Polo del Conocimiento

Pol. Con. (Edición núm. 63) Vol. 6, No 11
Noviembre 2021, pp. 354-369
ISSN: 2550 - 682X
DOI: 10.23857/pc.v6i11.3268

Comentarios de Instagram extraídos a una base no-relacional para uso en Tecnologías del Lenguaje Humano

Instagram comments extracted to a non-relational database for use in Human Language Technologies

Comentários do Instagram extraídos em uma base não relacional para uso em tecnologias de linguagem humana

César Espin-Riofrio [I]
cesar.espinr@ug.edu.ec
https://orcid.org/0000-0001-8864-756X

Verónica Mendoza-Morán [II]
veronica.mendozam@ug.edu.ec
https://orcid.org/0000-0001-7520-3505

Jorge L-Charco [III]
jorge.charcoa@ug.edu.ec
https://orcid.org/0000-0002-0099-0345

Correspondencia: cesar.espinr@ug.edu.ec

Ciencias Técnicas y Aplicadas
Artículo de investigación

*__Recibido:__ 30 de Septiembre de 2021 *__Aceptado:__ 31 de Octubre de 2021 __* Publicado:__ 11 de Noviembre de 2021

I. Magister en Sistemas de Información Gerencial, Universidad de Guayaquil, Guayaquil, Ecuador.
II. Magister Universitario en Software y Sistemas, Universidad de Guayaquil, Guayaquil, Ecuador.
III. Magister en Inteligencia Artificial, Reconocimiento de formas e Imagen Digital, Universidad de Guayaquil, Guayaquil, Ecuador.

Resumen

Hoy en día se puede acceder fácilmente a mucha información a través de Internet. Las aplicaciones de redes sociales proporcionan al usuario funciones sencillas para compartir y publicar información y, a su vez, esto permite a muchas instituciones conocer la opinión sobre un determinado tema o producto. El objetivo de este trabajo es investigar herramientas para la extracción de comentarios de la red social Instagram y hacer pruebas verificando su eficacia en la creación de un dataset. La metodología aplicada es la diagnóstica bibliográfica tomando como referencia artículos científicos sobre Tecnologías del Lenguaje Humano (TLH) y extracción de datos de redes sociales, identificando los factores en común de los artículos así como herramientas y procedimientos usados para extracción y almacenamiento. Se analizaron diversos documentos científicos sobre el tema logrando determinar herramientas de extracción de texto de Instagram así como verificar su eficacia realizando pruebas de extracción y almacenamiento usando Python y MongoBD como base no relacional. Es posible extraer texto publicado en la red social Instagram y llevarlo a una base de datos no relacional para formar un corpus o dataset que pueda ser analizado en tareas de TLH.

Palabras Clave: Instagram; Corpus; Tecnologías de Lenguaje Humano.

Abstract

Today, a lot of information is easily accessible through the Internet. Social networking applications provide the user with simple functions to share and publish information and, in turn, this allows many institutions to know the opinion about a certain topic or product. The objective of this work is to investigate tools for the extraction of comments from the social network Instagram and to test their effectiveness in the creation of a dataset. The methodology applied is the bibliographic diagnosis taking as reference scientific articles on Human Language Technologies (HLT) and data extraction from social networks, identifying the common factors of the articles as well as tools and procedures used for extraction and storage. Several scientific papers on the subject were analyzed managing to determine Instagram text extraction tools as well as to verify their effectiveness by performing extraction and storage tests using Python and MongoBD as a non-relational base. It is possible to extract text published on the Instagram social network and take it to a non-relational database to form a corpus or dataset that can be analyzed in TLH tasks.

Keywords: Instagram; Corpus; Human Language Technologies.

Resumo

Hoje, muitas informações são facilmente acessíveis pela Internet. Os aplicativos de redes sociais oferecem ao usuário funções simples para compartilhar e publicar informações e, por sua vez, permite que muitas instituições conheçam a opinião sobre um determinado tema ou produto. O objetivo deste trabalho é investigar ferramentas para a extração de comentários da rede social Instagram e realizar testes verificando sua eficácia na criação de um conjunto de dados. A metodologia aplicada é o diagnóstico bibliográfico tendo como referência artigos científicos em Tecnologias da Linguagem Humana (TLH) e extração de dados de redes sociais, identificando os fatores comuns dos artigos, bem como as ferramentas e procedimentos utilizados para extração e armazenamento. Vários documentos científicos sobre o assunto foram analisados, conseguindo determinar as ferramentas de extração de textos do Instagram, bem como verificar sua eficácia por meio da realização de testes de extração e armazenamento usando Python e MongoBD como base não relacional. É possível extrair textos publicados na rede social Instagram e levá-los a um banco de dados não relacional para formar um corpus ou conjunto de dados que podem ser analisados em tarefas TLH.

