TUIA NLP 2023

TRABAJO PRÁCTICO 1

INTEGRANTES:

- LO MENZO, ALEJO
- PONCE, DANIEL
- YAÑEZ, MIRIAN

Url del repositorio:

https://github.com/Mirian2550/TP_NLP

Descargar este archivo:

https://cs.famaf.unc.edu.ar/~ccardellino/SBWCE/SBW-vectors-300-min5.bin.gz

Clasificador de noticias

Para llevar a cabo este trabajo práctico, implementamos un enfoque de web scraping para recopilar noticias de diversas categorías, incluyendo seguridad informática, bebés, deportes y comida. Utilizamos un scrapper multifunción que implementa multihilos, ejecutando simultáneamente scrappers específicos para cada categoría. La implementación se beneficia del módulo multiprocessing de Python, lo que permite una ejecución paralela eficiente de los scrappers. Adaptamos el código para garantizar su compatibilidad tanto con sistemas operativos Linux como Windows, incorporando la condición **if __name__ == "__main__"** para manejar las diferencias en la creación de procesos entre estos sistemas.

Para respetar las directrices de los sitios web y evitar posibles restricciones, consultamos el archivo robots.txt de cada sitio web objetivo antes de realizar el web scraping. Este enfoque garantiza la conformidad con las políticas de los sitios y ayuda a mantener la integridad del proceso.

Dentro de ese archivo revisamos cada uno de los links que tiene el archivo sitemap.xml y armamos la lista para extraer título y texto de cada uno de esos.

La limpieza y extracción de texto se realiza de manera meticulosa, centrándonos en los elementos **<p>** dentro de la estructura HTML y considerando la clase del div que los contiene. Eliminamos elementos innecesarios y posteriormente concatenamos el texto de los párrafos resultantes. Esto nos permite obtener únicamente el contenido textual relevante de las noticias que necesitamos para este trabajo.

En la etapa de ejecución de los scrappers, gestionamos posibles errores y eventos inesperados, registrando información detallada en un archivo de registro (**scraper.log**). Esta práctica facilita la identificación y resolución de problemas durante el proceso de web scraping.

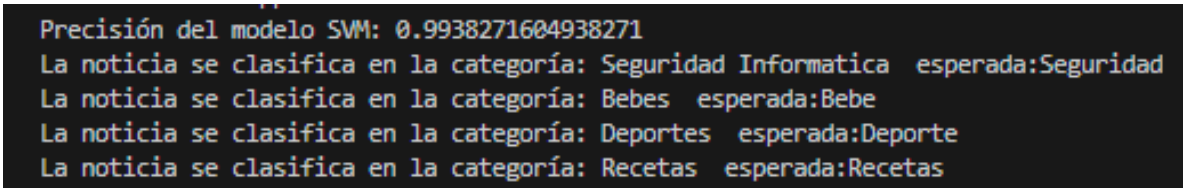
En la etapa de clasificación, optamos por utilizar una máquina de soporte de vectores (SVM), un método de clasificación que permite una subdivisión eficiente de clases. Decidimos entrenar el modelo con 50 noticias en lugar de 10 para mejorar la precisión de la clasificación. Enfrentamos la limitación del bloqueo después de un

número de intentos al consultar ciertos sitios web y mitigamos este problema utilizando User-Agents aleatorios.

Con el objetivo de eludir la detección automatizada durante las solicitudes HTTP, implementamos una estrategia eficaz mediante el uso de User-Agents aleatorios. Esto simula la variedad de navegadores que podrían realizar solicitudes y contribuye a evitar bloqueos por parte de los sitios web.

Una vez creado el modelo, realizamos pruebas con noticias específicas para evaluar su desempeño en la clasificación.

