ANÁLISIS DE ENTIDADES RELACIONADAS CON EL COVID-19 PARA LA IDENTIFICACIÓN DE TENDENCIAS EN REDES SOCIALES MEDIANTE NLP, CASO DE ESTUDIO TWITTER EN COLOMBIA

FUNDACIÓN
UNIVERSITARIA
DE POPAYÁN
35 ANIVERSARIO

JINNETH XILENA CHAMIZO AGREDO
YAMID ANCISAR QUINTERO NARVAEZ

INFORME FINAL DE SEMINARIO COMO OPCIÓN DE GRADO PARA OPTAR AL TÍTULO DE: INGENIERO DE SISTEMAS

PRESENTADO A:
PHD. JOSÉ ARMANDO ORDOÑEZ

FUNDACIÓN UNIVERSITARIA DE POPAYÁN
FACULTAD DE INGENIERÍA
PROGRAMA INGENIERÍA DE SISTEMAS
GRUPO DE INVESTIGACIÓN IMS
POPAYÁN, AGOSTO DE 2020

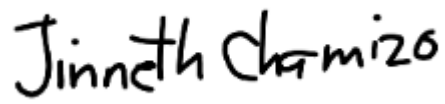
CERTIFICACIÓN DE AUTORÍA

Certifico que conozco el concepto de plagiar según la Real Académica de la lengua(“Copiar en lo sustancial obras ajenas, dándolas como propias.”)

Y certifico que el contenido de este documento son de mi autoría, no hay contenido que haya sido copiado directamente y al pie de la letra de ninguna fuente. En el caso de ideas, teorías, conceptos, resultados y otros contenidos tomados de otros autores se menciona explícitamente la fuente original, y solo en unos pocos casos se ha mantenido el mismo texto, colocándole entre comillas.

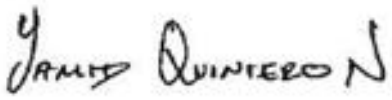
Reconozco las consecuencias académicas, jurídicas, y económicas que conlleva el plagio.

Firma



Jinneth Xilena Chamizo Agredo

CC. 1.061.748.267



Firma

Yamid Ancisar Quintero Narvaez

CC. 1.061.753.003

CONTENIDO

TABLAS

FIGURAS

RESUMEN

Twitter puede identificarse como una de las plataformas más grandes en las redes sociales, un gran número de usuarios hace uso de Twitter como una herramienta o plataforma universal para difundir noticias, compartir artículos y socializar con otras personas a nivel mundial mediante mensajes denominados tweets, por esa razón crecen los estudios relacionados a fin de extraer y analizar la gran cantidad de información que se genera en esta red social. Al realizar análisis de entidades en Twitter se puede obtener/ encontrar tendencias o necesidades sobre un tema en cuestión, en este documento se desarrolla un código en python en el cual se utilizan técnicas de Natural Language Processing para la detección de tendencias en el idioma español, sobre la problemática actual de Covid-19 en Colombia, utilizando API de streaming para obtener Tweets en tiempo real, API de Twitter con el fin obtener las credenciales para extraer los datos de los tweets y las diferentes librerías de python para NLP y para graficar los resultados.

Palabras Claves: Entidades, Tendencias,Twitter, NLP, API, Streaming.

1 CAPITULO I. ASPECTOS GENERALES DE LA INVESTIGACIÓN

1.1. DESCRIPCIÓN DEL PROBLEMA

El análisis de tendencias en las redes sociales ha sido recientemente un importante foco de interés entre los investigadores que los estudian desde perspectivas como necesidades o expresiones. Actualmente las principales fuentes de comunicación son las redes sociales, donde cada día crece su nivel de importancia a nivel mundial por la facilidad de obtener rápidamente información de problemáticas sociales, culturales, ambientales, entre otras. Teniendo en cuenta los diferentes puntos de vista de los usuarios, por esta razón crecen los estudios sobre estas con el fin de extraer y analizar la información. Twitter es un servicio en linea de noticias y redes sociales donde los usuarios publican e interactúan sus mensajes o “tweets” por lo que se convierte en una herramienta eficaz para el análisis de tendencias [1].

Con base a esto se ha relacionado con la problemática actual en salud a nivel mundial Covid-19, los gobernantes de cada país de los diferentes continentes han creado políticas de protección para la población, teniendo en cuenta aspectos económico, educativo, cultural y social, es por ello que se hace necesario identificar y analizar tendencias de entidades de las necesidades de los usuarios activos en Twitter sobre la pandemia.

Actualmente se han realizado diferentes investigaciones y estudios de identificación de tendencias por medio de la extracción de entidades en las diferentes redes sociales, en la mayor parte de las investigaciones utilizaron Twitter para analizar los tweets e identificar tendencias utilizando Interfaces de Programación de Aplicaciones(API) y bibliotecas de Natural Language Processing (NLP) para la creación de algoritmos que ayudan al análisis de texto escrito o estructurado, en lenguaje humano [2].

En Colombia no existen estudios en cuanto a la identificación de tendencias mediante la extracción de entidades en cuanto a necesidades o expresiones relacionadas con el Covid-19 en Twitter, por este motivo es importante desarrollar este tipo de investigación para obtener una herramienta que permita analizar las tendencias de forma rápida y eficiente a partir de la información de las redes sociales.

1.2. FORMULACIÓN DEL PROBLEMA

¿Como analizar tendencias de entidades sobre la problemática del Covid-19 utilizando los tweets en el idioma español de la red social Twitter durante un determinado lapso de tiempo en Colombia?

1.3. OBJETIVOS

1.3..1 OBJETIVO GENERAL

Realizar análisis de tendencias de entidades en Twitter relacionado con el Covid-19 mediante Natural Language Processing en Colombia.

1.3..2 OBJETIVOS ESPECÍFICOS

- Realizar una revisión del estado del arte sobre análisis de las tendencias de entidades y métodos de Natural Language Processing en Twitter.
- Definir un modelo de análisis mediante Natural Language Processing para la identificación de tendencias de entidades en Twitter sobre el Covid-19 en Colombia.
- Implementar una estrategia de socialización de los resultados obtenidos por el modelo de análisis propuesto.

1.4. JUSTIFICACIÓN

De acuerdo con la obra de Roberto Hernández Sampieri [3] de la cual se enfatiza en algunos criterios para evaluar la importancia potencial de una investigación, estos fueron adoptados para justificar la presente investigación:

- **Valor metodológico de la investigación:** en el desarrollo de este trabajo se utilizará el análisis de datos para poder inspeccionar, limpiar, transformar y modelar los datos, con el objetivo de descubrir información útil para el análisis de tendencias de entidades en tweets que reflejan necesidades o pensamientos sobre el Covid-19. Para la evaluación y análisis de esta

problemática se estudiará e implementará un modelo de análisis que nos brinde el mínimo porcentaje de error.

- **Valor práctico de la investigación:** brindar a la comunidad en general la información de la identificación de tendencias de entidades con base al modelo de análisis mediante Natural Language Processing de los usuarios de Twitter en Colombia.
- **Valor Tecnológico:** identificar tendencias de entidades mediante Natural Language Processing utilizando la API de Twitter que permitirá conocer rápidamente la reacción de las personas ante un evento o medida.
- **Valor de emprendimiento e innovación:** implementar un modelo de análisis para la identificación de tendencias de entidades mediante los comentarios que reflejen necesidades o expresiones de usuarios en la red social Twitter sobre el Covid-19, con base a esta investigación se pueden realizar nuevos estudios teniendo en cuenta las diferentes redes sociales que pueden aportar datos reales de una situación o suceso con el fin de analizar tendencias de entidades.

2 CAPITULO II. MARCO REFERENCIAL

2.1. MARCO CONCEPTUAL

Para este apartado se presenta una descripción sobre los principales conceptos, con el fin de tener un mayor conocimiento sobre el presente trabajo investigativo.

2.1..1 Análisis de tendencias:

Proceso de recolección de información con el que se puede determinar comportamientos durante un cierto periodo de tiempo, por medio del tratamiento de dicha información. ”El comportamiento de la información(tendencias) permite tomar decisiones estratégicas ante las vulnerabilidades, amenazas y oportunidades”[4].

2.1..2 Natural Language Processing(NLP):

Natural lenguaje processing o NLP que al español traduce procesamiento de lenguaje natural, es uno de los procesos mas significativos dentro de la Inteligencia Artificial, “basándose en la comunicación entre el hombre y la computadora mediante el lenguaje humano a través de sistemas informáticos, por medio de la voz o del texto” [5]. Los pasos mas comunes para realizar NLP son [6]:

- Tokenización (Tokenize)
- Sentence Splitting (ssplit)
- Part-of-speech Tagging (pos)
- Morphological Analysis (lemma)
- Named Entity Recognition (NER)
- Syntactic Parsing (parse)
- Coreference Resolution (dcoref)
- otros (gender, sentiment)

2.1..3 Datos no estructurados:

Son aquellos datos que no pueden ser guardados de forma relacional al no poseer un tipo de dato fijo, es decir que “no tienen un formato normalizado determinado”. Entre estos datos se encuentran correos electrónicos, imágenes, archivos de audio y video, blogs, chats y demás. [7]

2.1..4 Covid-19

Enfermedad causada por coronavirus que dio su primer brote en la ciudad de Wuhan en China, en diciembre del año 2019 y el cual llega a Colombia registrándose el primer contagio el día 6 de marzo del año 2020[8].

2.1..5 Twitter:

Es una aplicación de microblogging que permite a sus usuarios activos publicar mensajes cortos de hasta 140 caracteres; “En Twitter, los mensajes son públicos por defecto, por lo que los datos son totalmente analizables. Además, cada tweet suele contener opiniones personales de los usuarios, lo cual es bastante interesante en el análisis de datos”[9].

2.2. Estado del arte

considerando los avances de la investigación se utiliza la metodología de mapeo sistemático, que permitiría tener una visión mas amplia acerca de los trabajos relacionados a este proyecto. [10]

2.2..1 Pregunta de investigación

En esta parte se indican preguntas de investigación a las cuales se pretende dar solución en el transcurso del desarrollo de este trabajo.

- ¿Que estudios se han realizado en cuanto a el análisis de tendencias de entidades en Twitter sobre la problemática del Covid-19 en Colombia?

2.2..2 Fuentes de datos y estrategia de búsqueda

En esta investigación se hace uso de motores de búsqueda: bases de datos que pueden contener estudios relacionados con el análisis de tendencias de entidades con NLP, como ELSEVIER, IEEE o EBSCO.

2.2..3 Selección de estudios

En esta sección se definen las cadenas de búsqueda necesarias para determinar si los estudios ya realizados son o no pertinentes.

- **Inclusión:** se tendrán en cuenta aquellos estudios relacionados con el análisis de tendencias de entidades sobre el Covid-19 en la red social Twitter, en periodos comprendidos desde el año 2015 hasta la fecha actual.
- **Exclusión:** aquellos estudios que no tengan relación alguna con el análisis de tendencias de entidades sobre el Covid-19 en la red social Twitter no serán considerados pertinentes para esta investigación.

Cadena de búsqueda	No. de trabajos ELSEVIER	No. de trabajos EBSCO	No. de trabajos IEEE
“Twitter" AND "Natural Language Processing" AND “Colombia”	10	0	1
"Twitter" AND “Trends" AND "analysis" AND "Natural Language Processing" AND “Colombia”	10	0	38
“Twitter” AND “Entity” AND “analysis” AND “Natural Language Processing”	10	1	65

Tabla 1: Cadenas de búsqueda

2.2..4 Clasificación de artículos

Una vez realizada la búsqueda con las cadenas definidas, se continua con la clasificación de estudios y su posterior lectura de en total cuatro artículos relacionados con el tema definido.

2.2..5 Extracción de datos

A continuación se realiza un breve resumen de cada uno de los artículos seleccionados sobre la identificación de tendencias de entidades mediante NLP en Twitter, realizando una tabla donde se especifican titulo, objetivos, algoritmos/técnicas o características y conclusiones.

Articulo		Objetivo	Algoritmos/ técnicas/ características	Conclusiones
[2]	Análisis social aplicando técnicas de lenguaje natural a información extraída de Twitter. **Año: 2019**	Presentar el desarrollo de un modelo para el "análisis y explotación de la información encontrada en redes sociales". en el que utilizan técnicas de procesamiento de lenguaje natural.	**-** Algoritmo de crawling para extraer tweets. **-** PLN para identificar la estructura sintáctica, léxica y semántica. **-T**okenización, lematización, segmentación de palabras e indexación semántica latente - I+D+i -Red social Twitter	- Concluyen que existe bastante información sin analizar al no encontrarse herramientas capaces de lograr un análisis correcto de texto, por ende se sugiere que al existir un algoritmo apoyado en PLN y análisis léxico que se base en el análisis de información en el idioma español ayudara en el análisis de la suficiente información que se encuentra en internet.
[11]	Detección de incidentes de transito en Twitter. **Año: 2016**	Proponen un modelo que al combinar técnicas de machine learning y procesamiento de lenguaje	- stop-words Eliminar estas palabras del texto normalizado. - SVM: Clasificadores Support Vector Machine.	Con este sistema se logró identificar incidentes en el 85.71% de los casos y erróneamente sólo el 0.33 % de las publicaciones. "Estos resultados demuestran la viabilidad de utilizar Twitter como fuente de información para la detección de incidentes de

		natural ayuden en la identificación de incidentes de transito publicado en la red social Twitter.	- SMO: Sequential Minimal Optimization. -Naive Bayes.	tránsito".
[12]	Efficient Natural Language Pre-processing for Analyzing Large Data Sets **Año: 2016**	"se propone una canalización de pre-procesamiento para tweets que consiste en filtrar parte del discurso, reconocimiento de entidades con nombre, segmentación de hashtag y desambiguación	- teoría de grafos - palabras grupales de tweets que usan relaciones semánticas de WordNet. -Tweet Normalization -Hashtag Segmentation -NER: Named Entity Recognition -Twitter API -Principal Component Analysis (PCA).	-"La canalización de pre-procesamiento propuesta integra varios procesos de PNL que ayudan a analizar los grandes datos para cualquier aplicación de aprendizaje automático y de PLN".

