INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO

Prototipo de análisis ciudadanos relacionados a contaminación ambiental.

Trabajo Terminal: No. 2021-B051

Integrantes:

Hernández Clemente Samantha

Medina Flores Susana

Olivares Conchillos Leonel

Director de TT:

Zagal Flores Eswart Roberto.

Índice.

RESUMEN

Las redes sociales generan una gran cantidad de datos no estructurados que se pueden utilizar como una fuente de información, en este caso tomaremos en cuenta las publicaciones referentes a eventos contaminantes que afectan la calidad del aire en Ciudad de México. En el presente Trabajo Terminal se planea realizar un prototipo de análisis de datos ciudadanos relacionados a contaminación ambiental, como puede ser tráfico ó pirotecnia, extrayendo información de redes sociales con un robot de extracción e implementando un modelo de minería de datos para su análisis, y así mostrar los resultados en un dashboard.

Palabras clave - Contaminación del aire, Minería de datos, Redes sociales.

CAPÍTULO 1: INTRODUCCIÓN

1.1 Generalidades.

El término contaminación se refiere a todo elemento o sustancia que no está originalmente presente en el entorno, afectando el equilibrio del ecosistema y causando daño hacia los seres vivos.

La contaminación hoy en día está presente en la mayor parte del mundo, debido a la variedad del cómo se presenta en el ambiente, se divide en tres categorías [1]:

- Contaminación según la extensión de la fuente: Se presentan la contaminación difusa, puntual y lineal.
- Contaminación dependiente del contaminante: Existen nueve tipos, entre ellas son: la contaminación química, acústica, radioactiva, térmica, visual, lumínica, electromagnética, microbiológica y genética.
- Contaminación según el medio afectado: Las tres más conocidas por la población, que son la contaminación atmosférica, del agua (hídrica) y del suelo.

La contaminación puntual es aquella polución o alteración específica por contaminación del aire, agua, suelo, lumínica, térmica, acústica y otros agentes contaminantes desde una zona o lugar fijo, haciendo que sea sencilla de localizar, identificar y controlar en un lugar determinado [2].

En el caso de la contaminación difusa es aquella polución cuya fuente no se localiza en un punto fijo o determinado, sino que el contaminante impacta principalmente sobre el aire y el agua desde variados puntos dispersos o una gran zona que hace que la misma no sea fácil de identificar. Mientras tanto, la contaminación de fuente lineal es la que se produce a lo largo de una línea como por ejemplo el tráfico originado en una autopista produciéndose la contaminación química y sónica [2]. Algunas de las causas de estas fuentes son las actividades humanas, los desastres naturales y el tráfico.

La contaminación dependiente del contaminante se refiere a los elementos que perjudican negativamente al medio y pueden presentarse en cualquier medio (agua, aire, suelo). Como se mencionó anteriormente, en este apartado incluye:

- Contaminación química: Proviene principalmente de los usos industriales.
- Contaminación radiactiva: Se deriva de la emisión de materiales radiactivos producto de accidentes en centrales nucleares o abandono deliberado de residuos radiactivos [1].
- Contaminación térmica: Se refiere a la emisión de fluidos a elevadas temperaturas en el lugar.
- Contaminación acústica. Es la producción de ruido excesivo por encima de los niveles naturales afectando la calidad de vida de los seres vivos en la zona.
- Contaminación visual. Aquella que destruye de forma visual un paisaje natural, como las torres de energía eléctrica, vallas publicitarias, vertederos, etcétera [1].
- Contaminación lumínica. Es toda aquella iluminación artificial que altera la oscuridad natural de la noche, provocada por luz desaprovechada, innecesaria o inadecuada, generada por el alumbrado de exteriores, el cual genera impactos en la salud y en la vida de los seres vivos [3].
- Contaminación electromagnética. Las radiaciones que generan los equipos electrónicos.
- Contaminación microbiológica. Es una alteración debido a un agente contaminante (microorganismos) ajeno al medio. Se da sobre todo en aguas servidas, subterráneas y terrestres [1].
- Contaminación genética. Afecta ante todo a las plantas cuando se produce una transferencia incontrolada de material genético en ellas [1].

Enfocándonos en la contaminación atmosférica, una de las causas principales es por sustancias químicas en la atmósfera que afectan a la calidad del aire del entorno. Algunos ejemplos de estas sustancias nocivas son el monóxido de carbono, el dióxido de azufre, CFCs (clorofluorocarbonos) y óxidos de nitrógeno [4].

Finalmente, en la categoría de contaminación según el medio afectado. La contaminación del agua o hídrica afecta directamente a las especies animales, vegetales y también al ser humano ya que convierte el agua potable en un recurso no apto para su consumo [4]. Por otro lado, la contaminación del suelo, es provocada principalmente por el aumento del uso de compuestos químicos en diferentes productos y la mala gestión de los residuos son algunas de las causas más evidentes de la contaminación del suelo [5]. Todas ellas tienen al ser humano como uno de los principales causantes provocando la alteración de estos medios.

Según la Organización Mundial de la Salud (OMS), en 2012 la contaminación del aire fue responsable de 3.7 millones de muertes en el planeta (11% por enfermedad pulmonar obstructiva crónica, 6% por cáncer de pulmón; 40% por enfermedad isquémica del corazón, 40% por accidente cerebrovascular y alrededor de 3% por infección respiratoria aguda) [6]. En particular, la contaminación del aire puede causar problemas cardiovasculares, alergias, ataques de asma, conjuntivitis, enfermedades bronquiales, cáncer de pulmón o piel, problemas de visión,

problemas sanguíneos en el desarrollo mental del niño, entre otros. Los más vulnerables son los niños, los ancianos, las mujeres embarazadas y los enfermos [7].

Además de los efectos sobre la salud de las personas, la contaminación atmosférica también afecta a los bosques y ecosistemas acuáticos, debido a la presencia de contaminantes como los óxidos de nitrógeno y de azufre, los cuales se producen por la quema de combustibles fósiles y que, al combinarse con el agua presente en la atmósfera, provocan el fenómeno conocido como lluvia o deposición ácida [6].

En la actualidad, la contaminación del aire es un gran problema y cada vez la calidad del aire va disminuyendo. En la Ciudad de México, la contaminación ha crecido año tras año, afectando la calidad de vida y el entorno en el que vivimos, nos volvemos más vulnerables a cierto tipo de enfermedades y síntomas consecuentes del estado actual en que se encuentra el medio ambiente.

Una de las principales fuentes que afecta a la calidad del aire son los vehículos. Comprende de una serie de contaminantes tales como: el monóxido y bióxido de carbono, los hidrocarburos, los óxidos de nitrógeno y las partículas, además, contaminantes como el azufre y, hasta hace algunos años, el plomo se liberaba en el ambiente por el proceso de combustión de los automóviles [8].

De acuerdo con el primer Inventario Nacional de Emisiones de México, 1999, los vehículos automotores contribuyen con el 31% de la emisión de óxidos de nitrógeno, 62% de monóxido de carbono y 22% de las emisiones totales estimadas de compuestos orgánicos volátiles [8].

Otra fuente contaminante son los incendios. El día 13 de mayo de 2019 se habían registrado 23 incendios en once alcaldías, de los cuales 16 correspondían a incendios forestales y 7 a incendios urbanos [9].

Una vez que se presenta un incendio se pueden presentar emisiones significativas de distintos contaminantes como monóxido de carbono, óxidos de nitrógeno, dióxido de azufre y partículas (PM10, PM2.5) [9]. La materia particulada (PM), son partículas muy pequeñas que también se atribuyen a la contaminación del aire. Son las más dañinas, además que han sido catalogadas por la Organización Mundial de la Salud como carcinógenas [9].

La PM2.5 se trata de una mezcla que puede incluir sustancias químicas orgánicas, polvo, hollín y metales, y regularmente estas partículas pueden provenir de los automóviles, camiones, fábricas, quema de madera y otras actividades [9].

Como hemos visto, existen múltiples fuentes contaminantes hacia la atmósfera, unos más conocidos que otros, y hay uno en particular que, a pesar de no aparecer en gran medida como las demás, se presenta en meses concretos que, de igual forma, influyen en la calidad del aire: La pirotecnia.

Neutralizantes, oxidantes y aglomerantes se mezclan en la pirotecnia, además del perclorato de sodio que da propulsión al cohete, los metales pesados que aportan el color y los aerosoles que producen la detonación [10].

Ya en los aires, esa mezcla libera, entre otros, monóxido de carbono (CO) y partículas suspendidas (PM2.5), y junto con las emisiones del transporte, fábricas, fogatas, calentones y quema de llantas o basura, genera, sobre todo los días 12 y 25 de diciembre, y en la primera semana del mes de enero, alta contaminación, escasa visibilidad y sensación de neblina [10].

Por otro lado, la manera de comunicarnos ha evolucionado en grandes pasos, en la actualidad con un simple clic podemos conversar en cualquier parte, asimismo informarnos de lo que sucede a nuestros alrededores. Todo esto es posible gracias a las redes sociales, las cuales no solo nos permiten comunicarnos independientemente de la distancia, sino que también nos deja compartir y crear contenido de cualquier tipo, esto permite, por un lado, informar a las personas de lo que está sucediendo por su zona, por ejemplo, el tan solo subir una foto del cielo grisáceo ya da a entender que en ese sitio se presenta contaminación atmosférica.

Tanto las autoridades como los ciudadanos son conscientes de que la situación es grave, por lo que, las organizaciones correspondientes buscan medidas para contrarrestar, y a la vez, los ciudadanos buscan aportar información relacionada con este tema a través de distintos medios de comunicación, principalmente en las redes sociales.

1.2 Planteamiento del Problema.

La comunicación a través de las redes sociales se ha vuelto de vital importancia en los últimos años, no solo como forma de entretenimiento sino como una alternativa para transmitir información, que toma los elementos, recursos y características de los medios tradicionales pero que incorpora un nivel de interacción más grande [11].

Actualmente podemos notar un incremento de publicaciones mediante redes sociales donde la comunidad denuncia o alerta a los demás sobre diferentes temas.

Como hemos mencionado en las generalidades, hoy en día la calidad del aire es bastante mala y por lo regular podemos encontrar sitios web donde podemos visualizar la calidad del aire en tiempo real, dicha información se basa en el monitoreo de calidad del aire y de emisiones atmosféricas, pero en redes sociales existen datos que pueden ayudar a identificar denuncias de carácter social relacionadas al medio ambiente, por ejemplo, existen ciudadanos que reportan a las autoridades o comunidades vecinales situaciones que afectan la calidad del aire y a la convivencia social; esta información no ha sido analizada para encontrar una caracterización de estos eventos contaminantes que pudiera ayudar a encontrar situaciones de riesgo al ambiente u otros seres vivos. Esta información es una fuente alternativa de denuncias ciudadanas que, al tratarse de datos no estructurados, no se toman en cuenta en el monitoreo del impacto ambiental por parte de las autoridades correspondientes.

El reto técnico es obtener, limpiar y analizar una gran cantidad de información sobre la calidad del aire que nos arroje un valor sobre la contaminación en la Ciudad de México y nos permita generar datos sobre la contaminación a partir de las denuncias ciudadanas en las redes sociales.

1.3 Objetivos.

1.3.1. Objetivo General.

Desarrollar un prototipo de software que permita analizar descriptivamente publicaciones extraídas de Twitter relacionadas a situaciones de contaminación del medio ambiente en Ciudad de México a fin de obtener una caracterización espacial y temporal sobre eventos que generan contaminación del aire como es el caso de la pirotecnia, el tráfico y los incendios.

1.3.2. Objetivos Específicos.

- Construir un proceso de extracción, transformación y carga de publicaciones en Twitter.
- Generar análisis para categorizar los datos extraídos de Twitter.
- Generar un proceso de análisis de datos ciudadanos.
- Generar un proceso semiautomático de análisis de datos sociales para generar modelos de temas dentro del conjunto de datos extraídos.
- Detectar y caracterizar publicaciones en twitter relacionado a eventos o situaciones de contaminación del ambiente
- Identificar, analizar y delimitar los eventos que describen situaciones que generen contaminación.
- Desarrollar un tablero o dashboard de datos para presentar los resultados del análisis de los datos.

1.4. Justificación.

En 2015, la OMS (Organización Mundial de la Salud) reconoció que la contaminación atmosférica es el riesgo ambiental más perjudicial para la salud humana a nivel mundial [12].

Por encima de la contaminación del agua y del suelo, la contaminación del aire, a pesar de que las tres tienen un gran peso dañino para el ecosistema, la contaminación atmosférica es la que más destaca y tiene mayores puntos de contacto hacia el ser humano y en México no es la excepción.

En ese mismo año, la contaminación del aire fue el noveno factor de riesgo de muerte y discapacidad en México. Con un total de 14 666 muertes y 150 771 años de vida ajustados por discapacidad [12].

En el año 2018, la Secretaría del Medio Ambiente de la Ciudad de México (SEDEMA), en el informe anual de calidad de aire publicó que el índice de la calidad del aire, en la mayoría del año se presentó una mala calidad del aire, con 276 días que superaron los 100 puntos, el valor máximo indicado por la Norma Oficial Mexicana de salud (NOM) y sólo 89 días donde ningún contaminante superó dicha cifra [13].

La Procuraduría Ambiental y del Ordenamiento Territorial de la Ciudad de México (PAOT) es un organismo público descentralizado de la Administración Pública que una de sus funciones es defender los derechos de la población referente al medio ambiente [13]. La ciudadanía puede presentar una denuncia referente a daños o incumplimiento de las normas al medio ambiente. En el año 2021, recibió más de 9,000 denuncias, donde apenas el 1.3% son acerca de las emisiones del aire [14].

Algunas de las fuentes más conocidas de compuestos orgánicos volátiles (COV) y óxidos de nitrógeno que afectan a la calidad del aire son las emisiones de vehículos e industrias, aerosoles, aromatizantes, entre otros [15]. Estas sustancias son comunes en la vida diaria, pero también hay días específicos en los cuales se generan una gran cantidad de ozono dañino: en los eventos sociales. Hay eventos o festividades que ocurren a lo largo del año en los cuales la contaminación se incrementa de forma alarmante, por ejemplo, el 15 de septiembre, en la Ciudad de México se celebra el grito de la Independencia, desde hace varios años se acostumbra a utilizar grandes cantidades de pirotecnia que provocan la liberación de sustancias tóxicas al aire, incluso hoy en día, terminando en incendios forestales o daño a infraestructura.

Esta clase de eventos también aumentan considerablemente el tráfico por varias rutas, ocasionando una fuerte concentración de sustancias perjudiciales para la calidad del aire. Sucede de forma similar en algunas marchas, que se apoyan de pirotecnia o productos similares que aumentan el factor contaminante en el área. La mayoría de esta información no se anuncia de forma oficial o no se toma en cuenta, por lo que, los ciudadanos buscan otros medios para dar a conocer e informarse acerca de estos eventos que dañan al medio ambiente: las redes sociales.

En cuanto al año anterior del 2021, México se localizó en el lugar número 11 en la categoría de los países más contaminados por PM2.5, uno de los distintos elementos nocivos al medio ambiente y en el lugar número 16 con índice de muertes atribuibles con 340 mil muertes [16].

Es importante concientizar a las personas sobre la contaminación al medio ambiente ya que esto trae repercusiones en el cambio climático y el calentamiento global. En los últimos 50 años, el aumento en la formación de contaminantes y gases invernadero ha propiciado el alza de temperaturas y el cambio climático. Este aumento de temperatura ha sido acompañado con la degradación de la calidad del aire en distintas partes del mundo. En particular, la Organización Mundial de la Salud (OMS) reportó que mueren alrededor de 4.2 millones de personas al año debido a la contaminación en el exterior [17].

Como ya hemos observado, las personas están proporcionando de forma indirecta información del impacto real de eventos contaminantes los cuales no han sido recopilados, integrados, y analizados de forma específica en la Ciudad de México, de acuerdo con la revisión del estado del arte.

Por ello, nosotros vamos a desarrollar un prototipo de software que analiza de forma descriptiva publicaciones de las redes sociales referentes a eventos que dañan al medio ambiente como pueden ser el tráfico y el uso de pirotecnia para determinar el impacto del lado de la ciudadanía en la Ciudad de México.

Obtendremos los datos no estructurados de una de las redes sociales en México: Twitter. Esta red social no necesariamente es la más popular ya que Twitter ocupa el quinto lugar en preferencia [18], sin embargo, esta red social nos permite obtener la información de diferentes hashtags clave para identificar denuncias sobre contaminación al medio ambiente. Por otro lado, las características de esta red social nos permiten obtener de manera más sencilla la información, por ejemplo, la principal característica de Twitter es su sencillez y capacidad de sintetizar, pues el tope de escritura es de alrededor de 140 caracteres [19]. Además, de cada 10 usuarios de Internet (más de 80 millones), 3 están en Twitter. La mayor actividad en la twittósfera mexicana se realiza en la Ciudad de México (60%), en Monterrey (17%) y en Guadalajara (10%) [19].

1.5. Alcances.

Debido a que nuestra disposición de recursos es limitada, solamente recolectaremos publicaciones de Twitter. De igual forma, las publicaciones que se tomarán en cuenta serán tweets de los años 2019, 2020, 2021 y 2022, con el motivo de hacer una comparativa más adelante entre la información obtenida por cada año.

Los resultados obtenidos después del análisis de información serán plasmados en un dashboard. Cabe destacar que nuestro punto de interés es la información que el tweet puede brindar referente a la contaminación atmosférica, es decir, no nos estamos enfocando a la persona que creó dicho tweet, por lo cual no se llevará a cabo un análisis de sentimientos sobre estas publicaciones.

En cuanto a la información de los tweets, se recopilarán los que entran en la categoría de contaminación del aire y únicamente referente a los tres temas seleccionados: pirotecnia, incendios y tráfico, los cuáles son los que identificamos como los más presentes o posibles de ocurrir en consecuencia de los eventos sociales que se presentan en la Ciudad de México, además de que son fuentes altas de emisión de partículas nocivas mencionadas como el PM2.5. Con el fin de que queremos descubrir si dichos eventos son realmente un gran factor influyente en la contaminación atmosférica.

La razón por la que nos concentraremos en un tipo de contaminación por el medio, es decir, la contaminación del aire es porque, al menos en México, es el medio más contaminado de todos en base a la investigación realizada previamente.

CAPÍTULO 2: MARCO TEÓRICO

A continuación, se presentarán los conceptos clave y una breve explicación de estos para la comprensión del contenido de este documento.

API

Una API o interfaz de programación de aplicaciones es un conjunto de definiciones y protocolos que se usa para diseñar e integrar el software de las aplicaciones.[20]

Permite a las empresas abrir los datos y la funcionalidad de sus aplicaciones a desarrolladores externos, socios comerciales y departamentos internos dentro de sus empresas. Esto permite que los servicios y productos se comuniquen entre sí y aprovechen los datos como la funcionalidad de los demás a través de una interfaz documentada. Los desarrolladores no necesitan saber cómo se implementa una API; simplemente usan la interfaz para comunicarse con otros productos y servicios.

Una API es un conjunto de reglas definidas que explican cómo las computadoras o las aplicaciones se comunican entre sí. Las API se ubican entre una aplicación y el servidor web, actuando como una capa intermediaria que procesa la transferencia de datos entre sistemas. [21]

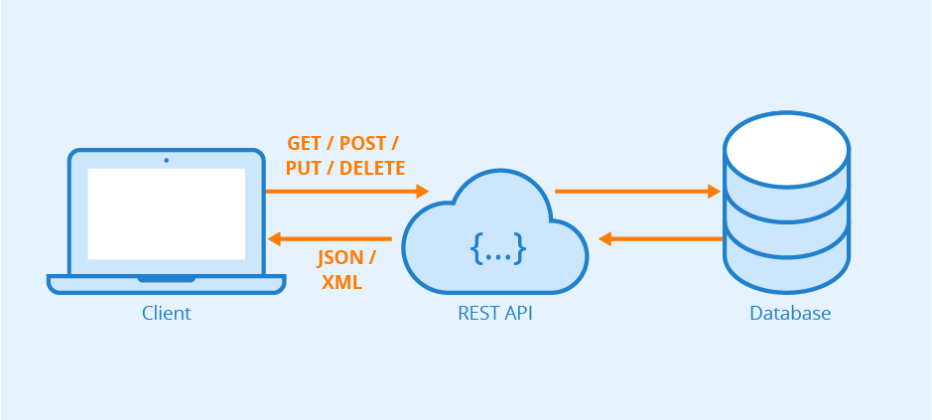

Figura 1. Diagrama lógico de una API.

Data Mining

La minería de datos, también conocida como descubrimiento de conocimiento en datos (KDD), es el proceso de descubrir patrones y otra información valiosa de grandes conjuntos de datos. Dada la evolución de la tecnología de almacenamiento de datos y el crecimiento de big data, la adopción de técnicas de minería de datos se ha acelerado rápidamente en las últimas dos décadas, ayudando a las empresas a transformar sus datos sin procesar en conocimiento útil. Sin embargo, a pesar de que esa tecnología evoluciona continuamente para manejar datos a gran escala, los líderes aún enfrentan desafíos con la escalabilidad y la automatización.

La minería de datos ha mejorado la toma de decisiones organizacionales a través de análisis de datos perspicaces. Las técnicas de minería de datos que sustentan estos análisis se pueden dividir en dos propósitos principales; pueden describir el conjunto de datos de destino o pueden predecir resultados mediante el uso de algoritmos de aprendizaje automático.

El proceso de minería de datos implica una serie de pasos desde la recopilación de datos hasta la visualización para extraer información valiosa de grandes conjuntos de datos. Los científicos de datos describen los datos a través de sus observaciones de patrones, asociaciones y correlaciones. También clasifican y agrupan datos a través de métodos de clasificación y regresión, e identifican valores atípicos para casos de uso, como la detección de spam.

ETL

Extracción, transformación y carga (ETL) es una canalización de datos que se usa para recopilar datos de varios orígenes. A continuación, transforma los datos según las reglas de negocio y los carga en un almacén de datos de destino. El trabajo de transformación en ETL tiene lugar en un motor especializado y, a menudo, implica el uso de

tablas de almacenamiento provisional para conservar los datos temporalmente a medida que estos se transforman y, finalmente, se cargan en su destino.

La transformación de datos que tiene lugar a menudo conlleva varias operaciones como filtrado, ordenación, agregación, combinación de datos, limpieza de datos, desduplicación y validación de datos.

Frecuentemente, las tres fases del proceso ETL se ejecutan en paralelo para ahorrar tiempo. Por ejemplo, mientras se extraen datos, puede que esté funcionando un proceso de transformación sobre los datos ya recibidos y de preparación para la carga, y puede que empiece a funcionar un proceso de carga sobre los datos preparados, en lugar de tener que esperar a que termine todo el proceso de extracción. [22]

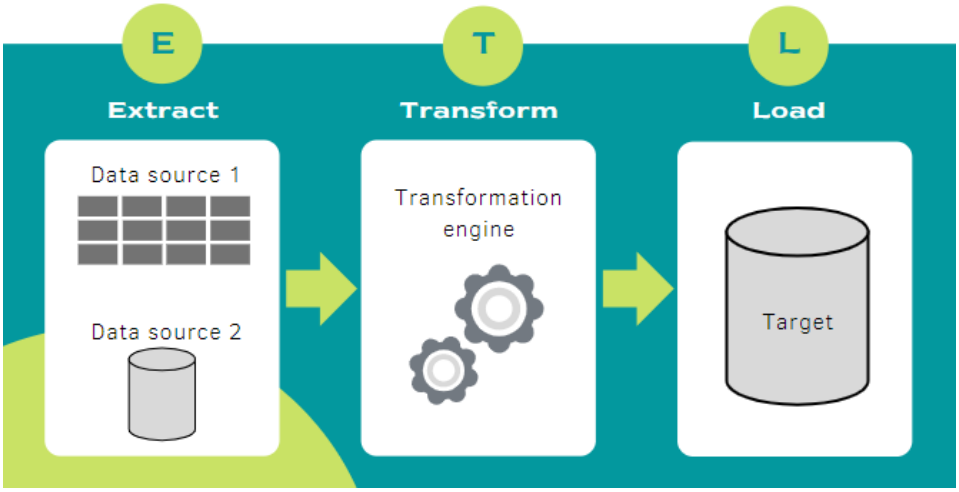

Figura 2. Diagrama de fases de un proceso ETL.

Procesamiento de lenguaje natural (NLP)

El procesamiento del lenguaje natural (NLP, por sus siglas en inglés) se refiere a la rama de la informática, y más específicamente, la rama de la inteligencia artificial o IA, que se ocupa de brindar a las computadoras la capacidad de comprender textos y palabras habladas de la misma manera que los seres humanos. [23]

El procesamiento de lenguaje natural (NLP) se utiliza para tareas como el análisis de opiniones, la detección de temas, la detección de idioma, la extracción de frases clave y la clasificación de documentos. [24] Estos enfoques utilizan muchas técnicas del procesamiento de lenguaje natural, como:

- **Tokenizador**. Dividir el texto en palabras o frases.
- **Lematización**. Normalizar las palabras para asignar distintas formas a la palabra canónica con el mismo significado. Por ejemplo, "corriendo" y "corrió" se asignan a "correr".
- **Extracción de entidades**. Identificación de sujetos en el texto.
- **Detección de partes de la oración**. Identifica el texto como un verbo, nombre, participio, frase verbal, etc.
- **Detección del límite de las frases**. Detectar frases completas en párrafos de texto.

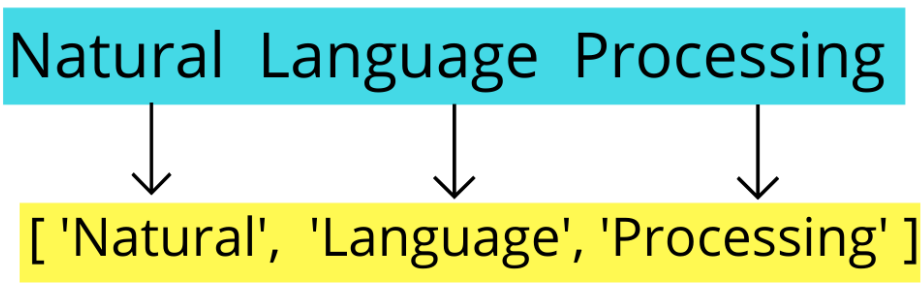

Figura 3. Resultado de un proceso de tokenización. [24]

Data Warehouse

Es un repositorio centralizado de datos integrados procedentes de uno o varios orígenes dispares. Los almacenamientos de datos almacenan datos históricos y actuales, y se utilizan para realizar informes y análisis de los datos.