Palavras-chave: Instagram; Corpus; Tecnologias da linguagem humana.

Introducción

Internet ha permitido que emerjan opciones a las encuestas tradicionales ya que a través de las redes sociales brinda un conglomerado de información que engloba desde los gustos de las personas hacia cualquier tipo de producto como la satisfacción que ellos proveen. Las discusiones de las personas por medio de la Web 2.0 son una fuente de información relevante para las empresas debido a la gran cantidad de datos que es generada a partir del contenido que es publicado cada día (Li et al., 2020).

Sin embargo, la información se presenta en forma de línea de tiempo o feed, que a veces no es relevante para el usuario o es bastante difícil de acceder por el usuario debido a la redundancia de la información (Dewi et al., 2019)

Sin un dataset debidamente administrado, la gran cantidad de información recolectada por las diferentes herramientas existentes en la actualidad no estaría debidamente ordenada y categorizada dando paso a posibles errores en su posterior análisis a través del uso de TLH.

Este proyecto busca analizar alternativas para la extracción de comentarios de la red social Instagram y de cómo crear un corpus o dataset llevado a una base no relacional. Dicha problemática se da ya que las empresas necesitan un dataset de datos para analizar e identificar las preferencias de los usuarios, tendencias, nichos de mercado que permitan popularizar la marca de sus productos. Se mostrará también cómo representar dicho corpus mediante un modelo de base de datos no relacional que permitirá facilitar el análisis y la comprensión de la información extraída para que pueda ser utilizada como base de conocimiento para el uso en Tecnología de Lenguaje Humano (TLH).

Metodología

Se utiliza la metodología de investigación deductiva y bibliográfica. Se puntualizan las herramientas empleadas para la recolección de información, se consideran los artículos o documentos científicos de revistas de alto impacto, elegidos empleando las palabras clave como "Minería de opiniones", "Parse", "TLH y sus aplicaciones" y "Herramientas para extracción de comentarios".

Según (Ocampo, 2019), paradójicamente, en los tiempos de ahora, el gran reto de buscar información no está en lograr hallarla, sino, en organizar el material encontrado y tener la sinceridad necesaria para poder asegurarnos de su validez. Esta clase de investigación llega a obtener gran relevancia en el proceso investigativo, por motivo de que, para efectuar una investigación, sea con punto de vista cuantitativo o cualitativo, la investigación bibliográfica tiene que estar presente.

Se definió como objetivo un total de 46 artículos científicos de las siguientes bibliotecas virtuales: Scopus, Scielo, Springer, ScienceDirect, como se observa en la Tabla 1.

El subconjunto de la población objetivo se obtuvo mediante la búsqueda de artículos que contengan embebidos temas relacionados con corpus, explotación de información, TLH, utilización de una API con integración entre Python y base de datos no relacionales.

Los instrumentos para la obtención de la información o datos empíricos, fue realizada mediante la lectura y el análisis del contenido de cada artículo seleccionado en las diferentes bibliotecas virtuales.

Tabla 1 Cantidad de artículos revisados según bases bibliográficas

Población	**N°**
Scielo	15
ScienceDirect	25
SpringerLink	3
Scopus	3
Total	**46**

Durante el proceso de investigación se registraron las citas bibliográficas de los artículos de las diferentes fuentes de información, en la Tabla 2 se presenta una muestra de la recolección de información para su posterior análisis.