[5]	Analítica de datos en Twitter. **Año: 2015**	descubrir la rivalidad entre las marcas Adidas y Nike en la red social de Twitter. utilizando técnicas de Machine Learning y al extraer tendencias a con el uso de métodos de NLP	-Lenguaje de programación Python. - Twitter API -Machine learning y NLP. - K-means - Algoritmos ML - Algoritmos PCA: Principal Component Analysis. - Metodo Elbow	Con este trabajo se pudo conocer el potencial de la analítica en Twitter y de la API. -”En la fase experimental se han conseguido resultados interesantes como la detección y clasificación de los seguidores más relevantes e influyentes, la extracción de los temas twiteados más frecuentes y la gran cantidad de seguidores fake desde el punto de vista del mercado objetivo de ambas cuentas”. -”Se han cumplido los objetivos marcados aunque todavía hay mucho margen de mejora y unas posibilidades enormes en materia de analytics, propuestas para un trabajo futuro”.

Tabla 2: Trabajos relacionados

2.3. ANÁLISIS DE TRABAJOS RELACIONADOS

- En Febrero de 2015, Dani Mir Montserrat, presenta un proyecto de grado en la universidad Autónoma de Barcelona, España, titulado Analítica de Datos en Twitter, donde hace la comparación entre dos grandes marcas como son, Adidas y Nike, en la red social Twitter para lo cual implementa, el algoritmo K-means que realiza la segmentación de datos, un método llamado Elbow para encontrar el número de clusters adecuado, el algoritmo PCA y técnicas de procesamiento de lenguaje natural, con lo cual consigue la detección y clasificación de los seguidores mas relevantes e influyentes, algunos beneficios de la analítica de datos y de lo que es capaz la API en Twitter, entre otros.[5]

  A diferencia en nuestra investigación en la que se hace una agrupación con un mapreduce creando pares clave valor que contienen las entidades más mencionadas durante el periodo de tiempo y en la que no se hace comparaciones entre marcas, a pesar de que el enfoque en común en ambas investigaciones son los tweets de la red social Twitter, nuestro proyecto se enfoca en obtener los tweets en un determinado lapso de tiempo de los usuarios activos para así poder analizar entidades y tendencias de la red social Twitter en cuanto a la situación del Covid-19 en Colombia.

- En 2016, Billal, Fonseca y Sadat, proponen un modelo de canalización eficiente para el pre-procesamiento de Big Data, usando técnicas de procesamiento de lenguaje natural y técnicas comunes de aprendizaje automático, para analizar grandes conjuntos de fechas extraídos de Twitter, el cual consiste en filtrar parte del discurso, reconocimiento de entidades con nombre, segmentación de hashtag y desambiguación; también presentan una revisión de algunos trabajos relacionados con el análisis de Big Data para aplicaciones de aprendizaje automático y procesamiento de lenguaje natural, detallan los principales pasos de pre-procesamiento: normalización de tweets, descomposición de hashtag y reconocimiento de entidades con nombre; en relación con nuestro proyecto es el reconocimiento de entidades para identificar tendencias y en nuestra investigación se realiza tokenizacion por medio de análisis de entidades con

Sparkml en los Tweets publicados durante la pandemia del Covid-19 en Colombia [12]

- En 2016, Caimmi, Vallejos, Berdun, Soria, Amandi y Campo; presentan un estudio llamado Detección de Incidentes de Transito en Twitter, para el cual combinan técnicas de machine learning y procesamiento de lenguaje natural para detectar publicaciones en Twitter acerca de incidentes de transito de usuarios en la región de la Ciudad Autónoma de Buenos Aires (CABA), en el cual se hace un filtro para encontrar publicaciones relacionadas al tema entrenando individualmente con la técnica Cross-Validationa los clasificadores de texto SVM (Kernel Linear), SVM (Kernel RBF), SMO y Naive Bayes; y entre muchos otros extraer su ubicación, para la recolección de las publicaciones durante una semana hicieron monitoreo a varias cuentas de Twitter que por lo general publican sobre incidentes de transito, “obteniendo un conjunto de 6667 publicaciones”, las cuales analizan manualmente y le indica una categoría como ‘relevante’ si en su frase contiene las palabras ‘accidente’, ‘manifestación’, etc y considerando irrelevante al resto. El sistema logró identificar los incidentes en el 85.71 % de los casos y malinterpretó sólo el 0.33 % de las publicaciones[11].

  A diferencia de esta investigación, se hace la captura de datos en tiempo real sobre una ventana de tiempo apuntando a unas tendencias definidas, en Colombia y en lenguaje español. Una vez se obtienen los tweets se tokeniza utilizando Sparkml , se utiliza tweepy y se manejan expresiones regulares para la limpieza de los datos.

- En 2019, Diaz-Mendivelso y Suarez-Baron presentan un estudio denominado Análisis Social Aplicando Técnicas de Lenguaje Natural a Información Extraída de Twitter; donde realizan una revisión completa de estado del arte, encontrando la necesidad de desarrollar un modelo computacional que ayude al análisis o explotación de información extraída de Twitter, utilizan Crawling para el rastreo y extracción de información de páginas web o entornos virtuales, técnicas de PLN e implementando métodos de aprendizaje a nivel de análisis morfológico, sintáctico y léxico en la información extraída, el modelo también aplica normas, políticas y marcos de referencia de investigación, desarrollo e

innovación (I+D+i – R&D) para la gestión de calidad, lo que genera un nivel de confianza para los usuarios. En el desarrollo del crawling usando el API de Twitter, se logró notar que, por parte del estudio del análisis del lenguaje, es necesario que otras plataformas entren en la tarea de desarrollar sus propias herramientas o permitan la conexión con sus procesos para capturar información y de esta manera lograr hacer ejercicios investigativos como el que se elaboró. Todo esto para lograr crear un enlace directo con la gran cantidad de información que se procesa o transmite en la red en el día de hoy.

En la actual investigación no se propone un modelo, simplemente se implementa varias herramientas incluidas en el lenguaje de programación Python con las cuales se logra obtener los tweets en tiempo real, idioma español, locación de Colombia y así analizar entidades encontrar tendencias o necesidades relacionadas al Covid-19.

Al momento de realizar la búsqueda de bibliografía relacionada no fue posible encontrar algún trabajo que se haya realizado en Colombia donde se analicen las entidades de los tweets en el idioma español y que encuentren tendencias o necesidades sobre la pandemia del Covid-19, es por esto que surge la necesidad de analizar los tweets para determinar que tipo de necesidades publican los usuarios colombianos durante la pandemia y en los estudios encontrados sobre análisis de tendencias y/o entidades en los tweets de la red social Twitter tienen en común el uso de técnicas de Procesamiento de lenguaje natural, lo cual aporta bases de conocimiento para entender el tema y para que pueda ser desarrollado por medio de la implementación de modelos o su combinación y así lograr buenos resultados.

3 CAPITULO III. METODOLOGÍA

Dentro del marco de desarrollo del proyecto, con el objetivo de utilizar un modelo de análisis que permita identificar tendencias de entidades en Twitter que reflejen necesidades o pensamientos sobre el Covid-19 en Colombia mediante técnicas de Natural Language Processing, se hace necesario aplicar diferentes fases de la metodología CRISP-DM [13] para la construcción y evaluación de los modelos relacionados en el estado del arte.

3.1. Fase I. Comprensión del negocio

Esta fase se relaciona con la comprensión de los objetivos del proyecto y los requisitos desde una perspectiva de negocio, teniendo en cuenta el análisis de tendencias de entidades mediante NLP, de acuerdo a las referencias consultadas, para que el modelo identifique las necesidades relacionadas al Covid-19 en a red social Twitter en Colombia.

3.2. Fase II. Comprensión de los datos

En el desarrollo de esta fase se utilizó el análisis de datos para poder recoger y almacenar los datos, generando así el DataSet, el cual se obtiene de los comentarios en Twitter(Tweets) mediante el hashtag(#) referentes al Covid-19 en Colombia, esta información de los datos fue exportada en un archivo con extensión .json

Para continuar con la recolección de los datos fue necesario tener una cuenta activa en Twitter, para solicitar una cuenta modo desarrollador(Developer) y en el cual se obtiene cuatro credenciales necesarias para tener acceso y extraer el DataSet de la información que contiene Twitter.

3.2..1 Solicitud creación App de Twitter

En este link https://developer.twitter.com/en/labs encontramos la pagina web de modo desarrollador, en el cual debemos iniciar sesión con una cuenta activa de Twitter ver (Figura 1.)

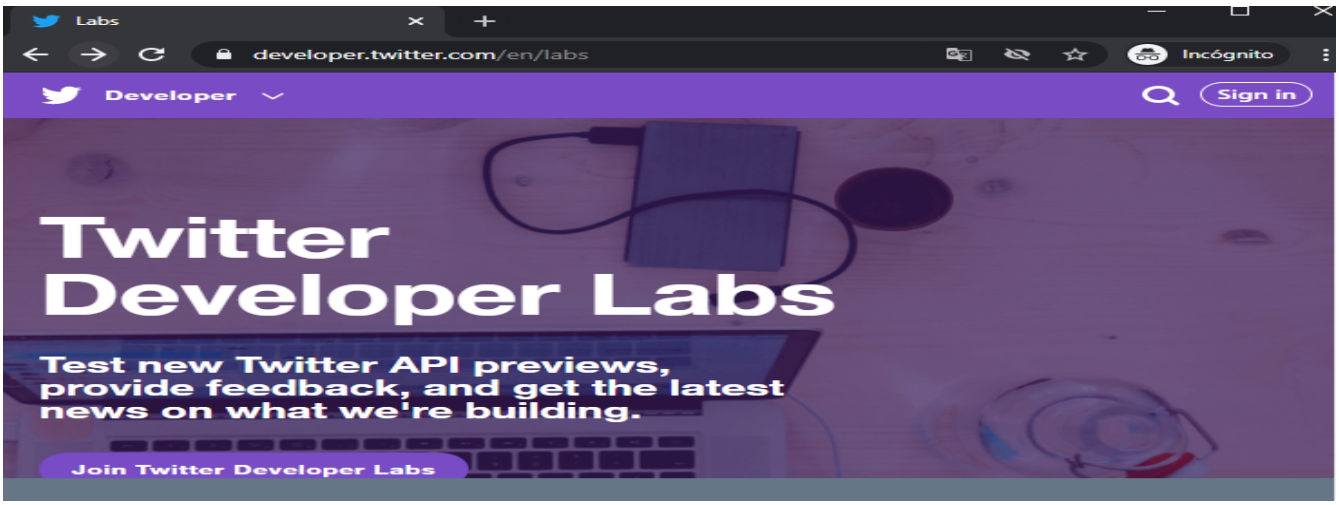


Figura 1: URL y pagina web modo Desarrollador

Una vez que se inicia sesión, se muestra la pagina de bienvenida, ver (Figura 2.)

Luego hacer clic en la opción de **Crear una aplicación** para dar inicio a la solicitud de la creación de la API a Twitter.