Para mover los datos a un almacenamiento de datos, estos se extraen de forma periódica de diversos orígenes que contienen información empresarial de importancia. Cuando se mueven los datos se puede dar formato, limpiar, validar, resumir y reorganizar. Como alternativa, los datos pueden almacenarse en el nivel más bajo de detalle, con vistas agregadas proporcionadas por el almacenamiento para realizar informes. En cualquier caso, el almacenamiento de datos se convierte en un almacén de datos permanente para informes, análisis e inteligencia empresarial (BI). [25]

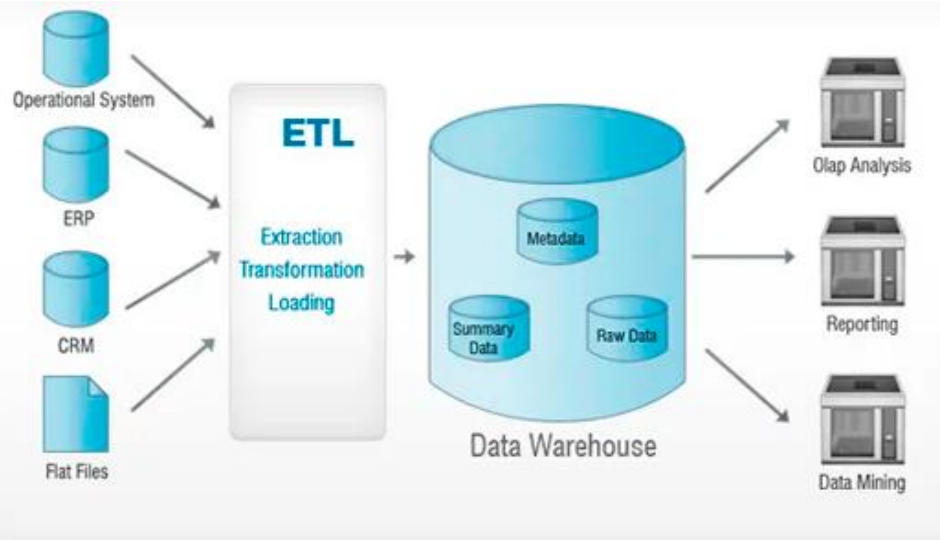

Figura 4. Diagrama de un proceso de análisis de datos con un Data Warehouse.

Modelo de Estrella

Un esquema de estrella es un tipo de esquema de base de datos relacional que consta de una sola tabla de hechos central rodeada de tablas de dimensiones.

En la siguiente figura se muestra un esquema de estrella con una sola tabla de hechos y cuatro tablas de dimensiones. Un esquema de estrella puede tener cualquier número de tablas de dimensiones. Las ramas situadas al final de los enlaces que conectan las tablas indican una relación de muchos a uno entre la tabla de hechos y cada tabla de dimensiones. [26]

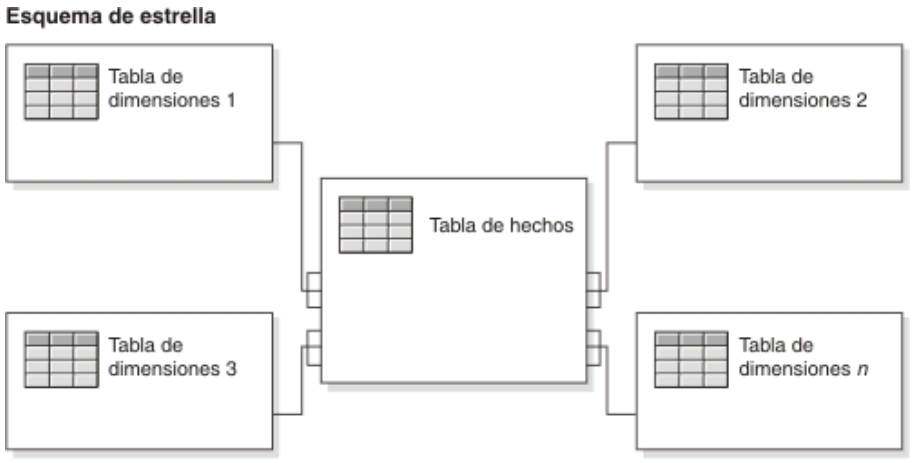

Figura 5. Esquema de estrella con una sola tabla de hechos con enlaces a varias tablas de dimensiones.

Aprendizaje supervisado

Los algoritmos de aprendizaje supervisado basan su aprendizaje en un juego de datos de entrenamiento previamente etiquetados. Por etiquetado entendemos que para cada ocurrencia del juego de datos de entrenamiento conocemos el valor de su atributo objetivo. Esto le permitirá al algoritmo poder "aprender" una función capaz de predecir el atributo objetivo para un juego de datos nuevo. Las dos grandes familias de algoritmos supervisados son:

- Los algoritmos de regresión cuando el resultado a predecir es un atributo numérico.
- Los algoritmos de clasificación cuando el resultado a predecir es un atributo categórico.

Aprendizaje no supervisado

Los métodos no supervisados (unsupervised methods en inglés) son algoritmos que basan su proceso de entrenamiento en un juego de datos sin etiquetas o clases previamente definidas. Es decir, a priori no se conoce ningún valor objetivo o de clase, ya sea categórico o numérico. El aprendizaje no supervisado está dedicado a las tareas de agrupamiento, también llamadas clustering o segmentación, donde su objetivo es encontrar grupos similares en el conjunto de datos.

Existen dos grupos principales de métodos o algoritmos de agrupamiento:

1. Los métodos jerárquicos, que producen una organización jerárquica de las instancias que forman el conjunto de datos, posibilitando de esta forma distintos niveles de agrupación.
2. Los métodos particionales o no jerárquicos, que generan grupos de instancias que no responden a ningún tipo de organización jerárquica.

Modelado de tópicos (Topic Modelling)

El modelado de tópicos o temas es una técnica de aprendizaje automático que analiza los datos de texto para categorizar las palabras de un conjunto de documentos en grupos. Esto se conoce como aprendizaje automático "no supervisado" porque no requiere una lista predefinida de etiquetas o datos de entrenamiento que hayan sido previamente clasificados por humanos.

Dado que el modelado de temas no requiere capacitación, es una manera rápida y fácil de comenzar a analizar sus datos. Sin embargo, no puede garantizar que recibirá resultados precisos, razón por la cual muchas empresas optan por invertir tiempo en la capacitación de un modelo de clasificación de temas. [27]

En realidad, sólo observamos los documentos (no los temas). La otra estructura es subyacente o latente. El objetivo es inferir las variables ocultas (temas), dado lo que observamos (documentos). [28]

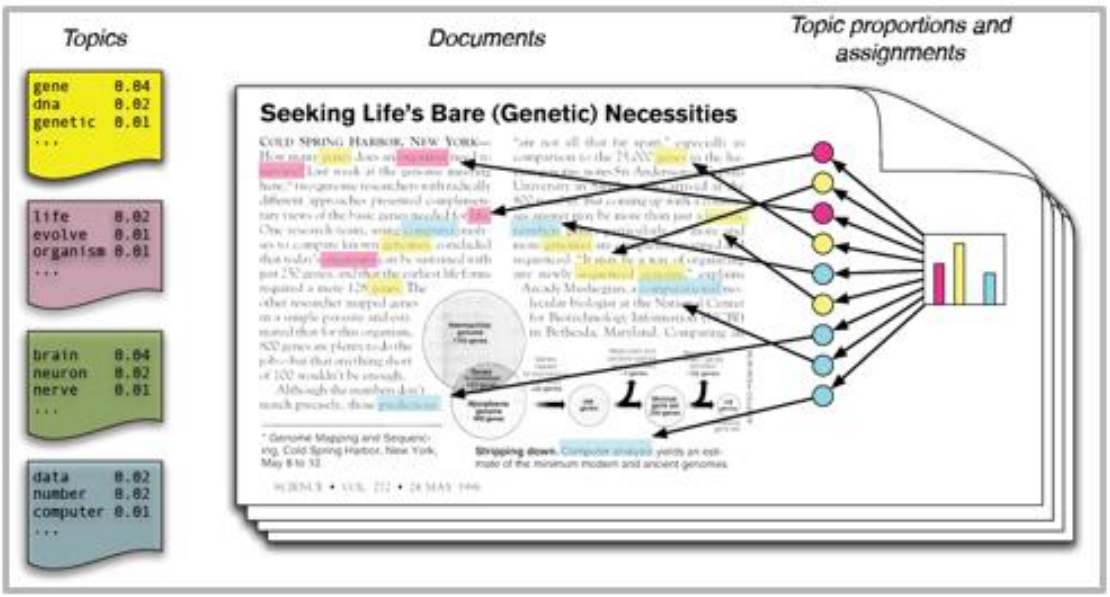

Figura 6. Fases de un Modelado de tópicos.

Asignación latente de Dirichlet (LDA)

LDA es una de las técnicas de NLP más populares para la detección de temas. Extrae temas de un corpus de texto basado en probabilidades de palabras: para cada tema latente, extrae la distribución de probabilidad de una combinación de palabras, lo que ayuda a identificar los temas principales. [29]

LDA funciona de la siguiente manera:

1. Primero, LDA requiere que la investigación especifique un valor de k o el número de temas en el corpus. En la práctica, esta es una decisión muy difícil y consecuente.
2. Cada palabra que aparece en el corpus se asigna aleatoriamente a uno de los k temas. Si es estricto con los detalles, técnicamente esta asignación no es aleatoria, ya que involucra una distribución de Dirichlet que emplea una probabilidad simplex en lugar de números reales (esto simplemente significa que los números asignados en los k temas suman 1).
3. Las asignaciones de temas para cada palabra se actualizan de forma iterativa actualizando la prevalencia de la palabra en los k temas, así como la prevalencia de los temas en el documento. Las asignaciones de temas se actualizan hasta un umbral especificado por el usuario, o cuando las iteraciones comienzan a tener poco impacto en las probabilidades asignadas a cada palabra en el corpus.

LDA, y la mayoría de las otras formas de modelado de temas, producen dos tipos de resultados. En primer lugar, se pueden identificar las palabras que se asocian con mayor frecuencia a cada uno de los k temas especificados por el usuario. En segundo lugar, LDA produce la probabilidad de que cada documento dentro del corpus esté asociado también con cada uno de los k temas especificados por el usuario. [31][30]

CAPÍTULO 3: ESTADO DEL ARTE

Los proyectos similares a nuestro trabajo terminal que se han desarrollado se muestran en la Tabla 1:

SOFTWARE	CARACTERÍSTICAS	PRECIO EN EL MERCADO
[32] Buscador geosocial para monitoreo de polución del aire urbano.	Un mecanismo geosocial de búsqueda basado en palabras clave para registrar patrones espaciales de denuncias a la calidad del aire a partir de publicaciones de Twitter en las regiones de Francia, Brasil y China.	No tiene un precio definido por el momento.
[33] Análisis de correlación entre el índice de calidad del aire y el impacto en Twitter para la ciudad de Bucaramanga (Colombia) aplicando análisis de series temporales, extracción y procesamiento de lenguaje natural.	Aplicación de una serie de modelos y algoritmos que predicen el índice de la calidad del aire a partir de las publicaciones realizadas en la red social twitter de los usuarios habitantes de una ciudad de Colombia, Bucaramanga.	No tiene un precio definido en el mercado.
[34] Análisis del efecto en las redes sociales de las crisis de contaminación en Madrid.	Análisis de las publicaciones de la red social twitter para determinar si es un medio indicativo que permita inferir el nivel de contaminación en una zona de Madrid específica.	No tiene precio en el mercado.
[35] De tweets a trayectorias semánticas: Minería de patrones urbanos anómalos de movilidad.	Análisis de trayectorias extraídas por referencias geográficas proporcionadas por redes sociales (mayormente Twiter) para la detección de anomalías y patrones de movilidad en Barcelona.	Tiene un precio de 29.95 USD.

Tabla 1. Resumen de proyectos similares.

En el primer trabajo, los autores desarrollaron una implementación de un mecanismo geosocial de búsqueda basado en palabras clave para registrar patrones espaciales de denuncias referentes a la calidad del aire en publicaciones de Twitter [32].

Los resultados que se obtuvieron mostraron una significante correlación a lo largo del tiempo en una serie de ciudades pertenecientes a los países Francia, Brasil y China.

Para poder realizar su trabajo se apoyaron de un diccionario sobre términos de polución, lo cual ayudó a identificar las publicaciones relevantes que se habían hecho en twitter, una vez identificadas las categorizaron y posteriormente las mapearon en diferentes vecindarios urbanos para poder realizar un comparativa sociocultural.

En la Figura 7, además de mostrar la visualización de la distribución de publicaciones, también marca los puntos dónde se localizan las estaciones que se encargan de detectar el valor de la calidad del aire.

São Paulo presenta más puntos críticos de tweets, principalmente por la densidad de población que tiene el país, por ello, presenta mayores concentraciones de publicaciones que París.

A continuación, se presentarán los resultados obtenidos, mediante mapas de calor de los países de París y São Paulo:

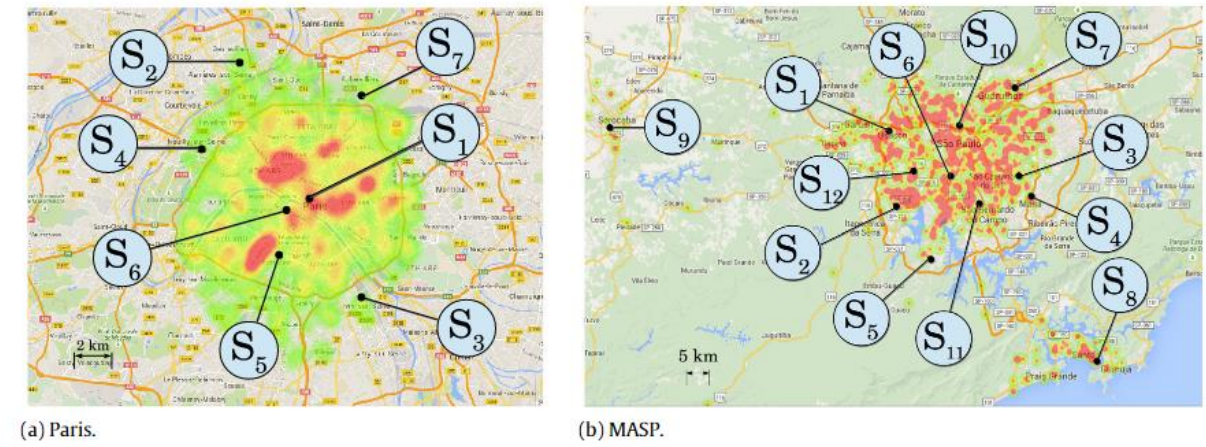

Figura 7. Mapas de calor de tweets publicados de los países de París y São Paulo.

El siguiente trabajo se trata de una tesis acerca de un análisis entre datos sacados de una empresa dedicada a recopilar los datos referentes a la calidad del aire con la información recolectada de las publicaciones de la red social Twitter [33]. Consistió en comparar la información recolectada por parte de las estaciones que tiene esta empresa con los datos recopilados a través de Twitter.

La técnica de recolección de datos aplicada fue Web Scraping (extracción de información en sitios web), debido a que no tenía tantas restricciones como la API de Twitter y permite acceder a cualquier publicación existente en la plataforma. En el caso de las técnicas de procesamiento de datos, se utilizaron Machine Learning para el procesamiento de Lenguaje Natural y herramientas de análisis estadístico.

Posteriormente, realizar un análisis exhaustivo de la información obtenida sobre la correlación con el índice de la calidad del aire en un país de Colombia. Por ejemplo, al llevar a cabo el análisis, se detectó que en las estaciones de monitoreo tenían datos particulados históricos de PM10 y PM2.5 incompletos como se muestra en la Figura 8. Para completar la información faltante se emplearon técnicas de Deep Learning y varios modelos de regresión de redes neuronales.

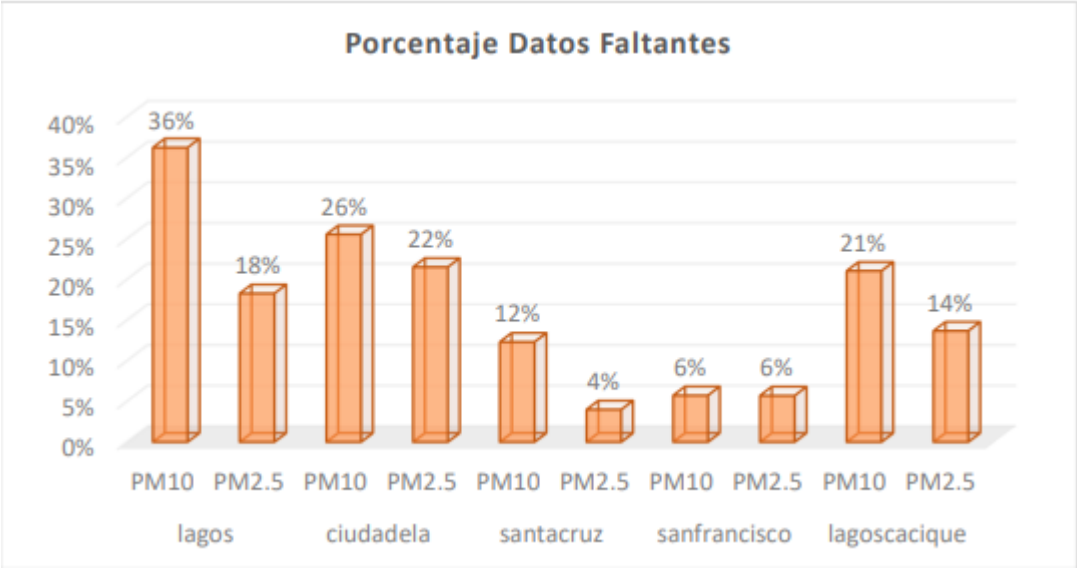

Figura 8. Gráfica de contraste de datos faltantes sobre la calidad de aire sobre estaciones de Colombia.

Otro trabajo similar, es un proyecto de titulación referente al efecto en las redes sociales en momentos críticos de contaminación en la zona de Madrid [34]. En este trabajo se llevó a cabo un análisis de las publicaciones de Twitter con la finalidad de identificar si es un medio confiable para detectar el nivel de la contaminación de un país, en este caso Madrid, para esto, compararon los datos que obtuvieron con otras publicaciones que ocurrieron fuera de la alerta de contaminación, y se evaluó la probabilidad de ocurrencia de esos eventos utilizando distribución de Poisson, de este mismo modo llevaron a cabo la evaluación del impacto ambiental con los resultados previamente obtenidos.

El último que vamos a mencionar es un artículo científico que habla acerca de las técnicas empleadas para detectar patrones urbanos de movilidad mediante información de referencia geográfica de redes sociales (una de ellas Twitter). Al recolectar una gran cantidad de tweets con geolocalización de un área específica a través del tiempo, semánticamente enriquece las publicaciones disponibles con datos sobre el autor (si es residente o turista) y el propósito de su trayectoria [35].

Entre estas técnicas se utilizó minería de datos y análisis de redes sociales. En los resultados mostrados se señalan los puntos con mayor flujo de personas y variaciones en el flujo del tiempo.

CAPÍTULO 4: ANÁLISIS

4.1 Necesidades básicas.

El propósito de este trabajo es ampliar el alcance de los sistemas que reportan cantidades de denuncias ciudadanas utilizando como fuente principal una de las redes sociales más utilizadas para hacer denuncias o reportes de lo que pasa día a día en la vida de los usuarios, Twitter.

Con la minería de datos realizaremos una exploración de los tweets que contengan las palabras clave de contaminación que analizamos anteriormente.

El principal problema que requerimos resolver para llevar a cabo este trabajo fue encontrar la manera de contribuir y compartir los cambios como avances de cada uno durante su desarrollo, asimismo tener dicho proyecto en un solo sitio y que pueda accederse por cada uno de nosotros en cualquier momento. Una forma de solucionarlo al principio fue subir todos nuestros avances a un repositorio al cual todos tuviéramos acceso, una vez que ya teníamos nuestro robot de extracción decidimos alojar todo en una máquina virtual, para que de esta manera pudiéramos acceder a nuestro trabajo desde cualquier equipo.

El siguiente problema que se nos presentó fue buscar una alternativa para extraer tweets de años anteriores, debido a que la cuenta de desarrollador que Twitter nos otorgó solo permitía acceder a publicaciones con no más de una semana de antigüedad, además de que algunas características necesarias estaban desactivadas en la versión actual, por lo tanto, decidimos utilizar una librería externa capaz de sobrepasar dichas restricciones y desbloquear tales funciones. Cabe destacar que a la hora de extracción ocupamos ambos métodos, el método original donde diariamente extraímos publicaciones con una antigüedad de una semana y el método donde podíamos extraer en un rango de fechas sin importar la antigüedad que tuvieran las publicaciones.

Después, se requirió la implementación de un sistema automático para el robot de extracción, con el fin de que la extracción de tweets ya no fuera de forma manual y se ejecutara cada día sin la necesidad que de una persona diera la instrucción. Al principio comenzamos a hacer la extracción de los datos diariamente, pero de forma manual, es decir, todos los días teníamos que ejecutar nuestro robot y extraer los datos, pero una vez que adquirimos la máquina virtual, alojamos nuestro robot y logramos automatizarlo para que la extracción se hiciera día con día sin la necesidad de que nosotros estuviéramos ejecutando el robot diariamente.

Una necesidad que detectamos es la visualización de los diagramas o gráficas resultantes del análisis de datos, esto lo resolveremos integrando todos los resultados a un Dashboard en un sitio web, con ayuda de las herramientas de Tableau y Python.

4.2. Requisitos de software.

El primer paso y primera necesidad es obtener la base histórica delimitada anteriormente de tweets con palabras clave como incendio, tráfico y pirotecnia.

Con un módulo ETL podemos realizar la extracción de tweets, identificar y limpiar las irregularidades en cada uno de los metadatos extraídos y crear nuestra data warehouse, de donde vamos a extraer los datos específicos para realizar nuestro análisis de tendencias (tiempo y espacio) y nuestro modelado de tópicos. Al final tendremos un dashboard que nos muestre todas las gráficas obtenidas y resultados encontrados a partir de nuestro análisis.

4.2.1 Requisitos funcionales.

- El proceso de extracción se ejecutará de manera automática cada noche a las 10:30 para extraer todos los tweets de cada 7 días.
- El proceso de extracción de tweets abarcará los siguientes pasos: Lectura de las querys, consulta de las querys, extracción de datos del tweet, almacenamiento de los datos en archivos CSV.
- Los tweets pueden o no tener una referencia geográfica (longitud y latitud).
- El proceso de limpieza abarcará los siguientes pasos: Lectura de los archivos CSV, unión de todos los tweets en una sola lista, estructurar la información de acuerdo con el modelo de base de datos, realizar cambio de formatos en los datos y realizar la tokenización y lematización de los tweets para su futuro análisis de modelado de temas.
- Al finalizar el proceso de limpieza, inmediatamente empezará el proceso de carga a la base de datos.

- El proceso de análisis de los datos sociales abarcará los siguientes pasos: Conexión entre la base de datos con software de análisis de datos , filtrados de temas para obtener gráficas de tendencias.
- Creación de un software que realice el modelado de tópicos.
- Generar una interfaz para poder consultar la información.
- La interfaz debe tener una conexión directa con el proceso de análisis de datos.
- En la interfaz el usuario podrá seleccionar los filtros para hacer una consulta.
- En la interfaz el usuario podrá visualizar las gráficas de la información que solicito.

4.2.2 Requisitos no funcionales.

- El sistema debe soportar los sistemas operativos Linux o Windows.
- Se debe implementar el proceso de ETL en un ambiente virtual.
- Se usará Python 3.9 en adelante.
- Se usará Pip y virtualvenv para Python 3.
- Se usará un entorno virtual creado para Python 3 en el directorio de trabajo.
- Para realizar la extracción se usarán los módulos Tweepy y Snscrape.
- Para poder manejar los datos se usarán pandas y numpy.
- Para el proceso de tokenización se usará nltk.
- Para el proceso de lematización se utilizara spacy.
- Los datos extraídos y limpios se almacenarán en una base de datos MySQL.
- Se usará el software Tableau para analizar los datos y generar las gráficas de tendencias.
- Se usará python con la biblioteca gensim para realizar el topic modeling.
- La interfaz de consulta de resultados del análisis estará construida a partir de HTML y CSS.

4.3 Casos de Uso

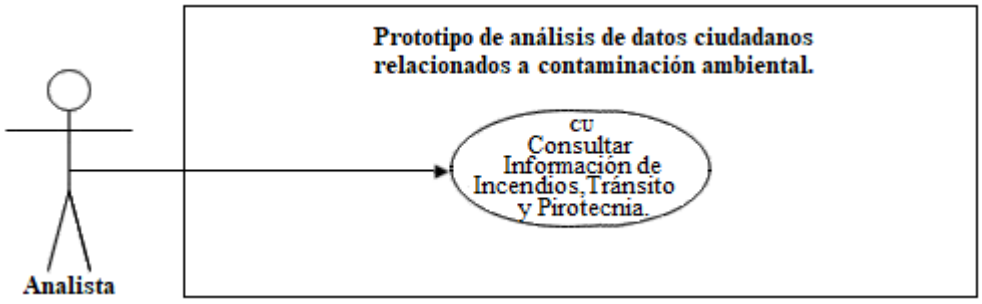

Figura 9. Diagrama de casos de uso.

Nombre: Consultar información de Incendios, Tránsito y Pirotecnia.
Identificador: CU
Descripción: Se mostrará la consulta de información de Incendios, Tránsito y Pirotecnia que haya solicitado el usuario de interés.
Precondiciones: El usuario debe solicitar una consulta de información de Incendios, Tránsito y Pirotecnia.
Postcondiciones: Se muestra la consulta que haya solicitado el usuario.

Trayectoria principal:

1) El usuario solicita hacer una consulta de la información de Incendios, Tránsito y Pirotecnia.
2) Si el usuario desea hacer una consulta de tiempo, da clic en conteo temporal.
3) Si el usuario desea hacer una consulta de espacio, da clic en conteo espacial.
4) Si el usuario desea hacer una consulta de nubes de palabras, da clic en Nube de palabras.
5) Si el usuario desea hacer una consulta de meses destacados, da clic en conteo de tweets por día en los meses más destacados.
6) Se muestra la consulta que el usuario haya solicitado.
7) El caso de uso termina.

4.3.1 Mockups del diagrama de caso de uso.

Primero tendremos una pantalla de inicio donde mostraremos información general del proyecto, mencionando los tres temas que estamos analizando, tránsito, incendio y pirotecnia.

Figura 10. Mockup de pantalla de inicio.

En la siguiente pantalla se podrán observar las gráficas correspondientes al conteo temporal, es decir se podrá visualizar la información por rango de tiempo, ya sea por mes o año.

Figura 11. Mockup de Conteo temporal.