Tabla 2 Muestra de artículos seleccionados

N°	Nombre del artículo
1	Web scraping technologies in an API world (Glez-Pen‹ et al., n.d.)
2	Social Media Web Scraping using Social Media Developers API and Regex. (Dewi et al., 2019)
3	Mining user preferences, page content and usage to personalize website navigation (Flesca et al., 2005)

Para efectos prácticos se realizó un análisis específico de las herramientas y técnicas utilizadas para la extracción de información y tratamiento, se consideraron 2 artículos por indexación de las bibliotecas virtuales de los cuales se menciona:

(Dewi et al., 2019) utilizó el método de extracción de información web scraping implementado por Facebook Developers API y Twitter Developers API. La información extraída se comparó

con preferencias del usuario mediante el uso de expresiones regulares (o Regex), que es una construcción de lenguaje que se puede utilizar para hacer coincidir el texto usando algunos patrones. Este proceso inició por la elección de la red social ya que es necesario token de seguridad de la red social para las peticiones de información en el proceso determinando la cantidad de datos a analizarse estableciendo rangos mínimos y máximos.

(Fernandes et al., 2020) analizó la expresión de la intolerancia racial en Facebook. El proceso de elección se llevó a cabo seleccionando 5 sitios abiertos, entre páginas y grupos, en el sitio de Facebook mediante el descriptor "racismo" y términos relacionados. De estos, se recopilaron comentarios de las 5 publicaciones más relevantes. Los datos se transcribieron con el fin de componer un corpus textual que se analizó mediante el software Iramuteq que reproduce el método de clasificación descrito por (Reinert, 1998) (clasificación jerárquica descendente en una tabla cruzando formas completas y segmentos de texto) y análisis estadísticos en corpus de texto y en tablas individuales de caracteres.

Por otro lado, en (Ichau et al., 2019), se utilizó un API de Instagram e integración con librerías de Python que permitieron analizar las representaciones en red de los judíos y el judaísmo en las redes sociales, también se analizaron las redes de co-ocurrencia para examinar patrones en los hashtag que los usuarios publicaban al pie de las imágenes, la muestra final utilizada fue un conjunto de datos de 1500 publicaciones de Instagram marcadas con los hashtags #jew , #jewish y #jews, y combinamos análisis de contenido cualitativo y análisis de redes de co-ocurrencia para explorar temas y prácticas de representación en el contenido de Instagram sobre judíos y judaísmo.

El preprocesamiento de los datos obtenidos se realizó en tres fases. 1), depuración de publicaciones que contenían datos duplicados, de este proceso se identificaron y eliminaron 6744 URL, 2) Para evitar una distorsión de los datos, se eliminó todas las secuencias de hashtag duplicadas siguiendo el mismo procedimiento, 3) Se utilizó el método aleatorio para extraer una muestra simple de los hashtags recopilados. Los métodos de extracción utilizados fueron local browser, local software, online, copiado, los resultados fueron.

(Canós, 2017), como parte del proyecto de grado, desarrolló un sistema de seguimiento de los posts que publican los usuarios en Instagram. El sistema daba seguimiento por etiqueta de post para luego, en un proceso paralelo, realizar el análisis del contenido de los comentarios

extraídos. Se utilizó la Instagram API platform con la cual se inicia sesión en la aplicación por medio de la generación de un token único que brinda Instagram para poder recibir peticiones.

Se utilizó el sistema sentiment analysis que permite extraer del lenguaje la opinión subjetiva subyacente del usuario, Python por el dinamismo y sencillez en su estructura de código fuente y MongoDB por su capacidad de almacenar datos en formato Json. Como resultado del sistema se lograron capturar opiniones por medio de publicaciones que contenían una etiqueta específica o por usuario midiendo el nivel de likes obtenidos por publicaciones.

Para el resultado de la investigación referente a las API más utilizadas para la captura automatizada de información se tomó en consideración el top 5 de acuerdo con lo indicado por la página oficial Octopus Data Inc.

El trabajo de (Hamada & Naizabayeva, 2020), desarrolló un sistema de apoyo a la decisión basado en el algoritmo de agrupación de K-means que realiza una clasificación minimizando la suma de distancias entre cada objeto y el centroide de su grupo o cluster. Se suele usar la distancia cuadrática de objetos en k grupos basándose en sus características para detectar la ubicación óptima de la tienda a través de eventos de redes sociales. También explica cómo extraer datos de un canal de red social "Instagram" utilizando la "API de Octoparse" como herramienta de extracción de datos web, esto fue realizado mediante el uso de su configuración incorporada de Regex y XPath para localizar elementos con precisión. Como resultado analizaron 12754 publicaciones iniciadas el 1 de enero de 2019 que fueron depurados utilizando algoritmos Minimax y K-means en formato json con centros que se colocaron en un mapa para proporcionar una mejor comprensión.