Resultado obtenido:



Nube de palabras

En el archivo nube_palabras.py, utilizando las bibliotecas NLTK y wordcloud, implementamos una clase encargada de realizar un conteo de frecuencia de palabras, evaluando cuántas veces aparece cada palabra en el texto. Antes de esto, tuvimos que llevar a cabo un análisis del lenguaje para eliminar las palabras vacías del mismo.

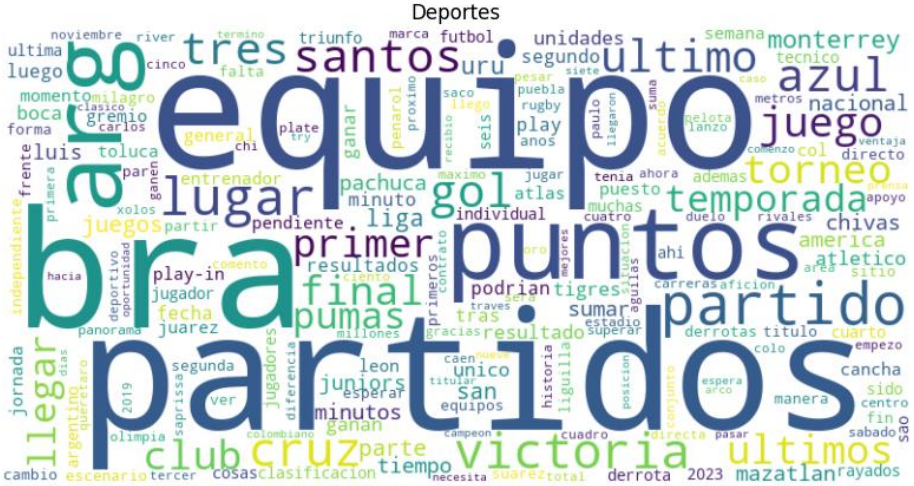
Con el fin de optimizar el tiempo de ejecución, decidimos emplear procesos, siguiendo la misma estrategia utilizada en los scrappers. En el archivo normalizador.py se encuentra un método llamado procesar_texto, el cual recibe como parámetro una categoría y realiza un resumen, convierte el texto a minúsculas y utiliza la librería Unicode para eliminar las tildes de las palabras.

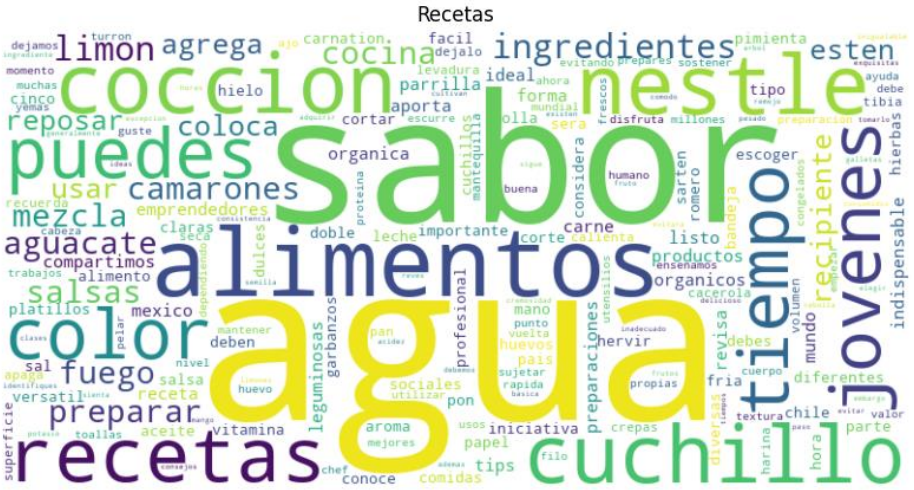
Además, se emplea la herramienta wordtokenize para dividir el texto en palabras. Posteriormente, se utiliza el método stopword de NLTK para eliminar las palabras vacías del lenguaje. Es importante mencionar que modificamos la lista de palabras vacías, ya que la predeterminada nos proporcionaba palabras en inglés y omitía algunas palabras en español.