CAPÍTULO 5: DISEÑO

5.1. Arquitectura del sistema.

5.1.1 Arquitectura Física.

A continuación, se presenta la arquitectura física de nuestro sistema:

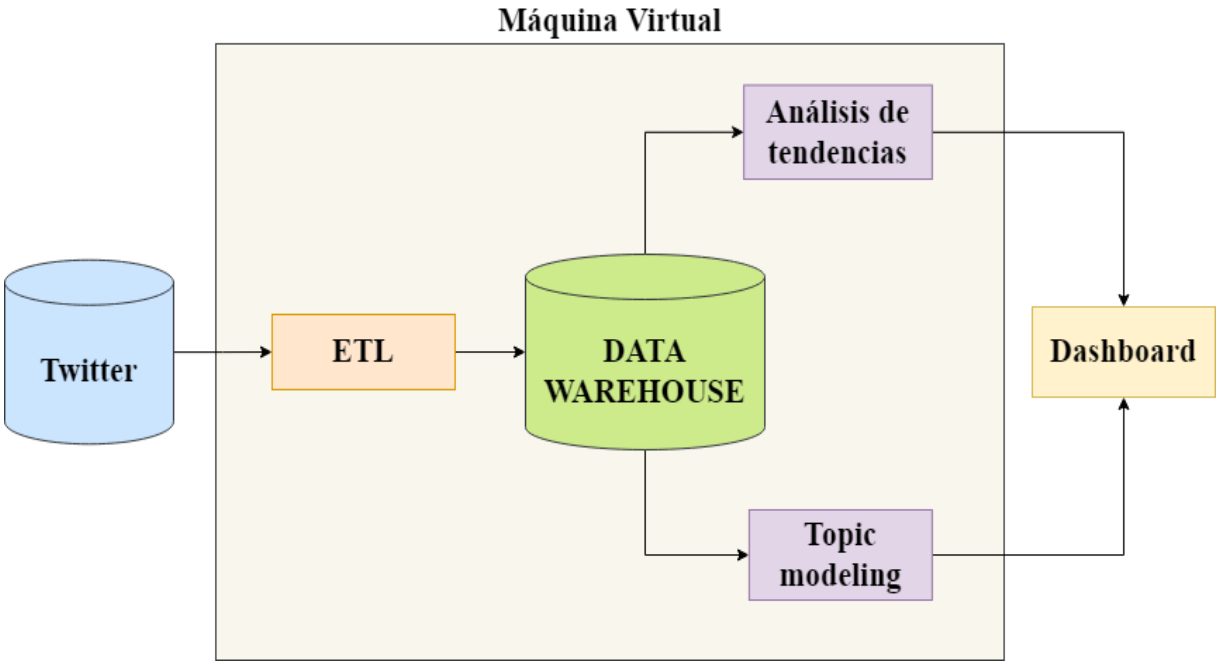

Figura 12. Arquitectura física del sistema.

5.1.2 Arquitectura Lógica.

A continuación, se presenta la arquitectura lógica de nuestro sistema:

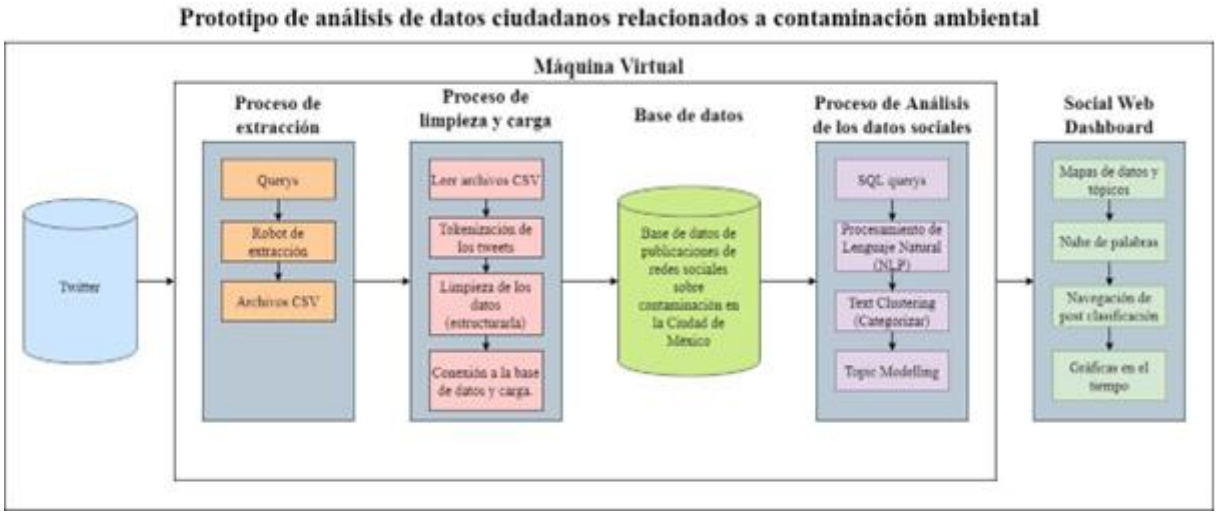

Figura 13. Arquitectura lógica del sistema.

5.2. Comportamiento de módulos o componentes.

5.2.1 Proceso de extracción de datos.

La base de datos de Twitter funciona como nuestra principal fuente de datos sobre contaminación en el aire, toda esa información no está organizada y el encargado de obtenerla es el proceso de ETL, como se mencionó anteriormente, está conformado de 3 fases que es la Extracción, Transformación y Carga, los cuales son pequeños subsistemas o procesos que definimos a continuación.

El subsistema o proceso de extracción tiene la responsabilidad de realizar la extracción de tweets de la base de datos histórica de Twitter utilizando como consulta las palabras clave de nuestro interés, como lo es tráfico, pirotecnia, incendios. Utiliza 2 módulos para realizar esta tarea, Tweepy y Snscrape, el primero nos ayuda a obtener los tweets de la última semana, ya que esta es una API de Twitter, no nos proporciona una exploración a profundidad en la base histórica; el segundo es un módulo que utiliza Web Scraping para realizar la búsqueda de tweets directamente desde la plataforma web de Twitter y así poder acceder a un histórico, con el fin de obtener mucha más información de los años anteriores. Cada tweet tiene información importante como fecha de publicación, geolocalización, metadatos de likes (me gusta), comentarios y retweets, al final toda la información es extraída también y todo es almacenado en archivos CSV, estos son utilizados por el programa de limpieza como entrada.

Después de nuestra primera extracción se analizó de manera manual los tweets para poder identificar los que tenían la estructura que buscamos para el análisis, esto implica que los tweets debe indicar el tiempo, ubicación y evento de la denuncia. Una vez que examinamos esos tweets realizamos una segunda extracción de datos junto con nuevas palabras claves que estaban relacionadas con los tweets que cumplen con las expectativas de la investigación y de esta manera filtrar aquellos tweets que no sean denuncias ciudadanas.

5.2.2 Proceso de limpieza y carga.

Normalmente el proceso de ETL se realiza de manera paralela, es decir cada vez que un dato es extraído, pasa directamente al proceso de limpieza y una vez finalizado este proceso se realiza el proceso de carga a la base de datos. Eso pasa siempre cuando ya se tiene desde un principio el dataset con toda la información, en este caso desde un principio no se tenía una noción de como Twitter iba a regresar los datos y la extracción se realizará hasta el último día posible en la que se realiza esta investigación, por lo que no se planeó paralelamente el desarrollo del proceso de limpieza y carga, es decir, el ETL se dividió en 2 partes, el proceso de extracción y el proceso de limpieza y carga.

El subsistema o proceso de *limpieza y carga* tiene la responsabilidad de leer todos los archivos CSV, en esencia, se encarga de transformar los datos anteriormente extraídos en datos limpios y sin inconsistencias que no puedan afectar al análisis de los datos. Algunas de las inconsistencias que serán tratadas son:

- Transformar valores booleanos a 0 y 1 para poder cargarlos en el servidor de base de datos (MySQL).
- Eliminar registros que contengan valores nulos en columnas importantes para el análisis.
- Eliminar espacios en blanco en las cadenas de texto.
- Cambiar el formato de las fechas al formato DD-MM-AAAA HH:mm:SS.
- Ajustar la hora de los tweets dentro de los horarios de verano dentro de cada año.
- Asignación de un tema a cada tweet a partir de las querys de extracción.
- Asignación de la alcaldía y coordenadas geográficas (Latitud y Longitud) a los tweets que no los incluían en sus metadatos.

También para posteriormente realizar el análisis de tendencias y el modelado de tópicos, necesitamos convertir nuestros datos y para eso realizamos las siguientes tecnicas.

- Tekenización de los tweets.
- Eliminar stopwords.
- Lematización de los tweets.
- Agregar bigramas a los tokens
- Listar las hashtags y usuarios mencionados en los tweets.

Una vez realizado el proceso de limpieza, se realiza la conexión con el servidor de base de datos, en este caso utilizamos MySQL como administrador de base de datos, para poder cargar todos los datos de los tweets extraídos y limpios.

5.2.3 Proceso de Servidor de base de datos.

El subsistema Servidor de base de datos solo se encargará de almacenar los datos anteriormente procesados y limpios, para poder realizar peticiones para su futuro análisis. En este servidor también haremos cubos de datos que nos permitan acceder rápidamente a un grupo de información para un análisis específico.

5.2.4 Proceso análisis de los datos sociales.

El subsistema Proceso de Análisis de Datos tiene la responsabilidad de extraer los datos limpios de la base de datos para su análisis, este proceso se divide en dos tipos de análisis, el primero es un análisis exploratorio donde utilizaremos herramientas como Tableau para poder obtener estadísticas de tendencias como gráficas de barras, circulares, gráficas de calor y nubes de palabras, donde exploraremos las tendencias en tiempo y espacio de todo el conjunto de datos históricos almacenados en la base de datos.

El segundo es un análisis de modelación de tópicos utilizando python, donde se realizará una técnica de aprendizaje automático para analizar los tokens de los tweets y generar temas o tópicos basados en la frecuencia de los tokens en el set de tweets extraídos. El modelo más utilizado para esta técnica es el Latent Dirichlet Allocation (LDA).

5.2.5 Social web dashboard.

El subsistema Dashboard tiene la responsabilidad de convertir todos los datos estadísticos y gráficas obtenidas del proceso de análisis en un pequeño dashboard para su fácil visualización y comprensión.

5.3. Modelo de Datos

Utilizamos el esquema de estrella para nuestra base de datos porque tenemos una tabla de hechos con múltiples dimensiones como lo son autores, fechas de publicación, ubicación y este esquema nos permite:

- Tener un mejor y más sencillo entendimiento del modelo.
- Un menor número de tablas a comparación de modelos como el copo de nieve y de constelaciones.
- La complejidad de las consultas es baja.
- El procesamiento de los cubos de datos y el desempeño de las consultas son mucho más rápidos.

5.3.1 Diagrama de la Base de datos.

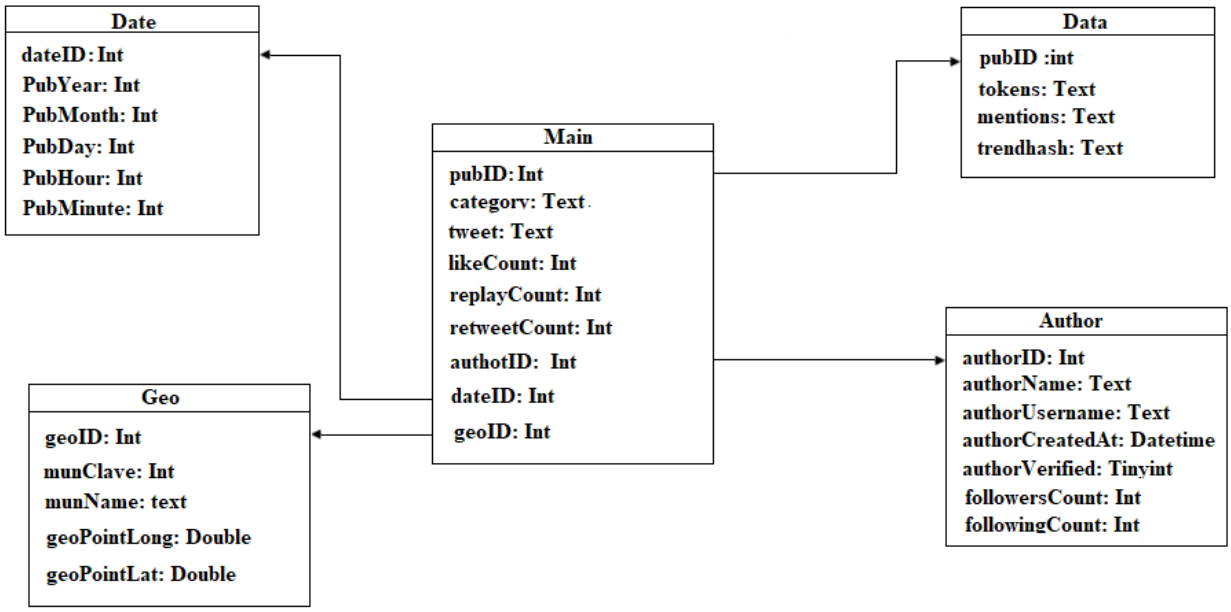

Figura 14. Diagrama de la base de datos.

5.3.2 Descripción de tablas

Tabla Main

Objetivo: Guardar información acerca del tweet que se ha publicado.

Main		
Nombre	**Tipo de Dato**	**Descripción**
pubID	Int	Id del tweet.
category:	Text	Categoría del Tweet.
tweet	Text	Tweet que se publicó.
likeCount	Int	Número de me gusta que obtuvo la publicación.
replayCount	Int	Número de respuestas de la publicación
retweetCount	Int	Número de retweets de la publicación.
authorID	Int	Id del autor del tweet.
dateID	Int	Id de la fecha.
geoID	Int	Id de la geolocalización.

Tabla 2. Atributos de la tabla Main.

Tabla Data

Objetivo: Guardar información relacionada a los tweets, tales como las menciones o conteos de hashtagas más usados.

Data		
Nombre	**Tipo de Dato**	**Descripción**
pubID	Int	Id del tweet.
tokens	Text	Palabras tokenizadas del tweet. (Palabras con el mayor valor semántico).
mentions	Text	Menciones que se hacen en el tweet.
trendhash	Text	Hashtags usados dentro del tweet publicado

Tabla 3. Atributos de la tabla Data.

Tabla Author

Objetivo: Guardar información acerca del autor que publicó el tweet.

Author		
Nombre	**Tipo de Dato**	**Descripción**
authorID	Int	Id del autor del tweet.
authorName	Text	Nombre del autor del tweet.
authorUsername	Text	Nombre de usuario del autor del tweet.
authorCreatedAt	Datetime	Hora y Fecha de creación del tweet.
authorVerified	Tinyint	En este campo colocamos si el usuario es verificado o no.
followersCount:	Int	Número de personas que siguen la cuenta del autor.

followingCount:	Int	Número de personas a las que sigue la cuenta del autor.

Tabla 4. Atributos de la tabla Author.

Tabla Geo

Objetivo: Guardar información acerca de la geolocalización del tweet.

Author		
Nombre	**Tipo de Dato**	**Descripción**
geoID	Int	Id de la geolocalización.
munclave	Int	Clave de la alcaldía.
numName	Text	Nombre de la alcaldía.
geoPointLong	Double	Longitud de la geolocalización del tweet.
geoPointLat	Double	Latitud de la geolocalización del tweet.

Tabla 5. Atributos de la tabla Geo.

Tabla Date

Objetivo: Guardar información de la fecha y hora de publicación del tweet.

Author		
Nombre	**Tipo de Dato**	**Descripción**
dateID	Int	Id del autor de la fecha
pubYear	Text	Año de publicación del tweet.
pubMonth	Text	Mes de publicación del tweet.
pubDay	Datetime	Día de publicación del tweet.
pubHour	Tinyint	Hora de publicación del tweet.
pubMinute	Int	Minuto de publicación del tweet.

Tabla 6. Atributos de la tabla Date.

5.4. Algoritmos usados.

ETL

Este proceso ya fue definido anteriormente, pero este es un algoritmo muy utilizado en la ciencia de datos para manejar grandes conjuntos de información (BIG DATA) para realizar análisis de datos con la finalidad de encontrar nueva información o responder a preguntas de negocio y tomar mejores decisiones en base a la información.

Tokenización

La tokenización es un proceso para dividir una cadena de texto en una lista de tokens, estos se pueden entender como el símbolo de una oración. Es un proceso fundamental en el procesamiento de lenguaje Natural (NLP) para poder entender de manera computacional qué es lo que quiere decir el mensaje.

Utilizamos este algoritmo para poder dividir los tweets en tokens (palabras) utilizando expresiones regulares para omitir los caracteres como signos de puntuación, espacios extra u otro tipo de expresiones que no nos sirvan, así mismo en el lenguaje hay muchos conectores que no nos importa mantener, por lo que también se hace un proceso de limpieza para omitir dichos conectores (conocidos como stopwords) ya que no tienen ningún valor semántico.

Lematización.

La lematización es un proceso para convertir las palabras conjugadas a su significado base, la intención de realizar este proceso es que no tengamos una misma palabra conjugada de diferentes formas y al realizar nuestro análisis se tome como palabras diferentes cuando su valor semántico es el mismo.

Este proceso se realiza usando un diccionario proporcionado por una biblioteca llamada spacy donde se encuentran las relaciones entre palabras conjugas con su base para poder convertirlas.

Latent Dirichlet Allocation (STM)

LDA o Latent Dirichlet Allocation es el modelo más famoso de Topic Modeling. Su premisa es modelar los documentos como distribuciones de temas, es decir, en un conjunto de temas qué porcentaje tiene cada uno en el conjunto de documentos (prevalencia) y los temas como distribuciones de palabras (contenido). Su diagrama es el que se muestra en la figura 15.

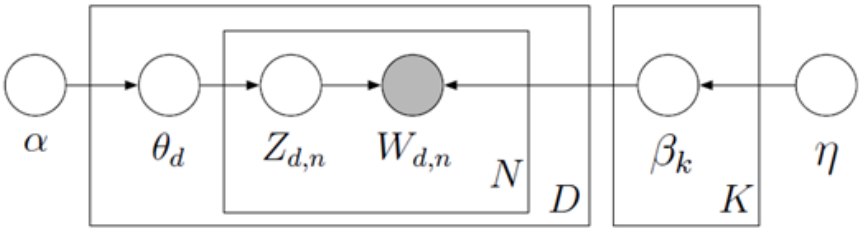

Figura 15. Diagrama de LDA.

El corpus es un conjunto de M documentos y hay N términos en el corpus. Hay k distribuciones de temas que contiene N términos y β es una distribución multinomial de cada uno de los temas. El proceso LDA generativo consiste en, primero, para cada documento, obtener una distribución de tema, θd∽Dir(α), donde Dir(α) se obtiene de la distribución de Dirichlet y α es un parámetro escalable. La distribución de Dirichlet es una generalización multivariada continua de la distribución β con el vector de parámetros α, que controla la forma promedio y la escasez de las proporciones del tema. Luego, para cada palabra del documento, obtenemos un tema z(d,n)∽multi(θd), dónde multi(θd) es un multinomial y obtiene una palabra w(d,n)∽β(zd,n). [36]

Este modelo nos permitirá obtener gráficas de top de temas y gráficas de diagnóstico para obtener el valor de número de temas.

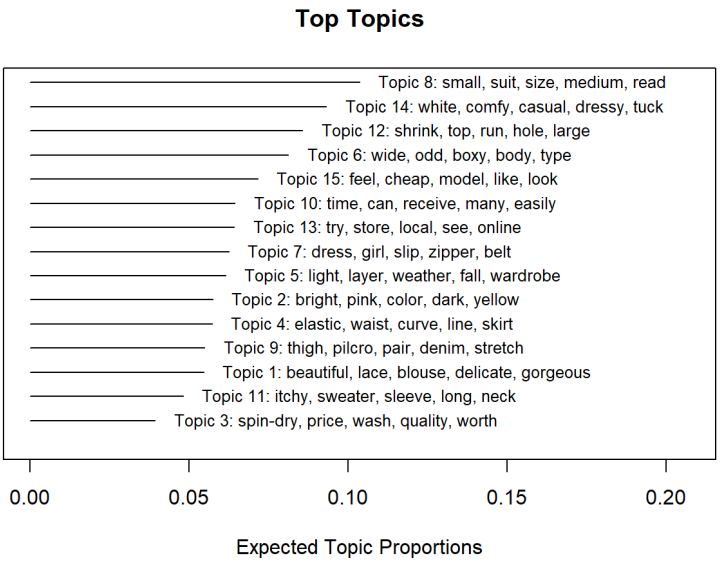

Figura 16. Diagrama de top de temas

5.5. Metodología.

La metodología que usamos es el modelo incremental, el cual nos permitió construir el proyecto en etapas, que reciben el nombre de “incremento”, en donde cada una de ellas agrega una funcionalidad hasta llegar al sistema final.

Este modelo nos permite reducir el riesgo de errores a través de la visibilidad de las nuevas versiones y obtener retroalimentación de las funcionalidades generadas en cada incremento. Nos evitará realizar un proyecto muy largo por la alta y cuidadosa planeación y nos permitirá generar valor al proyecto con cierta frecuencia.

El proyecto tiene la característica de poder ser dividido en diferentes subsistemas y cada incremento cuenta con etapas que nos permitirán definir los requisitos de cada uno de los subsistemas y evaluar el avance del proyecto. En seguida se muestra la definición de procesos que se realizan en cada una de las etapas de cada incremento.

Etapa	Proceso
Comunicación	Se realiza una reunión con el equipo para establecer los requerimientos.
Planeación	Se plantea la iteración correspondiente.
Modelado	Se realiza el diseño del subsistema.
Construcción	Se construye el subsistema propuesto.
Despliegue	Se presenta el subsistema y se evalúa por los participantes.

Tabla 7. Etapas de cada incremento del modelo incremental.

En la elaboración de nuestro proyecto tendremos un total de 7 incrementos, los cuales mostraremos a continuación:

- **Incremento 1:** Extracción de datos por método Tweepy.
- **Incremento 2:** Extracción de datos por método Snscrape.
- **Incremento 3:** Programa de Limpieza de datos.
- **Incremento 4:** Llenado de la base de datos.
- **Incremento 5:** Gráficas en Tableau.
- **Incremento 6:** Gráficas de Topic Modeling.
- **Incremento 7:** Dashboard.

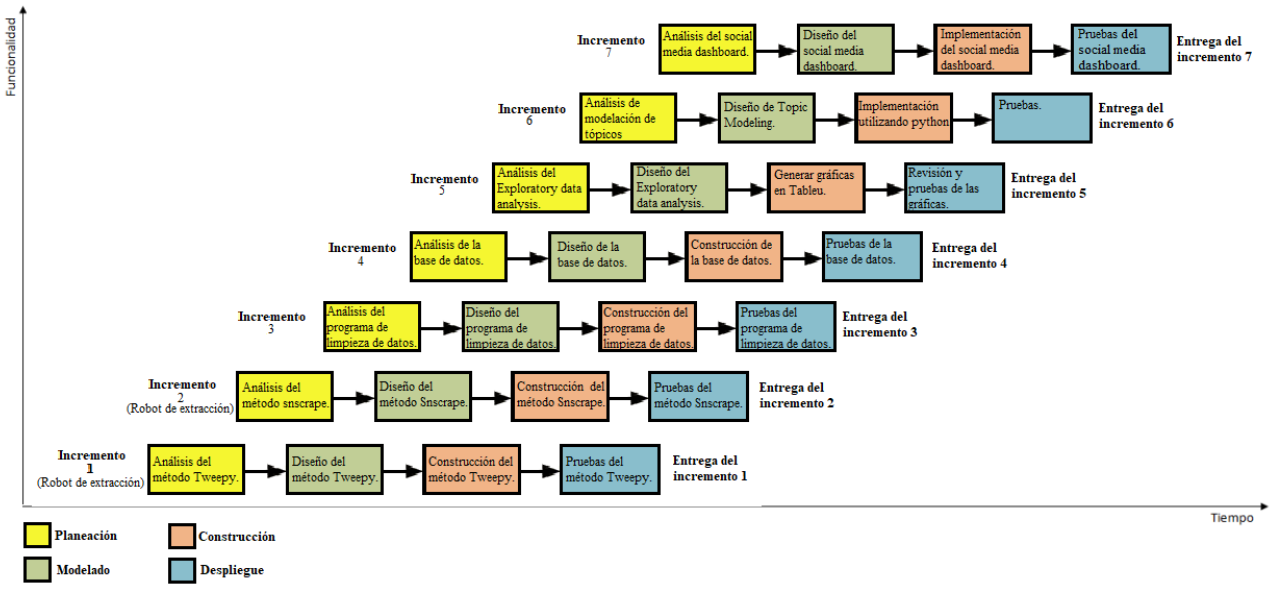

Figura 17. Modelo incremental.

5.5.1. Fase 1: Recolección de datos sociales

5.5.1.1. Definición de criterios o parámetros de búsqueda.

En un inicio, se necesita definir los temas de interés que queremos abarcar para el análisis. Las fuentes contaminantes son variadas, unas más presentes que otras, sin embargo, nuestro punto de interés está en aquellas que presentan mayor impacto respecto a la calidad del aire en la Ciudad de México.

Los temas principales que se decidieron tomar en cuenta para la realización de este proyecto son: pirotecnia, tráfico e incendios.

¿Por qué estos temas? A continuación, se dará una explicación individual de la razón de la selección de estos tres temas.

Primero hablemos de la pirotecnia. Para muchos de nosotros, la pirotecnia simboliza celebración y fiestas. Sin embargo, a pesar de que es modo de festejo, su uso trae graves consecuencias para la salud de los seres vivos y al medio ambiente.

Por ejemplo, la Secretaría del Medio Ambiente (SEDEMA) de la Ciudad de México, a través de la Dirección General de Calidad del Aire, alerta a las y los capitalinos sobre el daño que provoca la quema de pirotecnia, debido a que es la principal generadora de contingencias ambientales atmosféricas en época invernal [37].

Quiere decir, que en la Ciudad de México debido al exceso uso de pirotecnia en época invernal, concretamente en los días 24 y 25 de diciembre, al igual que en año nuevo, las concentraciones de partículas nocivas en la atmósfera se incrementan enormemente, ante dicha situación las autoridades tienen que advertir sobre la mala calidad del aire y pedirles a los ciudadanos quedarse en sus casas por su salud y bienestar.

Se debe recalcar que los fuegos artificiales no solamente se conforman de pólvora. Para conseguir los distintos efectos y colores, se requieren mezclas con múltiples compuestos químicos: bario para los tonos verdes, estroncio para los rojos, sodio para los dorados, aluminio para chispas plateadas y blancas y antimonio para destellos [38].

Otros de sus ingredientes que son señalados por algunos estudios por su peligrosidad, son el perclorato de potasio o de amonio como oxidantes, los cuales pueden causar desde irritación de la nariz, la garganta y los pulmones, ocasionando estornudo, tos y dolor de garganta y, en niveles más altos, puede causar dificultades respiratorias, colapso e incluso la muerte. Asimismo, la exposición repetida podría afectar al riñón y al sistema nervioso [38].

Además de todos los químicos que se utilizan para crear la pirotecnia, igualmente crea basura, el cual, si no se desecha de forma correcta estaría afectando al medio ambiente incluso después de usarse.

Por ello, en base a nuestra investigación, decidimos incluir el tema de la pirotecnia, aunque este sea el menos frecuente que los otros dos temas que abordaremos, en las temporadas que surge genera igual o mayor cantidad de elementos dañinos a la atmósfera, provocando que las autoridades tengan que implementar medidas como contingencias ambientales para intentar manejar las altas concentraciones a causa de la pirotecnia.

En otro tema, de alguna forma somos conscientes del problema de tránsito que sucede en la Ciudad de México. A pesar de las modificaciones y programas para reducir, asimismo controlar el uso vehicular de los ciudadanos, el país sigue presentando los mismos problemas sin una pizca significante de mejora a largo plazo. Esto se debe a que no se han considerado todos los factores que engloban la existencia de este problema.

Aún si programas como el Hoy No Circula tienen como objetivo restringir y controlar los automóviles que pueden circular en el día a día, no logra combatir del todo el problema del tránsito vehicular. La necesidad latente de controlar el número de autos que circulan a diario por la zona metropolitana, esta medida obedece más a un tema de movilidad, pues en términos de contaminación ambiental, el verdadero problema radica en los huecos en la legislación actual que permite circular a más de 600 mil autos en condiciones adversas [39].