Cyotek WebCopy es otra herramienta utilizada para la extracción/descarga del contenido de un sitio web, (Villabona et al., n.d.) caracterizó el contenido de sitios web de agencias turísticas a través de scraping y minería web para contribuir a la satisfacción de turistas, utilizando una metodología cualitativa con diseño no experimental, transeccional, descriptivo cualitativa, porque no se manipuló la información consultada ni se procesaron datos medibles. Como resultado se obtuvo un registro de 77 agencias de viajes certificadas, de las cuales 80,5% cuenta con un sitio web y con registro nacional de turismo activo, estas técnicas scraping y minería web permiten a las agencias turísticas reducir dificultades considerando aspectos técnicos y publicitarios.

ScrapingHub es un sistema de escaneo web configurado con inteligencia artificial y detección de objetos de imagen utilizado por (KOROBOV & LOPUKHIN, 2020). Este sistema procesa una página web con una red neuronal para realizar la localización de objetos a obtener datos estructurados, incluidos imágenes, texto y otros tipos de datos, de páginas web. La red neuronal tolera que el sistema procese de manera eficaz información visual, estructura HTML y contenido de texto para lograr buena calidad y disminución del tiempo de extracción. Es un software que sustrae información estructurada de la web de manera automática se maneja de forma convencional creando rastreadores personalizados (arañas) para cada sitio web que se indaga utilizando reglas especificadas manualmente. Es creador y mantenedor del marco de código abierto más popular para crear estas arañas (scrapy).

La investigación de (Milev, 2017) fue orientada a los problemas del desarrollo de aplicaciones de web scraping. El artículo se basó en el web scraping como parte de la minería de datos donde trató algunos posibles enfoques del web scraping en términos del desarrollo de soluciones de software, también hizo una descripción general de la investigación sobre web scraping y las características funcionales de varias soluciones tradicionales para web scraping que han demostrado ser exitosas, el trabajo muestra algunas ventajas funcionales de la concepción propuesta sobre las soluciones de software tradicionales en el campo.

Outwit Hub es una extensión de Firefox que se puede descargar fácilmente desde la tienda de complementos de Firefox. Esta herramienta sirve para navegar automáticamente por las páginas y almacenar la información extraída en un formato adecuado, también ofrece una única interfaz para extraer cantidades pequeñas o enormes de datos según las necesidades.

Web Scraper es una alternativa a Outwit Hub y es una extensión de Google Chrome, que se puede utilizar para web scraping. Puede extraer información de varias páginas simultáneamente e incluso tiene capacidades de extracción de datos dinámicas. Web Scraper también puede manejar páginas con JavaScript y AJAX. La desventaja de esta solución es que no tiene muchas funciones de automatización integradas.

Por otro lado en (Feed RSS: ¿qué Es, Para Qué Sirve y Cómo Crear Uno?, n.d.), se indica que Spinn3r es una aplicación para extraer datos completos de blogs, sitios de noticias, redes sociales y feeds RSS. Puede filtrar los datos que extrae utilizando palabras clave, lo que ayuda a eliminar el contenido irrelevante. Spinn3r funciona escaneando continuamente la web y

actualizando sus conjuntos de datos. Tiene una consola de administración repleta de funciones que pueden realizar búsquedas en los datos sin procesar.

Extracción de datos

Para recolectar los datos se realizó observación de los intereses mencionados en los artículos recogidos en la investigación como se muestra en la Tabla 3

Tabla 3 Registro de observación

Palabras Claves	Intereses	Año de publicación
Corpus, Instagram, TLH, web scraping, redes sociales, minería de opinión, Python, base de datos no relacionales.	Tratamiento de información obtenida de la extracción de comentarios de redes sociales. Conocer la utilización de los datos no estructurados de un corpus.	Rango entre 2015 a 2020

Para el desarrollo de la extracción se consideró utilizar las herramientas utilizadas en el trabajo realizado por (Dewi et al., 2019) el cual consistió en la extracción de comentarios de la red social Facebook utilizando el lenguaje de programación Python y el empleo de Regex que es una secuencia de caracteres que conforma un patrón de búsqueda (Canós, 2017). Se utilizan principalmente para la búsqueda de patrones de cadenas de caracteres u operaciones de sustituciones.