Resultado obtenido:

Seguridad Informatica

Deportes





Comparación de Titulos

Para realizar la comparación de similitud entre los títulos de noticias dentro de una categoría específica, implementamos la función **title_compare**. En esta función, utilizamos tres modelos de embedding diferentes: Word2Vec (representado por KeyedVectors), BERT (Bidirectional Encoder Representations from Transformers), y SBERT (Sentence-BERT).

Primero, cargamos los modelos preentrenados, que incluyen SBW-vectors-300-min5.bin.gz para Word2Vec, bert-base-uncased para BERT, y paraphrase-MiniLM-L6-v2 para SBERT. Posteriormente, obtenemos los títulos de noticias de la categoría seleccionada.

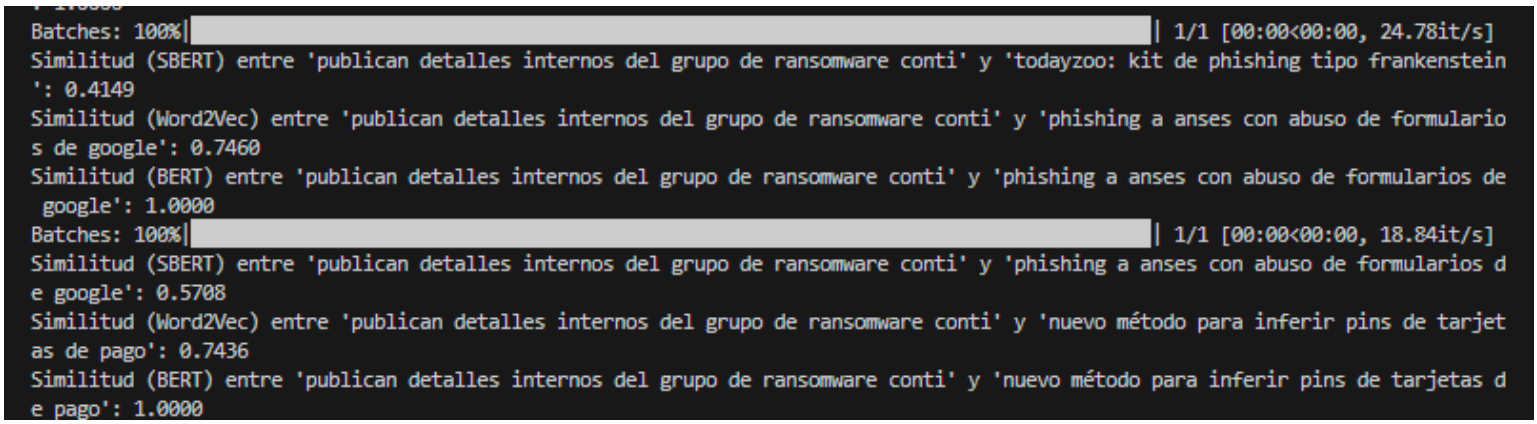
Luego, para cada par de títulos en la categoría, realizamos la comparación de similitud utilizando los tres modelos:

1. **Word2Vec:** Utilizamos el modelo Word2Vec para obtener la similitud coseno entre los tokens de los títulos. Se filtran los tokens que no están presentes en el modelo.
2. **BERT:** Empleamos el modelo BERT para generar embeddings de los títulos y calcular la similitud coseno entre los vectores resultantes.
3. **SBERT:** Usamos el modelo SBERT para obtener embeddings de los títulos y calcular la similitud coseno entre estos vectores.

Los resultados de cada comparación se imprimen, proporcionando el porcentaje de similitud entre los títulos correspondientes. Este enfoque nos permite evaluar la similitud semántica entre los títulos utilizando distintas técnicas de embedding.