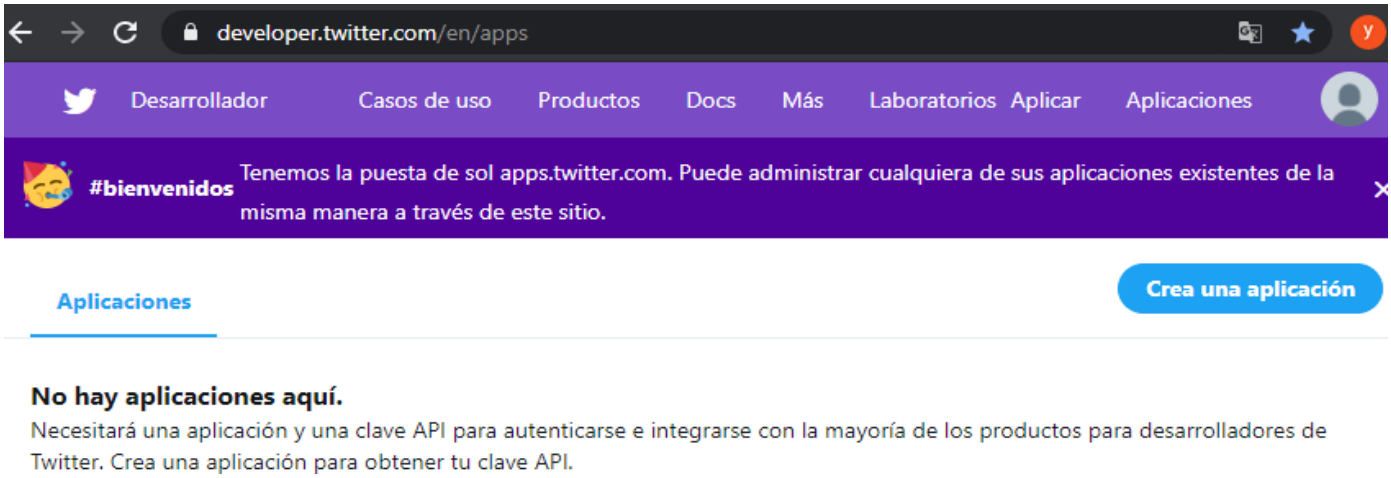


Figura 2: Bienvenida pagina modo desarrollador

La solicitud para la cuenta de la API como desarrollador en Twitter brinda un primer mensaje informativo el cual hace referencia a que es necesario ingresar algunos criterios o requisitos, para continuar clic en la opción **Aplicar**. Ver (Figura 3)

Figura 3: Ventana de confirmación de solicitud cuenta desarrollador

Es necesario en este paso seleccionar una opción; para este trabajo investigativo se solicitó la cuenta como **Estudiante,** click en la opción **próximo.** Ver (Figura 4)

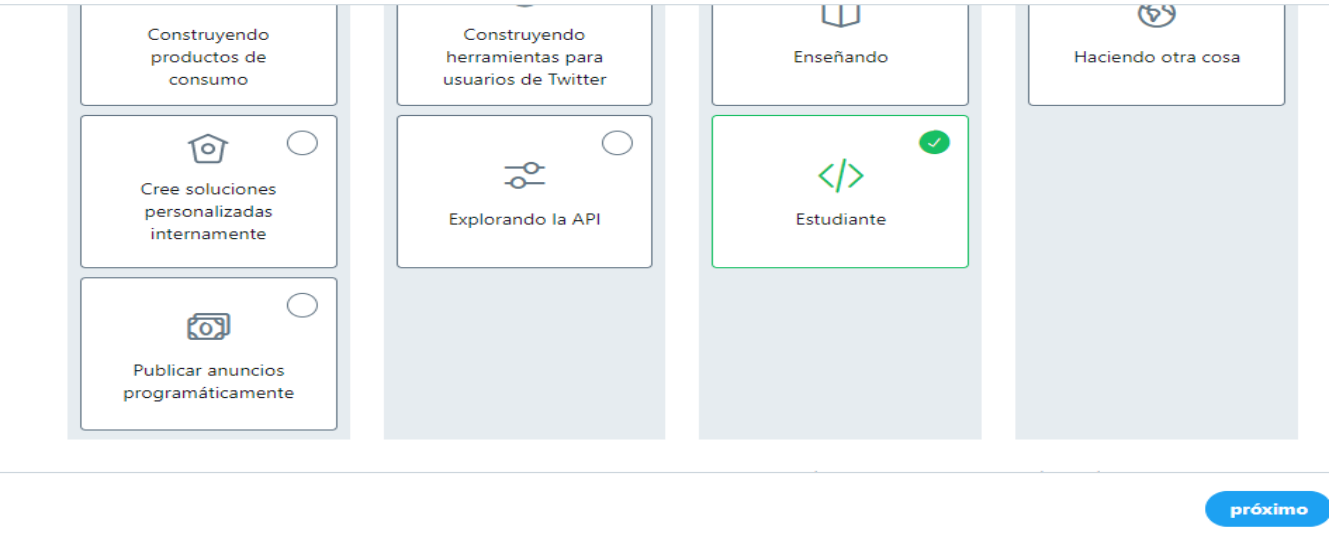


Figura 4: Se define cuenta como estudiante desarrollador

Para continuar con la solicitud de la creación de la API de Twitter es necesario diligenciar los campos solicitados. Click en **próximo**. Ver (Figura 5)

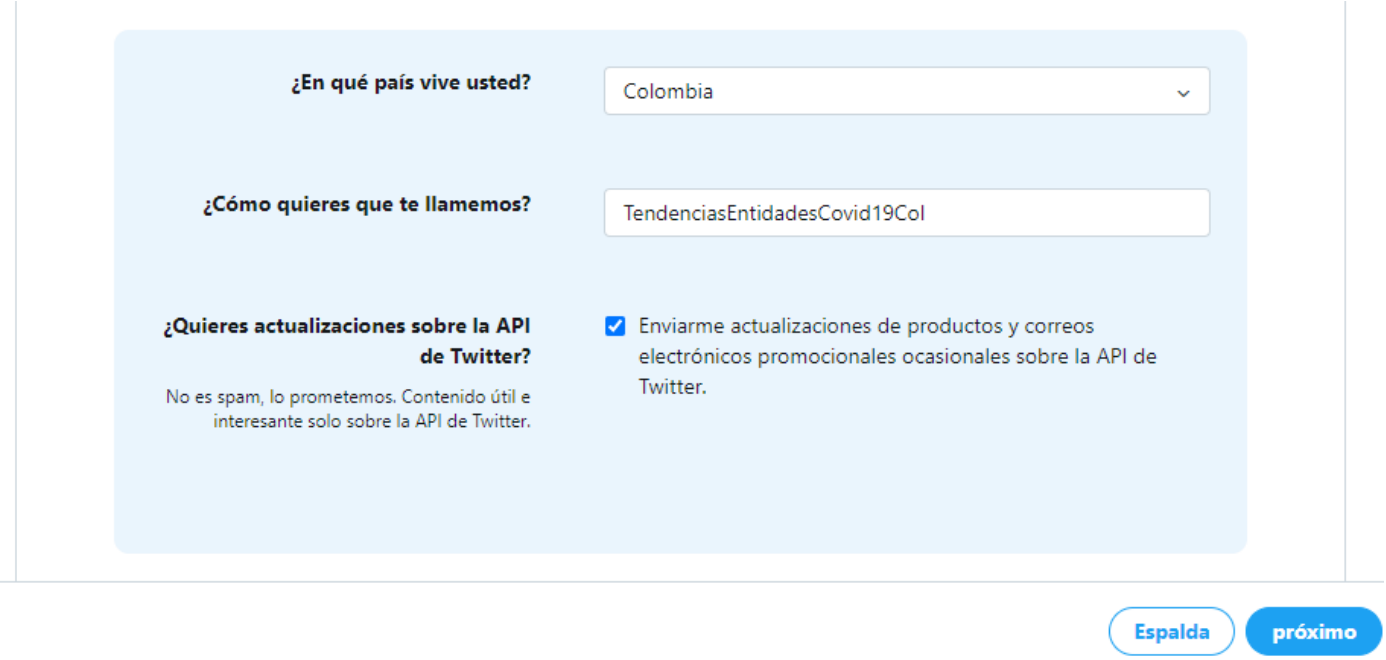


Figura 5: Ingreso de información solicitada

A continuación se describe en un mínimo de 200 caracteres, qué uso se le van a dar a los datos extraídos. Ver(Figura 6.), click en **próximo.**

Figura 6: Describir finalidad de la cuenta y uso de los datos

Seguidamente se despliegan las políticas de desarrollador de Twitter, se debe marcar la casilla de aceptar políticas y luego click en la opción **Presentar la Solicitud.** Ver (Figura 7.)

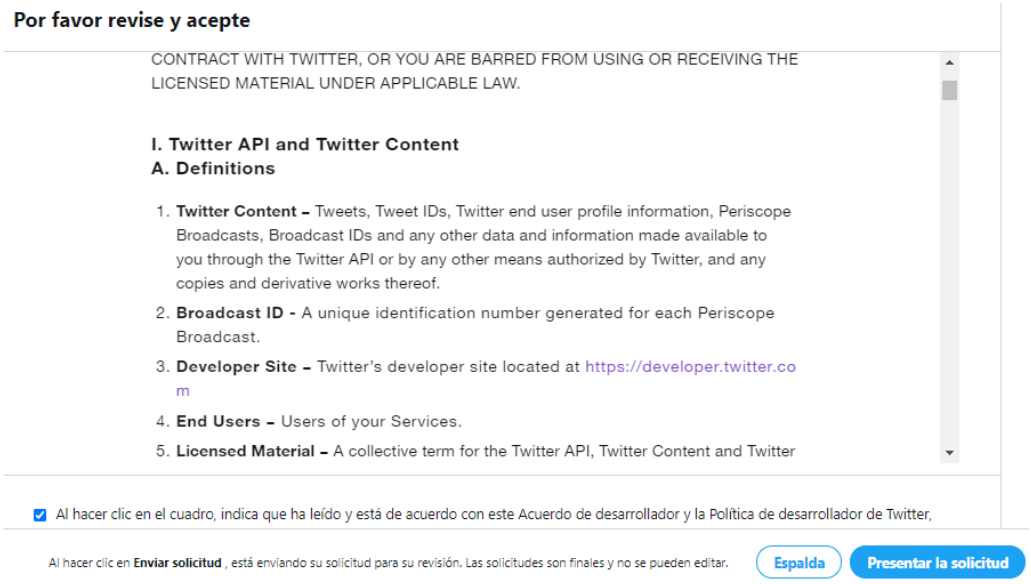


Figura 7: Políticas de desarrollador de Twitter

Confirmar si los datos que muestra la ventana son correctos. Ver (Figura 8.)

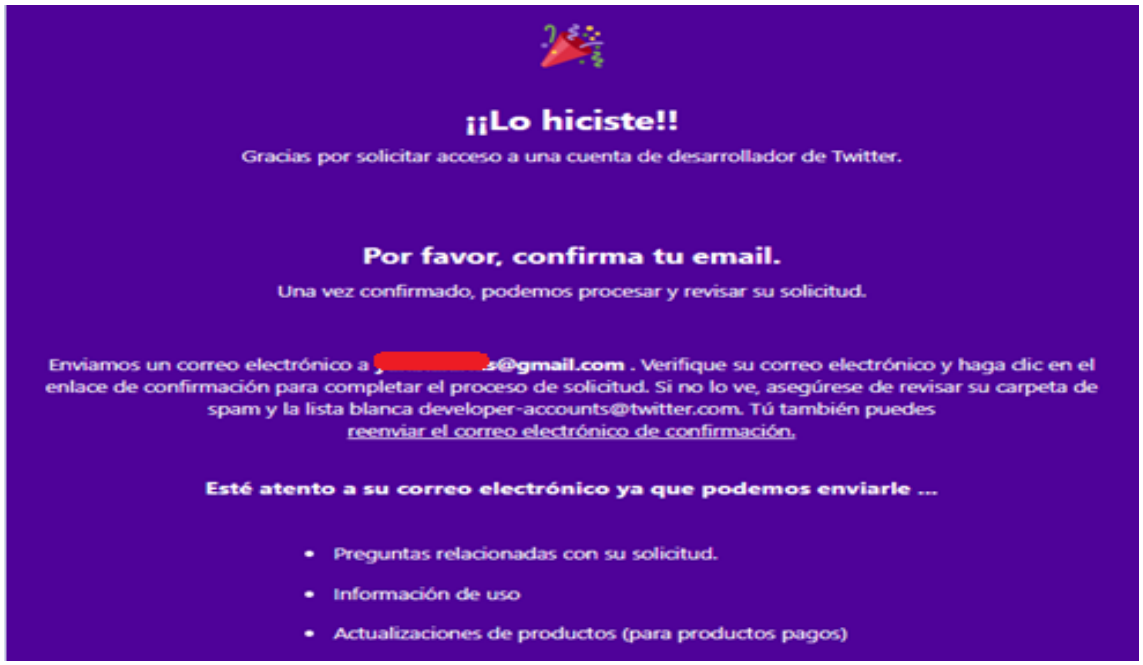


Figura 8: Verificación de información

Y por ultimo, Twitter envía un correo electrónico como medida de confirmación para la solicitud de la APP como desarrollador.

Una vez Twitter verifique y acepte la solicitud de la creación de la APP, se mostrara una interfaz con pagina inicial de Dashboard del desarrollador. ver (figura 9).

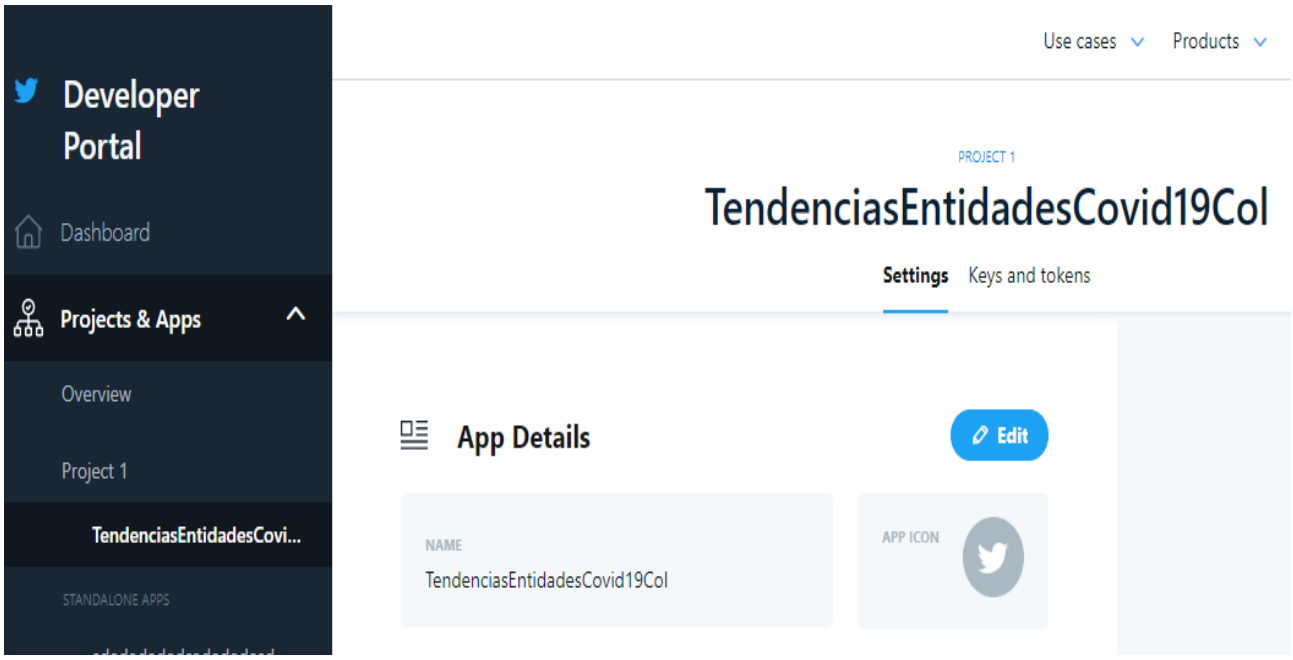


Figura 9: Pagina inicial modo desarrollador

Para observar nuestras primeras credenciales, ingresamos a nuestra app creada; en la opción Proyectos y Apps, seguidamente damos click en la opción **Keys and Tokens** y luego en **View Keys,** ver (Figura 10).