Noventa por ciento del problema lo genera la incorporación de los 600 mil vehículos y 10 por ciento, las complicaciones a la movilidad derivadas de la gran cantidad de autos e incluso del nuevo reglamento de tránsito. Un vehículo cuyo convertidor catalítico no funciona contamina 35 veces más que uno que sí funciona. Ahí está el verdadero problema, 20 por ciento de los vehículos en la Ciudad de México genera 80 por ciento de la contaminación [39].

Sumado a ello, con la intención de mitigar el uso del vehículo, el Reglamento de Tránsito para la Ciudad de México fue modificado para fomentar el uso de la bicicleta como medio de transporte; no obstante, esta medida no reflexionó en torno a la emisión de gases contaminantes, ya que se redujo la velocidad drásticamente en diversas arterias principales y se suprimieron las vueltas a la derecha, medidas que contribuyen al congestionamiento vial e impactan directamente en la emisión de gases contaminantes [39].

Como se mencionó anteriormente, el tráfico es un problema que seguirá presente y se irá empeorando al paso de los años si no se llevan a cabo mejores maneras de contrarrestar los factores detrás de la problemática. Actualmente, la Ciudad de México se encuentra en el segundo lugar como la ciudad con el peor tráfico a nivel mundial. De acuerdo con el Índice TomTom, la Ciudad de México alcanza un nivel de congestionamiento de 93 por ciento durante las mañanas y hasta de 89 por ciento durante las tardes, superada solo por Estambul (Turquía), ciudad que experimenta un nivel de congestionamiento de 76 por ciento en las mañanas y de 109 por ciento en las tardes [39].

Con todo lo anterior, se decidió agregar el tema del tráfico vehicular como segundo tema a abarcar en nuestro proyecto. Es uno de los focos más presentes y principales fuentes que se atribuyen a la contaminación atmosférica en la Ciudad de México.

Abarcando el último tema, los incendios, ya sea forestales o que suceden en las zonas urbanas, ocasionan un grave daño al ambiente por igual. En el caso de los incendios forestales. En el país la temporada más crítica de incendios forestales ocurre justamente en los meses de marzo, abril y mayo durante el estiaje, es decir, cuando los ríos, lagunas y acuíferos alcanzan su nivel mínimo [28][9]. Si no fuera poco, las temporadas influyen, por ejemplo, en temporadas de calor la humedad del medio decrece, provocando el incremento de la vegetación seca siendo más propensa a iniciar un incendio o extender las llamas de una ya existente.

En tan sólo en los primeros cuatros meses e inicio del quinto del año 2019, se habían registrado 445 incendios de las cuales 2501 hectáreas corresponden a pastizales en la Ciudad de México. Específicamente en el caso de los incendios ocurridos el fin de semana del 10 al 13 de mayo de ese año en la CDMX, las altas emisiones de partículas (PM2.5) y su ubicación dentro de la zona metropolitana, fueron las que desencadenaron la contingencia ambiental que se está viviendo actualmente [9].

En México, los incendios, a diferencia de los sismos y huracanes, se presentan con más frecuencia de lo que sabemos y sus impactos son mayores, solo que se presentan de manera aislada y no tienen un impacto mediático. Anualmente se registran más de 95 mil incendios urbanos y no urbanos, son 260 al día, de acuerdo con el Instituto Nacional de Estadística y Geografía [40].

Al igual que los temas anteriores hablados, los incendios son una gran fuente de humo y de partículas nocivas que perjudican la salud de los seres vivos, incluyendo animales como personas, asimismo dichas partículas pueden aterrizar en fuentes de agua descubiertas y contaminarlas.

Sin duda, los incendios son un factor que no debe pasar desapercibido, por ello, también se tomó la decisión de incluirlo, con el fin de determinar el impacto que presenta entre la ciudadanía a través de la plataforma de Twitter referente a la calidad del aire.

Estableciendo nuestros temas de interés, el siguiente paso es determinar los parámetros que debe de contener cada uno.

De antemano, lo primero que se llevó a cabo fue una búsqueda manual de publicaciones de Twitter, con el fin de determinar los criterios que se requieren para los diferentes temas de interés para posteriormente integrarlas al robot de extracción y este por su cuenta recolecta los tweets que cumplen dichos parámetros.

Para poder obtener nuestros parámetros fue necesario buscar tweets ideales que nos aporten información necesaria para poder visualizar esas palabras clave que caracterizan los 3 temas que estaremos analizando, cabe mencionar que los tweets ideales deben contener información como el tema, hashtag y menciones relacionados al tema y mención de la alcaldía desde donde se está denunciando.

A continuación, se presentarán los resultados obtenidos de la búsqueda manual.

Pirotecnia.

- Sinónimos detectados: fuegos artificiales, cohete y petardo.
- Hashtags comunes detectados: #pirotecnia, #CDMX, #contaminar.
- Menciones comunes: @SEDEMA_CDMX, @SSC_CDMX, @UGS_GCDMX, @C5_CDMX
- Palabras clave: Detonar, ilegal, uso, excesivo, aventar, …

Ejemplo de las muestras obtenidas al realizar la búsqueda manual:

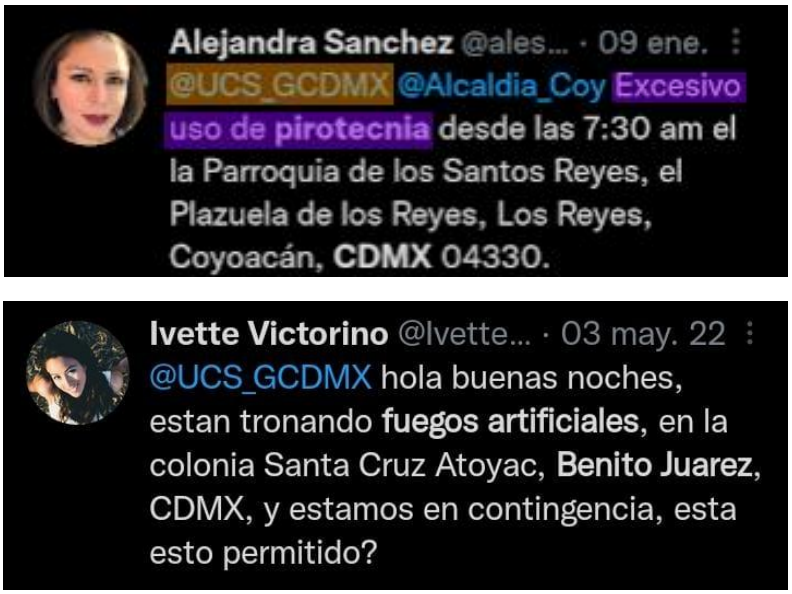

Figura 18. Ejemplo de denuncia ciudadana sobre pirotecnia en Twitter.

Tráfico.

- Sinónimos: tránsito, circulación y circulación rodada.
- Hashtags comunes detectados: #trafico, #tráfico o #Trafico; #CDMX, #caosvial, entre otras.
- Menciones comunes: @SSC_CDMX, @UGS_GCDMX, @C5_CDMX.
- Palabras clave: tráfico, cdmx, avenida o calle, caos, manifestación, accidente, sobre ruedas, etcétera.

Ejemplo de las muestras obtenidas al realizar la búsqueda manual:

Figura 19. Ejemplo de denuncia ciudadana sobre tránsito en Twitter.

Incendios.

- Sinónimos: quema, humo y fuego.
- Hashtags comunes detectados: #incendio, #CDMX, #humo, #contaminación.
- Menciones comunes: @SSC_CDMX, @Bomberos_CDMX
- Palabras clave: Humo, sofocar, incendio, forestal, quema.

Ejemplos de las muestras obtenidas al realizar la búsqueda manual:

Figura 20. Ejemplos de denuncia ciudadana sobre incendios en Twitter.

Una vez que se determinaron las palabras clave para cada tema, se procedió a construir las querys en base a los resultados de nuestra investigación previa, los cuales se mostrarán más adelante. Para construir una query tiene un gran catálogo de operadores que nos permiten delimitar las características requeridas para disminuir la cantidad de publicaciones no deseadas y recolectar la mayor cantidad de tweets para nuestra investigación.

Para poder obtener un mayor número de tweets realizamos búsquedas con diferentes querys, esto nos ayudó a poder delimitar la extracción de los tweets para que nos arrojará la mayor cantidad de tweets ideales relacionados a los temas de tránsito, pirotecnia e incendios

En primer lugar se generó una búsqueda general donde incluimos palabras clave relacionadas al tema de tránsito, pirotecnia e incendios.

```
topics = [
        "(pirotecnia OR cohete OR 'fuegos artificiales')",
        "(tránsito OR tráfico)",
        "(incendio OR humo OR fuego)"
    ]
```

Figura 21. Querys generales.

Una vez que tenemos una búsqueda general procedemos a delimitar aún más para poder obtener solo los tweets relacionados a los temas de denuncias o reportes sobre actos que generan contaminación ambiental.

```
    meta_search = '-is:retweet lang:es'
    delimitadores = '(reportar OR denuncia OR denunciar
OR contaminación OR ilegal OR ruido OR estrés OR contin
gencia OR precaución OR prohibido OR detengan OR alto)'
```

Figura 22. Querys por cada tema

En esta búsqueda tenemos la expresión '-is: retweet lang:es' donde el símbolo "- ", nos permite determinar qué palabras o características no queremos. En este caso, la query no va a recolectar ninguna publicación que sea tipo retweet, limitándose solo a las publicaciones originales.

Por otra parte, "lang:es" determina el idioma en el que fue escrito el tweet. En este caso, la delimitación es que la query solo se concentrará en dichas publicaciones que hayan sido escritas en el idioma español.

Posteriormente dentro del paréntesis podemos observar las palabras generales para delimitar los tweets y así solo obtener los que se refieran a una denuncia de contaminación ambiental.

En el contenido del primer paréntesis en la figura 23, en el caso de pirotecnia contiene: "pirotecnia OR cohete OR fuegos artificiales". Con estas palabras estamos indicando a nuestro robot de extracción que solamente recolecte tweets que en su contenido tengan las palabras especificadas, al igual que los hashtags #pirotecnia, #cohete, etcétera. Esto mismo sucede, tanto para la query del tema de tráfico como la query del tema incendios. Con hashtag y sin él.

Además de obtener nuestras búsquedas de manera general y de manera más específica, también nos interesa realizar una búsqueda por localización; esto con la finalidad de poder encontrar y extraer tweets que pertenezcan solamente a Ciudad de México. Para esto utilizamos el método geocode, el cual nos permite hacer una búsqueda de tweets por coordenadas y en un rango de zona, así que con esto es posible saber con más precisión si el tweet se está publicando desde cdmx y sobre todo saber desde que alcaldía se publicó.

La estructura de este método es la siguiente:

geocode:(x,y),[z]km

Un ejemplo es:

geocode:40.4167750,-3.7037900,25km

En nuestro caso para cada una de las temáticas colocamos el método geocode con las coordenadas céntricas de cada alcaldía y colocando un radio de búsqueda.

Las coordenadas son extraídas de un dataset que se encuentra en la página de datos abiertos de la CDMX.

```
geo_municipios = [
    (2, '19.4853286147,-99.1821069423', '5km', 'Azcapotzalco', 'azcapotzalco'),
    (3, '19.3266672536,-99.1503763525', '6km', 'Coyoacán', 'coyoacan'),
    (4, '19.3246343001,-99.3107285253', '10km', 'Cuajimalpa de Morelos', '("cuajimalpa de morelos" OR cuajimalpa)'),
    (5, '19.5040652077,-99.1158642087', '9km', 'Gustavo A. Madero', '("gustavo a. madero" OR gam)'),
    (6, '19.3969118970,-99.0943297970', '5km', 'Iztacalco', 'iztacalco'),
    (7, '19.3491663204,-99.0567989642', '9km', 'Iztapalapa' , 'iztapalapa'),
    (8, '19.2689765031,-99.2684129061', '10km', 'La Magdalena Contreras', '"magdalena contreras"'),
    (9, '19.1394565999,-99.0510954218', '11km', 'Milpa Alta', '"milpa alta"'),
    (10, '19.3361755620,-99.2468197120', '9km', 'Álvaro Obregón' , '"alvaro obregon"'),
    (11, '19.2769983772,-99.0028216137', '7km', 'Tláhuac', 'tlahuac'),
    (12, '19.1983396763,-99.2062207957', '11km', 'Tlalpan', 'tlalpan'),
    (13, '19.2451450458,-99.0903636045', '7km', 'Xochimilco', 'xochimilco'),
    (14, '19.3806424162,-99.1611346584', '3km', 'Benito Juárez', '"benito juárez"'),
    (15, '19.4313734294,-99.1490557562', '4km', 'Cuauhtémoc', 'cuauhtemoc'),
    (16, '19.4280623649,-99.2045669144', '5km', 'Miguel Hidalgo', '"miguel hidalgo"'),
    (17, '19.4304954545,-99.0931057959', '5km', 'Venustiano Carranza', '"venustiano carranza"')
]
```

Figura 23. Coordenadas geográficas de cada alcaldía de la Ciudad de México.

Utilizando estas coordenadas obtuvimos las siguientes querys:

Pirotecnia.

querys > querys3.txt

```
(pirotecnia OR cohete OR "fuego artificial" OR petardo) geocode:19.4853286147,-99.1821069423,5km -is:retweet lang:es
(pirotecnia OR cohete OR "fuego artificial" OR petardo) geocode:19.3266672536,-99.1503763525,6km -is:retweet lang:es
(pirotecnia OR cohete OR "fuego artificial" OR petardo) geocode:19.3246343001,-99.3107285253,10km -is:retweet lang:es
(pirotecnia OR cohete OR "fuego artificial" OR petardo) geocode:19.5040652077,-99.1158642087,9km -is:retweet lang:es
(pirotecnia OR cohete OR "fuego artificial" OR petardo) geocode:19.3969118970,-99.0943297970,5km -is:retweet lang:es
(pirotecnia OR cohete OR "fuego artificial" OR petardo) geocode:19.3491663204,-99.0567989642,9km -is:retweet lang:es
(pirotecnia OR cohete OR "fuego artificial" OR petardo) geocode:19.2689765031,-99.2684129061,10km -is:retweet lang:es
(pirotecnia OR cohete OR "fuego artificial" OR petardo) geocode:19.1394565999,-99.0510954218,11km -is:retweet lang:es
(pirotecnia OR cohete OR "fuego artificial" OR petardo) geocode:19.3361755620,-99.2468197120,9km -is:retweet lang:es
(pirotecnia OR cohete OR "fuego artificial" OR petardo) geocode:19.2769983772,-99.0028216137,7km -is:retweet lang:es
(pirotecnia OR cohete OR "fuego artificial" OR petardo) geocode:19.1983396763,-99.2062207957,11km -is:retweet lang:es
(pirotecnia OR cohete OR "fuego artificial" OR petardo) geocode:19.2451450458,-99.0903636045,7km -is:retweet lang:es
(pirotecnia OR cohete OR "fuego artificial" OR petardo) geocode:19.3806424162,-99.1611346584,3km -is:retweet lang:es
(pirotecnia OR cohete OR "fuego artificial" OR petardo) geocode:19.4313734294,-99.1490557562,4km -is:retweet lang:es
(pirotecnia OR cohete OR "fuego artificial" OR petardo) geocode:19.4280623649,-99.2045669144,5km -is:retweet lang:es
(pirotecnia OR cohete OR "fuego artificial" OR petardo) geocode:19.4304954545,-99.0931057959,5km -is:retweet lang:es
```

Figura 24. Querys de Twitter con coordenadas de la temática Pirotecnia.

Tráfico.

querys > querys3.txt

```
(transito OR tráfico OR circulación) geocode:19.4853286147,-99.1821069423,5km -is:retweet lang:es
(transito OR tráfico OR circulación) geocode:19.3266672536,-99.1503763525,6km -is:retweet lang:es
(transito OR tráfico OR circulación) geocode:19.3246343001,-99.3107285253,10km -is:retweet lang:es
(transito OR tráfico OR circulación) geocode:19.5040652077,-99.1158642087,9km -is:retweet lang:es
(transito OR tráfico OR circulación) geocode:19.3969118970,-99.0943297970,5km -is:retweet lang:es
(transito OR tráfico OR circulación) geocode:19.3491663204,-99.0567989642,9km -is:retweet lang:es
(transito OR tráfico OR circulación) geocode:19.2689765031,-99.2684129061,10km -is:retweet lang:es
(transito OR tráfico OR circulación) geocode:19.1394565999,-99.0510954218,11km -is:retweet lang:es
(transito OR tráfico OR circulación) geocode:19.3361755620,-99.2468197120,9km -is:retweet lang:es
(transito OR tráfico OR circulación) geocode:19.2769983772,-99.0028216137,7km -is:retweet lang:es
(transito OR tráfico OR circulación) geocode:19.1983396763,-99.2062207957,11km -is:retweet lang:es
(transito OR tráfico OR circulación) geocode:19.2451450458,-99.0903636045,7km -is:retweet lang:es
(transito OR tráfico OR circulación) geocode:19.3806424162,-99.1611346584,3km -is:retweet lang:es
(transito OR tráfico OR circulación) geocode:19.4313734294,-99.1490557562,4km -is:retweet lang:es
(transito OR tráfico OR circulación) geocode:19.4280623649,-99.2045669144,5km -is:retweet lang:es
(transito OR tráfico OR circulación) geocode:19.4304954545,-99.0931057959,5km -is:retweet lang:es
```

Figura 25. Querys de Twitter con coordenadas de la temática Tráfico.

Incendios.

querys > querys3.txt

```
(incendio OR quema OR humo OR fuego) geocode:19.4853286147,-99.1821069423,5km -is:retweet lang:es
(incendio OR quema OR humo OR fuego) geocode:19.3266672536,-99.1503763525,6km -is:retweet lang:es
(incendio OR quema OR humo OR fuego) geocode:19.3246343001,-99.3107285253,10km -is:retweet lang:es
(incendio OR quema OR humo OR fuego) geocode:19.5040652077,-99.1158642087,9km -is:retweet lang:es
(incendio OR quema OR humo OR fuego) geocode:19.3969118970,-99.0943297970,5km -is:retweet lang:es
(incendio OR quema OR humo OR fuego) geocode:19.3491663204,-99.0567989642,9km -is:retweet lang:es
(incendio OR quema OR humo OR fuego) geocode:19.2689765031,-99.2684129061,10km -is:retweet lang:es
(incendio OR quema OR humo OR fuego) geocode:19.1394565999,-99.0510954218,11km -is:retweet lang:es
(incendio OR quema OR humo OR fuego) geocode:19.3361755620,-99.2468197120,9km -is:retweet lang:es
(incendio OR quema OR humo OR fuego) geocode:19.2769983772,-99.0028216137,7km -is:retweet lang:es
(incendio OR quema OR humo OR fuego) geocode:19.1983396763,-99.2062207957,11km -is:retweet lang:es
(incendio OR quema OR humo OR fuego) geocode:19.2451450458,-99.0903636045,7km -is:retweet lang:es
(incendio OR quema OR humo OR fuego) geocode:19.3806424162,-99.1611346584,3km -is:retweet lang:es
(incendio OR quema OR humo OR fuego) geocode:19.4313734294,-99.1490557562,4km -is:retweet lang:es
(incendio OR quema OR humo OR fuego) geocode:19.4280623649,-99.2045669144,5km -is:retweet lang:es
(incendio OR quema OR humo OR fuego) geocode:19.4304954545,-99.0931057959,5km -is:retweet lang:es
```

Figura 26. Querys de Twitter con coordenadas de la temática Incendios.

5.5.1.2 Extracción semi-automática de datos.

Una vez que tenemos las querys de búsqueda para encontrar los tweets de interés para nuestro análisis, creamos el robot de extracción de tweets, este será un proceso independiente del ETL ya que no tenemos certeza de cuál es el

formato de cada dato que nos regresa Twitter y no podemos trabajar en paralelo para crear el proceso de limpieza de datos.

Para realizar el robot de extracción utilizamos Python, el cual nos proporciona 2 herramientas muy útiles para la extracción de tweets, el primero es el módulo Tweepy, este módulo nos permite utilizar el API de twitter de manera más intuitiva y escoger los metadatos que más nos interesen. Pero para utilizarlo debemos tener credenciales de desarrollador que Twitter nos proporciona para utilizar dicha API.

El primer paso fue solicitar dichas credenciales, lo cual fue un proceso tardado ya que nos pidieron muchas explicaciones sobre el uso de la información que vamos a extraer, pero una vez explicado creamos una cuenta de desarrollador en el segundo nivel de importancia de desarrolladores de Twitter, este es el nivel elevado, el cual nos permite tener un proyecto, con 2 aplicaciones y cada una puede extraer hasta 2 millones de tweets por mes. El dashboard de la cuenta de desarrollador es el que se muestra en la siguiente figura.

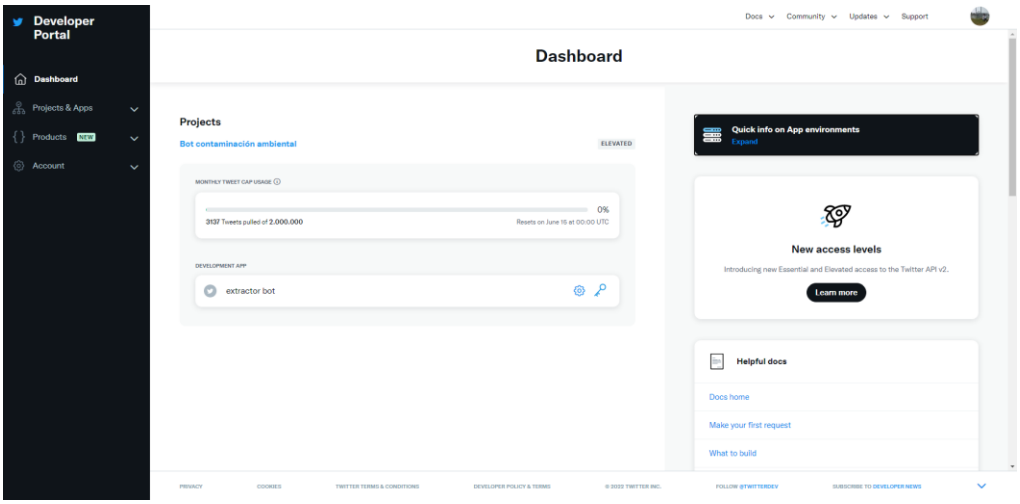

Figura 27. Dashboard del Twitter Developer Account.

Aquí podemos visualizar la cantidad de tweets que hemos descargado el mes y cuando es la fecha de corte, pero lo que realmente nos importa son las credenciales, las llaves de acceso. Estas solo se muestran una vez que se crea la aplicación, si queremos verlas de nuevo, se borran las credenciales anteriores y se crean unas credenciales nuevas, esto para mayor seguridad. Estas credenciales las guardamos en un archivo ".py" para poder importarlas al programa.

El programa que se encarga de la extracción de tweets se llama "01_extract.py" los módulos que utiliza para funcionar son los siguientes:

```
1  import csv
2  import datetime
3  import calendar
4  import os
5  import sys
6  import tweepy
7  import snscrape.modules.twitter as sntwitter
8  import credentials
```

Figura 28. Módulos utilizados para el programa de extracción.

La función que se encarga de la extracción de tweets es "search_tweets_tweepy".

```python
def search_tweets_tweepy(query, start_time, end_time, file_name):
    client = get_client()
    tweets = client.search_recent_tweets(query=query,
                                         max_results=100,
                                         tweet_fields=[
                                             'created_at', 'geo', 'public_metrics', 'author_id'],
                                         user_fields=[
                                             'created_at', 'public_metrics', 'verified'],
                                         place_fields=[
                                             'country', 'geo', 'name', 'place_type'],
                                         expansions=[
                                             'geo.place_id', 'author_id'],
                                         start_time=start_time +
                                         datetime.timedelta(hours=6),
                                         end_time=end_time +
                                         datetime.timedelta(hours=6)
                                         )
    ext_date = datetime.datetime.now().strftime('%Y-%m-%d %H:%M:%S')
    tweets_data = tweets.data
```

Figura 29. Función de búsqueda de tweets por el método Tweepy (parte 1).

Primero vamos por partes, lo que la función recibe como parámetros son:

- La query de búsqueda de tweets.
- Las fechas de inicio y fin de la búsqueda.
- El nombre del archivo donde se almacenarán los tweets.

Empezamos obteniendo un objeto de tipo “Client” que nos permitirá conectarnos a la API de Twitter, esta se encuentra en una función llamada “get_client” y es aquí donde introducimos las credenciales antes solicitadas en la cuenta de desarrollador de Twitter.

```python
def get_client():
    client = tweepy.Client(bearer_token=credentials.BEARER_TOKEN,
                           consumer_key=credentials.CONSUMER_KEY,
                           consumer_secret=credentials.CONSUMER_SECRET,
                           access_token=credentials.ACCESS_TOKEN,
                           access_token_secret=credentials.ACCESS_TOKEN_SECRET
                           )
    return client
```

Figura 30. Función para obtener objeto cliente.

Para obtener los tweets en formato Json utilizamos la función “search_recent_tweets” del objeto “client”. Esta función solicita como parámetros:

- La query de búsqueda.
- Número máximo de resultados (El máximo es 100).
- Metadatos principales.
- Metadatos de usuarios.
- Metadatos de lugares.
- Expansión de los metadatos de lugares.
- Fechas de inicio y fin de la extracción.

Los tweets los extraemos directamente de la variable “tweets.data”.

```
if tweets_data:
        tweets_includes = tweets.includes
        if tweets_includes:
            user_objects = {}
            place_objects = {}
            for expansion in tweets_includes.keys():
                if expansion == 'users':
                    user_objects = {
                        user.id: {
                            'authorName': user.name,
                            'authorUsername': user.username,
                            'authorCreatedAt': user.created_at,
                            'authorVerified': user.verified,
                            'followersCount': user.public_metrics['followers_count'],
                            'followingCount': user.public_metrics['following_count']
                        } for user in tweets_includes['users']}
                if expansion == 'places':
                    place_objects = {
                        place.id: {
                            'geoCountry': place.country,
                            'geoFullname': place.full_name,
                            'geoName': place.name,
                            'geoType': place.place_type,
                            'geoBbox': place.geo['bbox']
                        } for place in tweets_includes['places']}
```

Figura 31. Función de búsqueda de tweets por el método Tweepy (parte 2).