Es importante destacar que, a pesar de la utilidad de estos modelos, existen limitaciones en la evaluación de la similitud entre títulos de noticias. Estas limitaciones pueden incluir la falta de contexto más amplio, la sensibilidad a errores en el preprocesamiento del texto, y la dependencia de la calidad y representatividad del conjunto de datos de entrenamiento. Además, la elección del modelo también puede afectar los resultados, ya que cada uno tiene sus propias fortalezas y debilidades en términos de captura de significado semántico.

Resultado obtenido:



Chatbot de telegram

Para la implementación del chatbot de Telegram, elegimos utilizar el modelo BART (Bidirectional and Auto-Regressive Transformers) para generar resúmenes de noticias. La elección de BART se basa en su capacidad para realizar resúmenes coherentes y contextualmente ricos, lo que lo convierte en una opción adecuada para proporcionar información condensada pero significativa.

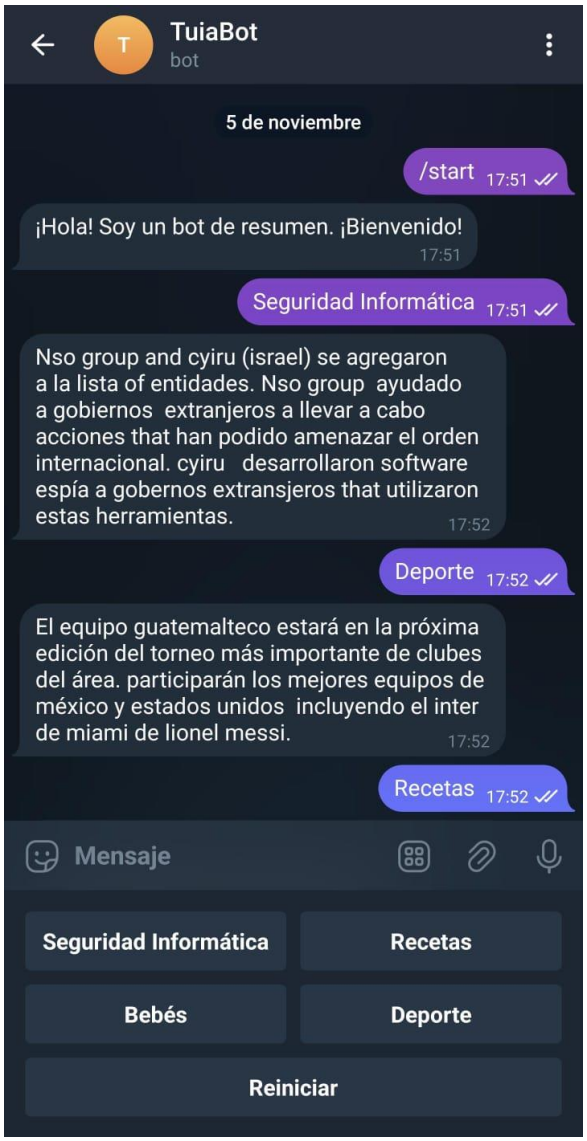
Justificamos el uso de BART sobre otras opciones como NLTK y Spacy debido a su capacidad para generar resúmenes más detallados y con una mejor estructura gramatical. Mientras que NLTK y Spacy son excelentes para tareas de procesamiento de lenguaje natural, BART destaca específicamente en la generación de resúmenes, lo que lo hace más idóneo para nuestro propósito.