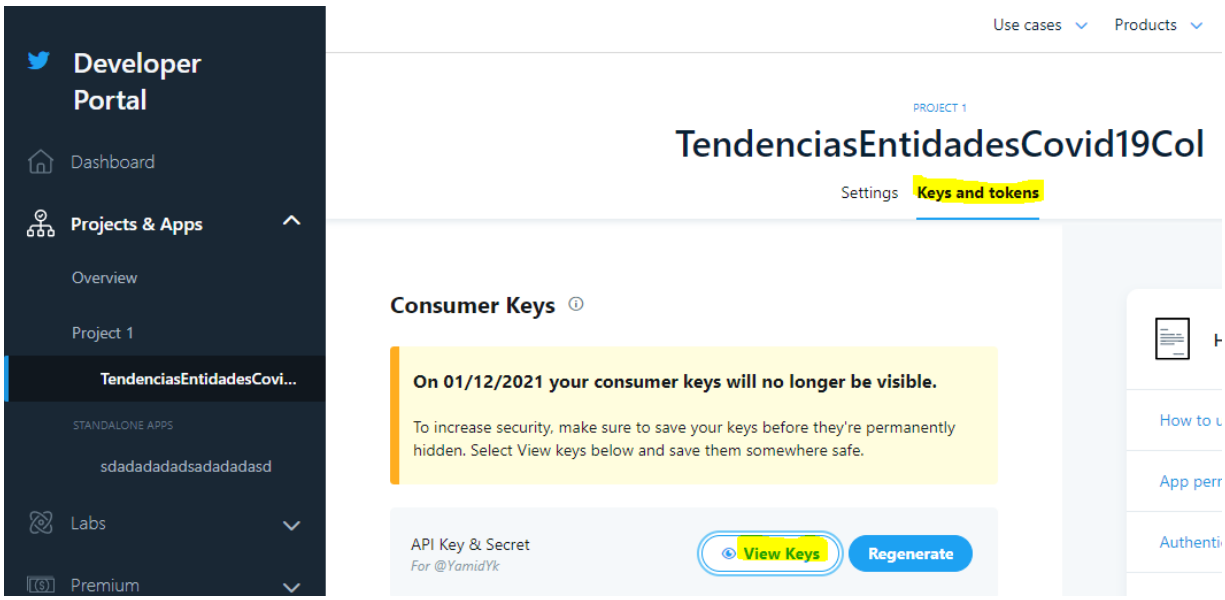


Figura 10: Obtención primeras dos credenciales

Seguidamente se muestran en una ventana las dos primeras claves denominadas **API key** y **API key secret** ver (Figura 11).

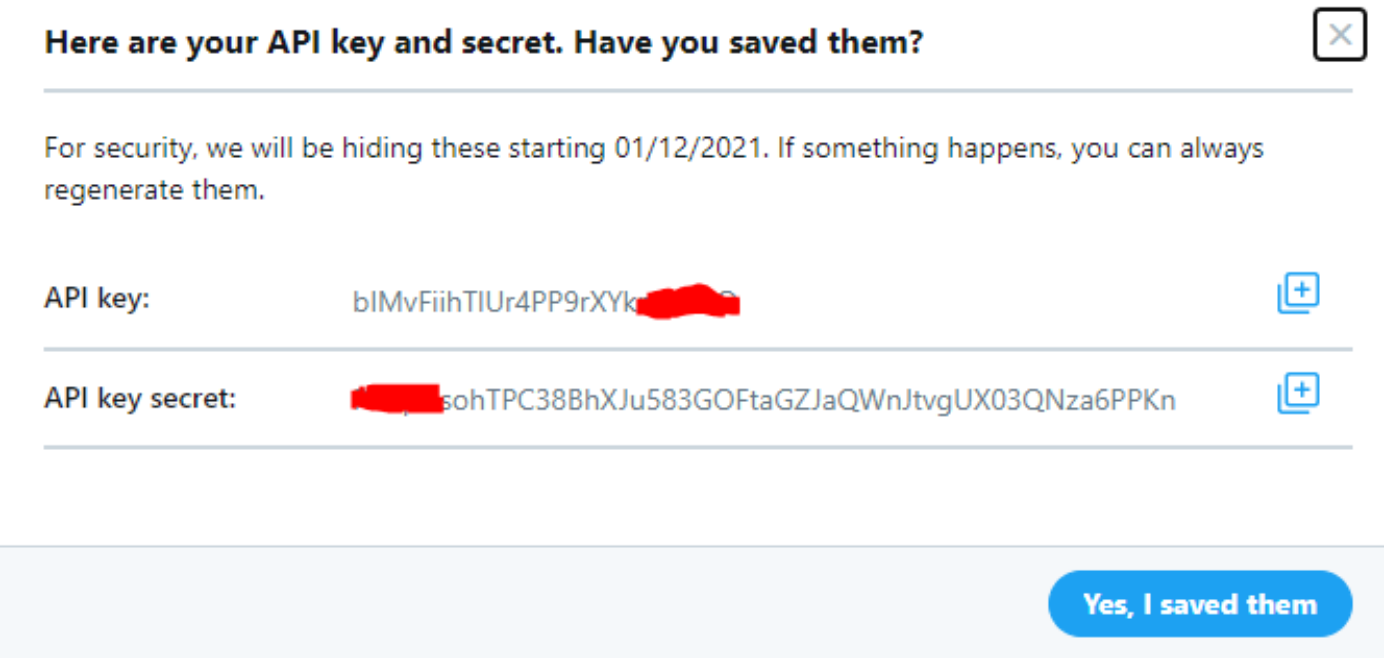


Figura 11: API key, API key secret

Luego, para obtener las otras dos claves, hay que dar click en la opción **Regenerate,** de la sección de **Acces token y Secret,** ver (Figura 12)

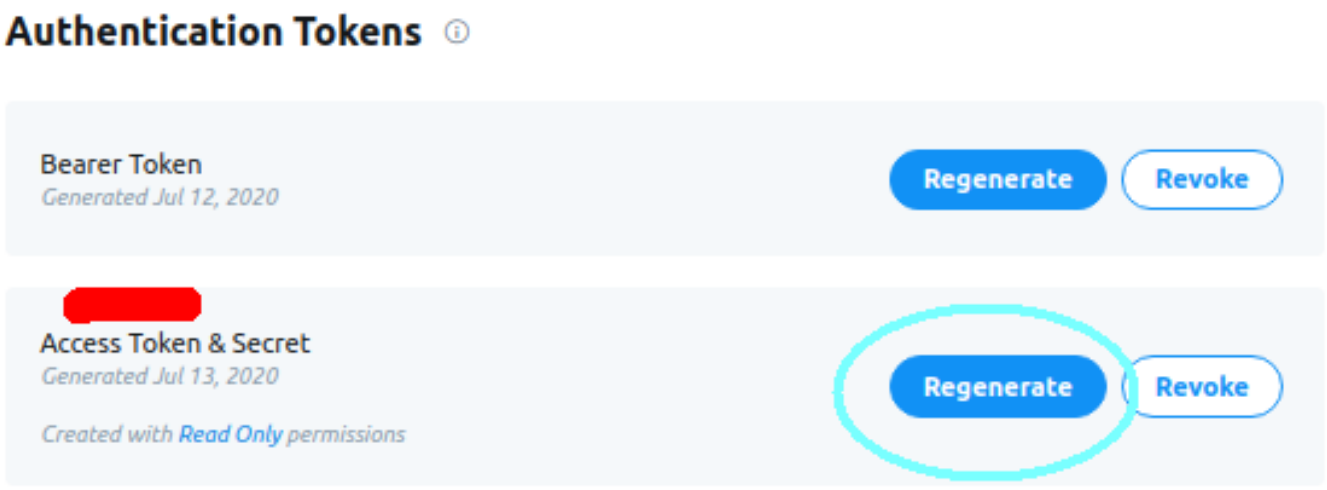


Figura 12: Solicitud de claves faltantes

Y por ultimo se muestra la ventana con las claves de **Access token y Access token secret.** Ver (Figura 13.)

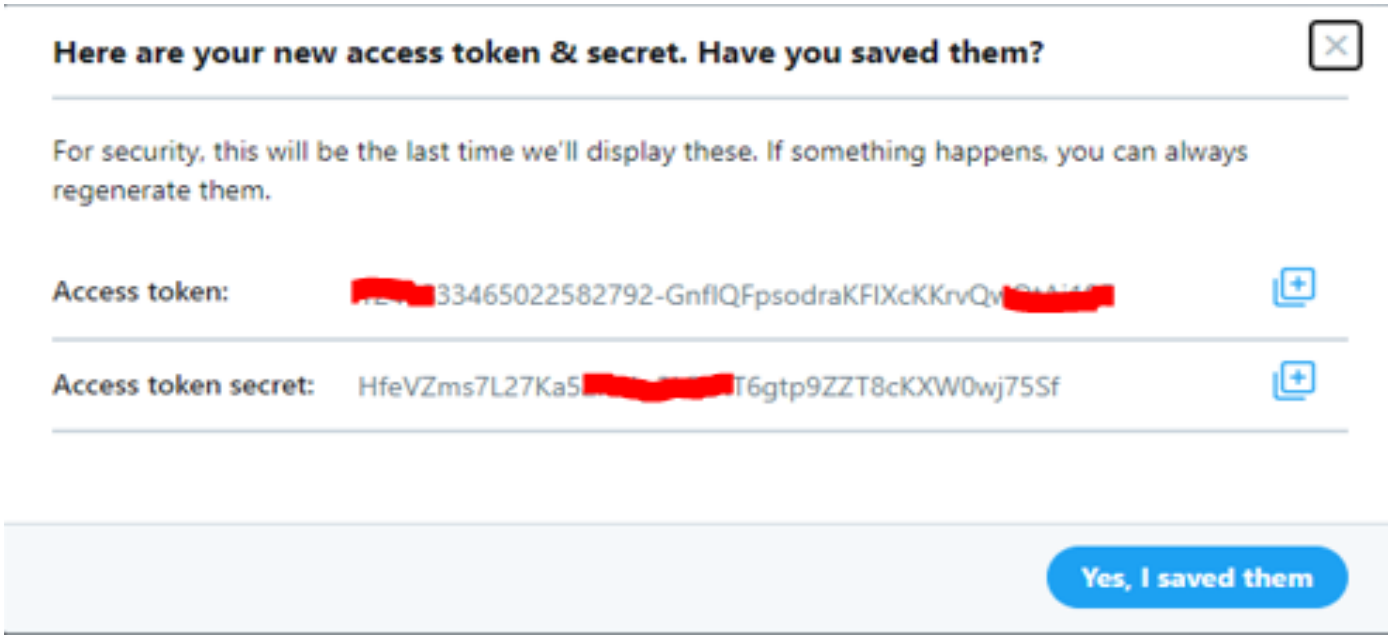


Figura 13: Access Token y Access Token Secret

Al poseer las cuatro claves de API se procede a relacionarlas en el código del proyecto que permite la extracción de DataSet.

3.2..2 Extracción de los datos con las credenciales de Twitter

Después de solicitar y que por parte de Twitter sea aceptada la creación de la app de desarrollo, además de contar con las cuatro credenciales necesarias para la conexión y posterior extracción de DataSet; se utilizó el IDE Visual Studio Code ver (Figura 14,15), el lenguaje de programación python, un entorno virtual en python, diferentes librerías de Python como: Tweepy, TextBlob, Pandas, y Spark para NLP.

Figura 14: Logo de IDE Visual Studio Code

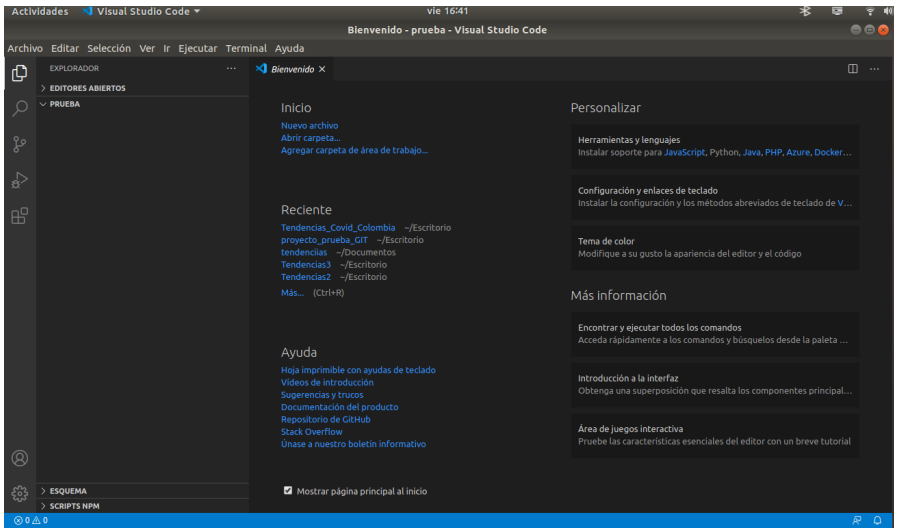


Figura 15: Entorno de Visual Studio Code

Para iniciar, es necesario crear un entorno virtual para python, para ello, dentro de Visual Studio Code, se abre una terminal integrada ver (Figura 16 y 17), que permite la ejecución de los diferentes comandos necesarios.