Sí existen tweets, entonces extraemos los metadatos y si hay metadatos creamos diccionarios para almacenar los metadatos específicos de los usuarios y las ubicaciones, identificándolos por el user id y el place id.

```
results = []
for tweet in tweets_data:
    tweet_row = []
    tweet_row.append(tweet.id)
    tweet_row.append(query)
    tweet_row.append(tweet.text)
    tweet_row.append(tweet.public_metrics['like_count'])
    tweet_row.append(tweet.public_metrics['reply_count'])
    tweet_row.append(tweet.public_metrics['retweet_count'])
    tweet_row.append(tweet.author_id)
    tweet_row.append(user_objects[tweet.author_id]['authorName'])
    tweet_row.append(user_objects[tweet.author_id]['authorUsername'])
    tweet_row.append((user_objects[tweet.author_id]['authorCreatedAt']
                      - datetime.timedelta(hours=6)).strftime('%Y-%m-%d %H:%M:%S'))
    tweet_row.append(user_objects[tweet.author_id]['authorVerified'])
    tweet_row.append(user_objects[tweet.author_id]['followersCount'])
    tweet_row.append(user_objects[tweet.author_id]['followingCount'])
    tweet_date = tweet.created_at-datetime.timedelta(hours=6)
    tweet_row.append(tweet_date.strftime('%Y-%m-%d %H:%M:%S'))
    tweet_row.append(tweet_date.strftime('%Y'))
    tweet_row.append(tweet_date.strftime('%m'))
    tweet_row.append(tweet_date.strftime('%d'))
    tweet_row.append(tweet_date.strftime('%H'))
    tweet_row.append(tweet_date.strftime('%M'))
    tweet_row.append(ext_date)
    try:
        tweet_row.append(tweet.geo['place_id'])
        tweet_row.append(
            place_objects[tweet.geo['place_id']]['geoCountry'])
        tweet_row.append(
            place_objects[tweet.geo['place_id']]['geoFullname'])
        tweet_row.append(
            place_objects[tweet.geo['place_id']]['geoName'])
        tweet_row.append(
            place_objects[tweet.geo['place_id']]['geoType'])
        tweet_row.append(
            place_objects[tweet.geo['place_id']]['geoBbox'])
    except:
        for i in range(6):
            tweet_row.append(None)
    try:
        tweet_row.append(tweet.geo['coordinates'].coordinates)
    except:
        tweet_row.append(None)

    results.append(tweet_row)

fill_dataset(results, file_name)
```

Figura 32. Función de búsqueda de tweets por el método Tweepy (parte 3).

Creamos una lista de resultados, que contendrá listas con todos los datos y metadatos de cada tweet en el orden en la que los almacenaremos en archivos CSV. En este proceso también hacemos un poco de limpieza por lo que al final el proceso de limpieza se enfocará más en tratar al tweet para su análisis para procesamiento de lenguaje natural.

Ahora el siguiente módulo que utilizamos fue snscrape. Este módulo no utiliza credenciales ya que su interacción con twitter es por medio de web scraping. La función que se encarga de la extracción de tweets es "search_tweets_snscrape".

```
def search_tweets_snscrape(query, start, end, file_name):
    start_date_time = datetime.datetime.strptime(
        start, '%Y-%m-%d') + datetime.timedelta(hours=6)
    end_date_time = datetime.datetime.strptime(
        end, '%Y-%m-%d') + datetime.timedelta(hours=6)

    full_query = query + \
        f' since_time:{calendar.timegm(start_date_time.utctimetuple())} until_time:{calendar.timegm(end_date_time.utctimetuple())}'
```

Figura 33. Función de búsqueda de tweets por el método Snscrape (parte 1).

Este módulo no necesita tantos detalles sobre lo que buscamos, lo que necesita es solamente la query y que se defina dentro de la misma query las fechas de inicio y fin de la búsqueda.

```
results = []
tweets_found = 0
ext_date = datetime.datetime.now().strftime('%Y-%m-%d %H:%M:%S')
for tweets_found, tweet in enumerate(sntwitter.TwitterSearchScraper(full_query).get_items()):
    tweet_row = []
    tweet_row.append(tweet.id)
    tweet_row.append(query)
    tweet_row.append(tweet.content)
    tweet_row.append(tweet.likeCount)
    tweet_row.append(tweet.replyCount)
    tweet_row.append(tweet.retweetCount)
    tweet_row.append(tweet.user.id)
    tweet_row.append(tweet.user.displayname)
    tweet_row.append(tweet.user.username)
    tweet_row.append((tweet.user.created - datetime.timedelta(hours=6)).strftime('%Y-%m-%d %H:%M:%S'))
    tweet_row.append(tweet.user.verified)
    tweet_row.append(tweet.user.followersCount)
    tweet_row.append(tweet.user.friendsCount)
    tweet_date = tweet.date - datetime.timedelta(hours=6)
    tweet_row.append(tweet_date.strftime('%Y-%m-%d %H:%M:%S'))
    tweet_row.append(tweet_date.strftime('%Y'))
    tweet_row.append(tweet_date.strftime('%m'))
    tweet_row.append(tweet_date.strftime('%d'))
    tweet_row.append(tweet_date.strftime('%H'))
    tweet_row.append(tweet_date.strftime('%M'))
    tweet_row.append(ext_date)
    tweet_row.append(None)
    if tweet.place != None:
        tweet_row.append(tweet.place.country)
        tweet_row.append(tweet.place.fullName)
        tweet_row.append(tweet.place.name)
        tweet_row.append(tweet.place.type)
    else:
        for i in range(4):
            tweet_row.append(None)
    tweet_row.append(None)
    if tweet.coordinates != None:
        coordinate = []
        coordinate.append(tweet.coordinates.longitude)
        coordinate.append(tweet.coordinates.latitude)
        tweet_row.append(coordinate)
    else:
        tweet_row.append(tweet.coordinates)

    results.append(tweet_row)

fill_dataset(results, file_name)
```

Figura 34. Función de búsqueda de tweets por el método Snscrape (parte 2).

Igualmente creamos una lista de resultados y una lista de metadatos para poder almacenar todos los metadatos de todos los tweets que encuentre la función "sntwitter.TwitterSearchScraper".

Hacemos el mismo procedimiento de limpieza y almacenamiento, para el final dentro de ambas funciones (tweepy y snscrape), la lista de resultados se pasa a una función llamada “fill_dataset”, junto con el nombre del archivo para que se almacenen todos los datos en archivos CSV.

```python
def fill_dataset(data, file_name):
    file = f'./data/{file_name}'
    if not os.path.exists(file):
        with open(file, 'w', newline='', encoding='utf-8') as csvfile:
            csv_writer = csv.writer(csvfile, delimiter=',')
            csv_writer.writerow(['pubID',
                                 'querySearch',
                                 'tweet',
                                 'likeCount',
                                 'replyCount',
                                 'retweetCount',
                                 'authorID',
                                 'authorName',
                                 'authorUsername',
                                 'authorCreatedAt',
                                 'authorVerified',
                                 'followersCount',
                                 'followingCount',
                                 'pubDate',
                                 'pubYear',
                                 'pubMonth',
                                 'pubDay',
                                 'pubHour',
                                 'pubMinute',
                                 'extDate',
                                 'geoID',
                                 'geoCountry',
                                 'geoFullname',
                                 'geoName',
                                 'geoType',
                                 'geoBbox',
                                 'geoCoordinates'])

    with open(file, 'a', newline='', encoding='utf-8') as csvfile:
        csv_writer = csv.writer(csvfile, delimiter=',')
        csv_writer.writerows(data)
```

Figura 35. Función de guardado de tweets en archivos CSV.

En la práctica, utilizamos ambos métodos porque uno cumple las funciones que el otro no puede. Tweepy es directamente la API de Twitter, pero es muy limitada ya que por extracción solo nos deja obtener 100 tweets, por lo que se tuvo que desarrollar un ciclo que llamara a la función cada hora para que extrajera solamente un día. También el usuario de desarrollador no nos dejaba explorar la base histórica de Twitter, solamente podíamos revisar los tweets de la última semana y no más.

Por otro lado, Snscrape no tiene limitaciones, no hay límite de tweets por ejecución, se puede explorar toda la base de datos de Twitter y además es más fácil de usar porque no tienes que definir los metadatos que necesitas, cada tweet trae todo lo que pueda extraer de metadatos. Además, la querys de búsqueda por geolocalización solo funciona en este módulo, ya que Tweepy también tiene restringido ese tipo de búsqueda para nuestro usuario de desarrollador.

El resultado son diferentes dataset separados por meses desde 2020 a la fecha en formato CSV con todos los tweets relacionados con el tema de interés.

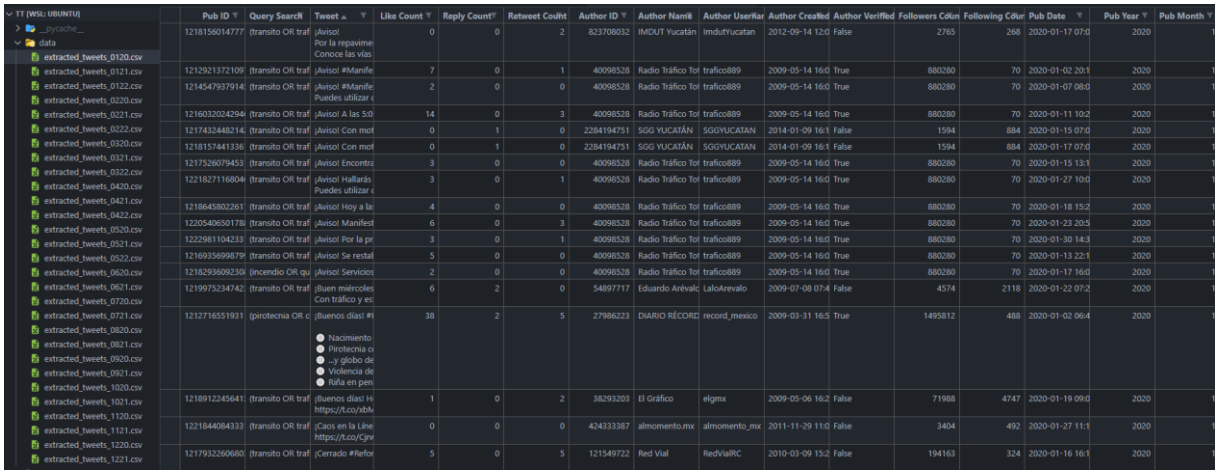

Figura 36. Captura de los archivos CSV extraídos a la fecha.

5.5.2 Fase 2. Limpieza de datos

Uno de los principales retos en la limpieza de datos es la limpieza del tweet para el procesamiento de lenguaje natural para el modelado de temas. Por ese motivo la primera función que creamos es la que se encarga de tokenizar cada uno de los tweets. Los módulos que utiliza para funcionar son los siguientes.

```
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
import spacy
from autocorrect import Speller
import requests
import json
import re
import datetime
import pandas as pd
import numpy as np
import credentials
```

Figura 37. Módulos utilizados para el programa de Limpieza.

La función tokens_tweets se encarga de convertir un texto en una lista de tokens (palabras) que lo conforman, pero omitiendo las palabras conectoras, que se conocen como stopwords. También corregimos las palabras mal escritas y procedemos a lematizarlas para regresarlas a su base.

```
def tokens_tweet(tweet):
    pattern = r'''(?x)          # set flag to allow verbose regexps
            (?:[A-Z]\.)+        # abbreviations, e.g. U.S.A.
            | \w+(?:-\w+)*      # words with optional internal hyphens
            | \$?\d+(?:\.\d+)?%? # currency and percentages, e.g. $12.40, 82%
            | [][.,;"'?():-_`]  # these are separate tokens; includes ], [
    '''
    # obtenemos la lista de stopwords
    stop_words = stopwords.words('spanish')
    # Convertir todo el texto en minúsculas
    tweet = tweet.lower()
    # Remover menciones, hashtags, links y saltos de linea
    tweet = re.sub("@[A-Za-z0-9_]+","", tweet)
    tweet = re.sub("#\S+","", tweet)
    tweet = re.sub(r"http\S+", "", tweet)
    tweet = re.sub(r"www.\S+", "", tweet)
    tweet = re.sub(r'\n', '', tweet)
    # Tokenización
    tokens_tweet = nltk.regexp_tokenize(tweet, pattern)
    tokens_tweet = [token for token in tokens_tweet if len(token) > 1]
    # Remover los stopwords
    interesting_tokens = [w for w in tokens_tweet if not w in stop_words]
    # corregir palabras
    spell = Speller(lang='es')
    interesting_tokens = [spell(w) for w in interesting_tokens]
    # Lematización
    nlp = spacy.load("es_dep_news_trf")
    doc = nlp(' '.join(interesting_tokens))
    interesting_tokens = [w.lemma_ for w in doc]

    return interesting_tokens
```

Figura 38. Función de tokenización y lematización.

Primero declaramos la expresión regular que nos permitirá filtrar algunos errores en la escritura y que pueda generar tokens vacíos o sin algún sentido semántico. Después descargamos la lista de stopwords para filtrar palabras como conectores, quitar menciones y links. Realizamos la tokenización, quitamos los stopwords y los signos de puntuación. Corregimos la ortografía de los tokens y hacemos la lematización para regresarlos a su estado base.

También definimos una función muy importante para la limpieza que utiliza la API de google maps para poder determinar el municipio correcto de un tweet que fue buscado por coordenadas.

```python
def mun_request(row):
    geo_mun = {
        'Azcapotzalco': 2,
        'Coyoacán': 3,
        'Cuajimalpa de Morelos': 4,
        'Gustavo A. Madero': 5,
        'Iztacalco': 6,
        'Iztapalapa': 7,
        'La Magdalena Contreras': 8,
        'Milpa Alta': 9,
        'Álvaro Obregón': 10,
        'Tláhuac': 11,
        'Tlalpan': 12,
        'Xochimilco': 13,
        'Benito Juárez': 14,
        'Cuauhtémoc': 15,
        'Miguel Hidalgo': 16,
        'Venustiano Carranza': 17
    }
    if row['typeQuery'] == 'coordenadas':
        url = f"https://maps.googleapis.com/maps/api/geocode/json?latlng={row['l
atitude']},{row['longitude']}&key={credentials.GOOGLE_MAPS_KEY}"
        res = requests.get(url)
        elements = res.json()
        mun = ''
        if  re.search("CDMX", elements['plus_code']['compound_code']):
            for i in elements['results']:
                if i['address_components'][0]['long_name'] in geo_mun.keys():
                    mun = i['address_components'][0]['long_name']
                    row['geoID'] = geo_mun[mun]
                    row['geoName'] = mun
                    break
        else:
            row['geoID'] = 0
    return row
```

Figura 39. Función de asignación de coordenadas.

Una vez explicado esto, la función principal (main) se encarga de realizar la limpieza de los datos y también la dicha transformación de los tweets en tokens. Aquí incluimos todos los métodos mencionados anteriormente en este documento como lo son la limpieza de tweets con campos nulos, espacios en blanco y formateo de algunos campos.

```python
for año in range(19,23):
        print(año)
        for mes in range (1,13):
            print(f'\t{mes}')
            # abrimos un archivo de cada mes para hacer la limpieza
            if mes < 10:
                df = pd.read_csv(f'./datos_filtrados/tweets_0{mes}{año}.csv')
            else:
                df = pd.read_csv(f'./datos_filtrados/tweets_{mes}{año}.csv')
```

Figura 40. Abrir archivos CSV.

Lo primero que hacemos es abrir cada uno de los archivos para poder explorarlos y realizar todos los procesos de limpieza, esto ayudado de la biblioteca pandas para el manejo de archivos csv.

```python
# eliminamos registros repetidos
df = df.drop_duplicates(['pubID'])
# obtener el municipio correcto de los registros extraídos por coordenadas
df = df.apply(mun_request, axis=1)
# eliminar registros que están fuera de la CDMX
df = df.drop(df[df['geoID'] == 0].index)
# cambiaremos los tipos de datos bool a 0 y 1 para su uso en MySQL
df['authorVerified'] = df['authorVerified'].apply(lambda x: 1 if x else 0)
# eliminamos registros con posibles datos nulos de los campos importantes.
cabeceras = list(df.columns)
cabeceras.remove('likeCount')
cabeceras.remove('replyCount')
cabeceras.remove('retweetCount')
cabeceras.remove('followersCount')
cabeceras.remove('followingCount')
for columna in cabeceras:
    df = df[df[columna].notna()]
```

Figura 41. Limpieza parte 1.

Lo primero que realizamos es eliminar registros repetidos, seguido de buscar las coordenadas correctas de cada tweet, en caso de que un tweet está ubicado fuera de la Ciudad de México se le asigna un 0 y es eliminado, convertimos los valores True y False a 1 y 0 para poder manejarlos dentro de MySQL y borramos aquellos registros que tengan valores nulos.

Figura 42. Limpieza parte 2.

La siguiente parte se encarga de corregir las fechas y horas de los tweets que se encuentran dentro del horario de verano de cada uno de los años que analizamos.

```
# ordenar por fecha
df['pubDate'] = pd.to_datetime(df['pubDate'])
df = df.sort_values(by='pubDate')
# creamos los tokens de cada tweet
if "tokens" in df.columns:
    df['tokens'] = df['tweet'].apply(tokens_tweet)
else:
    df.insert(4, "tokens", df['tweet'].apply(tokens_tweet))
# crear lista de hashtags
if "hashtags" in df.columns:
    df["hashtags"] = df['tweet'].apply(lambda x: re.findall("\B#([\w-]+)", x))
else:
    df.insert(5, "hashtags", df['tweet'].apply(lambda x: re.findall("\B#([\w-]+)", x)))
# crear lista de menciones
if "mentions" in df.columns:
    df["mentions"] = df['tweet'].apply(lambda x: re.findall("\B@([\w-]+)", x))
else:
    df.insert(6, "mentions", df['tweet'].apply(lambda x: re.findall("\B@([\w-]+)", x)))
# guardar dataset limpio
if mes < 10:
    df.to_csv(f'./data_clean/clean_tweets_0{mes}{año}.csv', index=False)
else:
    df.to_csv(f'./data_clean/clean_tweets_{mes}{año}.csv', index=False)
```

Figura 43. Limpieza parte 2.

Por último, ordenamos los tweets de manera ascendente por fechas y creamos las columnas de tokens, hashtags y menciones para los análisis de tendencias y el topic modeling.

5.5.2. Fase 3. Creación de las tablas para la base de datos.

Lo que hicimos para la carga de las bases de datos fue crear las diferentes tablas en archivos csv para poder cargarlas manualmente, las bibliotecas que utilizamos para este proceso son las siguientes:

```
import numpy as np
import pandas as pd
import emoji
```

Figura 44. Módulos utilizados para la carga de datos.

Lo que hicimos fue utilizar pandas para abrir cada unos de los archivos limpios y agregarlos todos a un solo dataframe, de esta manera ya teníamos en un solo archivos todos los datos y pudimos crear diferentes dataframes que actuaría como las tablas de nuestra base de datos y que tienen los campos de cada una de las tablas que definimos anteriormente.

```
df = pd.DataFrame()
for año in range(19,23):
    for mes in range (1,13):
        if mes < 10:
            df_aux = pd.read_csv(f'./csv_revisados/tweets_0{mes}{año}_c.csv')
        else:
            df_aux = pd.read_csv(f'./csv_revisados/tweets_{mes}{año}_c.csv')

        df = pd.concat([df,df_aux], sort=False, ignore_index=True)

main = df[['pubID', 'topicQuery', 'tweet', 'likeCount',
           'replyCount', 'retweetCount', 'authorID']]
data = df[['pubID', 'tokens', 'mentions', 'hashtags']]
author = df[['authorID', 'authorName', 'authorUsername',
           'authorCreatedAt', 'authorVerified', 'followersCount',
           'followingCount']]
date = df[['pubDate', 'pubYear', 'pubMonth', 'pubDay',
           'pubHour', 'pubMinute']]
coordinates = df[['geoID', 'geoName', 'longitude', 'latitude']]

idx = list(df.index)

main['tweet'] = main['tweet'].apply(lambda x: x.replace("\n", " "))
main['tweet'] = main['tweet'].apply(deEmojify)
main.insert(7, 'dateID', list(map(lambda x: 'd'+str(x+1), idx)))
main.insert(8, 'coordinateID', list(map(lambda x: 'c'+str(x+1), idx)))
author['authorName'] = author['authorName'].apply(deEmojify)
author['authorUsername'] = author['authorUsername'].apply(deEmojify)
date.insert(0, 'dateID', list(map(lambda x: 'd'+str(x+1), idx)))
coordinates.insert(0, 'coordinateID', list(map(lambda x: 'c'+str(x+1), idx)))

main.to_csv('./mysql/main.csv', index=False)
data.to_csv('./mysql/data.csv', index=False)
author.to_csv('./mysql/author.csv', index=False)
date.to_csv('./mysql/date.csv', index=False)
coordinates.to_csv('./mysql/coordinates.csv', index=False)
```

Figura 45. Código de la creación de tablas para la base de datos.

Al último tenemos el código para guardar los dataframes como archivos csv y posteriormente cargarlos a MySQL.

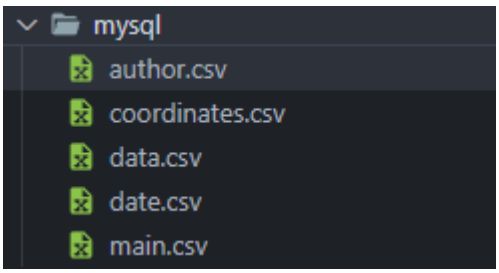

Figura 46. Tablas de MySQL de los datos extraídos y limpiados.

5.5.2. Fase 4. Análisis de datos - Topic Modeling.

Como este proceso se realiza por medio de Python entonces también podemos explicar el procedimiento y el código para obtener modelado de tópicos. Los módulos utilizados son los siguientes:

```
import ast
import numpy as np
import pandas as pd
from collections import defaultdict

from gensim import corpora
from gensim.models import Phrases
from gensim.models import CoherenceModel
from gensim.models.ldamodel import LdaModel

import pyLDAvis
import pyLDAvis.gensim_models as gensimvis

import matplotlib.pyplot as plt
from matplotlib.lines import Line2D
import seaborn as sns
import logging
```

Figura 47. Módulos utilizados para el topic modeling usando LDA.

Lo primero que realizamos es cargar la tabla data donde se encuentra los tokens de cada uno de los tweets para poder pasarlos a una lista a la cual le quitaremos las palabras más mencionadas y menos mencionadas, para esto quitaremos primeros las palabras que solo aparecen una vez en todo el conjunto de tokens y después procedemos a realizar los bigramas para el análisis.

```
df = pd.DataFrame()
for año in range(19,23):
    for mes in range (1,13):
        if mes < 10:
            df_aux = pd.read_csv(f'./csv_revisados/tweets_0{mes}{año}_c.csv')
        else:
            df_aux = pd.read_csv(f'./csv_revisados/tweets_{mes}{año}_c.csv')

        df = pd.concat([df,df_aux], sort=False, ignore_index=True)

texts = list(df['tokens'])
texts = [ast.literal_eval(tokens) for tokens in texts]

# remove words that appear only once
frequency = defaultdict(int)
for text in texts:
    for token in text:
        frequency[token] += 1
texts = [[token for token in text if frequency[token] > 1] for text in texts]
# Add bigrams to docs (only ones that appear 20 times or more).
bigram = Phrases(texts, min_count=20)
for idx in range(len(texts)):
    for token in bigram[texts[idx]]:
        if '_' in token:
            # Token is a bigram, add to document.
            texts[idx].append(token)
```

Figura 48. Topic modeling parte 1.

Procedemos a crear el diccionario de tokens y el corpus para el análisis, en el diccionario borramos los tokens menos mencionados y los más comunes, esto con el fin de no incluir todos los términos que usamos para las búsquedas.

```
# Create the dictionary
dictionary = corpora.Dictionary(texts)
# Filter out words that occur less than X documents,
# or more than X% of the documents.
dictionary.filter_extremes(no_below=65, no_above=0.5)
# Create the corpus.  This is a Term Frequency
# or Bag of Words representation.
corpus = [dictionary.doc2bow(text) for text in texts]
# print tokens and len of documents
print(f'Number of unique tokens: {len(dictionary)}')
print(f'Number of documents: {len(corpus)}')
```

Figura 49. Topic modeling parte 2.

El primer paso es entrenar el modelo hasta que todo los documentos converjan, para eso tenemos que jugar un poco con las variables del modelo, el chunksize es el número de documentos que puede analizar en cada iteración, este valor depende más que nada de la capacidad computacional para poder colocar un número grande, passes controla la frecuencia con la que entrenamos el modelo en todo el corpus y iterations es algo técnico, pero esencialmente controla la frecuencia con la que repetimos un ciclo particular sobre cada documento.

```
# Training the Model
NUM_TOPICS = 6
chunksize = 3500
passes = 10
iterations = 200
eval_every = None
temp = dictionary[0]
id2word = dictionary.id2token

model = LdaModel(
    corpus=corpus,
    id2word=id2word,
    chunksize=chunksize,
    alpha='auto',
    eta='auto',
    iterations=iterations,
    num_topics=NUM_TOPICS,
    passes=passes,
    eval_every=eval_every
)
```

Figura 50. Topic modeling parte 2.

Una vez que tengamos los valores correctos también podemos manipular el número de tópicos que hay dentro de un conjunto de documentos, para eso se debe correr el programa varias veces con múltiples valores para poder ver los resultados y ver si existe coherencia en los resultados de cada tema. Para poder visualizar los resultados usamos un graficador que nos muestra la frecuencia de cada palabra en cada uno de los temas y la relación que tiene cada tema entre sí.

```
# feed the LDA model into the pyLDAvis instance
lda_viz = gensimvis.prepare(model, corpus, dictionary, sort_topics=True)

pyLDAvis.save_html(lda_viz, './resultados_tm/lda.html')
```

Figura 51. Topic modeling parte 3.

5.7. Cronograma.

CRONOGRAMA Nombre del alumno(a): Hernández Clemente Samantha TT No.:2021-B051
Título del TT: Prototipo de análisis de datos ciudadanos relacionados a contaminación ambiental.

Actividad	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC	ENE
Análisis y diseño del robot de extracción.												
Revisión y pruebas del robot de extracción.												
Análisis y diseño de la base de datos.												
Creación de la base de datos en la máquina virtual.												
Análisis y diseño del programa de limpieza de datos.												
Avance de la construcción del programa de limpieza de datos.												
Evaluación de TT I.												
Análisis exploratorio utilizando Tableau.												
Análisis de modelación de tópicos.												
Análisis y diseño del dashboard.												
Implementación del social media dashboard.												
Pruebas de dasboard.												
Evaluación de TT II.												

CRONOGRAMA Nombre del alumno(a): Olivares Conchillos Leonel TT No.:2021-B051
Título del TT: Prototipo de análisis de datos ciudadanos relacionados a contaminación ambiental.

Actividad	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC	ENE
Análisis y diseño del robot de extracción.												
Revisión y pruebas del robot de extracción.												
Análisis y diseño de la base de datos.												
Creación de la base de datos en la nube.												
Análisis y Diseño del programa de limpieza de datos.												
Avance de la construcción del programa de limpieza de datos.												
Evaluación de TT I.												
Análisis exploratorio utilizando Tableau.												
Análisis de modelación de tópicos.												
Análisis y diseño del dashboard.												
Implementación del social media dashboard.												
Pruebas de dasboard.												
Evaluación de TT II.												

CRONOGRAMA Nombre del alumno(a): Medina Flores Susana TT No.:2021-B051
Título del TT: Prototipo de análisis de datos ciudadanos relacionados a contaminación ambiental.

Actividad	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC	ENE
Análisis y diseño del robot de extracción.												
Construcción del robot de extracción.												
Revisión y pruebas del robot de extracción.												
Análisis y diseño de la base de datos.												
Extracción de datos sociales (Twitter).												
Análisis y Diseño del programa de limpieza de datos.												
Evaluación de TT I.												
Análisis exploratorio utilizando Tableau.												
Análisis de modelación de tópicos.												
Análisis y diseño del dashboard.												
Implementación del social media dashboard.												
Pruebas de dasboard.												
Evaluación de TT II.												