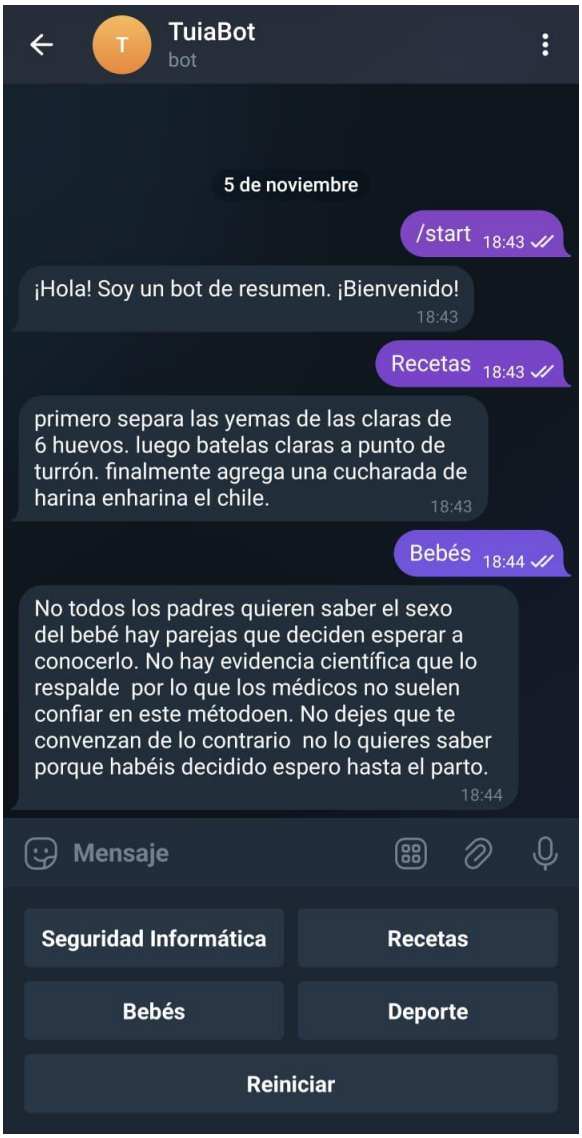
La interactividad del chatbot se logra mediante un menú interactivo que limita las opciones del usuario a las categorías predefinidas (Seguridad Informática, Recetas, Bebés y Deporte). Esto simplifica la interacción y permite al usuario seleccionar la categoría de su interés.

La implementación también incluye la opción de reiniciar la conversación, brindando al usuario la posibilidad de explorar diferentes categorías en una misma sesión. Además, se incorpora un manejo de errores que informa al usuario que solo puede interactuar a través de los botones disponibles, brindando una experiencia más guiada y amigable.

Una vez que se ejecuta el bot se puede utilizar este enlace de telegram para utilizarlo: http://t.me/TuiaNBBot

Resultado obtenido:


TuiaBot
bot
5 de noviembre
/start 17:51
¡Hola! Soy un bot de resumen. ¡Bienvenido! 17:51
Seguridad Informática 17:51
Nso group and cyiru (israel) se agregaron a la lista of entidades. Nso group ayudado a gobiernos extranjeros a llevar a cabo acciones that han podido amenazar el orden internacional. cyiru desarrollaron software espía a gobernos extransjeros that utilizaron estas herramientas. 17:52
Deporte 17:52
El equipo guatemalteco estará en la próxima edición del torneo más importante de clubes del área. participarán los mejores equipos de méxico y estados unidos incluyendo el inter de miami de lionel messi. 17:52
Recetas 17:52
Mensaje
Seguridad Informática
Recetas
Bebés
Deporte
Reiniciar


TuiaBot
bot
5 de noviembre
/start 18:43
¡Hola! Soy un bot de resumen. ¡Bienvenido! 18:43
Recetas 18:43
primero separa las yemas de las claras de 6 huevos. luego batelas claras a punto de turrón. finalmente agrega una cucharada de harina enharina el chile. 18:43
Bebés 18:44
No todos los padres quieren saber el sexo del bebé hay parejas que deciden esperar a conocerlo. No hay evidencia científica que lo respalde por lo que los médicos no suelen confiar en este métodoen. No dejes que te convenzan de lo contrario no lo quieres saber porque habéis decidido espero hasta el parto. 18:44
Mensaje
Seguridad Informática
Recetas
Bebés
Deporte
Reiniciar