Development of cyberweapons using Artificial Intelligence

Author. Yohanna Andrea Toro Duran , Author. Daniel Alberto Rosales Castro

I. Introducción

EL procesamiento de lenguaje natural los podemos encontrar en diferentes ámbitos de la vida diaria como por ejemplo lo son una simple búsqueda en Google, respuestas automáticas de Gmail, incluso traducciones a diferentes idiomas, pero en este documento nos centraremos en la construcción de una herramienta de seguridad ofensiva utilizando modelos de procesamiento de lenguaje natural para:

- Identificación de vulnerabilidades.
- Perfilamiento de adversarios.
- Predicción de entidades relacionadas
- Descubrimiento de fuentes de información
- Relación entre diferentes contenidos

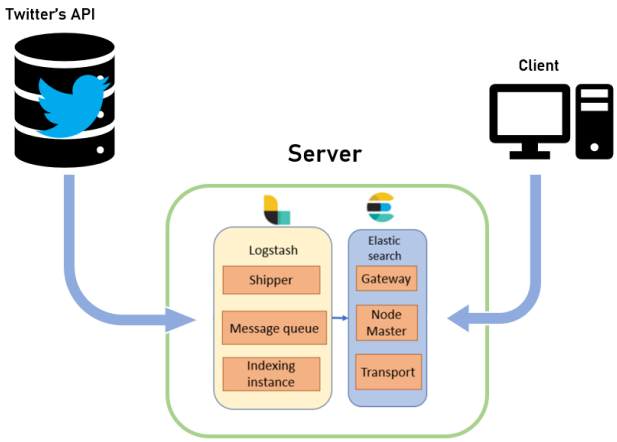
Todo esto en el ámbito del ciber terrorismo y analíticas de información de interés militar, para la prevención y contención de actividades terroristas futuras y recopilación de información para el perfilamiento de adversarios y entidades con fines dañinos para la sociedad.

Para esto usaremos 3 modelos concretamente para el análisis de información de redes sociales (Twitter) que son:

- Recolector
- Predicción de Hashtags
- Identificación de entidades
- Determinación de proximidad

II. Módulo de Recolección

Este modulo se encarga de Recolectar la información de la API de twitter para su futuro análisis y procesamiento para todos los modulos



A. *Twitter's Api*

Twitter posee una *API* que nos permite recolectar twitts directamente de Twitter, estos vienen de forma no estructurada de todo el mundo, Para la utilización de esta *API* es necesario una cuenta en Twitter y solicitar la conexión en su página web

B. *Servidor*

En el servidor se encuentran implementadas dos herramientas para la recolección y estructuración de los datos que se están recolectando desde la *API* de twitter. Estas herramientas son Logstash y Elasticsearch.

1) *Logstash:*

Logstash nos permite la conexión directa con el api y la recolección de los tweets que llegan de una forma no estructurada

2) *Elasticsearch:*

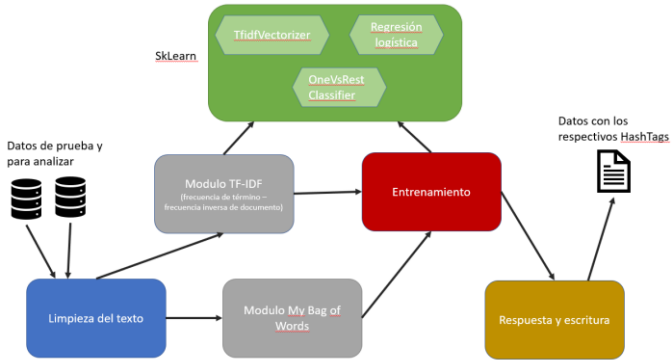
Elasticsearch nos permite la estructuración de los datos en Formato JSON para su correcto análisis mediante los 3 modelos, aparte esta herramienta nos deja especificar porque país queremos filtrar los datos, porque idioma y específicamente que palabras clave queremos la recolección y el filtro.

III. Módulo de predicción de hashtags

Este modulo nos permite asignarle una o más etiquetas a un texto. En este caso usaremos los tweets recolectados para

.

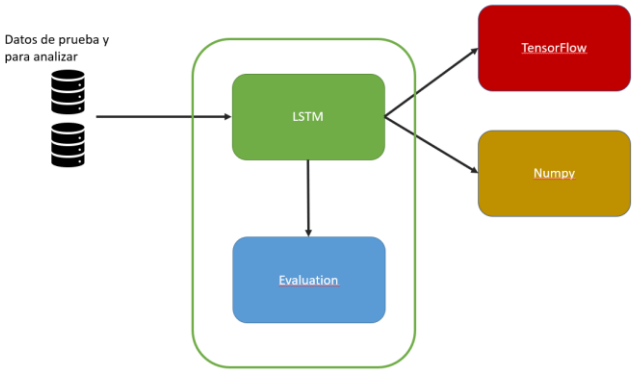
identificar en cada uno de ellos etiquetas correspondientes a terrorism



1. Limpieza de texto : Es neceseario preparar los datos para que los evitar tokens extraños en la extracción del texto.
2. Bag of words: la idea de este modulo es tener una las N palabras más populares siendo guardadas en un diccionario y tener un vector de tamaño N y poder identificar que palabras están en el corpus.
3. TF-IDF: este modulo permite determinar la frecuencia de las difernetes palabras en el diccionario lo cual permite sacar las palabras más comunes y generar un mayor filtro.
4. One vs rest classifier es una estrategia que consta en asociar un conjunto de ejemplos positivos para una clase determinada y un conjunto de ejemplos negativos que representan todas las demás clases.
5. Regresión logística: es un tipo de análisis de regresión que permite determinar el resultado de una variable categórica.
6. Paquetes necesarios que usa el modelo:
   - Numpy - un paquete para la computación científica.
   - scikit-learn - una herramienta para el análisis de datos.
   - NLTK – paquete para trabajar con lenguaje natural.