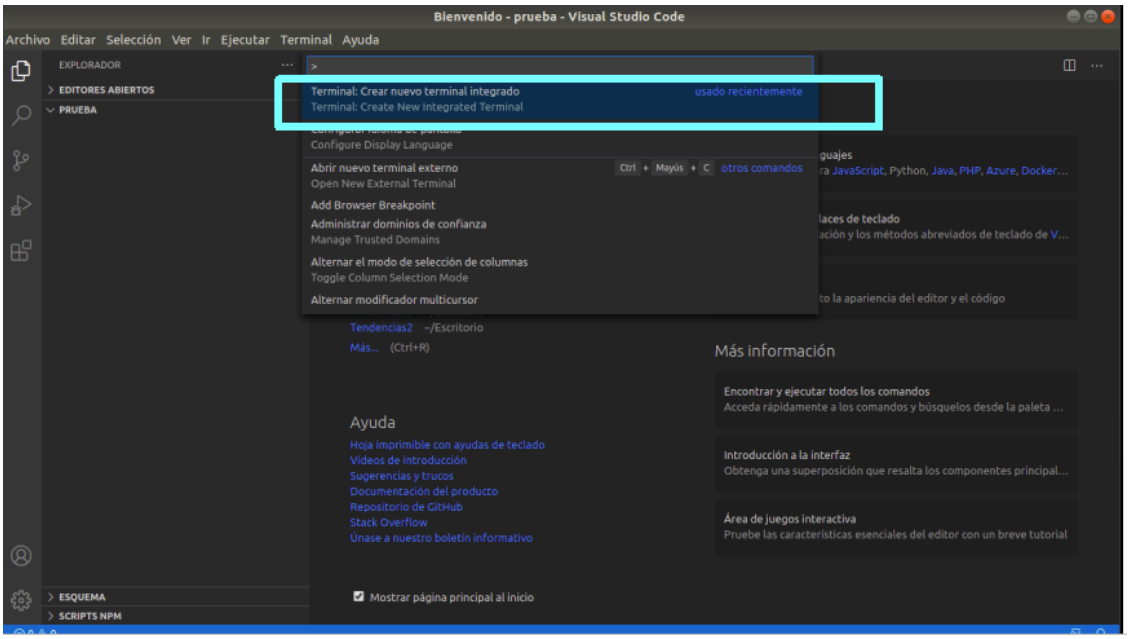


Figura 16: Búsqueda de la terminal de comandos

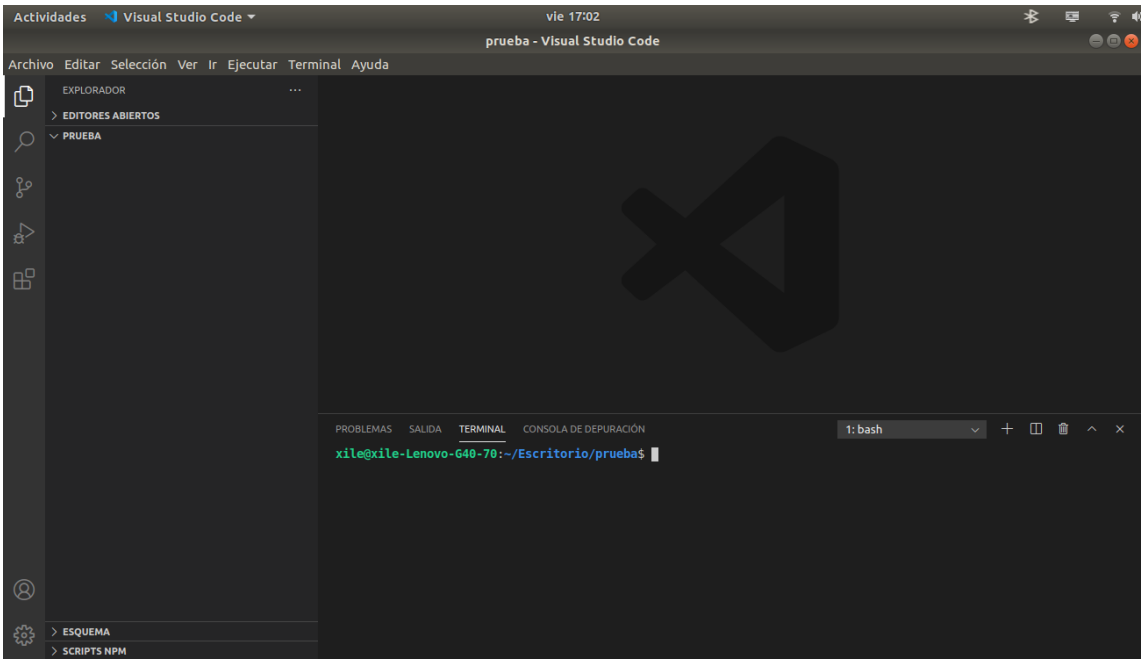


Figura 17: Terminal integrada de Visual Studio Code

Para la correcta instalación y posterior utilización del entorno virtual se debe verificar si python esta instalado, revisamos la versión instalada con el comando python --version si se utiliza sistema operativo windows o el comando python3 –version para sistemas operativos Linux ver (Figura 18).

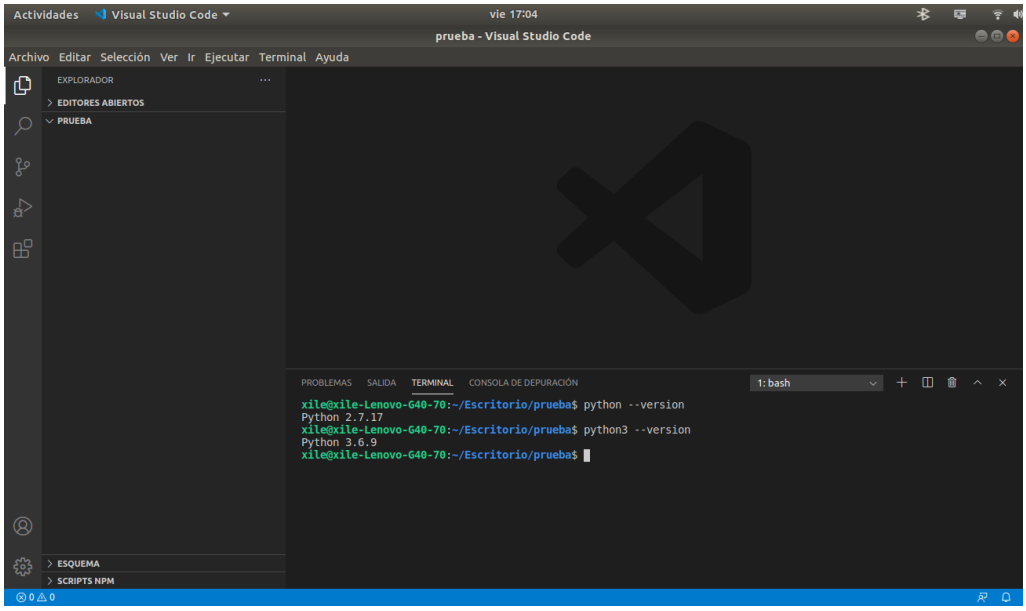


Figura 18: Versión de python

seguidamente se instala el gestor de entornos virtuales **virtualenv,** ver (Figura 19)**.**

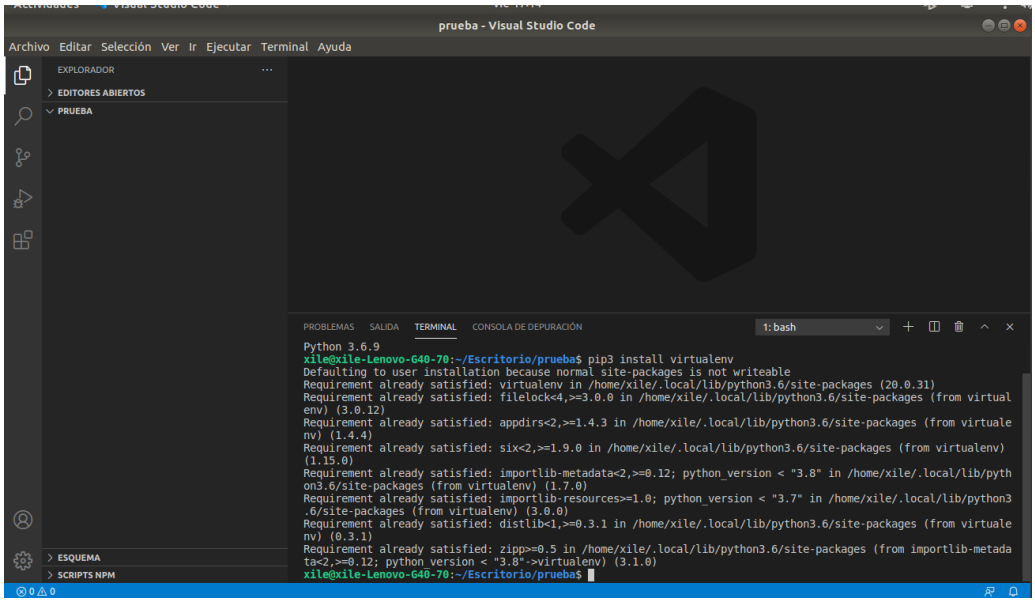


Figura 19: Instalación de Virtualenv

Creación del entorno virtual de nombre env1 con el comando **virtualenv env1.** Ver (Figura 20).

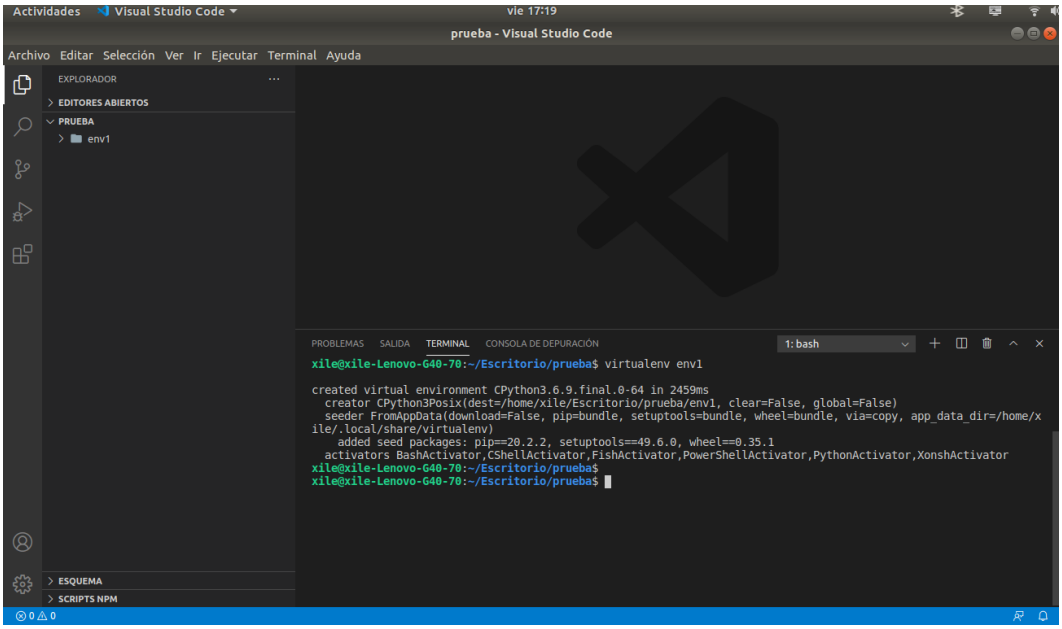


Figura 20: Creación del entorno virtual env1

Seguidamente, activar entrono virtual env1, con el comando **source env1/bin/activate.** Para verificar si el entorno esta activo se debe mirar al lado izquierdo (nombre del entorno creado anteriormente) ver (Figura 21).

Figura 21: Activación y verificación del entorno virtual creado

Luego se hace la instalación de el archivo requirements.txt, ver (Figura 22, 23), el cual permite automatizar la instalación de las diferentes librerías necesarias para el desarrollo del proyecto.

