Noticias CNN: Pandemia por COVID-19

Labastida Tolalpa Diego Octavio
11 de Abril 2022

Contexto

Actualmente vivimos una época difícil para todos, estamos en una pandemia que ha afectado a todos los sectores de la sociedad. Los problemas no solo se tratan de salud, también hay conflictos económicos, laborales, sociales y de otra índole.

Hoy afortunadamente contamos con el internet que nos mantiene actualizados e informados, pero también esto provoca casos de sobre información, es por eso que no debemos de perder de vista los medios de comunicación oficiales radio, periódicos y televisión. No todo lo que vemos en redes sociales es información verídica o con una investigación de respaldo como lo hacen los periodistas con experiencia.

La pandemia ha sido un tema de interés por todo el mundo y se ha visto envuelto en teorías de conspiración, noticias falsas, noticias alarmantes y muchas cosas más. Analizar todas esas posibles noticias podría darnos datos de comportamiento de la sociedad en medio de una pandemia o situación de estrés.

Descripción

El dataset adquirido por medio de la técnica de Web Scraping contiene información de las noticias difundidas en el canal “CNN en Español” dirigido principalmente a Latinoamérica, el caribe y el público hispanoparlante de Estados Unidos.

Las noticias son de la pandemia actual por la enfermedad Covid-19 o el virus Coronavirus, para esto hemos utilizado una lista de palabras clave relacionadas con la enfermedad para la búsqueda de noticias de interés. Estas palabras fueron:

- covid-19
- coronavirus
- sars-cov-2
- vacuna
- pandemia

Representación Gráfica

Figura 1, Noticia de último minuto

Contenido

Cada registro contiene la información más importante de las últimas noticias sobre casos de feminicidio en latinoamérica,

- **Clave**: Palabra clave para la búsqueda de noticias en la WEB de CNN en Español
- **Titulo** : Titulo principal de la noticia en CNN en Español
- **Autor** : El autor de la noticia
- **Título** Búsqueda : Título que se visualiza en la página de búsqueda, este no siempre coincide con el título original.
- **Etiqueta** : Es la palabra clave con la que se puede identificar la noticia como una categoría o clasificación de la misma. Ejemplo: Argentina, México o Crimen
- **Fecha** : Fecha de la publicación de la noticia en un formato día, hora y año. Ejemplo 11 abril 2022
- **Hora** : Hora de la publicación de la noticia en un formato hora y minuto GMT. Ejemplo: 15:00
- **Título Imagen** : La búsqueda contiene una imagen principal con la que se puede identificar la noticia en primera instancia, en esta columna almacenamos el título.
- **Imagen** URL: URL de la imagen principal.
- **Noticia** URL: URL de la noticia
- **Contenido** : Contenido o texto completo de la noticia en CNN en Español

La información de esta base de datos como se ha mencionado antes está formada por la búsqueda de 5 palabras clave relacionadas con la enfermedad Covid-19. Es importante mencionar que la programación realizada puede ser modificada para realizar la búsqueda de un tema de interés y no solo para el tema principal de esta práctica.

Agradecimientos

Los datos se han recolectado desde la página del noticiero “CNN en Español”. Les agradecemos las noticias publicadas que mantienen informado a todo el público y un especial agradecimiento a todos los autores de cada publicación que han realizado un gran trabajo investigando y escribiendo sobre los temas de actualidad.

Con el fin de no incumplir con los derechos de autor de las noticias o datos capturados es de importancia mencionar que la recolección de datos es para investigación académica y sin fines de lucro. También se a consultado y seguido el protocolo de exclusión de robots de la página web (https://cnnespanol.cnn.com/robots.txt) quien permite el acceso de cualquier agente y sólo excluye de acceso la ruta “/wp-admin/”.

Al realizar una búsqueda en Zenodo encontraremos dataset similares a la que se presenta en este documento. *(https://zenodo.org/record/4722470#.YlNZg3VByV6) “COVID-19 Real News Data of CBC news”* es un dataset de noticias de la emisora pública canadiense de radio y televisión “CBC/Radio Canada” y asi como este podemos encontrar otros estudios como “Covid-19 News Dataset Both Fake and Real” y “Covid Fake News Dataset”

Inspiración

En el siglo 20 antes de la revolución digital nuestros padres, abuelos o hasta algunos de nosotros vivimos en una época de desinformación por los pocos medios de difusión que existían en ese momento. Hoy en día vivimos el extremo opuesto una época de desinformación por la cantidad excesiva de datos, lo que ha creado un concepto llamado Fake News o en español Noticias Falsas que se ven originadas por las filtraciones de los medios o difusión de videos, fotografías y otro tipo de media por personas poco o nada calificadas.

Esto me ha inspirado como a los creadores de datasets antes mencionados, investigar sobre cómo identificar las Fake News y la evolución de estas. Muchas veces podemos notar patrones de escritura o aumento de noticias en fechas clave de interés político o social como distracción para el público.

Dicho lo anterior, ¿Podremos usar algoritmos de Lenguaje Natural para identificar patrones de escritura?, ¿Usar modelos supervisados o no supervisados para identificar cambios o evolución en la cantidad de noticias publicadas en fechas de interés público? O ¿Podríamos identificar noticias duplicadas o sin valor agregado?

Licencia

La licencia utilizada para la publicación del dataset fue Creative Commons Attribution 4.0 International o CC BY 4.0. Esta licencia es seleccionada por la conveniencia de los términos y condiciones que prestan al trabajo realizado.

Como se menciona en la **Sección 3.- Condiciones de la Licencia**, inciso **a.Atribución / Reconocimiento** en el siguiente link:

https://creativecommons.org/licenses/by/4.0/legalcode.es

Código

Repositorio GitHub con código fuente del Scraping realizado.

https://github.com/diegooclato/uoc-TyCD

Dataset

https://zenodo.org/record/6440655#.YlOKgXVByV5

Fecha de Publicación

11 de Abril del 2022

Digital Object Identifier

10.5281/zenodo.6440655

Recursos

Recursos utilizados para la realización de la práctica:

- Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.
- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
- Tutorial de Github https://guides.github.com/activities/hello-world.