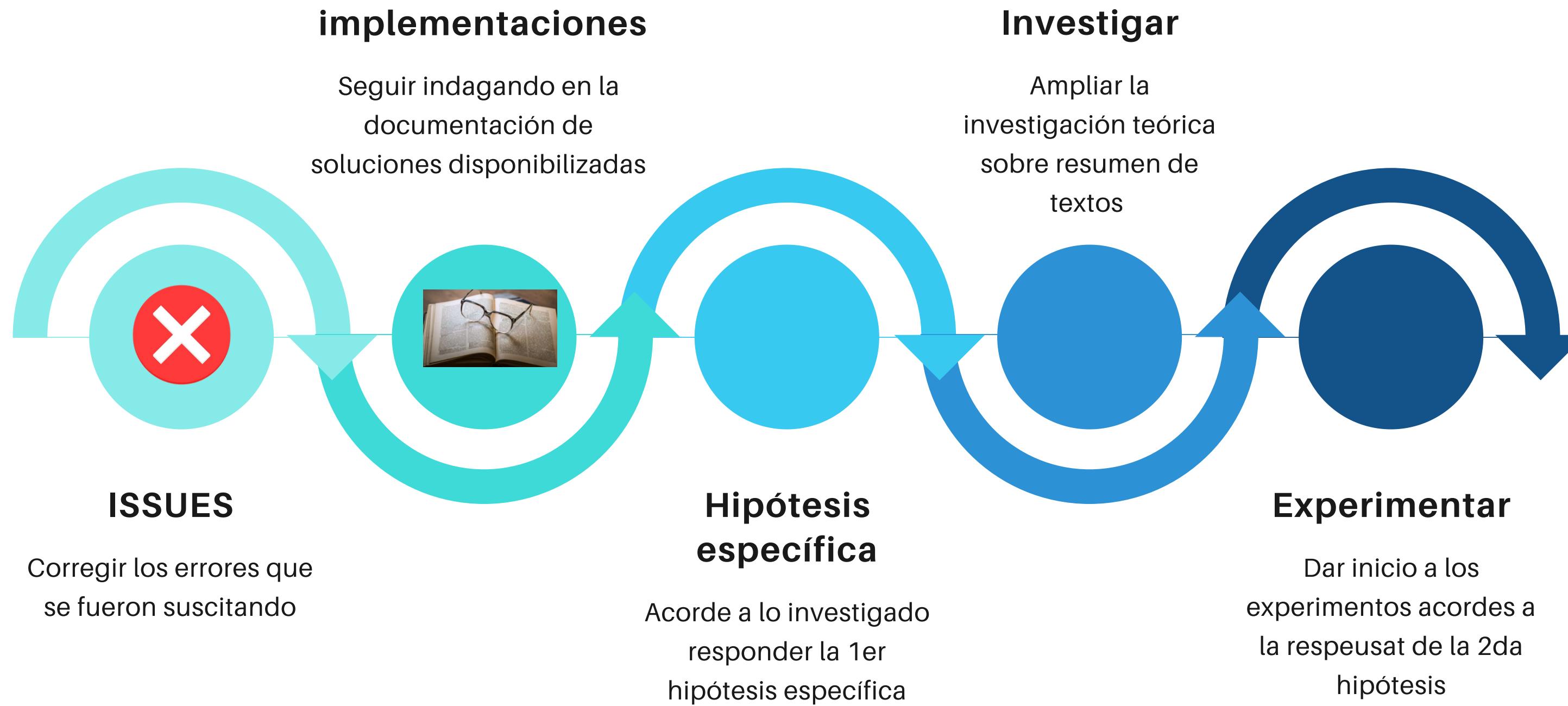
TEXT-RANK

f: 0.23

LSTM

aun no probado

SIGUIENTES PASOS



REPOSITORIO

https://github.com/kendalvictor/tesis_maestria_ciencia_datos

kendalvictor/
tesis_maestria_ciencia_...