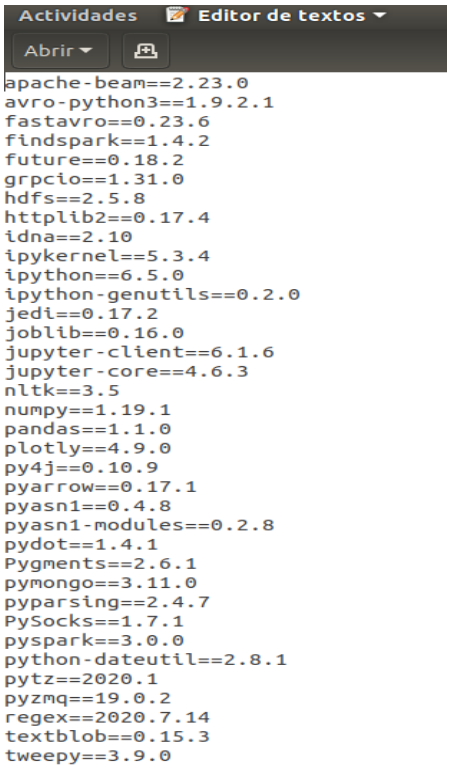


Figura 22: Archivo requiremets.txt

Para realizar la instalación del archivo requirements.txt, se debe utilizar el comando:

pip install -r requiremts.txt

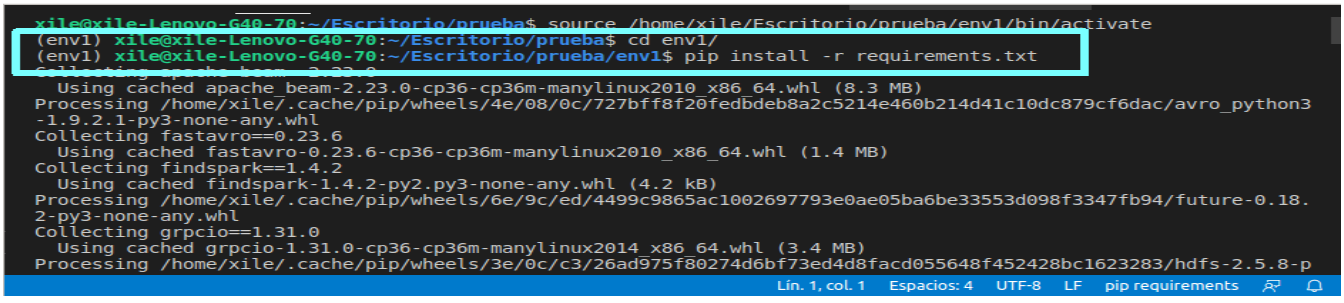


Figura 23: Instalación de archivo requirements.txt

Una vez instaladas las librerías necesarias, se continua con la importación de las mismas dentro del archivo file_te.py que contiene el código que permitirá la extracción de los tweets ver (Figura 24).

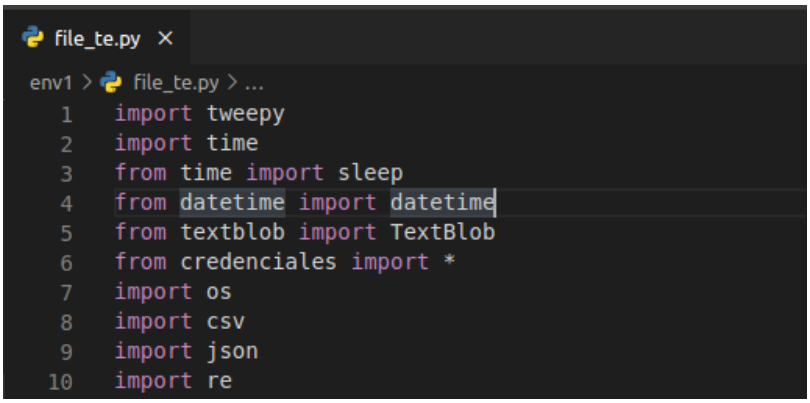


Figura 24: Librerías de python 3

A continuación se pone la ubicación geográfica de Colombia en coordenadas que genera la herramienta GEOBOX, (linea de código #12) se definen las tendencias para buscar los tweets (linea de código #13), luego se determina el idioma que es Español (linea de código #14), luego se hace autenticación de las claves de usuario que nos genera app developers de Twitter (linea de código #17,18,19); ver (Figura 25).

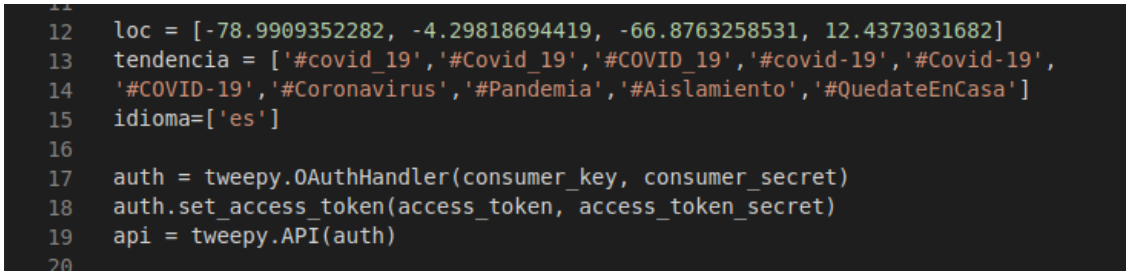


Figura 25: Localización, tendencia, idioma, autenticación

El archivo credenciales.py contiene las claves secretas que genera app developers de Twitter, ver(Figura 26).

```
file_te.py   credenciales.py ×
env1 > credenciales.py > consumer_key
1   consumer_key = "rThS[redacted]kejg8QDzfJPc"
2   consumer_secret = "JkwqiOqR7wBDI9W7ycBeeum[redacted]U5KMp4BTU"
3   access_token = "1245093235[redacted]GwezQ0C3XZiYqxSvlii5"
4   access_token_secret = "6oMWN0q2y1uJmuQLxd9KG2jvql6RH[redacted]sV"
```

Figura 26: Credenciales

Siguiendo en el archivo file_te.py, se utiliza el código de la api streaming de Twitter ver (Figura 27), la cual permite obtener los tweets en tiempo real tomando como fecha el día 24 de agosto del año 2020 por un periodo de 3 horas equivalentes a 10.800 segundos comenzando a las 8: 15 pm y finalizando a las 11:15 pm, donde se obtienen 19.023 tweets en total.

```
21  class MyStreamListener(tweepy.StreamListener):
22      def __init__(self, time_limit=10800):
23          self.start_time = time.time()
24          self.limit = time_limit
25          self.saveFile = open('tweets.json', 'a')
26          #self.saveFile.write('[')
27          super(MyStreamListener, self).__init__()
28
29      def on_status(self, status):
30          if ((time.time() - self.start_time) < self.limit):
31              data="""{{"fecha_creacion": "{fecha}","nombre":"{name}","texto":"{texto}","ubicacion":"{loc}"
32              .replace('.','').replace(',','').replace('\n','').replace('"',''),
33              name = status.user.screen_name,
34              loc=status.user.location,retweets = status.retweet_count)
35
36              self.saveFile.write(data)
37              self.saveFile.write(',')
38              self.saveFile.write('\n')
39              return True
40          else:
41              #self.saveFile.write(']')
42              self.saveFile.close()
43              return False
44
45      def on_error(self, status_code):
46          if status_code == 420:
47              return False
```

Figura 27: Código de Streaminng

Y por ultimo se crea la instancia del streaming y el filtro de: tendencia, idioma y locación geográfica. Ver (Figura 28).

```
49  stream_listener = MyStreamListener()
50  stream = tweepy.Stream(auth=api.auth, listener=stream_listener)
51  stream.filter(track=tendencia,languages=idioma,locations=loc)
```

Figura 28: Instancia de streaming y filtro

A continuación se muestra el Data set de tweets como resultado de la ejecución del archivo file_te.py. Ver (Figura 29).

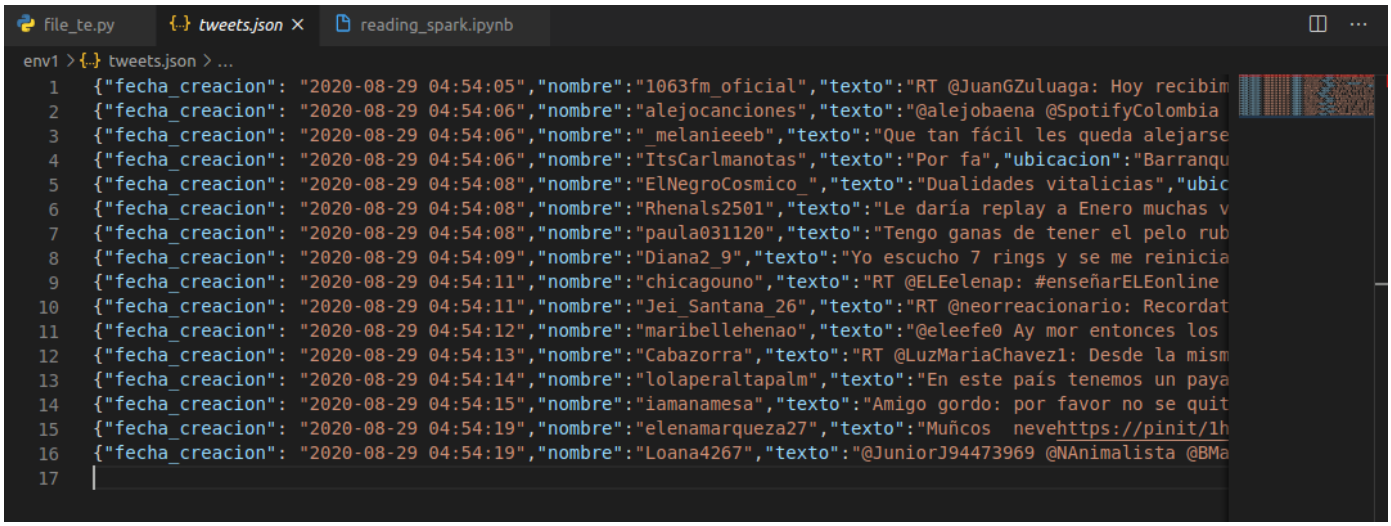


Figura 29: Archivo generado tweets.json

3.3. Fase III. Preparación de los datos existentes

Para esta actividad es necesario inspeccionar, limpiar, transformar y presentar los datos, con el objetivo de descubrir información útil para el análisis de tendencias de entidades de tweets que reflejan necesidades o pensamiento sobre el Covid-19.

3.3..1 Modelo: en esta actividad se pretende clasificar y calibrar los diferentes parámetros que se necesitan para la modelación del algoritmo y para ello se utiliza un notebook llamado reading_spark.ipynb, el cual lee el archivo tweets.json (se genera al ejecutar file_te.py), utiliza pandas, pyspark, pyspark.sql ver (Figura 30).

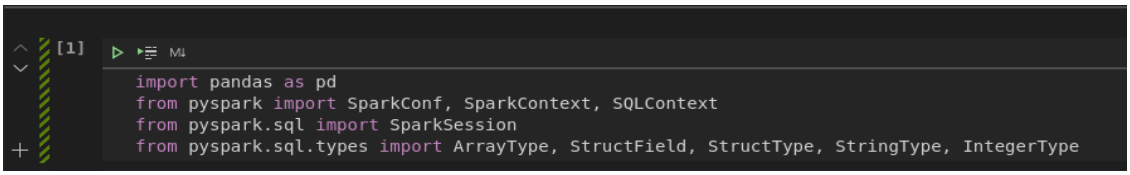


Figura 30: Importación de librerías en notebook

Seguidamente se indica la ruta donde se encuentra el ejecutable de pyspark ver(Figura 31)

```
[2]
import findspark
findspark.init('/home/xile/Escritorio/prueba/env1/lib/python3.6/site-packages/pyspark')
```

Figura 31: Ruta de ejecutable pyspark

Luego se crea una sesión de spark (se crea la app) ver (Figura 32).

```
[3]
appName = "Spark Detected Trends"
master = "local[1]"

# Create Spark session
spark = SparkSession.builder \
    .appName(appName) \
    .master(master) \
    .getOrCreate()
```

Figura 32: Sesión de Spark

Luego spark carga el archivo tweets.json generado anteriormente en la ruta. Ver (Figura 33)

```
json_file_path = './tweets.json'
df = spark.read.json(json_file_path)
```

Figura 33: Lectura de tweets.json

Seguidamente se muestra el esquema de los tweets, ver(Figura 34).

```
print(df.schema)
df.show()
```

```
StructType(List(StructField(RTs,StringType,true),StructField(fecha_creacion,StringType,true),StructField(nombre,St
ringType,true),StructField(texto,StringType,true),StructField(ubicacion,StringType,true)))
+---+-------------------+--------------+--------------------+--------------------+
|RTs|     fecha_creacion|        nombre|               texto|           ubicacion|
+---+-------------------+--------------+--------------------+--------------------+
|  0|2020-08-28 23:41:53|    JuanRhendon|Yo categorizo las...|  Medellín, Colombia|
|  0|2020-08-28 23:41:53|       ae_radio|  Ya está al air...|  Concepción, Chile.|
|  0|2020-08-28 23:41:54|  Delia01433863|RT @RoilTuiteros:...|                None|
|  0|2020-08-28 23:41:56|    Perezpuebla|RT @LauPerezCisne...|Puebla Puebla Méx...|
|  0|2020-08-28 23:41:56|   MarbelKarley|Que la luna sea t...|           Venezuela|
|  0|2020-08-28 23:41:56|     Clodyellow|RT @Cahora: 🏙️🏙️ ...| Birmingham, England|
|  0|2020-08-28 23:41:58|       Liz_pty|RT @Mitoz97: El h...|      Ángeles, Illea|
|  0|2020-08-28 23:41:59|AlejoSalazarIns|@Cindypramos Y ap...|Colombia, Barranq...|
|  0|2020-08-28 23:41:59|       nellya53|Iváncito no es Pr...|                None|
|  0|2020-08-28 23:41:59|     juanguirol|@reaImaIpa4ever @...|  Medellín, Colombia|
|  0|2020-08-28 23:42:00| prefectura_zea|👉Juramentada nue...|        Zea, Merida |
|  0|2020-08-28 23:42:00|diegobuenaventl|Puerco miserables...|   Cartago, Colombia|
```