IV. Módulo de identificación de entidades

Este modulo usa NER que por sus siglas indican es reconocimiento de entidades nombradas lo que consiste es que en un texto en este caso los tweets extraer entidades como personas, organizaciones, entre otras cosas.

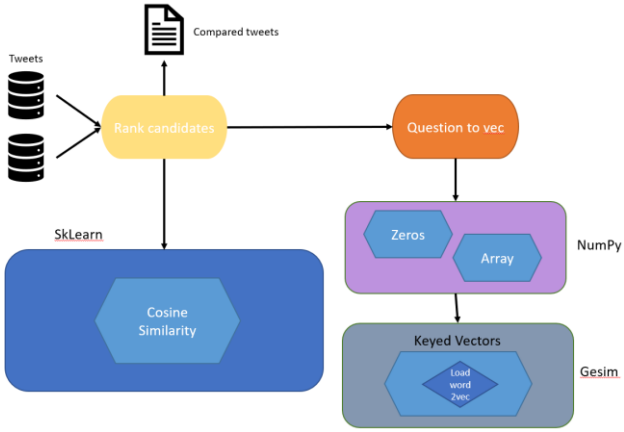


1. Bidirectional LSTMs: consiste en la construccion de bloque de redes neuronales el cual realiza un procesamiento sobre el texto para analizar los diferentes tokens de interés después del procesamiento y antes de este.

2. evaluation: este módulo se encarga de evaluar las diferentes entidades con métrica F1 que consiste en determinar la cantidad de entidades que son verdaderas y el número total de entidades que es la suma de entidades verdaderas con entidades falsas y mediante esto se puede determinar la precisión de dicha entidad.
3. Paquetes necesarios para el desarrollo del modelo:
   - Numpy un paquete para la computación científica.
   - Tensorflow una plataforma de código abierto de para el aprendizaje automático.

V. Módulo de determinación de proximidad

El modulo de determinación de proximidad nos permite la comparación de diferentes textos o contenidos y poder determinar que tan parecidos son estos textos, en este caso nosotros utilizaremos los contenidos de los tweets recolectados para identificar que otros diferentes tweets son los más similares entre sí.



A. *Rank candidates*

En esta parte es donde se reciben los tweets recolectados y el que nos da la salida de los tweets mas parecidos entre sí, Aquí es donde con la ayuda de las diferentes herramientas de vectorización de texto y la comparación mediante *Similitud del coseno* podemos determinar lo ya antes mencionado y rankear los diferentes entre si los tweets más similares.

B. *Question to vec*

Aquí se vectorizan los tweets para el correcto análisis futuro mediante métodos cuantitativos con la ayuda de diferentes librerias de Python como lo son Numpy y Gesim:

1) Numpy

Numpy es una librería que nos proporciona una mayor flexibilidad al trabajar con vectores y matrices, esta centrado en la computación científica usando Python. Además, lo podemos usar como un contenedor multidimensional de datos genéricos y así facilitar el análisis de los datos.

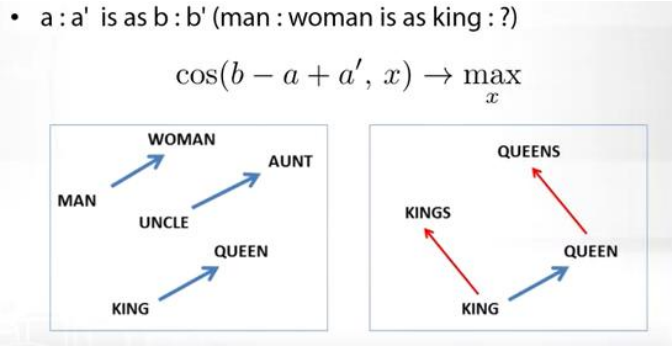
2) Gesim

Gensim es una biblioteca de código abierto para el modelado de temas no supervisados y el procesamiento del lenguaje natural, que utiliza el aprendizaje automático moderno de estadística.

Esta herramienta nos permite la estructuración y formación de diferentes fuentes de datos que se utilizaran para el entrenamiento del modelo. Estas fuentes las sacaremos en este caso de Google de una librería que posee embeddings especiales para el entrenamiento de este modulo

C. Cosine Similarity

La evaluación del conseno nos permite saber mediante métodos cuantitativos que tan similares son dos palabras diferentes



La cual nos permite por ejemplo con los datasets correctos decir si woman es similar a man como también king es similar a queen.

Para esto estamos usando una libreriar de Python que ya posee una implementación de la similitud del coseno que se llama SkLearn.

VI. CONCLUSIONES

- Por medio del modelo de predicción de hashtags se pueden identificar contenidos relacionados con terrorismo con el fin de agilizar el proceso de análisis de información y prevención del delito

- El modelo de identificación de entidades permitiría notar de una manera rápida las organizaciones o personas que están siendo mencionadas por parte de un conjunto de fuentes de información consideradas de interés

- El modelo de identificación de entidades es necesario empezar a crear un procesamiento de texto netamente relacionado a terrorismo para que el entrenamiento permita en el momento de hacer la prueba tener una mayor precisión y evitar falsos positivos.

- El modelo de identificación de proximidad permite descubrir nuevas fuentes de información que inicialmente no se habían considerado de interés por parte de un analista de inteligencia militar