CAPÍTULO 6: RESULTADOS PRELIMINARES.

6.1. Pruebas.

Para realizar las pruebas y revisiones de la programación utilizamos guiones de prueba, los cuales son una descripción de los pasos a seguir para realizar la prueba de una funcionalidad del programa. Finalmente, al hacer estas pruebas podemos verificar si el resultado es el esperado o si existe alguna falla, en caso de que estas existan, simplemente se buscará realizar las correcciones y se volverá a realizar la prueba.

6.1.1. Guión de prueba del robot de extracción

Métodos de prueba	Caso de Prueba	Condición para probar	Precondiciones	Guion de prueba	Prioridad	Resultados Esperados	¿Se cumplen los resultados?	Tiempo de Ejecución	Realizó	Revisó	ID incidencia.
Main()	1	La capacidad del programa de ejecutarse automáticamente en la máquina virtual.	-Tener el programa en la máquina virtual. -Contar con el archivo de query y archivo de resultados	1- Correr el programa. 2- Cumplir precondiciones. 3- Fin del guión de prueba.	Alta.	¿El programa se ejecutó de manera correcta?	SI	5 seg.	Susana	Samantha	
	2			1- Correr el programa. 2- No cumplir precondiciones. precondición. 3- Fin del guión de prueba.	Baja.	¿El programa muestra el mensaje "Introduce the query_file and file_name of the search in cosole's args."??	SI	5 seg.	Susana	Samantha	
	3	La capacidad del programa para abrir el archivo de query.	-Contar con el archivo query.	1- Correr el programa. 2- Cumplir precondición. 3- Fin del guión de prueba.	Alta.	¿El programa se ejecutó de manera correcta?	SI	5 seg.	Susana	Samantha	
	4			1- Correr el programa. 2- No cumplir precondición. 3- Fin del guión de prueba.	Baja.	El programa muestra el mensaje "The file with the querys doesn´t exist."?	SI	5 seg.	Susana	Samantha	
	5	La capacidad del programa para llamar al método Tweepy.	-Cumplir caso de prueba 1 y caso de prueba 2. -Elegir el método de extracción por Tweepy.	1- Correr el programa. 2- Cumplir precondiciones. 3-Elegir la opción t para realizar la extracción. 3- Fin del guión de prueba.	Alta.	¿El programa ejecuta el método Tweepy?	SI	8 seg.	Susana	Samantha	
	6	La capacidad del programa para mandar a llamar al método Snscrape.	-Cumplir caso de prueba 1 y caso de prueba 2. -Elegir el método de extracción por Snscrape.	1- Correr el programa. 2- Cumplir precondiciones. 3- Elegir la opción s para realizar la extracción. 5- Fin del guión de prueba.	Alta.	¿El programa ejecuta el método Snscrape?	SI	8 seg.	Susana	Samantha	

	7	La capacidad del programa para almacenar los tweets extraídos.	-Cumplir los casos de prueba anteriores. -Contar con el archivo file.	1- Correr el programa. 2- Cumplir precondiciones. 3- Extraer los tweets. 4- Almacenar los datos en la variable tweets_found 5-Fin del guión de prueba.	Alta.	¿El programa almacena la información de los tweets extraídos?	SI	20 seg	Susana	Samantha	
	8	La capacidad del programa para imprimir los tweets extraídos.	-Cumplir los casos de prueba anteriores. -Contar con el archivo file.	1- Correr el programa. 2- Cumplir precondiciones. . 3-Agregar la información de data en cada fila. 4-Mostrar la tabla con la información 5-Fin del guión de prueba.	Alta.	¿El programa muestra la información de los tweets extraídos?	SI	20 seg.	Susana	Samantha	
search_tweets_tweepy()	8	La capacidad del programa para convertir la fecha en UNIX TIME STAMP	-Contar con la función timedelta().	1- Correr el programa. 2- Cumplir precondiciones. 3- Convertir la fecha en UNIX TIMESTAMP. 4- Fin del guión de prueba.	Baja.	¿El programa convierte la fecha en UNIX TIME STAMP?	SI	10 seg.	Samantha	Leonel	1
	10	La capacidad del programa para hacer una extracción por día.	-Cumplir casos de prueba del método main(). -Contar con el archivo file.	1- Correr el programa. 2- Cumplir precondiciones. 3- Extraer los tweets del día 15 de marzo de 2022. 4- Guardar las extracciones durante 24 horas. 5- Fin del guión de prueba.	Alta.	¿El programa agrega los datos y metadatos del tweet por día de manera automática?	SI	4 min.	Samantha	Leonel	

search_tweets_snscrape()	11	La capacidad del programa para convertir la fecha en UNIX TIME STAMP	-Contar con la función timedelta().	1- Correr el programa. 2- Cumplir precondiciones. 3- Convertir la fecha en UNIX TIMESTAMP. 4- Fin del guión de prueba.	Baja.	¿El programa convierte la fecha en UNIX TIME STAMP?	SI	10 seg.	Leonel	Susana	2
	12	La capacidad del programa para hacer una extracción por un periodo de tiempo.	--Cumplir casos de prueba del método main(). -Contar con el archivo file.	1- Correr el programa. 2- Cumplir precondiciones. 3- Indicar el periodo de tiempo de extracción (30/11/2021 al 30/12/2021). 4- Extraer los tweets del periodo indicado. 4- Guardar las extracciones. 5- Fin del guión de prueba.	Alta.	¿El programa agrega los datos y metadatos del tweet por periodo de tiempo?	SI	min.	Leonel	Susana	

Análisis del guion de prueba.

	Casos satisfactorios.	**Casos fallidos.**	**Casos Especiales.**
Prioridad Baja.	2	0	2
Prioridad Alta.	8	0	0
TOTAL	10	0	2

Tabla 8. Análisis de casos satisfactorios, fallidos y especiales.

Podemos observar que tenemos un total de 12 casos de prueba, de los cuales 10 fueron satisfactorios y 2 especiales. Dentro de los casos satisfactorios 8 son de Prioridad Alta y 2 de Prioridad baja, mientras que nuestros 2 casos especiales son de prioridad baja.

Nota: Categorizamos a 2 de nuestros casos como especiales debido a que cumplen su funcionamiento, pero solo en un periodo de tiempo. Estos casos especiales, se refieren a los casos de prueba donde queremos probar la capacidad del programa para convertir la fecha y hora al formato UNIX TIME STAMP. En el programa de extracción de datos hicimos la conversión de UTC a UNIX restando la diferencia de 6 horas, y durante el periodo de extracción de febrero y marzo funcionó de manera correcta, pero a partir del día 2 de abril, cuando se realizó un cambio de horario el robot funcionaba, pero las horas que mostraba en la extracción eran incorrectas, y lo mismo pasaba en la extracción de años anteriores cuando el horario se encontraba en horario de verano.

Con la finalidad de enfocarnos en la limpieza de datos y la construcción de la base de datos, decidimos conservar nuestro robot como está hasta este momento debido a que los casos especiales que tenemos son de prioridad baja.

Estos casos de prueba son de prioridad baja porque en este momento estamos recolectando información de manera general, entonces esta situación no nos perjudica, y además podemos solucionarlo durante el proceso de limpieza de datos.

Periodo de Extracción	**Tiempo de Ejecución**	
	Método Tweepy	**Método Snscrape**
1 día (15/03/22)	4 minutos.	NO APLICA
1 semana (01/02/22 al 08/02/22)	NO APLICA	7 minutos
1 mes (30/11/2021 al 31/12/2021)	NO APLICA	18 minutos

1 año (01/09/20 al 01/09/21)	NO APLICA	34 minutos
2 años (04/01/20 al 04/01/22)	NO APLICA	80 minutos

Tabla 9. Análisis de tiempo de ejecución de la extracción de datos.

El tiempo de ejecución de los métodos para la extracción de datos es variable porque depende de la cantidad de datos que se encuentren dentro del periodo que se busca extraer, pero con fines de la realización de este guion de prueba utilizamos las fechas señaladas en la tabla 3.

6.2. Código Fuente.

Programa de extracción de tweets:

https://github.com/Leonelney/Trabajo-Terminal/blob/main/01_extract.py

Programa de limpieza:

https://github.com/Leonelney/Trabajo-Terminal/blob/main/02_cleaner.py

Programa de carga de datos:

https://github.com/Leonelney/Trabajo-Terminal/blob/main/03_load.py

Programa de topic modeling:

https://github.com/Leonelney/Trabajo-Terminal/blob/main/04_tm.py

6.3. Análisis de los datos sociales.

Una vez que se completó la fase de limpieza de datos, teniendo los datos limpios a nuestra disposición, se llevó a cabo el análisis de estos datos sociales. Este análisis se divide en dos tipos: El análisis exploratorio mediante la herramienta Tableau enfocándonos en los datos históricos y el segundo, el análisis de modelación con tópicos, en el cual se concentrará en los tokens de los tweets y los temas que fueron generados a través de la información obtenida.

6.3.1. Análisis exploratorio en Tableau.

En la siguiente figura, se muestran en conjunto cuatro gráficas creadas mediante los datos históricos que se obtuvieron de la última extracción, además de haber realizado el proceso de limpieza correspondiente.

En la primera gráfica (parte superior lado izquierdo), nos muestra la cantidad de tweets recolectados distribuidos por el tema que les corresponde: tránsito (color azul), incendio (color rojo) y pirotecnia (color naranja). Dicha gráfica resalta que la mayoría de tweets que obtuvimos son referentes al tema del tránsito y el tema de pirotecnia como el de menor de los tres.

Profundizando más, en la segunda gráfica (parte superior lado derecho), se puede apreciar las horas las cuales se presentaron las publicaciones de los respectivos temas siguiendo la misma simbología de colores para los temas de la gráfica anterior. En esta gráfica nos muestra que los de pirotecnia se concentran en altas horas de la noche (a partir de las 21:00 horas), en el caso de incendio e incendio no se muestra crecimiento casi lineal como el de pirotecnia, sin embargo, se puede apreciar la hora con más tweets. En el caso de incendio, entre las 12:00 a 15:00 horas, y entre las 18:00 a 20:00 horas presentan la mayor cantidad de tweets. Para el tema de tránsito, hay varias horas que sobresalen, por ejemplo las horas impares a partir de las 9:00 horas y antes de las 20:00 horas existen un mayor auge de tweets, esto se podría deber a que, por ejemplo a las 15:00 o 19:00 horas muchas personas terminan su horario laboral y se dirigen a sus casas, el cuál ocasiona que haya mucho vehículo en movimiento por las calles, saturando la circulación vehicular y provocando aglomeraciones que ocasionan tráfico en la zona.

En cuanto a la tercera gráfica (parte inferior lado izquierdo) son los tweets distribuidos esta vez por mes, clasificados con su respectivo tema al que corresponden. Para tránsito, los últimos meses son los que presenta una mayor cantidad de tweets similar a la pirotecnia, aunque para este el mayor valor lo tiene en el mes de mayo, nos da a entender que debido a que en la Ciudad de México se presentan diversos eventos como el día de muertos o la navidad, la gente suele viajar para ver a familiares pues el transporte común es mediante un vehículo, por lo que similar a la interpretación anterior con la segunda gráfica, ocasiona aglomeraciones creando tráfico, en el caso de pirotecnia, en estos eventos, es especial los religiosos y año nuevo se acostumbra el uso de fuegos artificiales para celebrar dicha festividad, con el fin de dar un bonito espectáculo visual. Por último, en el tema de incendio tiene mayores tweets en el mes de mayo y vuelve a incrementarse hasta el mes de noviembre, este último podría deberse a una consecuencia por descuido durante las celebraciones, ligandolo con la pirotecnia, si no se usa o limpia correctamente sus restos pueden llegar a ocasionar accidentes como son el caso de incendios.

Por último, en la cuarta gráfica (parte inferior lado derecho) nos muestra los tweets por tema y año. Se logra observar que el tránsito, pese a ser el de mayor cantidad de tweets presenta un decremento considerable entre el 2020 a 2022, los años donde se presentó el problema del COVID-19 y las restricciones para salir de casa durante ese tiempo. En el tema de incendio lo muestra en el año 2019 para después disminuir al siguiente año, en el año 2020 aumenta considerablemente. Para la pirotecnia, la mayor cantidad de tweets fueron en el año 2020.

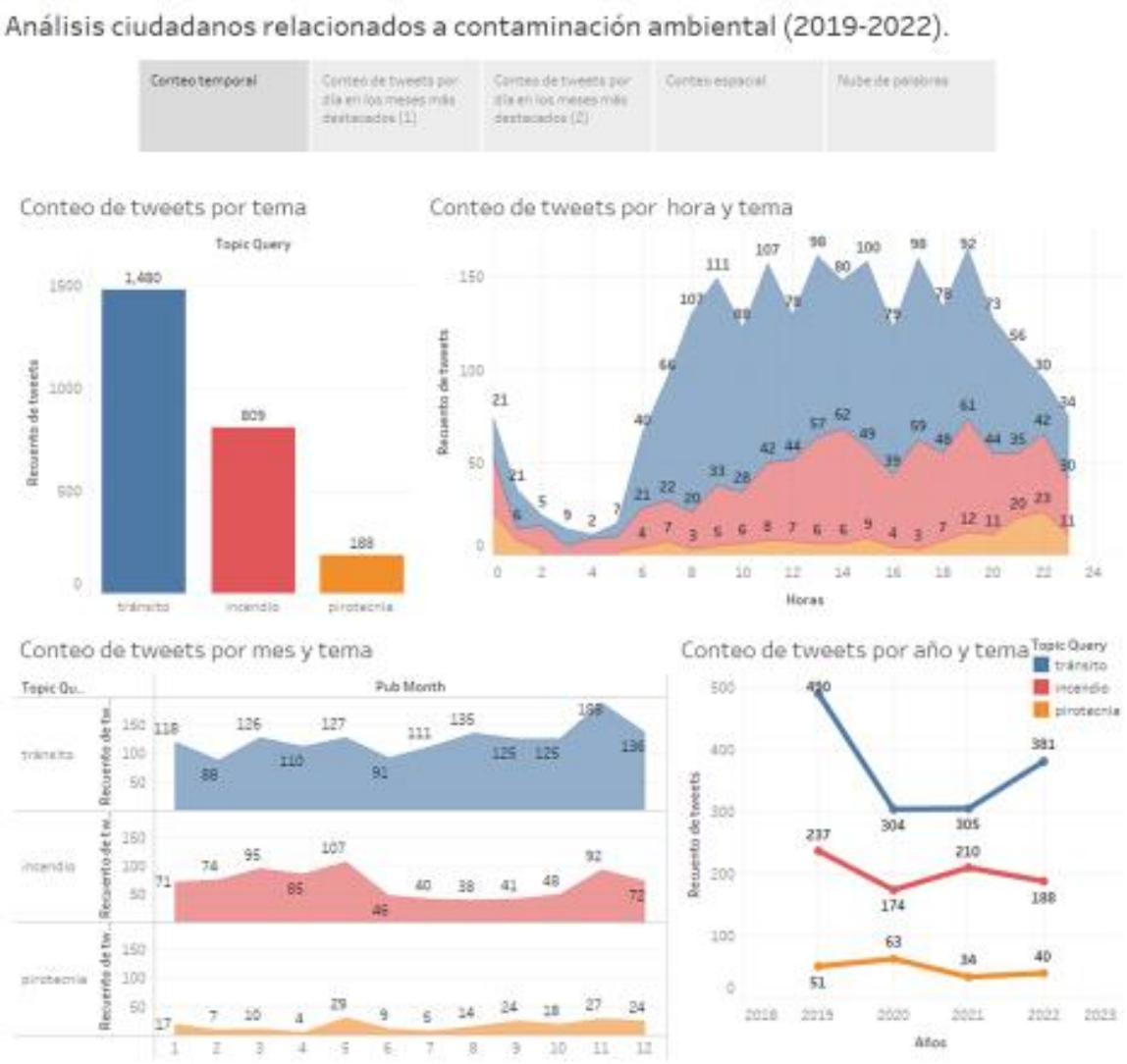

Figura 52. Gráficas de: Conteo de tweets por tema, por hora y tema, por mes y tema, y, por año y tema de los datos históricos.

En la figura que se muestra a continuación, se tienen dos gráficas, en la primera nos muestra un conteo de tweets por día en el mes de mayo y podemos observar que en este mes los temas que más predominan son Tránsito e incendios.

Podemos notar que el día con mayor número de denuncias relacionadas a tránsito fue el 15 de mayo con once denuncias mientras que los días con menor número de denuncias fueron el 1 y el 28 de mayo. De una manera lógica podemos deducir que el día 1 de mayo no hubo tantas denuncias de tránsito debido a que es un día festivo en el que alguna parte de la población permanece en casa o se encuentra de vacaciones en otro lugar. Hablando del tema de incendios podemos ver que el día con mayor número de denuncias fue el 11 de mayo con diez denuncias, aunque es notorio que este mes las denuncias de este tipo se mantienen en casi la misma cantidad en un periodo del 11 al 19 de mayo, y esto pasa porque son fechas en donde se presentan climas muy calurosos que a veces llegan a provocar incendios.

Finalmente, en el tema de pirotecnia el día con más denuncias es el 3 de mayo con 8 denuncias y de los 31 días que contiene el mes solo once días hubo denuncias de este tipo.

Con todo esto podemos decir que durante el mes de mayo se pudo observar que hay una mayor cantidad de denuncias referentes al tema de incendios.

Por otra parte en la segunda gráfica se observa el conteo de tweets por día en el mes de septiembre, y aquí algo que se nota a simple vista es que debido a la festividad del día de la independencia el día 16 de septiembre es el día con más denuncias relacionadas al tema de pirotecnia. De igual forma podemos ver que aproximadamente en esas fechas es cuando se presentan denuncias de incendios.

También podemos darnos cuenta de que en este mes las denuncias de tráfico se mantienen en un rango de 2 a 8.

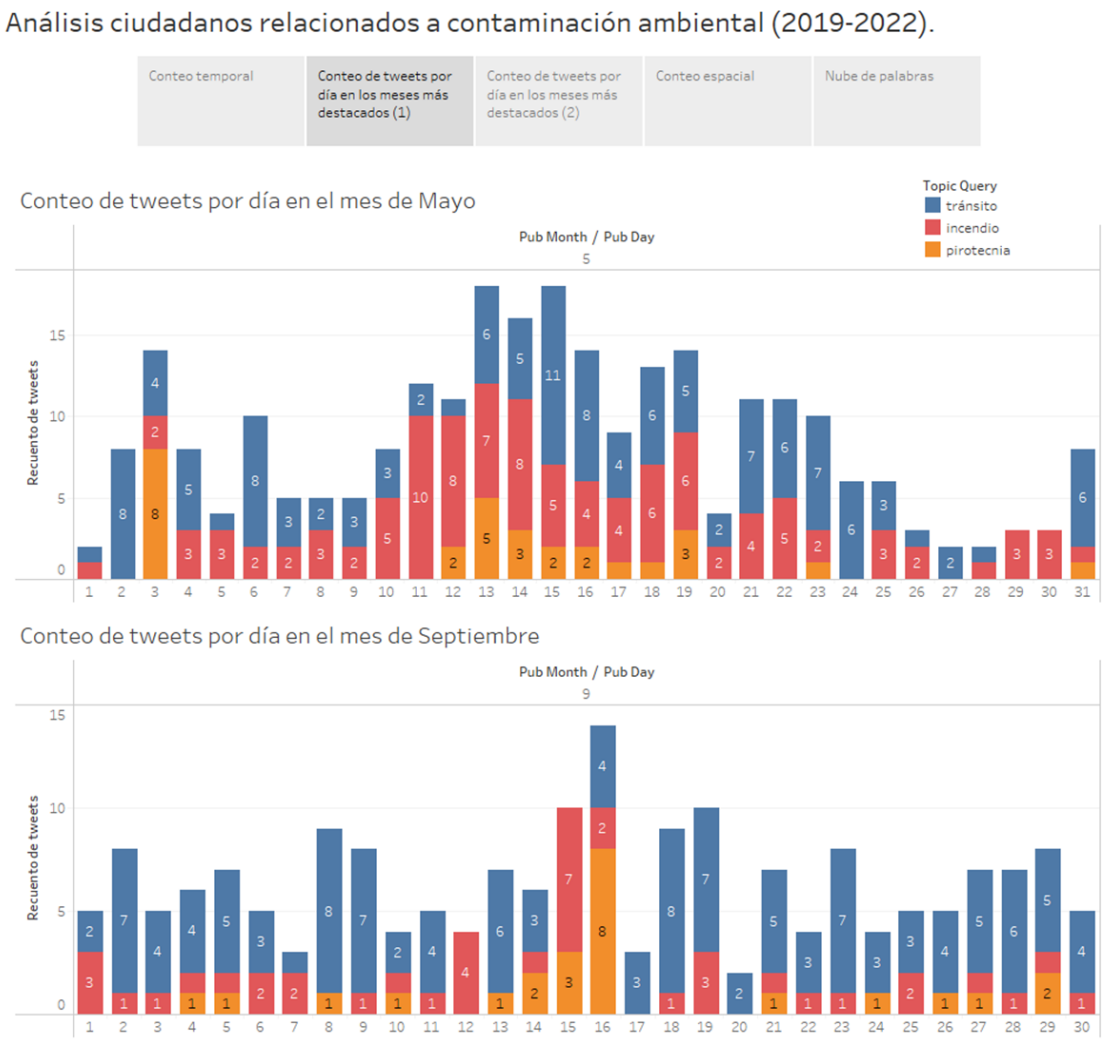

Figura 53. Gráficas de: Conteo de tweets por día en los meses más destacados (Mayo y Septiembre)

Al igual que las gráficas anteriores, en estas se muestra el conteo de tweets por día en un mes, en este caso es para el mes de noviembre y para el mes de Diciembre.

Para el mes de noviembre observamos que el mayor número de denuncias son sobre tránsito, siendo el día 13 de Noviembre el día con mayor número de denuncias.

En cuestión de incendios el día que más incendios se presentaron fue el 29 de noviembre.

En la segunda gráfica se muestra el conteo en el mes de diciembre y como era de esperar los días con mayores denuncias por pirotecnia son el 11 y 12 debido a las festividades religiosas.

Hablando de tránsito, podemos notar que los primeros días de diciembre hay más denuncias de este tipo, mientras que los últimos días de diciembre no hay tantas denuncias, podemos llegar a deducir que esto sucede porque es un periodo vacacional y la población sale de vacaciones o simplemente están en casa.

También se nota que, aunque son pocas, las denuncias sobre incendios permanecen casi todo el mes, siendo el día con más número de denuncias el 3, 19, 20 y 30 de diciembre.

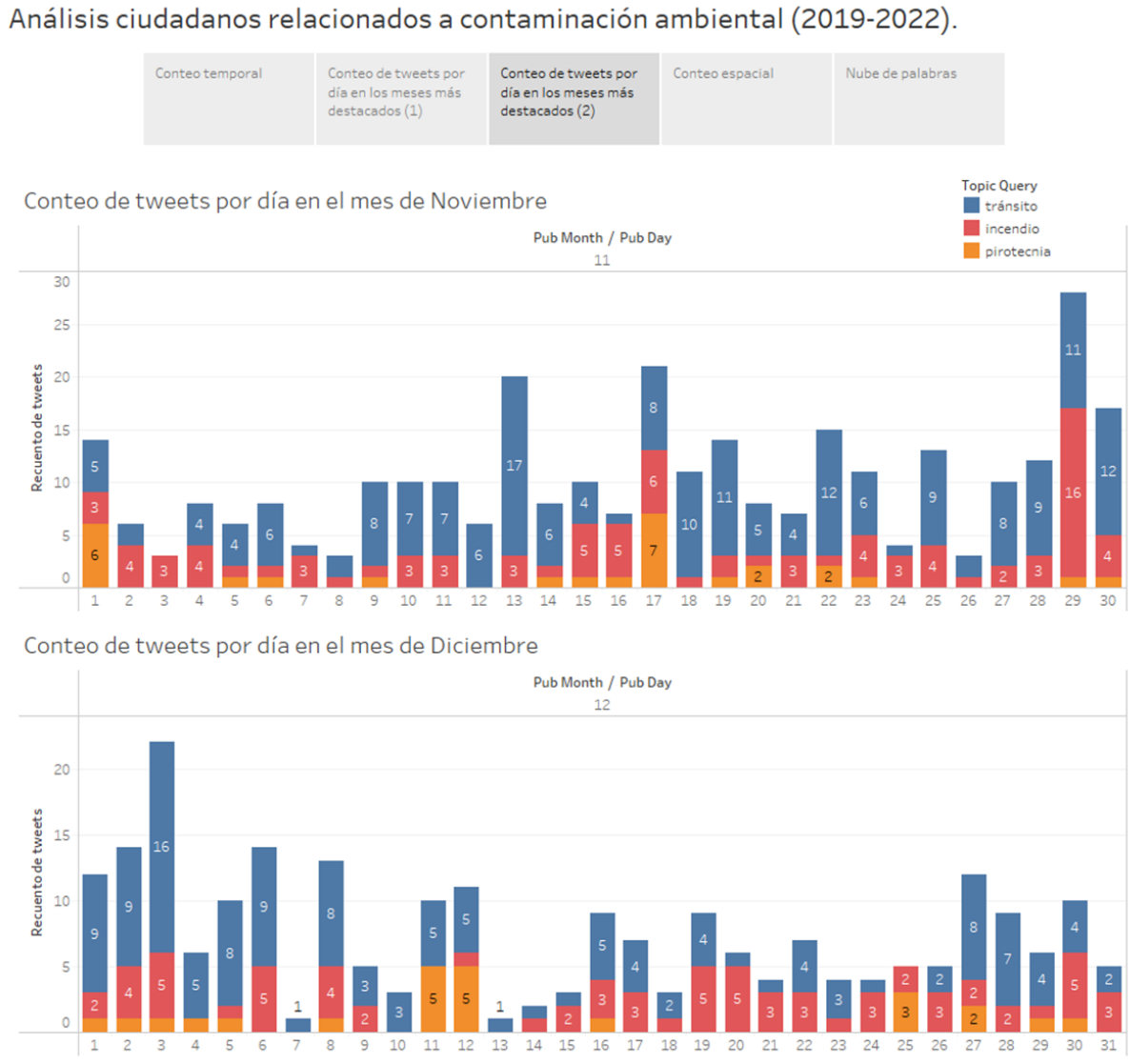

Figura 54. Gráficas de: Conteo de tweets por día en los meses más destacados (Noviembre y Diciembre)

En la siguiente figura podemos observar 3 gráficas que nos muestran la cantidad de tweets que están relacionadas con cada una de las alcaldías, en la primera gráfica (parte superior) tenemos un conteo de tweets en cada uno de los temas, donde podemos identificar rápidamente cuáles son las 3 alcaldías más relevantes para cada uno de los temas. Para tránsito tenemos que la alcandía Tlalpan, Cuauhtémoc y Azcapotzalco son las alcaldías donde más se realizan denuncias relacionadas con el tránsito, tráfico vehicular y accidentes automovilísticos. En Iztapalapa, Tlalpan y Álvaro Obregón son los lugares donde más se reportan incendios y accidentes con fuego, donde la alcaldía de Iztapalapa destaca entre las demás. Por último, tenemos que en pirotecnia se denuncian más casos de su uso en las alcaldías de Iztapalapa, Tlalpan y Miguel Hidalgo, donde la diferencia no es mucha ya que la cantidad de denuncias por pirotecnia no es mucha porque es situacional por las fechas de alto consumo.