Figura 34: Schema locación

Se hace consulta con lenguaje sql de spark.sql donde se muestran los tweets generados solamente en Colombia y lógicamente de idioma español, en un archivo denominado output ver(Figura 35).

```
df.createOrReplaceTempView("tweets")
colombiaDF = spark.sql("SELECT texto FROM tweets WHERE ubicacion like '%Colombia%' """)
colombiaDF.coalesce(1).write.format("text").option("header", "false").mode("append").save("output")
colombiaDF.show()
```

```
+--------------------+
|               texto|
+--------------------+
|Yo categorizo las...|
|@Cindypramos Y ap...|
|@reaImaIpa4ever @...|
|Puerco miserables...|
|@Paz_MiPez Tengo ...|
|Cuantos aguardien...|
|@margeryrb1 @Feli...|
|@SiendoLuz25 Amén...|
|@JaviDemocrata Sa...|
|@MartinSantosR @C...|
|@mcamilacch Noooo...|
|Me decido por ti ...|
|Tengo tres: Unico...|
+--------------------+
```

Figura 35: Consulta SQL de spark

Se realiza una consulta donde se muestra el conteo de tweets por ubicación, ver (Figura 36).

```
busqueda = spark.sql("SELECT COUNT(*) as cantidad_tweets,ubicacion FROM tweets WHERE texto like
'%Colombia%' GROUP BY ubicacion""")

busqueda.show()
```

```
+--------------+--------------------+
|cantidad_tweets|           ubicacion|
+--------------+--------------------+
|             1|                None|
|             1|              Bogota|
|             1|Bogotá, DC, Colombia|
+--------------+--------------------+
```

Figura 36: Conteo de tweet por ubicación

Luego, se realiza tokenización con sparkml, donde se separa el texto y el conteo de longitudes, ver (Figura 37) .

```
from pyspark.ml.feature import Tokenizer
from pyspark.sql.functions import col, udf
from pyspark.sql.types import IntegerType

tokenizer = Tokenizer(inputCol="texto", outputCol="words")
countTokens = udf(lambda words: len(words), IntegerType())

tokenized = tokenizer.transform(colombiaDF).withColumn("conteo",countTokens("words")).show()
```

```
+--------------------+--------------------+------+
|               texto|               words|conteo|
+--------------------+--------------------+------+
|Yo categorizo las...|[yo, categorizo, ...|    11|
|@Cindypramos Y ap...|[@cindypramos, y,...|    17|
|@reaImaIpa4ever @...|[@reaimaipa4ever,...|     6|
|Puerco miserables...|[puerco, miserabl...|     6|
|@Paz MiPez Tengo ...|[@paz mipez, teng...|    12|
|Cuantos aguardien...|[cuantos, aguardi...|    22|
|@margeryrb1 @Feli...|[@margeryrb1, @fe...|     9|
|@SiendoLuz25 Amén...|[@siendoluz25, am...|     3|
|@JaviDemocrata Sa...|[@javidemocrata, ...|    15|
|@MartinSantosR @C...|[@martinsantosr, ...|     8|
|@mcamilacch Noooo...|[@mcamilacch, noo...|     6|
|Me decido por ti ...|[me, decido, por,...|    26|
|Tengo tres: Unico...|[tengo, tres:, un...|    11|
+--------------------+--------------------+------+
```

Figura 37: Tokenización

Y por ultimo se generan claves objeto valor para saber cuales son las entidades mas nombradas, ver (Figura 38)

```
spark.stop()

sc = SparkContext("local[3]").getOrCreate()
words = sc.textFile("./output/*.txt").flatMap(lambda line: line.split(" "))
wordCounts = words.map(lambda word: (word, 1)).reduceByKey(lambda a,b:a +b)
wordCounts.saveAsTextFile("./conteo/")
```

Figura 38: Generar claves objeto-valor

3.3..2 Evaluación: una vez se realiza la limpieza de los datos y el análisis de entidades ver (Figura 39), se procede a realizar el análisis de los datos mediante un grafico de barras, ver (Figura 40) .

```
('@GachancipaCund', 4)
('presenta', 5)
('síntomas', 3)
('https://tco/PrZMJMiVMQ', 1)
('Amigos', 5)
('@redaupa', 1)
('agracemos', 1)
('buscamos', 2)
('https://tco/Uf0V5iFF3Q', 1)
('Albert', 2)
('Grupo', 1)
('brújilda', 1)
('colombianos', 2)
('mandar', 2)
('tormenta', 1)
('acerquen', 1)
('Doom!', 1)
('CLOACA', 1)
('cago', 1)
('gremio', 3)
('TOMA', 1)
('detestan', 1)
('amaran', 1)
('https://tco/RibOn8wlXg', 1)
('forma', 9)
('maluca', 1)
('arregla', 2)
('crespos', 1)
('iba', 8)
('juguito', 1)
('@profesorflavio1', 3)
('colegas', 2)
('muchos', 9)
('hermanos', 1)
('hermosa', 9)
('mundo:', 1)
('primeras', 2)
```

Figura 39: Archivo de entidades, part-0000

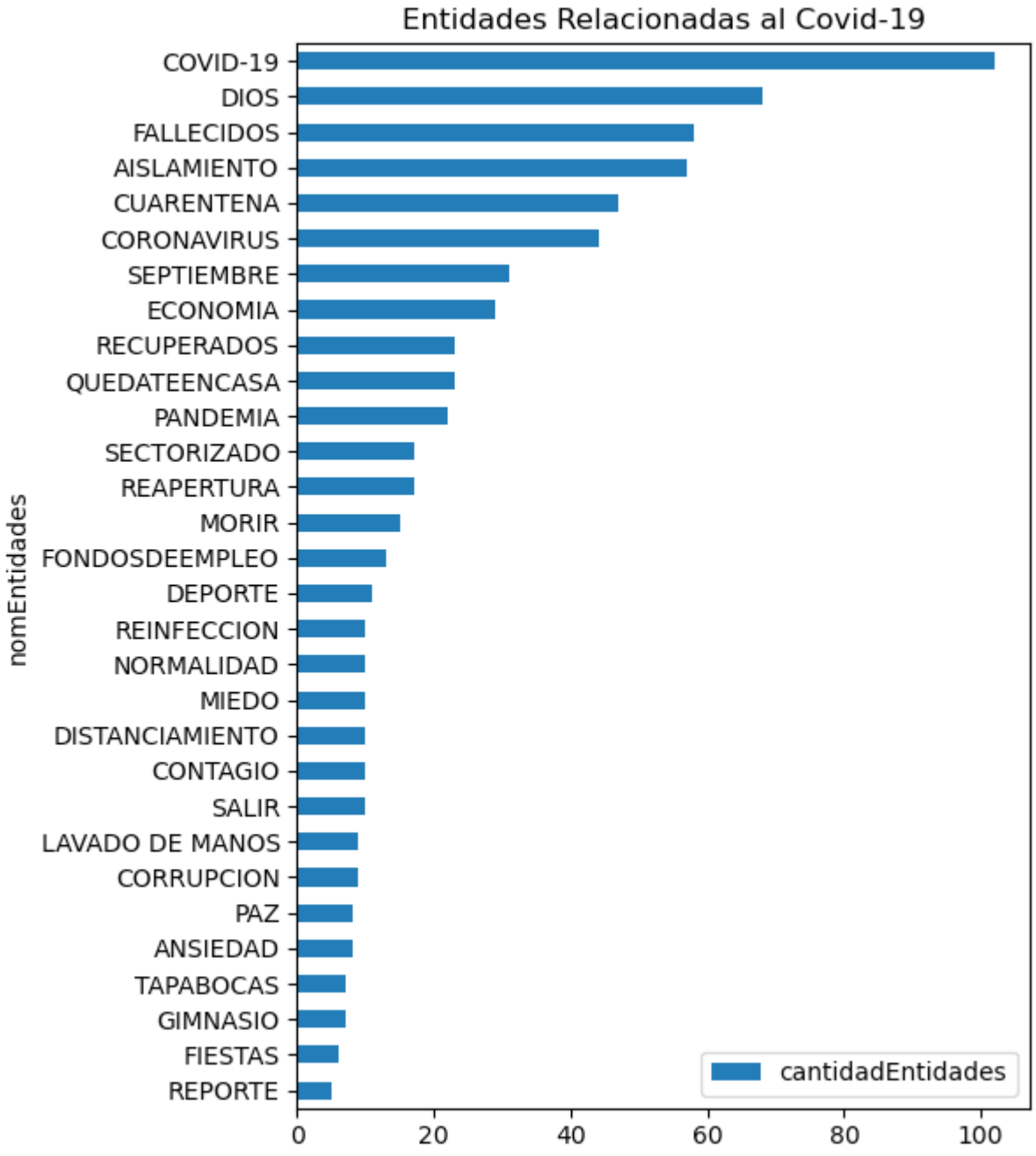


Figura 40: Resultado análisis de entidades

3.3..3 Implementación:

Los resultados obtenidos por el modelo se socializarán con el fin de que sirva de base para otros trabajos relacionados al tema.

3.4. Fase IV. Socialización de los resultados: En esta fase final se socializa a la comunidad en general por medio de una presentación la investigación realizada.

3.4..1 Limpieza:

Para la socialización de resultados se hace necesario utilizar el método de conocimiento del contexto donde se obtiene la limpieza de las entidades ver (Figura 41)

	A	B
1	**nomEntidades**	**cantidadEntidades**
2	REPORTE	5
3	FIESTAS	6
4	GIMNASIO	7
5	TAPABOCAS	7
6	ANSIEDAD	8
7	PAZ	8
8	CORRUPCION	9
9	LAVADO DE MANOS	9
10	SALIR	10
11	CONTAGIO	10
12	DISTANCIAMIENTO	10
13	MIEDO	10
14	NORMALIDAD	10
15	REINFECCION	10
16	DEPORTE	11
17	FONDOSDEEMPLEO	13
18	MORIR	15
19	REAPERTURA	17
20	SECTORIZADO	17
21	PANDEMIA	22
22	QUEDATEENCASA	23
23	RECUPERADOS	23
24	ECONOMIA	29
25	SEPTIEMBRE	31
26	CORONAVIRUS	44
27	CUARENTENA	47
28	AISLAMIENTO	57
29	FALLECIDOS	58
30	DIOS	68
31	COVID-19	102

Figura 41: Entidades

3.4..2 Implementación: mediante un código desarrollado en python, en el cual se utilizan las librerías de pandas y Matplotlib para realizar la grafica de los resultados donde recibe las entidades del archivo ENTIDADES.xlsx, asignando nombres de encabezado que se ven reflejados en los parámetros para la creación de la grafica, ver (Figura 42).

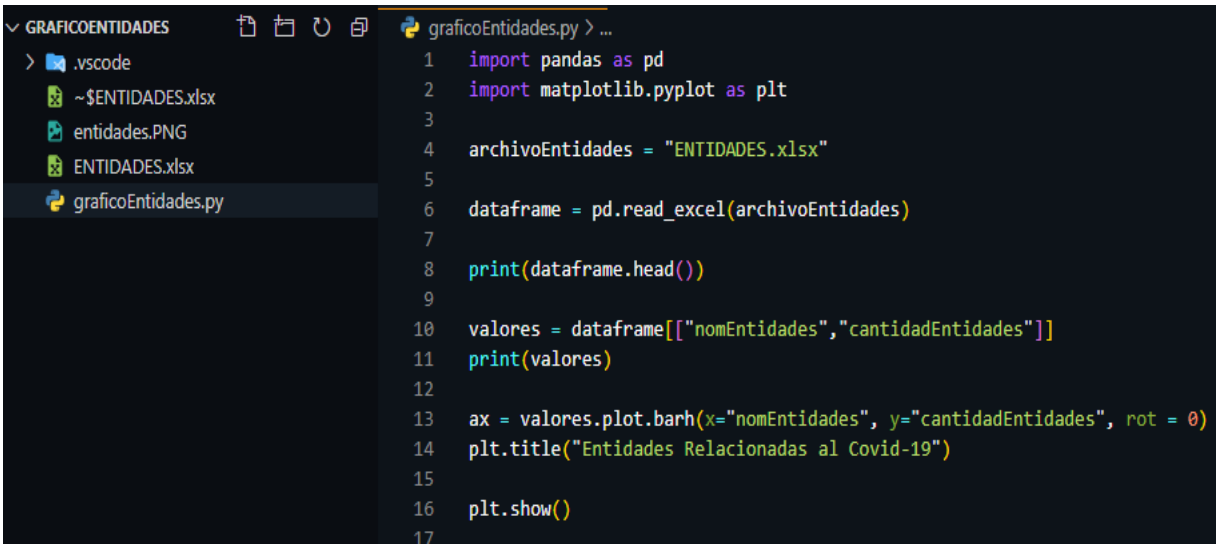


Figura 42: Código grafica

Al ejecutar el anterior código, ver (Figura 43), se despliega la grafica del análisis de entidades, ver (Figura 44).