Para usar la API de Instagram, se creó una cuenta de Instagram, luego en la página oficial de Facebook sitio web de desarrolladores plataforma de Instagram para registrar nuestra aplicación y adquirir la AplicaciónID (AppID) y token de la aplicación (AppSecret). Esta identificación se utiliza más en el desarrollo del web scraping para enviar varias solicitudes de datos a Instagram y a través del AppSecret que es utilizado para decodificar el cifrado de mensajes para que la información confidencial permanezca protegida.

El proceso de web scraping comenzó con la elección de la fuente (Instagram) ya que el API developers sirve también para Facebook, esto influye en el método de autenticación que recepta cada red social, la Figura 1 presenta un ejemplo del código fuente utilizado para autenticación para web scraping para Instagram.

```
user    =   'info.contigo.seguro'
passw   =   '#PWD_INSTAGRAM_BROWSER:0:1612460872:contigo2021'

# Despues descomentar
#session,head,csrftoken = login_Instagram_Session(user,passw)
#h={'X-CSRFToken':csrftoken}
#headers.update(h)

#Get Comments
re = getComments('2c4c2e343a8f64c625ba02b2aa12c7f8',"CK68oOvLSjv")
```

Figura 1 Autenticación API Instagram

Es necesario un inicio de sesión (login) como lo solicita la API de Instagram para lo cual se envía la contraseña y el usuario proporcionado en request.py

Se descargan e instalan las librerías necesarias para la implementación: el paquete pip install request, librería pymongo.

Al iniciar sesión, Instagram devuelve un código único que debemos poner en la cabecera de la petición http en la cookie que se llama CSRFToken, cabe recalcar que sin este token de autenticación no podríamos aplicar el web scraper.

Se crea la conexión hacia la base de datos con nombre <Instagram> y la colección “comments” de manera local.

```
import pymongo


def saveCollection(payload):
    try:
        db_name = 'instagram'
        db_collection = 'comments'


        myclient = pymongo.MongoClient("mongodb://localhost:27017/")
        db = myclient[db_name]

        col = db[db_collection]
        x = col.insert_one(payload)

        print(x.inserted_id)
        return True

    except Exception as e:
        print("saveCollection: "+str(e))
        return False
```

Figura 2 Arreglo MongoDB

Resultados

Creada la conexión hacia la base de datos se procede a consumir la URL del API de Instagram y el envío de solicitudes de información hacia la aplicación. Obtenida esa respuesta se presentará el total de comentarios contenidos en la publicación y el total de comentarios que presentaron coincidencia con los criterios ingresados para la extracción, se presenta el resultado de una extracción que contenía 781 comentarios de los cuales solo 90 cumplían con el criterio de extracción <<Arauz>>, se muestra en las Figuras 3 y 4.

```
#Save Mongo
dic          = re
process      = 0
errors       = 0
count        = dic['count']
parent       = 'Arauz'
```