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

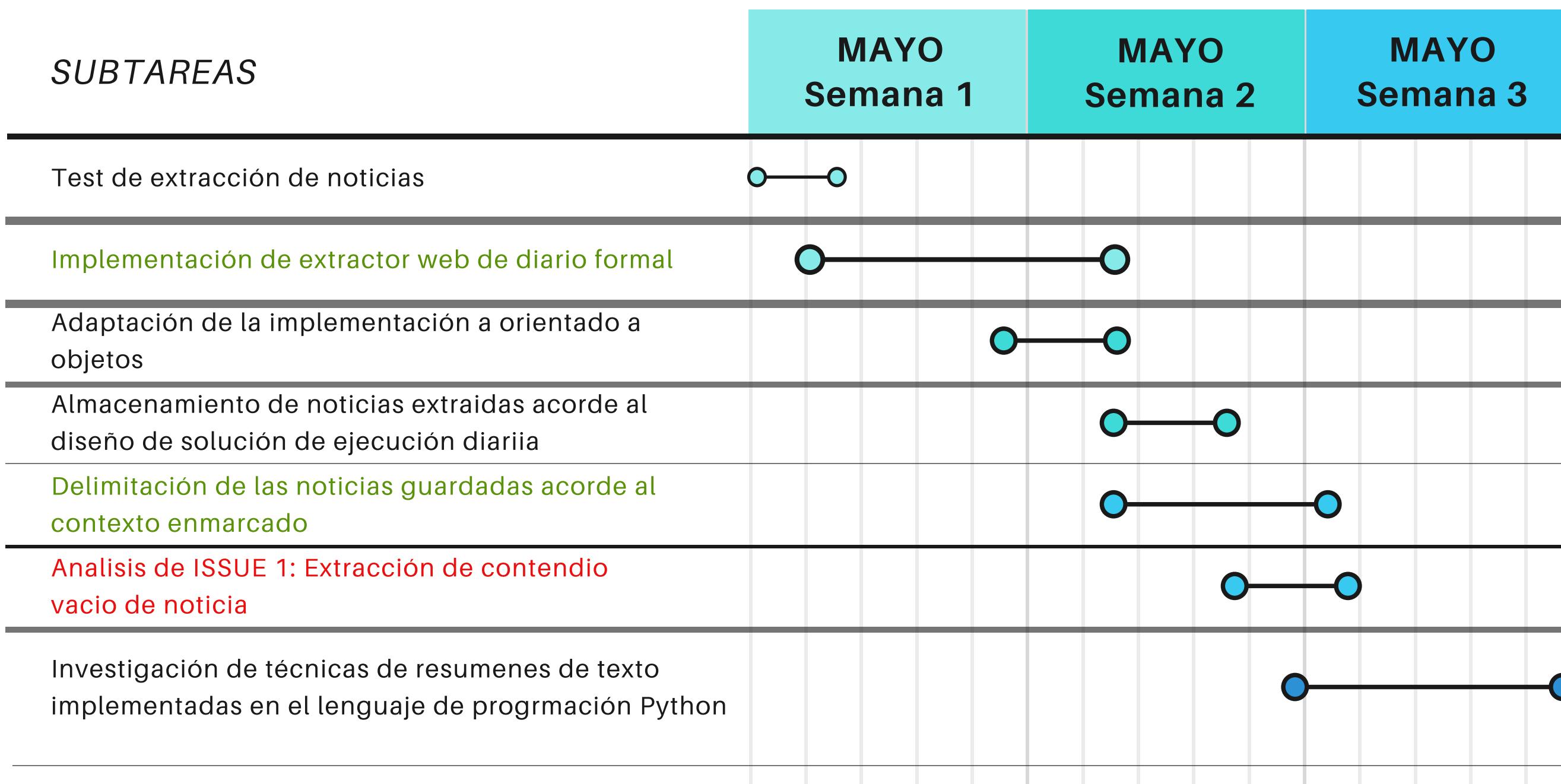
MATRIZ DE CONSISTENCIA

General

Preguntas	Objetivos	Hipótesis	Variable Independiente	Variable Dependiente
¿Afecta de manera positiva 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
¿ El uso de un resumen automático maduro podrá aportar en la elección de otros resúmenes automáticos válida para la elección de los mejores candidatos a integrar.	Crear una base de contraste de resúmenes automáticos válida para la elección de los mejores candidatos a integrar.	El uso de un resumen automático maduro aporta en la elección de otros resúmenes automáticos basados en LSTM y TEXT-RANK bajo el indicador ROUGE	DIMENSIÓN	
¿ La integración de las técnicas LSTM y Text Rank para resumen automático reflejará una mayor coherencia a criterio humano ?	Construcción del prototipo acorde a la factibilidad obtenida humano el efecto de integración de técnicas sobre el resumen de textos en español.	La integración de las técnicas LSTM y Text RANK para resumen automático tiene el efecto de reflejar mayor coherencia a criterio humano.	MÉTRICA	

FASE 1 - OBTENCIÓN DE NOTICIAS (STATUS 85%)

Extracción web de noticias disponibilizadas al público



FASE 2: PRE-PROCESAMIENTO (STATUS 85%)

Almacenamiento

SUBTAREAS	MAYO Semana 4	JUNIO Semana 1	JUNIO Semana 2	JUNIO Semana 3	JUNIO Semana 4
Análisis de ISSUE 2: Error en guardado de noticias nuevas	●	●			
Implementación de limpieza de textos	●	●			
Implementación de Traducción de español a inglés		●	●		
Test de flujo con implementaciones de resumen de texto automático disponibles		●	●		
Busqueda de implementaciones de métricas rouge-c y jensen-shanon		●	●		
Analisis de ISSUE 3: Incompatibilidad de versión de librerías de implementaciones existentes		●	●		
Analisis de ISSUE 4: Perdida de mensaje por parseado excesivo de caracteres		●	●		
Test Final del flujo de extracción y resumen en modo MBP		●	●		
Análisis descriptivo a nivel palabras / noticias		●	●		

FASE 3: PRE-EXPERIMENTO (STATUS 90%)

Investigación

SUBTAREAS	JULIO Sem. 1y2	JULIO Sem. 3 y4	AGOSTO Sem. 1 y 2	AGOSTO Sem. 3 y 4	SEPTIEMBRE Sem 1 y 2
Modificación de estructura para almacenar noticias descartadas	●	●			
Investigación de comunidades referentes al tema	●	●			
Investigación y aplicación de algoritmo TEXT RANK		●	●		
Investigación y aplicación de variantes de métrica Rouge (Rouge-c, Rouge-L)		●	●		
Análisis de Issue 4: Diferenciación de limpieza de datos acorde a la técnicas para no perder sentido en resultados		●	●		
Seguimiento curso de NLP español		●	●	●	●
Investigación y aplicación de redes recurrentes con memoria - LSTM		●	●		
Formalización de la generación de resultados de manera automática			●	●	