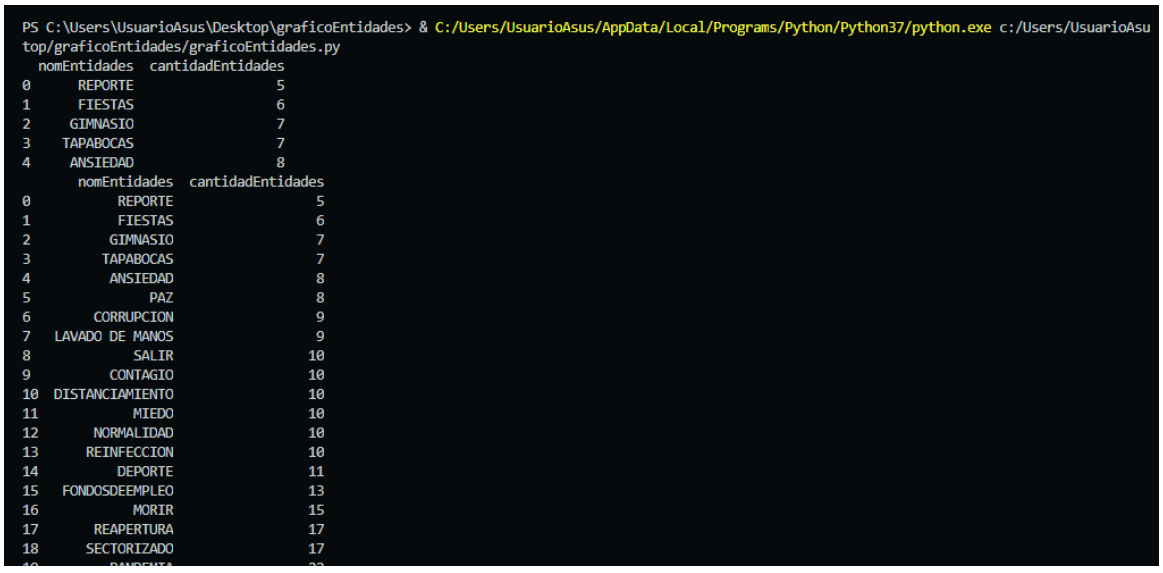


Figura 43: Comprobando ejecución de código

3.4..3 Visualización de grafica de resultados:

Mediante un grafico de barras como se muestra en la (Figura 44), donde se logra identificar las entidades mas relacionadas al Covid-19 de los tweets publicados en Colombia durante un lapso de tiempo de 3 horas durante la pandemia.

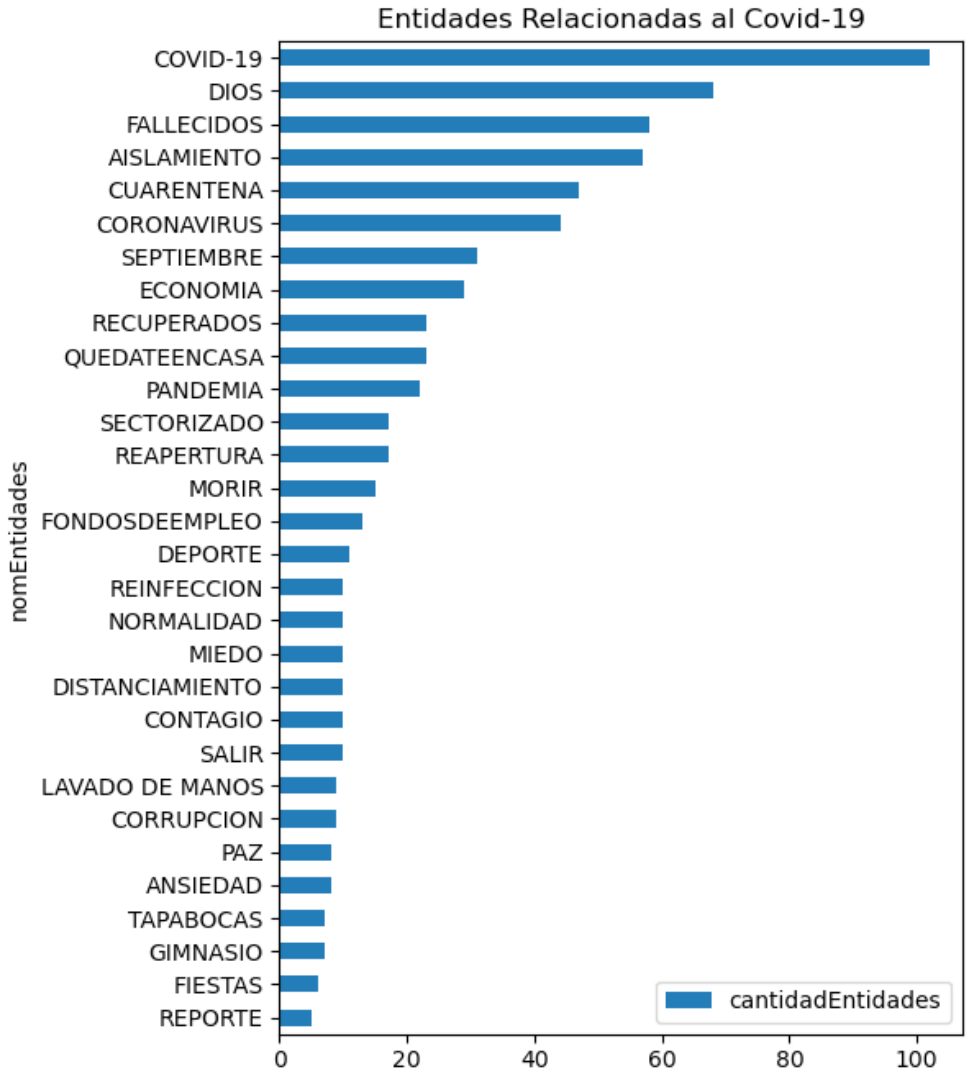


Figura 44: Resultado análisis de entidades

3.4..4 Comunicación:

Difundir mediante una presentación el resultado de la investigación.

4 CAPITULO IV. RESULTADOS

- En la revisión del estado del arte se pudo evidenciar la poca información que existe en cuanto investigaciones a tratar con análisis de entidades relacionadas al Covid-19 en Colombia utilizando tweets de la red social Twitter, puesto que es un tema relativamente nuevo.

- En cuanto al objetivo general se cumple con el uso de NLP mediante la herramienta de Spark NLP para realizar el análisis de entidades, tokenización permitiendo así, encontrar tendencias relacionadas con la problemática actual del Covid-19 en Colombia extrayendo los tweets de idioma español mediante el código de la API streaming de Twitter, la cual permite obtener los tweets en tiempo real tomando como fecha el día 24 de agosto del año 2020 por un periodo de 3 horas equivalentes a 10.800 segundos comenzando a las 8:15 pm y finalizando a las 11:15 pm, donde se obtienen 19.023 tweets en total.

- En cuanto a la pregunta de investigación, se implementa un código en lenguaje de programación de Python utilizando el entorno de desarrollo IDE Visual Studio Code, donde primeramente se pone la ubicación geográfica de Colombia en coordenadas que genera la herramienta GEOBOX, luego se determina el idioma que es Español, seguidamente se hace la autenticación de las claves de usuario que nos genera app developers de Twitter, también se definen las tendencias para buscar los tweets para esta investigación se utilizaron las tendencias: #covid_19, #Covid_19, #COVID_19, #covid-19, #Covid-19, #COVID-19, #Coronavirus, #Pandemia, #Aislamiento, #QuedateEnCasa, después utilizando un notebook llamado reading_spark.ipynb, el cual lee el archivo tweets.json que se genera al ejecutar file_te.py, utiliza pandas, pyspark, pyspark.sql. se indica la ruta donde se encuentra el ejecutable de pyspark, se crea una sesión de spark (se crea la app), luego spark carga el archivo tweets.json generado anteriormente en la ruta, se muestra el esquema de los tweets, se hace consulta con lenguaje sql de spark.sql donde se muestran los tweets generados solamente en Colombia y lógicamente de idioma español, en un archivo denominado output , se realiza una consulta donde se muestra el conteo de tweets por ubicación, luego, se realiza tokenización con

sparkml, donde se separa el texto y el conteo de longitudes, por ultimo se generan claves objeto valor para saber cuales son las entidades mas nombradas.

- La limpieza de los datos se realiza bajo conocimiento del contexto, logrando extraer las 30 entidades mas relacionadas a la problemática del Covid-19 en Tweets, donde la entidad mas nombrada es Covid-19 con 102 repeticiones y eliminando entidades no relacionadas al contexto y por ultimo se realiza la presentación de los resultados mediante un grafico de barras con el uso de la librería de pandas para python y matplotlib.

5 CONCLUSIONES

- Debido a la gran cantidad de datos extraidos mediante el uso de API Streaming que permite extraer tweets en tiempo real para esta investigación un lapso de 3 horas y mediante el uso de las diferentes librerías como: tweepy, pyspark, pandas, textblob, se hace necesario el uso de herramientas diseñadas para analizar Big Data como Apache Spark, dando como resultado la identificación de diferentes entidades en el idioma español relacionadas a la pandemia del Covid-19 en Colombia.
- Con la investigación desarrollada, se extrae un data set con 19.023 tweets, el día 24 de agosto del año 2020 durante un periodo de 3 horas equivalentes a 10.800 segundos, de los cuales se identificaron 30 principales entidades mediante el método de conocimiento del contexto, teniendo en cuenta que para el cumplimiento del objetivo general es satisfactorio dado a los datos obtenidos mediante el algoritmo.
- La extracción e identificación de entidades mediante NLP realizado en la presente investigación es eficiente para analizar gran cantidad de datos el cual pueden ser utilizados para trabajos de investigación de Big Data.

6 RECOMENDACIONES

A continuación, se describen algunas de las recomendaciones que pueden ser aplicadas en trabajos futuros con la intención de proporcionar mayor potencial y funcionamiento a la presente propuesta de investigación:

- Utilizar este modelo de extracción, identificación y análisis de tendencias con periodo de tiempo más largo que permita obtener una mayor cantidad de datos.
- Teniendo en cuenta que en la investigación realizada no se dispuso del suficiente tiempo para la recolección de la información es importante que se realicen estudios de identificación de identidades con diferentes temas: sociales, económicos, culturales, ambientales, tecnológicos entre otros, el cual permitan influir en toma de decisiones respecto a las diferentes problemáticas.

BIBLIOGRAFIA

[1] J. Jayadharshini, R. Sivapriya, y S. Abirami, «Trend square: An Android Application for Extracting Twitter Trends Based on Location», *Proc. 2018 Int. Conf. Curr. Trends Towar. Converging Technol. ICCTCT 2018*, pp. 1-5, 2018, doi: 10.1109/ICCTCT.2018.8551056.

[2] Diaz Mendivelso Johan David; Suarez Baron Marco Javier, «Análisis social aplicando técnicas de lenguaje natural a información extraída de twitter», *Sci. Tech.*, vol. 24, n.º 3, pp. 489-496, 2019, doi: 10.22517/23447214.21731.

[3] R. Hernandez Sampieri, C. Fernandez Collado, y M. del P. Baptista Lucio, *Metodologia de la investigación*, Quinta edi. Mc Graw HIll, 2010.

[4] O. F. Castellanos Domínguez, A. M. Fúquene Montañez, y D. C. Ramírez Martínez, *Análisis de tendencias: de la información hacia la innovación*, Primera ed. Bogotá DC.: Universidad Nacional de Colombia, 2011.

[5] D. Mir Montserrat y J. López Vicario, «Proyecto de grado-Analítica de datos en Twitter», Universidad Autónoma de Barcelona, 2015.

[6] C. D. Manning, J. Bauer, J. Finkel, y S. J. Bethard, «The Stanford CoreNLP Natural Language Processing Toolkit», *Aclweb.Org Proc. 52nd Annu. Meet. Assoc. Comput. Linguist. Syst. Demonstr.*, pp. 55-60, 2014, [En línea]. Disponible en: http://macopolo.cn/mkpl/products.asp.

[7] J. Camargo-Vega, J. Camargo-Ortega, y L. Joyanes-Aguilar, «Knowing the Big Data», *Fac. Ing. Univ. Pedagog. y Tecnol. Colomb.*, vol. 24, n.º 38, pp. 63-77, 2015.

[8] Ministerio de salud y protección social, «Abecé Nuevo Coronavirus (Covid-19)», *MinSalud*, 2020. https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/PP/ET/abece-coronavirus.pdf.

[9] N. R. González, «Proyecto Fin de Grado-DETECCION DE TENDENCIAS EN TWITTER UTILIZANDO MINERIA DE DATOS ADAPTATIVA», p. 47, 2014.

[10] C. Andres, O. Navia, D. Carrizo, y C. Ortiz, «Models of requirements elicitation process : A systematic mapping Modelos del proceso de educción de requisitos : Un mapeo sistemático Models of requirements elicitation process : A systematic mapping», n.º January, 2016, doi: 10.14482/inde.34.1.7958.

[11] B. Caimmi, S. Vallejos, L. Berdun, A. Soria, A. Amandi, y M. Campo, «Detección de incidentes de tránsito en Twitter», *IEEE Bienn. Congr. Argentina*, pp. 1-6, 2016, doi: 10.1109/argencon.2016.7585327.

[12] B. Billal, A. Fonseca, y F. Sadat, «Efficient Natural Language Pre-Processing for Analyzing Large Data Sets», *Proc. - 2016 IEEE Int. Conf. Big Data, Big Data 2016*, pp. 3864-3871, 2016, doi: 10.1109/BigData.2016.7841060.

[13] J. M. Moine, «Tesis de grado-Metodologías para el descubrimiento de conocimiento en bases de datos : un estudio comparativo», Universidad Nacional de la Plata, 2013.