En la segunda gráfica (parte inferior izquierda) tenemos un mapa de calor creado a partir de las coordenadas de cada uno de los tweets donde se puede observar la concentración de estos en la zona norte de la Ciudad de México, cubriendo zonas como la alcaldía Cuauhtémoc, Coyoacán, Benito Juárez, Miguel Hidalgo, Azcapotzalco, etc. esto como se muestra en la primera gráfica hace referencia a estas mismas alcaldías donde hubo muchas más denuncias por tema. Ya que de manera automática colocamos las coordenadas a los tweets que no las tenían, la mayor cantidad de tweets se concentran en un mismo lugar, por lo que en el caso de Iztapalapa o Tlalpan no se ve una concentración muy grande, pero la siguiente gráfica muestra lo contrario.

Para la última gráfica (parte inferior derecha) tenemos una nube de palabras con los nombres de todas las alcaldías de la Ciudad de México, donde notamos que la concentración de tweets es proveniente de las alcaldías Iztapalapa, Tlalpan, Cuauhtémoc y así sucesivamente, por lo que las alcaldías menos mencionadas o donde menos se hicieron denuncias son Milpa Alta, La Magdalena Contreras e Iztacalco.

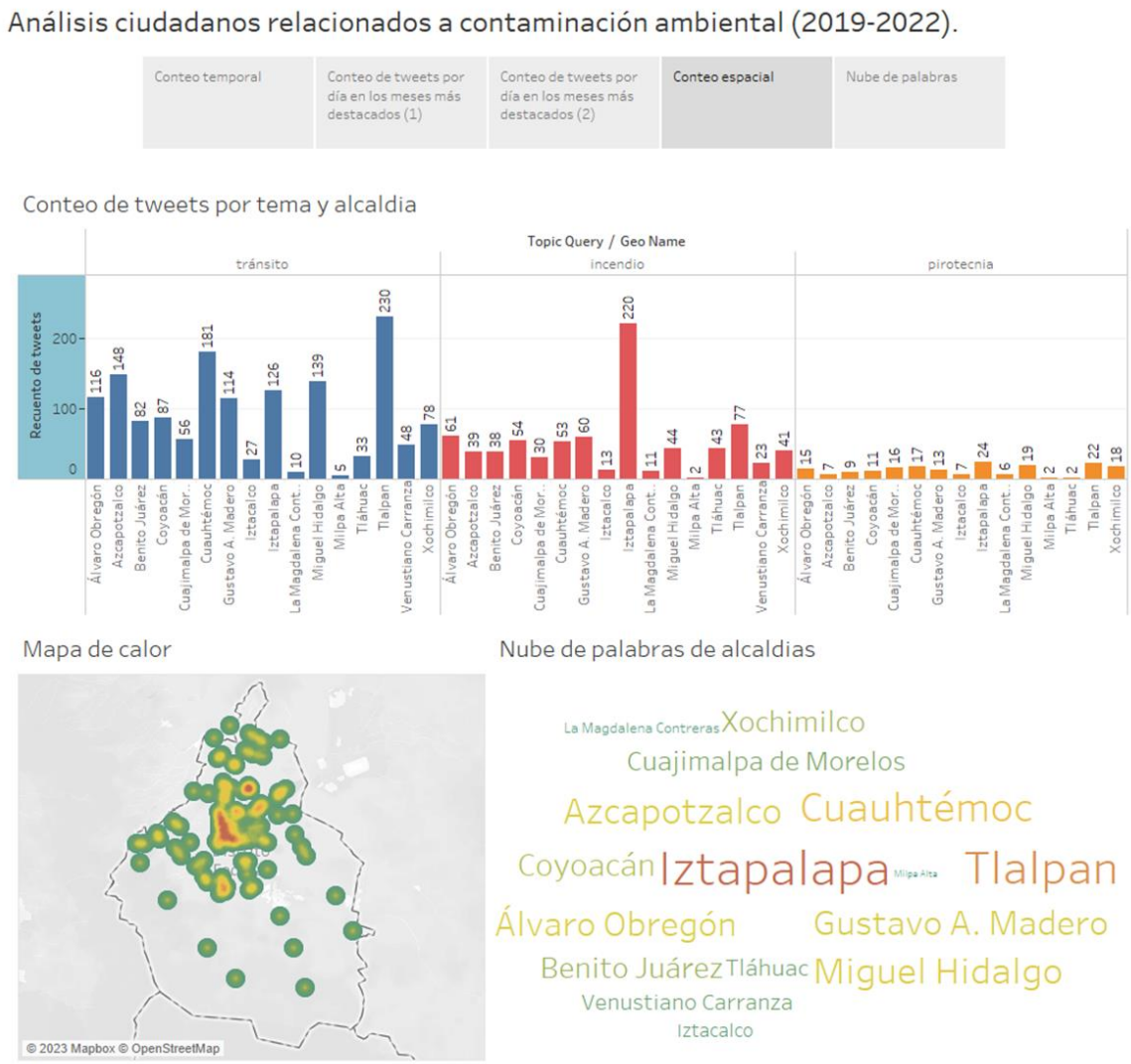

Figura 55. Gráficas de: Conteo de tweets por tema y alcaldía, mapa de calor y nube de palabras de las alcaldías.

Figura 56. Gráficas de: Nube de palabras de usuarios, menciones y tendencias.

En la figura 56, tenemos las nubes de palabras formadas a través de los datos de los tweets. En la primer nube (superior lado izquierdo) se enfoca en los usuarios que más realizaron denuncias que se representan con los nombres que tienen el mayor tamaño de letra, “trafico889”, “C5_CDMX”, “LXM2608” son algunos de los más destacados. Los que atribuyen menos se colocaron sus nombres con tipos de letras más pequeños.

Para nuestra segunda nube de palabras (parte superior, lado derecho), en este caso muestra a los usuarios más mencionados en las denuncias ciudadanas, “UGS_CDMX”, “Alc_Iztapalapa”, “Claudiashein” son algunos ejemplos de estos usuarios. “GobCDMX” o “SSC_CDMX” serían algunos de los menos mencionados en esta sección.

Llegando a nuestra tercera nube de palabras (parte inferior), aquí se muestran las mayores y menores tendencias que se mencionan en los tweets que son las palabras o términos más utilizados en dichas publicaciones de carácter de denuncia. Vemos que “precaución” es el más sobresaliente de todos seguido de “Incendio”, “accidente”, “vialidad” y así sucesivamente. Mencionando algunos poco frecuentes serían por ejemplo “Cuajimalpa” y “Iztapalapa”, dándonos a entender que en dichas alcaldías es poco frecuente encontrar denuncias referentes a los temas seleccionados.

6.3.2. Análisis de Modelado de Tópicos.

Durante el modelado de tópicos debemos hallar los principales temas o tópicos que nos ayudarán a interpretar la información que hemos recabado, para esto es necesario como primer punto obtener la cantidad de tópicos óptima. A continuación, se mostrará la gráfica de coherencia resultante de los tokens de los tweets.

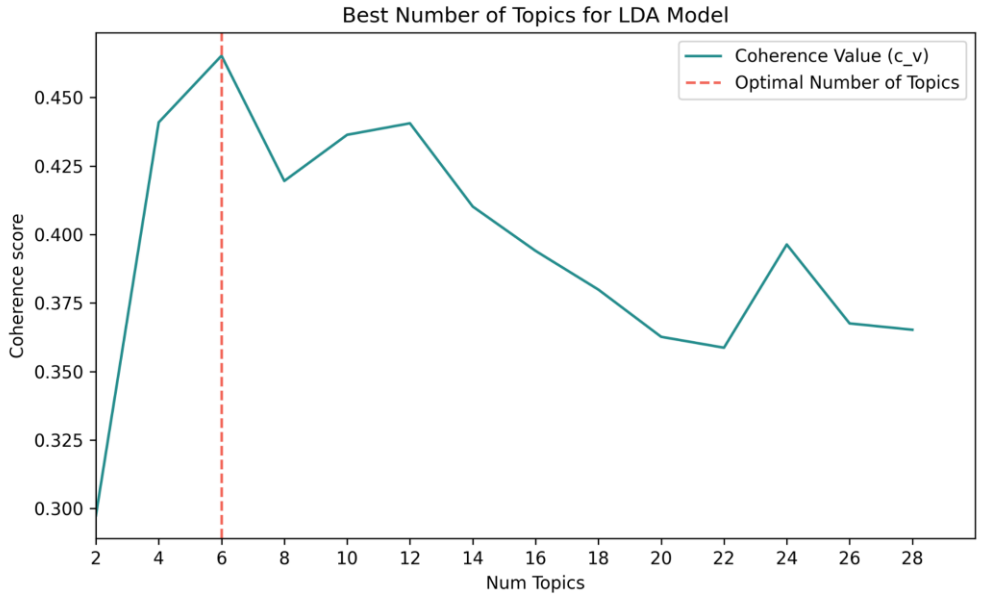

Figura 57. Gráfica de coherencia a través del modelo LDA.

Está gráfica, nos permite identificar la cantidad de tópicos que más nos favorece para llevar a cabo el modelo LDA. Como se puede observar en la figura anterior, en el eje vertical tenemos el valor de coherencia, que representa la ocurrencia que tienen de aparecer los distintos elementos encontrados a través de los tokens de los tweets. Con dicha gráfica se determinó, que la cantidad de tópicos que debemos considerar son un total de seis tópicos, los cuales se pueden apreciar en la siguiente figura.

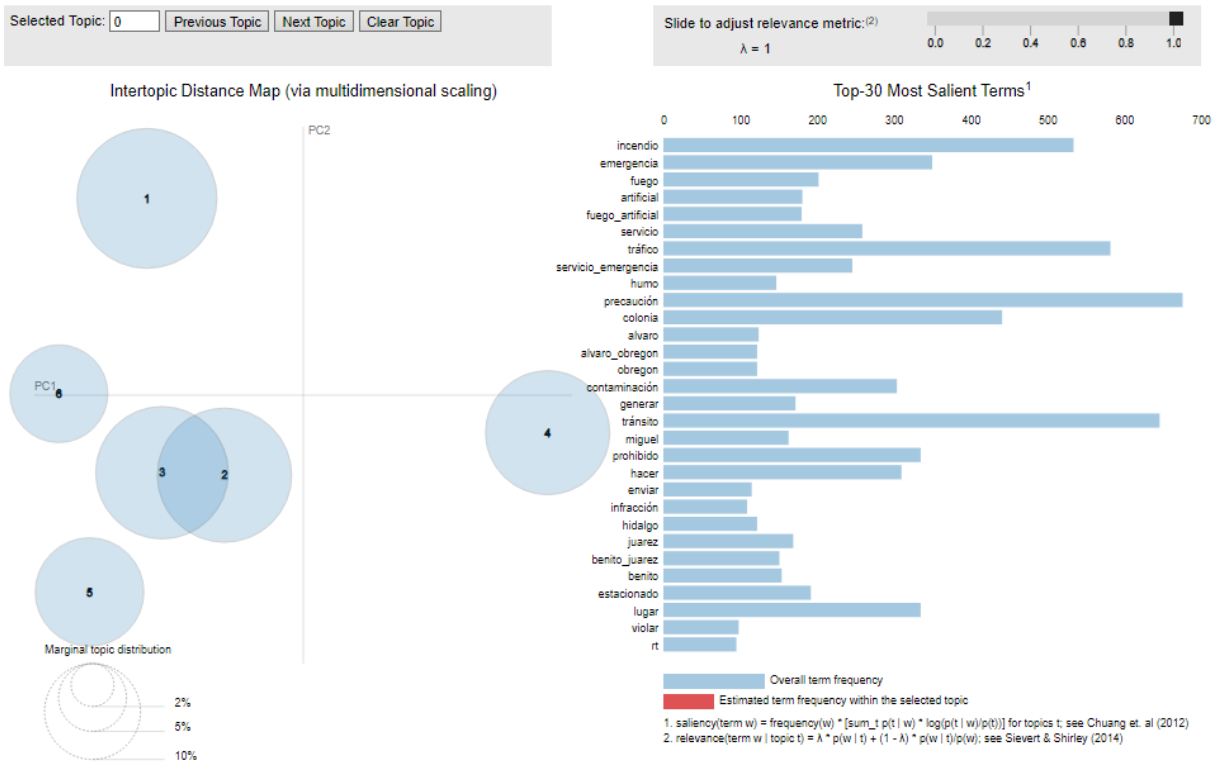

Figura 58. Visualización de tópicos y términos a través del modelo LDA.

En la figura 42, podemos ver en mayor medida la distribución de los tópicos generados por el modelo LDA que se encuentran del lado izquierdo, al igual que la visualización de los términos más frecuentes en los datos obtenidos de los tweets que están ubicados en la parte derecha. Estos valores de ocurrencia se pueden presentar de dos formas, una de ellas es, por así decirlo la frecuencia global de los términos como se muestra en la figura anterior, la otra es la estimación de esos términos en los distintos tópicos presentes que se mostrará más adelante.

Podemos ver que los términos que se muestran, las palabras "tráfico", "precaución" y "tránsito" son las que presentan una mayor frecuencia, siendo "precaución" el más ocurrente de todos. Con esto, de primeras, podemos determinar, que de los tres temas que seleccionamos: tráfico, incendios y pirotecnia, el tema del tráfico es el más común. Las palabras que se pueden apreciar, algunas con los valores más bajos serían "violar", "rt" e "infracción", el menor y más interesante sería el término "rt", una abreviación por parte de los usuarios que a simple vista no se puede identificar su significado y se tiene que recurrir a los datos completos, los cuales al analizarlos se detectó que se refiere en la mayoría de los casos al reglamento de tránsito de la CDMX, en otros es la forma acortada de "retweet".

En la sección de términos frecuentes, tenemos una barra que nos permite ajustar la métrica de relevancia, con el fin de mostrar los términos menos ocurrentes en los temas seleccionados.

Para la sección de los tópicos podemos seleccionar cualquiera de ellos para comenzar a ver los términos más frecuentes en dicho tópico o tema, también podemos hacerlo a través de los botones de la parte superior, asimismo un botón para limpiar la selección actual.

A continuación, mostraremos los términos más frecuentes del primer tema.

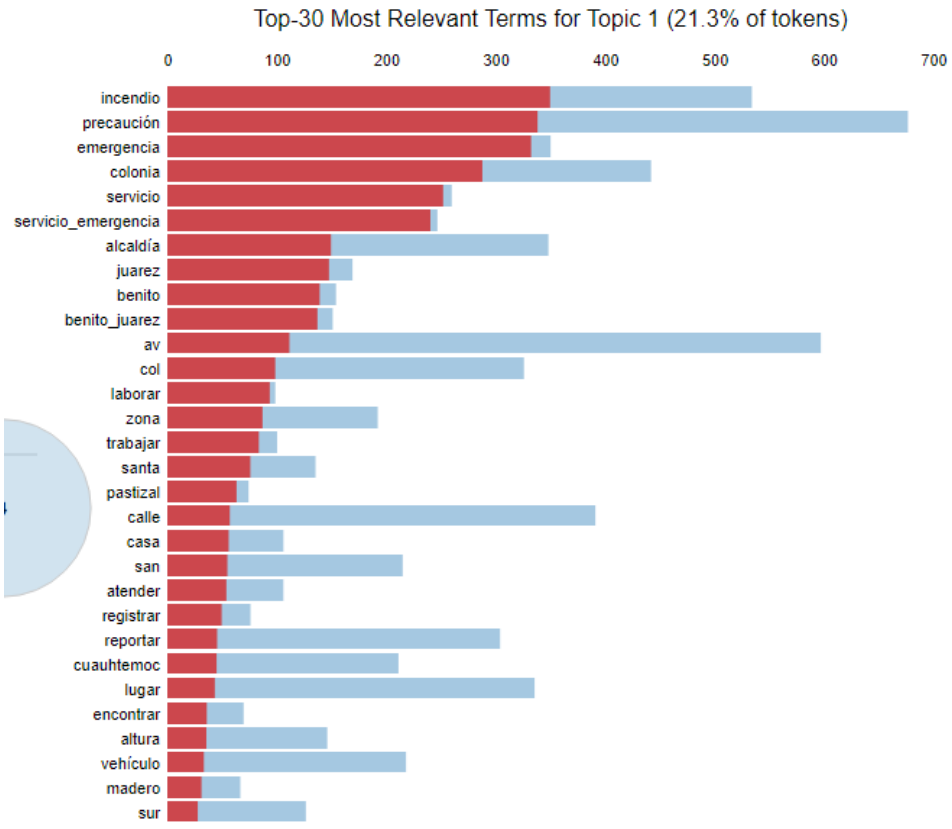

Figura 59. Gráfica de términos más relevantes del primer tópico a través del modelo LDA.

En la figura 43, podemos ver la representación de los valores del primer tema (las barras de color rojo) con los valores generales (barras de color azul). La palabra "precaución", es una de las más utilizadas de forma general, sin embargo, en este tópico, su frecuencia estimada es la mitad de dicho valor, esto quiere decir que es una de las

más relevantes y comunes en este tema, asimismo como término global. Sin embargo, en este tópico el más sobresaliente es la palabra “incendio” por muy poco.

Revisando más a fondo, encontramos términos como “pastizal”, “casa”, y “vehículo”, los cuales, si los enlazamos con el más destacado, podríamos interpretar que las denuncias que escribe la gente tratan comúnmente sobre incendios en pastizales, en viviendas o vehículos.

Las palabras “juarez”, “benito”, “cuauhtémoc”, “madero” y “alcaldía”, además de que nos dan a entender de qué se tratan de alcaldías, podemos deducir que esta clase de denuncias por parte de los ciudadanos se concentra en las alcaldías Cuauhtémoc, Benito Juárez y Gustavo A. Madero.

En cuánto el término “servicio_emergencia”, “laborar” y “trabajar” se refieren principalmente a los bomberos, que frecuentemente son mencionados cuando se reporta un incendio o accidente, una ubicación concisa que abarca el nombre de la colonia seguido del municipio, calle o avenida. Al final piden a los que se encuentran cerca de la zona de incendio que tengan mucha precaución o cuidado.

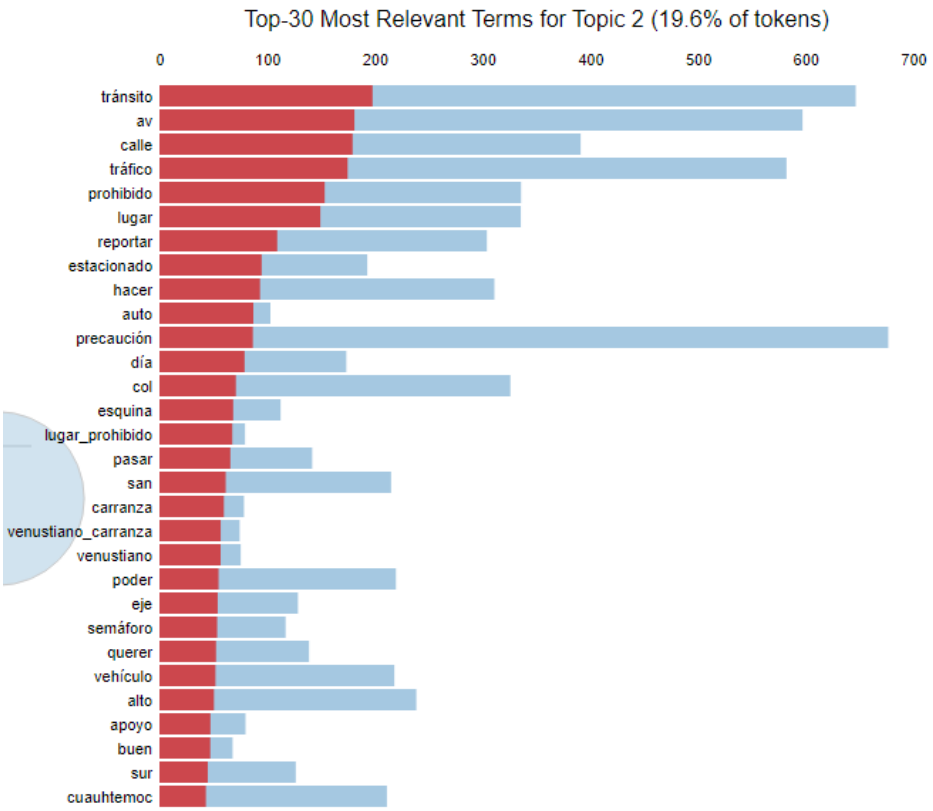

Figura 60. Gráfica de términos más relevantes del segundo tópico a través del modelo LDA.

En la figura 44, vemos los términos más ocurrentes del tópico o del segundo tema. En dicho tema, el término más frecuente es “tránsito”, después le siguen términos como “av”, “calle”, “esquina” y “tráfico” que interpretamos que este tópico abarca el tema del tráfico y los ciudadanos tienden a reportar utilizando “av”, la abreviación de avenida y la palabra calle para seguramente dar una dirección más concreta, junto a las palabra “reportar”, “semáforo”, lugar” y “vehículo” nos dan a comprender mejor qué se tratan de denuncias por parte de la gente, sobre semáforos o vehículos que entorpecen la circulación del lugar.

Hablando de las alcaldías, “carranza” y “cuauhtémoc”, o, dicho de otro modo, las alcaldías Venustiano Carranza y Cuauhtémoc serían las más relevantes o más comunes en las que pueden haber denuncias sobre tráfico a causa de un vehículo o avería de un semáforo.

El término “lugar_prohibido” y “estacionado”, se refiere a automóviles que se estacionan en sitios inapropiados o que concretamente está marcado por el reglamento como prohibido, sin embargo, al están analizando la información que nos proporcionaron los tweets, nos dimos cuenta que estás actividades suceden en gran medida y muchas veces estos actos incorrectos ayudan a entorpecer el tránsito de la zona.

La palabra “apoyo”, a diferencia de la palabra “reportar”, aquí los ciudadanos además de denunciar lo que está sucediendo solicitan a las autoridades venir al sitio para remediar la situación ya sea remover la unidad que está provocando el problema o manejar el tránsito con el fin de que sea más fluido.

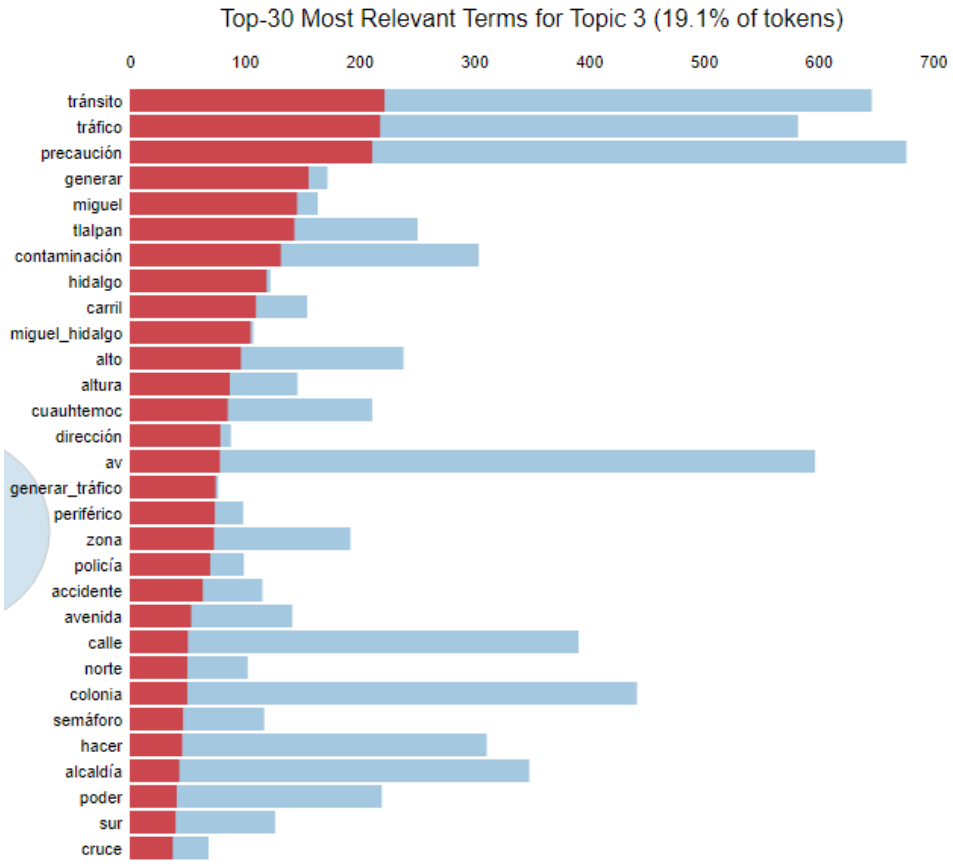

Figura 61. Gráfica de términos más relevantes del tercer tópico a través del modelo LDA.

En la figura 45 se visualizan los términos más relevantes del tercer tópico siendo el más destacable “tránsito” que, al igual que el tema anterior, este también abarca el tema del tráfico.

Aunque ambos enfocan el mismo tema, difieren en las palabras más comunes, por ejemplo, en este tópico tenemos el término “precaución” mientras que el anterior dicha palabra no se muestra como uno de los treinta más relevantes, también palabras como “carril” y “accidente” son más frecuentes en este tópico, esto sucede debido a que algunas causas de tráfico muy comunes que se denuncian en tweets son accidentes viales.

Sumando los términos “miguel_hidalgo”, “cuauhtémoc” y “tlalpan” se interpretaría que en las alcaldías Miguel Hidalgo, Cuauhtémoc y Tlalpan son frecuentes accidentes de tránsito que relaciona el término “generar_tráfico”.

También tenemos los términos “sur”, “norte” y “cruce” se utilizan para especificar si es en la zona norte o sur, en especial cuando se presenta “periférico”, ya que los ciudadanos tienden a especificar si se trata del periférico sur o periférico norte. Además, hablando de periférico encontramos tweets que pertenecían a la alcaldía Álvaro Obregón y también Tláhuac.

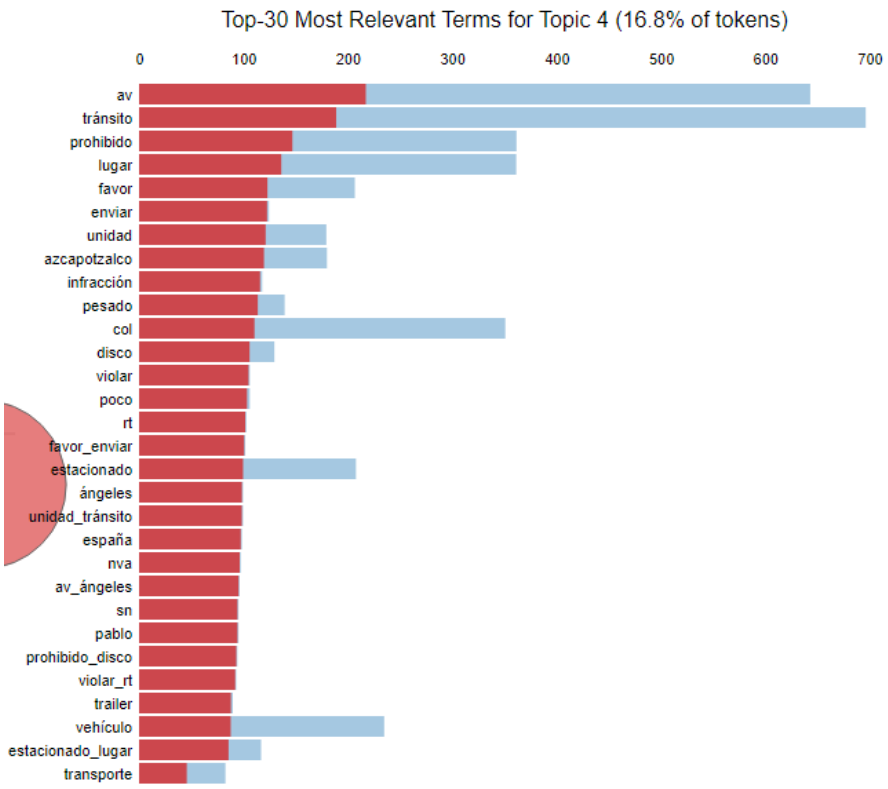

Figura 62. Gráfica de términos más relevantes del cuarto tópico a través del modelo LDA.