Figura 3 Búsqueda por criterio

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL
6023755993f3d1292fe180c6
6023755993f3d1292fe180c8
6023755993f3d1292fe180ca
6023755993f3d1292fe180cc
6023755993f3d1292fe180ce
6023755993f3d1292fe180d0
6023755993f3d1292fe180d2
6023755993f3d1292fe180d4
count= 781
process= 90
errors= 0
```

Figura 4 Resultados de extracción

Una vez contamos con el total de los comentarios, estos se proceden a almacenar en la base de datos la cual hemos llamado <Instagram> como se aprecia en la Figura 5.

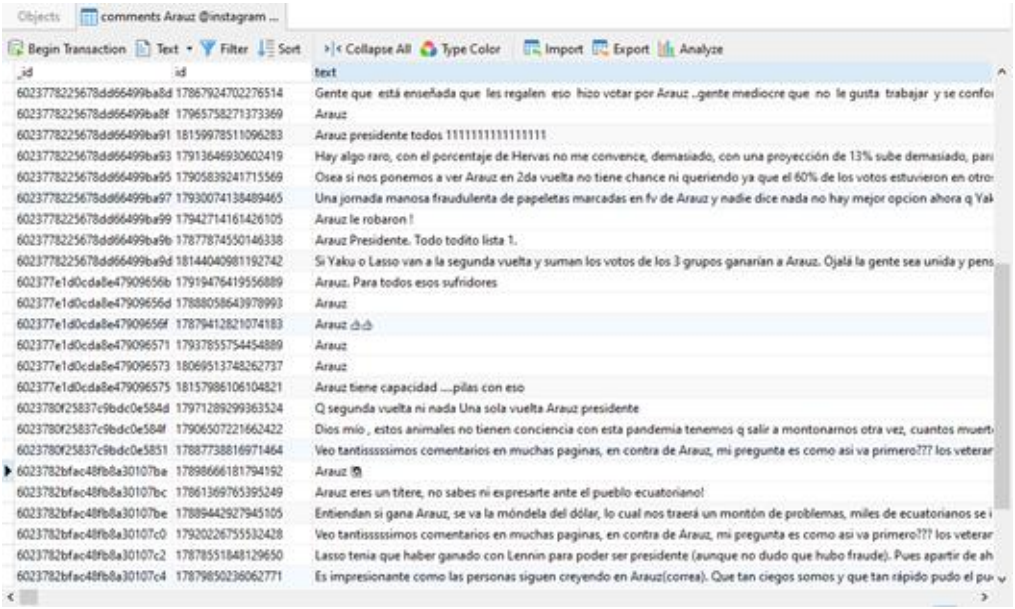


Figura 5 Almacén MongoDB

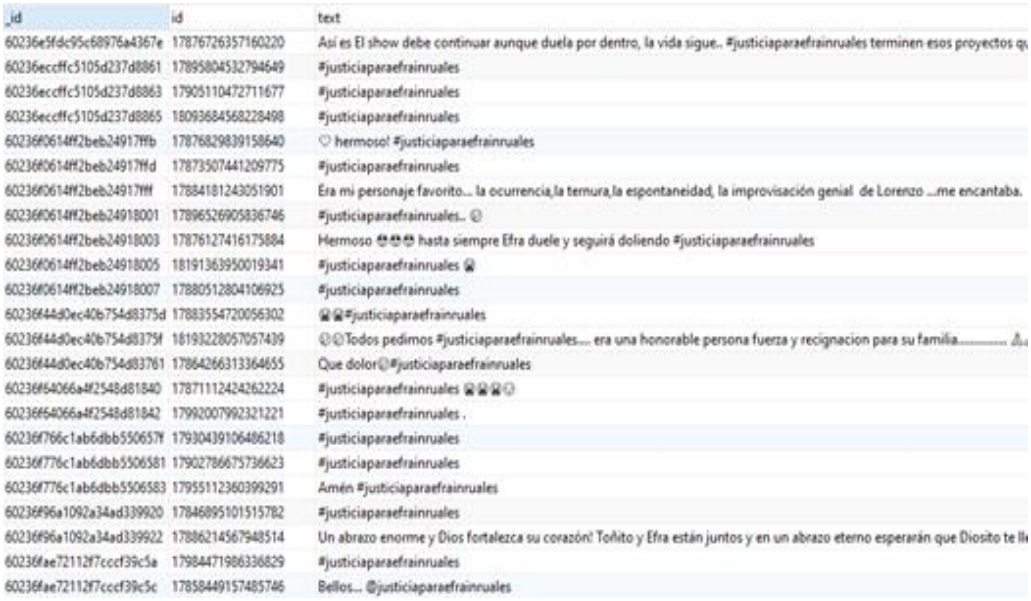
Otra de las formas de extracción que permite el prototipo desarrollado es a nivel de publicación sin el ingreso de una expresión específica, es decir se almacenan todos los comentarios contenidos en una publicación para su análisis, donde escogemos la publicación de Instagram, para este ejemplo se debe elegir un <usuario>

Se obtuvo un total de 1012 comentarios extraídos con éxito en su totalidad con diversas opiniones.

```
6023701d5185c42efcc4faf9
6023701d5185c42efcc4fafb
count= 1012
process= 1012
errors= 0
```

Figura 6 Extracción sin indicar criterio

En la Figura 7 se tiene una vista de MongoDB con el resultado de la extracción sin utilizar criterio.



_id	id	text
60236e5fdc95c68976a4367e	17876726357160220	Así es El show debe continuar aunque duela por dentro, la vida sigue.. #justiciaparaefrainruales terminen esos proyectos q
60236eccffc5105d237d8861	17895804532794649	#justiciaparaefrainruales
60236eccffc5105d237d8863	17905110472711677	#justiciaparaefrainruales
60236eccffc5105d237d8865	18093684568228498	#justiciaparaefrainruales
60236f0614ff2beb24917ffb	17876829839158640	♡ hermoso! #justiciaparaefrainruales
60236f0614ff2beb24917ffd	17873507441209775	#justiciaparaefrainruales
60236f0614ff2beb24917fff	17884181243051901	Era mi personaje favorito... la ocurrencia,la ternura,la espontaneidad, la improvisación genial de Lorenzo ...me encantaba.
60236f0614ff2beb24918001	17896526905836746	#justiciaparaefrainruales..
60236f0614ff2beb24918003	17876127416175884	Hermoso 😍😍😍 hasta siempre Efra duele y seguirá doliendo #justiciaparaefrainruales
60236f0614ff2beb24918005	18191363950019341	#justiciaparaefrainruales
60236f0614ff2beb24918007	17880512804106925	#justiciaparaefrainruales
60236f44d0ec40b754d8375d	17883554720056302	#justiciaparaefrainruales
60236f44d0ec40b754d8375f	18193228057057439	Todos pedimos #justiciaparaefrainruales.... era una honorable persona fuerza y recignacion para su familia.....
60236f44d0ec40b754d83761	17864266313364655	Que dolor#justiciaparaefrainruales
60236f64066a4f2548d81840	17871112424262224	#justiciaparaefrainruales
60236f64066a4f2548d81842	17992007992321221	#justiciaparaefrainruales .
60236f766c1ab6dbb550657f	17930439106486218	#justiciaparaefrainruales
60236f776c1ab6dbb5506581	17902786675736623	#justiciaparaefrainruales
60236f776c1ab6dbb5506583	17955112360399291	Amén #justiciaparaefrainruales
60236f96a1092a34ad339920	17846895101515782	#justiciaparaefrainruales
60236f96a1092a34ad339922	17886214567948514	Un abrazo enorme y Dios fortalezca su corazón! Tofito y Efra están juntos y en un abrazo eterno esperarán que Diosito te ll
60236fae72112f7cccf39c5a	17984471986336829	#justiciaparaefrainruales
60236fae72112f7cccf39c5c	17858449157485746	Bellos... @justiciaparaefrainruales

Figura 7 Resultado sin usar criterio en MongoDB

Información obtenida de esta forma puede ser utilizada para estudios de mercado en donde podrían evaluar la opinión sobre un tema o producto en específico. Se realizaron varios escenarios de extracción que a continuación se presenta.

Tabla 4 Temáticas usadas para extracción

Nro. Extracción	Temática	Expresión	Cantidad de comentarios	Comentarios procesados
1	Interés Social	justicia para Efrain Ruales	1012	212
2	Política	Arauz	781	90
3	Deportes	Messi	2000	150
4	Moda	gucci prada	300	115
6	Elecciones	Fraude	800	230
7	Sentimientos	Amor	420	52
8	Pandemia	Covid	487	6
9	Aborto	legalizar el aborto	546	254
10	Comedia	Monólogo	600	145

Se identificó un alto índice de uso de la API Facebook developers utilizada para la extracción de información que es publicada en la red social Instagram, el uso del lenguaje Python y Regex

para el tratamiento y clasificación de esta información se la utiliza para identificar comentarios suicidas por jóvenes a través de palabras claves en sus publicaciones, identificar el nivel de racismo a través de los hashtags en Instagram, etc.

Discusión

La información publicada y disponible en redes sociales va aumentando sustancialmente con el tiempo y es necesario contar con herramientas que extraigan automáticamente dicho texto de Instagram u otras redes sociales como Twitter, Facebook, etc.

Contando con un dataset se pueden emprender algunas tares relacionadas con las TLH y el Procesamiento del Lenguaje Natural PLN entre las que están el perfil de autor, análisis de sentimientos, preferencias, etc.