En la figura 46 podemos observar que los términos más relevantes son relacionados a tránsito, tenemos palabras como “trailer”, “vehículo” y “transporte” que podemos interpretar como los principales medios que se encuentran en el tránsito.

También podemos notar la presencia de palabras como “violar”, “infracción”, “prohibido”, “reglamento de tránsito”, “estacionado” la cuales pueden ser un indicador de posibles causas de tránsito. Por otra parte, tenemos el término “unidad_tránsito” así como “enviar” y “unidad” lo cual indica que en la mayoría de las denuncias que hace la población pide ayuda a unidades de tránsito para que éstas regulen la circulación automovilística y de esta manera se solucione el problema de tránsito. Otra interpretación que le podemos dar es que solicitan dichas unidades con la finalidad de infraccionar a alguien que ha violado el reglamento de tránsito; comúnmente encontramos tweets donde se indica que hay vehículos estacionados en lugares prohibidos evitando el paso a otros vehículos y por ende causando tráfico vehicular, también encontramos tweets donde la población indica que un vehículo dio vuelta prohibida creando problemas en la vialidad.

En este tópico podemos encontrar términos relacionados a datos de localización tales como “avenida”, “col”, “nva”, “españa”, “av_ángeles” y “azcapotzalco”. Después de analizar los tweets pudimos observar que una de las alcaldías donde se presentan reportes de tránsito es Azcapotzalco, y dentro de esta alcaldía la mayoría de las personas coinciden en que estos sucesos pasaban en la avenida ángeles que se encuentra dentro de la colonia nueva españa.

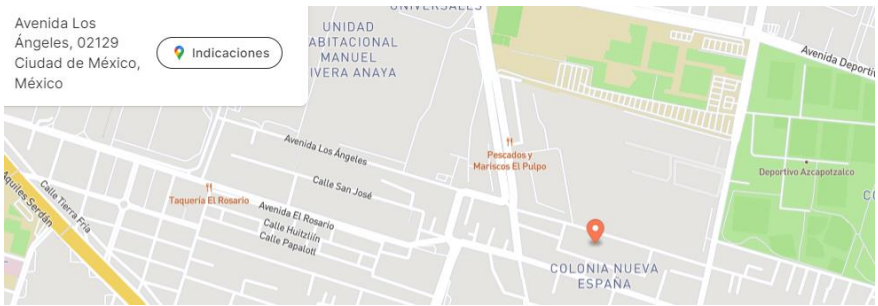

Figura 63. Mapa de la avenida ángeles, alcaldía Azcapotzalco.[41]

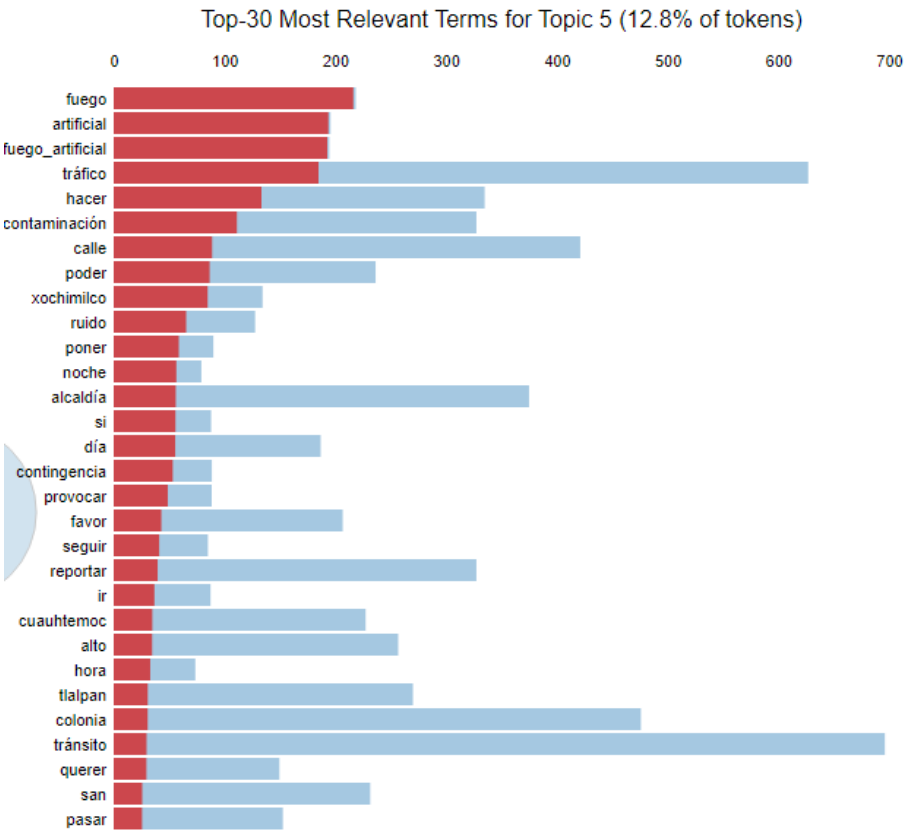

Figura 64. Gráfica de términos más relevantes del quinto tópico a través del modelo LDA.

En el quinto tópico notamos que los términos más relevantes son "fuego", seguido de "artificial" y "fuego_artificial" por lo que podemos decir que el tema predominante es sobre pirotecnia.

Otro de los términos asociados a pirotecnia es "ruido" ya que las personas se quejan por el ruido que causa la pirotecnia, tal como lo podemos ver en la figura 49.

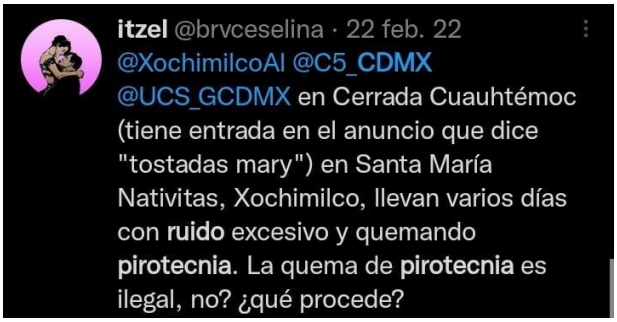

Figura 65. Ejemplo de denuncia sobre pirotecnia en Twitter.

De igual forma podemos ver que hay algunos términos generales como "contingencia" y "contaminación", esto sucede porque cuando la gente hace denuncias sobre pirotecnia suelen hacerlo cuando hay contingencia porque les parece molesto e inadecuado lanzar pirotecnia cuando hay contingencia en cdmx.

Aunque también tenemos el término "tránsito" y nuevamente encontramos términos asociados como "colonia".

Podemos observar que la mayor parte de denuncias por tráfico y pirotecnia se concentran en las alcaldías "cuauhtemoc", "tlalpan" y "xochimilco"

En este tópico también encontramos palabras como "día", "noche", "hora", ya que los usuarios en algunos tweets especifican el horario en el que están denunciado, aunque esto no es algo tan relevante. Usualmente la palabra noche es encontrada en los tweets relacionados a fuegos artificiales ya que es en este horario cuando se presentan más eventos de este tipo.

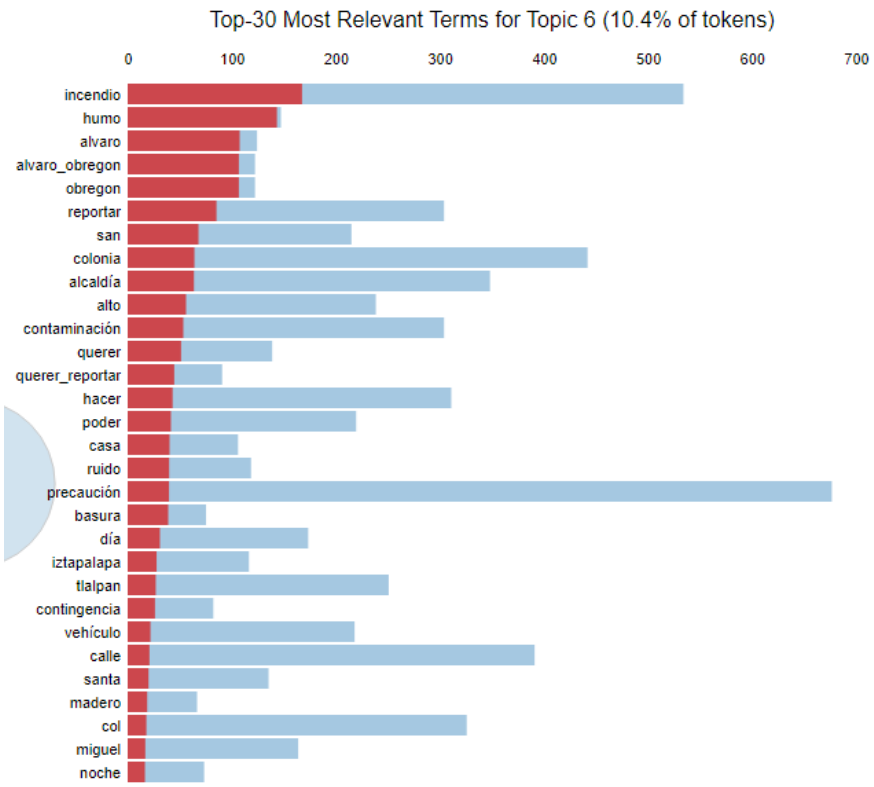

Figura 66. Gráfica de términos más relevantes del sexto tópico a través del modelo LDA.

Llegando al último tema, vemos que "incendio" es el más relevante, dándonos a entender que se enfoca al tema de los incendios en la Ciudad de México, consigo trae la palabra "humo", siendo la segunda más común, el cual puede significar que utilizan el término como un sinónimo de incendio o en otro escenario, los ciudadanos reportan presencia de humo inusual que pueden ver en sus cercanías, el cuál puede ser señal de un posible incendio.

Los términos "contaminación", "contingencia" y "ruido" son para referirse a lo que se estaba presentando en dicho momento que el ciudadano escribía el tweet, por ejemplo, si cerca de su ubicación un auto o cualquier clase de vehículo emitía una cantidad de humo anormal, agregaría que dicha unidad está aumentando la contaminación del lugar, si en dicho momento se había aplicado contingencia es posible que se haga referencia a ello.

Otra palabra que encontramos es "noche" que podría interpretarse que, en este tema, las presencia de humo o incendios sucedieron a altas horas de la tarde.

En cuanto a "alvaro_obregón", "tlalpan", "iztapalapa" o "madero" son las alcaldías más mencionadas en este tópico: Alvaro Obregón, Tlalpan, Iztapalapa y Gustavo A. Madero, que podemos deducir que son las cuales se presentaron más estos sucesos.

Finalmente de acuerdo a la interpretación que le hemos dado a cada uno de los 6 tópicos y con todos los términos que se han obtenido hasta el momento podemos decir que a grandes rasgos el tema más mencionado en las denuncias ciudadanas es el tránsito (el cual es mencionado en el tópico 2, 3 , 4 y parte del 5) , seguido de los incendios (mencionados en el tópico 1 y 6) y por último pirotecnia (mencionada en el tópico 5).

CAPÍTULO 7. CONCLUSIONES.

En este trabajo, a través de la implementación de distintas técnicas de minería de datos y módulos como Web Scrapping, nos permitieron tener acceso a las publicaciones de Twitter y llevar a cabo el proceso de extracción del ETL.

Las querys también tienen su gran peso para el proceso de extracción, con ellas podemos especificar al robot de extracción los criterios que deben de contener cada tweet respecto a los tres diferentes temas seleccionados: pirotecnia, tráfico e incendios.

Al llevar a la práctica el robot de extracción, se detectaron la frecuencia de que se habla de estos temas, siendo el tránsito vehicular el más popular, obteniendo una gran cantidad de tweets a diferencia de incendio y pirotecnia, siendo este último el menos frecuente. Sin embargo, demuestra un incremento considerable de tweets entre ciertos meses como diciembre y enero, aunque nunca llega estar al nivel de publicaciones referentes al tránsito e incendios.

Sin embargo, trabajar con los datos conseguidos de esa primera extracción, aunque podría darnos una mayor cantidad de tweets más para analizar de los que terminamos trabajando, existen una cantidad inmensurable de publicaciones que tendríamos que pasar para localizar los que nos interesan o contienen información relevante para nuestro proyecto, el cual, nos llevaría demasiado tiempo, por ende, se realizaron más extracciones, donde concretamos más los términos de búsqueda de nuestras querys para reducir la mayoría de los tweets innecesarios y en consecuencia posibles tweets útiles debido a que no cumplen con dichos parámetros.

A pesar de ello, con los tweets obtenidos, logramos obtener información relevante acerca de los tres temas seleccionados, por ejemplo, notamos que el tema de tráfico en los años donde se presentó la pandemia contra el COVID-19, sufrió un decremento de tweets considerable, exponiendo la posibilidad de que uno de los factores que influye el tráfico en la CDMX se debe a la gran cantidad de vehículos como conductores que la misma, sus calles no pueden manejar tanto flujo que en consecuencia las avenidas se saturan, agregando la información obtenida a través del análisis de modelado de datos que una parte de esos conductores violan las reglas establecidas de tránsito, estacionados en sitios prohibidos o yendo en sitio contrario al establecido provocando la disminución de carriles que los demás por seguridad propia bajen su velocidad con el fin de evitar un accidente que muchas veces no se pueden evitar, y, cuando ocurre un accidente, los servicios de emergencia y policías necesitan acudir al lugar donde ocurrieron los hechos creando caos vial y entorpeciendo más la circulación de la zona.

Con el tema de incendios, tienen una mayor frecuencia entre los primeros cinco meses y posteriormente al último mes, comúnmente incendios de pastizales, vehículos, hogares incluso la quema de basura por parte de empresas o de los mismos ciudadanos que son denunciados por los usuarios que se encuentran cerca del lugar o desde su ubicación a lo lejos presenciaron humo, los cuales estos indicios de humo detectados, posteriormente son confirmados como incendios o por otra fuente por cuentas oficiales de la CDMX como la de bomberos.

En el caso del tema de pirotecnia, pese a tener una cantidad diminuta en comparación a los demás, aun así, se logró adquirir información resultante. La pirotecnia es más común en épocas de celebración establecidas como son el grito de la independencia o año nuevo, como celebraciones religiosas, por ello, los meses donde se mostró un aumento en tweets fueron enero, mayo, septiembre, noviembre y diciembre. También cuando se realizaba el proceso de limpieza manual, se detectó que mucha gente no toma en cuenta la nocividad que la pirotecnia trae consigo al medio ambiente como a los seres vivos a pesar de que la misma CDMX expuso los percances que traen los fuegos artificiales, incluso imponiendo leyes que restringen su uso, sin embargo, debido a las tradiciones y costumbres, mucha gente termina obviando esas restricciones sin tener en cuenta las consecuencias que traerá en el futuro.

Enfocándonos en las alcaldías, logramos detectar los municipios más mencionados, o dicho en otras palabras, los que presentan mayores situaciones que perjudican a la calidad del aire. Con los resultados obtenidos, se ubicaron las alcaldías de Iztapalapa y Tlalpan que presentan mayor mención u publicaciones referentes. Ante dicha información, podemos identificar cuáles alcaldías son las que más necesitan atención o tienen la mayor fuente contaminante que recae en la calidad del aire de la CDMX.

Se podría extender más sobre todo los resultados que se consiguieron a través de los análisis que se llevaron a cabo, pero, el punto que se desea resaltar es pese a que se perdieron publicaciones que podían haber sido relevantes para este trabajo con el fin de ahorrar tiempo en limpiezas manuales, con los datos que se obtuvieron, se logró llevar a cabo un análisis exhaustivo de estos, asimismo de una interpretación de lo que se encontró en dichos análisis, refiriéndonos los realizados con la herramienta de Tableau y mediante el modelo de tópicos (LDA), los

cuales, el primero nos ayudó principalmente en apreciar las tendencias en tiempo y espacio de los datos históricos y mediante el modelado de tópicos nos permitió analizar en profundidad los tokens que se generaron de cada tweet para crear tópicos o temas que en base a la frecuencia de los términos de dichos tokens, ayudándonos a la interpretación de la información que contienen los tweets.

Para finalizar, con el desarrollo de este trabajo, con la implementación de diversos programas de software y múltiples conocimientos adquiridos en nuestra trayectoria escolar, se realizó un prototipo de software capaz de extraer publicaciones de la red social Twitter referentes a los tres temas propuestos: pirotecnia, tráfico e incendios que pertenezcan a la CDMX relacionadas a la contaminación del medio ambiente, en concreto a la calidad del aire y posteriormente realizar un análisis por el cual se obtendrá una caracterización temporal y espacial de los resultados obtenidos.

CAPÍTULO 8. REFERENCIAS.

[1] Fundación Aquae (2013). Los tipos de contaminación y sus principales consecuencias [En línea]. Disponible en: https://www.fundacionaquae.org/wiki/tipos-contaminacion/

[2] P. José. (2018). https://encolombia.com/medio-ambiente/interes-a/contaminacion-de-fuente-puntual-difusa-y-lineal/

[3] Ministerio del Medio Ambiente. (2018). [En línea]. Disponible en : https://luminica.mma.gob.cl/que-es-la-contaminacion-luminica/

[4] G. Carolina. (2021). Tipos de contaminación. [En línea]. Disponible en : https://ceupe.mx/blog/tipos-de-contaminacion.html

[5] Responsabilidad Social Empresarial y Sustentabilidad. (2021). [En línea]. Disponible en : https://responsabilidadsocial.net/contaminacion-que-es-tipos-y-causas/?amp

[6] Sistema Nacional de Información Ambiental y de Recursos Naturales. (2015). Informe del Medio Ambiente. [En línea]. Disponible en :https://apps1.semarnat.gob.mx:8443/dgeia/informe18/tema/cap5.html#:~:text=Además%20de%20los%20efectos%20sobre,combinarse%20con%20el%20agua%20presente

[7] Bicentenario Perú. (2021). Efectos de la contaminación del aire. [En línea]. Disponible en : https://infoaireperu.minam.gob.pe/efectos-de-la-contaminacion-del-aire/

[8] Instituto Nacional de Ecología y Cambio Climático. (2007). "Los vehículos automotores como fuentes de emisión," secretaria de Medio Ambiente y Recursos Naturales, Instituto Nacional de Ecología y Cambio Climático Ciudad de México. [En línea]. Disponible en: http://www2.inecc.gob.mx/publicaciones2/libros/618/vehiculos.pdf

[9] A. Lizeth. (2019). "Impacto de los Incendios Forestales en la Ciudad de México," Evaluación de Riesgos Naturales, Ciudad de México. [En línea]. Disponible en: https://ern.com.mx/boletines/NotadeInteres/ERNterate_Nota_IncendiosForestales_y_Contaminacion_CDMX.pdf

[10] Secretaría del Medio ambiente y Recursos Naturales. Contaminación por pirotecnia. [En línea]. Disponibles en: https://www.gob.mx/semarnat/articulos/contaminacion-por-pirotecnia

[11] C. Freire, "Las redes sociales trastocan los modelos de los medios de comunicación," Revista Latina de Comunicación Social, vol. 11, núm. 63, España, Canarias, 2008, pp. 287-293.

[12] Instituto Nacional de Salud Pública. (2019). "Contaminación del aire, ambiente y salud." Instituto Nacional de Salud Pública, Ciudad de México. Disponible en: https://insp.mx/assets/documents/webinars/2021/CISP_Contaminación%20del%20aire%20(19%20oct)%204.pdf

[13] Secretaría del Medio Ambiente de la Ciudad de México. (2020). "Calidad del aire en la Ciudad de México, Informe 2018," Dirección General de Calidad del Aire, Dirección de Monitoreo de Calidad de Aire, Ciudad de México. [En línea]. Disponible en: http://www.aire.cdmx.gob.mx/descargas/publicaciones/informe-anual-calidad-del-aire-2018.pdf

[14] Procuraduría Ambiental y del Ordenamiento Territorial. (2002). Estadísticas Generales. [En línea]. Disponible en: http://www.paot.org.mx/contenidos_graficas/delegaciones/graficas_gral.php

[15] D. R. Andrés, "Cambio climático, la penalidad del ozono y la mortalidad asociada en la Ciudad de México," Tesina Licenciatura de Econ., C. de I. y Docencia Econ. A.C, CIDE, CDMX, 2020.

[16] Xataka México (2021). México tiene el lugar 11 en países más contaminantes por PM2.5, las partículas culpables de cuatro millones de muertes anuales en el mundo [En línea]. Disponible en: https://www.xataka.com.mx/ecologia-y-naturaleza/mexico-tiene-lugar-11-paises-contaminantes-pm2-5-particulas-culpables-cuatro-millones-muertes-anuales-mundo#comments-close

[17] Camargo, R. (2020). Cambio climático, la penalidad del ozono y la mortalidad asociada en la Ciudad de México. [En línea]. Extraído de: http://revistas.sena.edu.co/index.php/rnt/article/view/3517/3953

[18] Hootsuite. (2021). Informe Global sobre el entorno digital 2021 [En línea]. Disponible en: https://www.hootsuite.com/es/recursos/tendencias-digitales-2021

[19] B. Carlos. (2020). Solo 1.3 de cada 10 mexicanos tiene una cuenta de Twitter y .8 son usuarios activos [En línea]. Disponible en: https://www.abestudiodecomunicacion.com.mx/solo-1-3-de-cada-10-mexicanos-tiene-una-cuenta-en-twitter-y-8-son-usuarios-activo

[20] Red Hat. (2017). ¿Qué es una API ? [En línea]. Disponible en : https://www.redhat.com/es/topics/api/what-are-application-programming-interfaces#:~:text=Una%20API%20o%20interfaz%20de,el%20software%20de%20las%20aplicaciones

[21] IBM Cloud Education. (2020). Application Programming Interface (API). [En línea]. Disponible en : https://www.ibm.com/cloud/learn/api#:~:text=Application%20programming%20interfaces%2C%20or%20APIs, and%20functionality%20easily%20and%20securely

[22] Microsoft. (2019). Extracción, transformación y carga de datos (ETL). [En línea]. Disponible en: https://docs.microsoft.com/es-mx/azure/architecture/data-guide/relational-data/etl

[23] IBM Cloud Education. (2020). Natural Language Processing (NLP). [En línea]. Disponible en: https://www.ibm.com/cloud/learn/natural-language-processing [24] Microsoft. (2019). Tecnología de procesamiento de lenguaje natural. [En línea]. Disponible en: https://docs.microsoft.com/es-mx/azure/architecture/data-guide/technology-choices/natural-language-processing

[25] Microsoft. (2019). Almacenamiento de datos en Microsoft Azure. [En línea]. Disponible en: https://docs.microsoft.com/es-mx/azure/architecture/data-guide/relational-data/data-warehousing

[26] IBM. (2021). Esquemas de Estrella. [En línea]. Disponible en: https://www.ibm.com/docs/es/ida/9.1.2?topic=schemas-star

[27] P. Federico. (2019). Topic Modeling: An introduction [En línea]. Disponible en: https://monkeylearn.com/blog/introduction-to-topic-modeling/

[28] H. Chelsey. (2020). Topic Models. [En línea]. Disponible en: https://rpubs.com/chelseyhill/672546

[29] Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, Zhao L (2019) Latent Dirichlet allocation (lda) and topic modeling: models, applications, a survey. Multimed Tools Appl 78(11):15169–15211

[30] B. Chris. (2019). Topic Modeling. [En línea]. Disponible en: https://cbail.github.io/SICSS_Topic_Modeling.html

[31] L. Theo. (2021). Introduction to The Structural Topic Model (STM). [En línea] Disponible en: https://towardsdatascience.com/introduction-to-the-structural-topic-model-stm-34ec4bd5383

[32] M. Sammarco, R. Tse, G. Pau, G. Marfia, "Using geosocial search for urban air pollution monitoring," Pervasive Mob. Comput., pp. 15-31, 2017. [En línea]. Disponible en: https://doi.org/10.1016/j.pmcj.2016.07.001

[33] K. A. Naranjo Análisis de correlación entre el índice de calidad del aire y el impacto en Twitter para la ciudad de Bucaramanga aplicando análisis de series temporales, extracción y procesamiento de lenguaje natural. [En línea]. Disponible en: http://hdl.handle.net/20.500.12749/15350.

[34] Porras Fresneda, Antonio Luis. (2019). Análisis del efecto en las redes sociales de las crisis de contaminación en Madrid. [En línea]. Disponible en: https://oa.upm.es/56734/

[35] G. Lorenzo, R. Salvatore, R. Francesco, V. Daniel, "From Tweets to Semantic Trajectories: Mining Anomalous Urban Mobility Patterns," Citizen in Sensor Networks, pp. 26–35, 2013. [En línea]. Disponible en: https://doi.org/10.1007/978-3-319-04178-0_3

[36] L. Theo. (2021). Introduction to The Structural Topic Model (STM). [En línea] Disponible en: https://towardsdatascience.com/intuitive-guide-to-correlated-topic-models-76d5baef03d3

[37]SEDEMA. (2021). Pide Sedema no quemar pirotecnia para evitar contingencias ambientales. [En línea]. Disponible en: https://www.sedema.cdmx.gob.mx/comunicacion/nota/pide-sedema-no-quemar-pirotecnia-para-evitar-contingencias-ambientales

[38] R. Alessandra. (2020) "Proyecto de decreto por el que se reforma la fraccion v¡i del artículo 28 de la ley de cultura civica de la Ciudad de México," Partido Verde Ecologista de México. [En línea]. Disponible en: https://congresocdmx.gob.mx/archivos/parlamentarios/IN_298_09_20_02_2020.pdf

[39] B. Armando. ¿Cuál es la relación de la contaminación, el tráfico y el Hoy No Circula? [En línea]. Disponible en: http://www.cienciamx.com/index.php/ciencia/ambiente/6435-el-trafico-el-hoy-no-circula-y-su-relacion-real-con-la-contaminacion-en-la-cdmx

[40] CONAPCI. Incendios urbanos en México. [En línea]. Disponible en: https://conapci.org/incendios-urbanos-en-mexico/

[41] MAPS. Coordenadas de avenida ángeles, Azcapotzalco [En línea]. Disponible en: www.ecosia.org/map?q=av%20angeles%20azcapozalco&coordinates=19.4988311,-99.1875765&placeName=Avenida%20Los%20Ángeles