El uso de los Transformers se propone para entrenar modelos que resuelven las tareas mencionadas y muchas más en el propósito de analizar el texto escrito extraído de redes sociales u otras fuentes.

Conclusiones

Analizadas las diferentes herramientas para la extracción de comentarios de la red social Instagram, tomando como referencia artículos científicos y lo investigado en el presente proyecto sobre Tecnologías del Lenguaje Humano, se determinó las herramientas usadas para realizar Web Scraping haciendo uso demostrativo de ellas.

Extraídos los comentarios de la red social Instagram, se almacenó en una base no relacional creada en el gestor de base de datos MongoDB, se formó un dataset de prueba con toda la información almacenada para su posterior análisis de las distintas variables del español utilizando Tecnologías del Lenguaje Humano.

El corpus de datos recopilado es de suma importancia para un posterior análisis en Tecnologías del Lenguaje Humano, independiente de las diferentes áreas o temas a escoger, debido a que la gran cantidad de información recopilada en el corpus puede someterse a diferentes estudios, como por ejemplo estudio de mercados.

Referencias

1. Canós, J. S. (2017). Desarrollo de un sistema de seguimiento para Instagram. 1–33. https://riunet.upv.es/handle/10251/87106
2. Dewi, L. C., Meiliana, & Chandra, A. (2019). Social media web scraping using social media developers API and regex. Procedia Computer Science, 157, 444–449. https://doi.org/10.1016/j.procs.2019.08.237
3. Feed RSS: ¿qué es, para qué sirve y cómo crear uno? (n.d.). Retrieved April 28, 2021, from https://rockcontent.com/es/blog/feed-rss/
4. Fernandes, S., Nascimento, M., Pereira, A., Melo, E., & Carlos, K. (2020). RELAÇÕES RACIAIS NO FACEBOOK: ANÁLISE DE COMENTÁRIOS ACERCA DE CONTEÚDOS RACIAIS DIGITAIS (pp. 317–329). https://doi.org/10.36367/ntqr.4.2020.317-329
5. Flesca, S., Greco, S., Tagarelli, A., & Zumpano, E. (2005). Mining user preferences, page content and usage to personalize website navigation. World Wide Web, 8(3), 317–345. https://doi.org/10.1007/s11280-005-1315-9
6. Glez-Pen‹, D., Lourenc¸o, L., Lo¤ Pez-Ferna¤ Ndez, H., Reboiro-Jato, M., & Fdez-Riverola, F. (n.d.). Web scraping technologies in an API world. https://doi.org/10.1093/bib/bbt026
7. Hamada, M. A., & Naizabayeva, L. (2020). Decision Support System with K-Means Clustering Algorithm for Detecting the Optimal Store Location Based on Social Network Events. 2020 IEEE European Technology and Engineering Management Summit, E-TEMS 2020, 1–4. https://doi.org/10.1109/E-TEMS46250.2020.9111758
8. Ichau, E., Frissen, T., & d'Haenens, L. (2019). From #selfie to #edgy. Hashtag networks and images associated with the hashtag #jews on Instagram. Telematics and Informatics, 44, 101275. https://doi.org/10.1016/j.tele.2019.101275
9. KOROBOV, M., & LOPUKHIN, K. (2020). SYSTEM AND METHOD FOR A WEB SCRAPING TOOL AND CLASSIFICATION ENGINE.
10. Li, W., Zhou, Q., Ren, J., & Spector, S. (2020). Data mining optimization model for financial management information system based on improved genetic algorithm.

Information Systems and E-Business Management, 18(4), 747–765. https://doi.org/10.1007/s10257-018-00394-4

11. Milev, P. (2017). Conceptual Approach for Development of Web Scraping Application for Tracking Information. Economic Alternatives, 3, 475–485.
12. Ocampo, D. S. (2019). Investigación bibliográfica - Investigalia. Investigalia. https://investigaliacr.com/investigacion/investigacion-bibliografica/
13. Reinert, M. (1998). QUEL "OBJET" POUR UNE "ANALYSE STATISTIQUE DU DISCOURS" ? http://lexicometrica.univ-paris3.fr/jadt/jadt1998/reinert.htm
14. Villabona, N., Garcés, D. J., & Martelo, R. J. (n.d.). Caracterización de contenido de sitios web turísticos mediante scraping y minería web para contribuir a la satisfacción de turista Characterization of content of tourism websites through web scraping and web mining to contribute to tourist satisfaction. 41(36), 2020. Retrieved April 29, 2021, from https://www.revistaespacios.com