Universidad de Alcalá

*Trabajo Fin de Master Data Science*

# Servicio de Quejas y Sugerencias Ayuntamiento de Zaragoza: análisis y predicción.

**Julián García Begué (riglos_jgb@yahoo.es)**

## 1. Introducción. Motivación y objetivos.

El Ayuntamiento de Zaragoza ofrece a sus conciudadanos su servicio de Quejas y Sugerencias via web (1) en el que pueden solicitar la realización de obras y prestación de servicios o la mejora de los existentes, expresar su descontento por las actuaciones del Consistorio, informar de problemas puntuales en la infraestructura de la ciudad…

Se trata de aprovechar una fuente de información desestructurada aportada por la ciudadanía para extraer conclusiones acerca de los problemas que cada uno de esos ciudadanos perciben como importantes (se toman la molestia de comunicarlos por la web) . Esta información surge por un canal completamente diferente a los canales formales de interacción entre el ciudadano y el gobierno local, regulados por la legislación y los procedimientos administrativos, y proporcionan la posibilidad de descubrir problemas latentes cuya solución mediante políticas adecuadas permita la mejora del bienestar de los ciudadanos de Zaragoza.

Los objetivos de este Trabajo son, por un lado, analizar los datos subyacentes referentes a las quejas emitidas durante los años 2017 a 2023, y por otro lado mejorar el tratamiento realizado sobre la información que representan , extrayendo información que de otra manera se encuentra implícita en dicho conjunto de datos y que puede mejorar su gestión tanto a nivel externo, hacia los usuarios, como interno, perfeccionando la herramienta a disposición de los funcionarios del Ayuntamiento.

Dichos objetivos se concretan de la siguiente manera:

-Correcta clasificación de la queja o sugerencia : Aunque al ciudadano se le propone inicialmente un servicio municipal como destinatario de su reclamación mediante un desplegable en el formulario web, es posible y ocurre, que el contenido de la queja tenga nada o poco que ver con el departamento al que la haya dirigido. Esta situación puede suponer una carga adicional de trabajo debido a la redirección de la misma al órgano municipal adecuado. Mediante un algoritmo de clasificación

puede corregirse automáticamente este problema optimizando la utilización de los recursos humanos del Ayuntamiento.

-Evaluación de las políticas municipales: por medio de técnicas de topic modelling, podemos agrupar y cuantificar las quejas recibidas, identificando las áreas de servicio susceptibles de ser mejoradas.

-Racionalización del conjunto de datos preexistente, reestructuración de la base de datos subyacente y visualización geográfica de la información según diversos criterios.

## 2. Estado del arte

Existen multitud de investigaciones en las que, por medio de la gran diversidad de herramientas que ofrece el Natural Language Processing , se ha abordado la clasificación, predicción y modelado de tópicos de datasets compuestos por textos de cualquier índole. Desde compilaciones de tweets, calificaciones de servicio por parte de clientes, valoraciones de películas… En todos ellos se utilizan técnicas de preprocesamiento de los textos incluidos (tokenización, lemmatización…), a las que se aplicarán diversas herramientas (Gensim, Scikit Learn Tfidfvectorizer y otras) con las cuales se consigue traducirlos a vectores numéricos. Por último, estos vectores alimentarán a la gran cantidad de modelos predictivos disponibles en Machine Learning para arrojar predicciones sobre conjuntos de nuevos datos.

En el caso que nos ocupa, podemos aplicar estas técnicas de igual manera, pero dadas las características de la información, en la que las reclamaciones superan ampliamente el número de agradecimientos, utilizaremos la predicción con otros fines que iré detallando más adelante, aunque no dejaré de lado el Topic Modelling, que nos dará una idea general de los contenidos analizados, ni la generación de visualizaciones que pueden mejorar si cabe la comprensión de los Topics generados.

Si bien, como he comentado antes existen infinidad de trabajos que se han centrado en el análisis de comentarios y opiniones, no he encontrado ninguno relacionado con la materia que nos ocupa.

## 3. Materiales y métodos

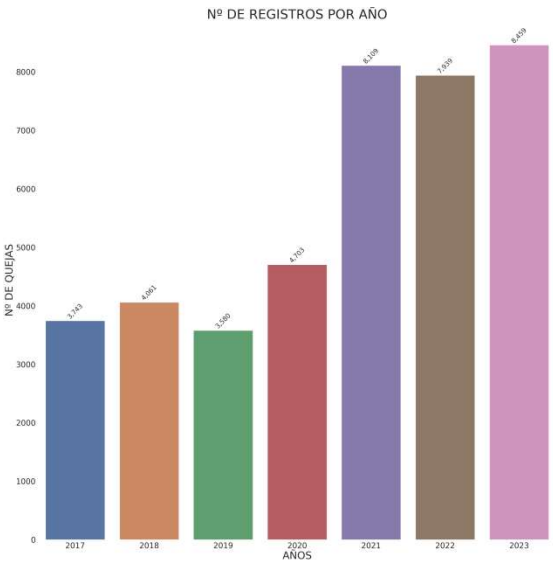
La página web del Ayto ofrece un formulario sencillo en la que aparte de datos personales , en el apartado “Descripción de la queja / sugerencia”, el ciudadano puede detallar su contenido. Además, puede optar por dirigirla al destinatario que entiende competente mediante un desplegable (“Categoria”) que ofrece 57 departamentos. Para 5 de ellos se sugieren a su vez los problemas que los responsables de la

implementación de esta herramienta consideran que son más frecuentes, elevando la cifra de opciones a 84 . Se ha de introducir un título introductorio (“Asunto”) y el detalle de la incidencia (“Descripción”). Se pueden adjuntar archivos siempre que no superen 1 Mb de información. Es posible también ubicar la dirección pertinente por medio de la navegación sobre un mapa de la ciudad. Por último se le sugiere al usuario la posibilidad de publicar el contenido de su reclamación (salvo datos personales) por la Oficina de Participación.

El dataset sobre el que se trabaja se ha descargado por medio de un script de Python utilizando la API proporcionada por el Ayuntamiento (2). Se dividen los registros obtenidos en dos subconjuntos que abarcan desde el 1-1-2017 hasta el 31-8-2023 (“train”) por un lado dejando los restantes desde esa fecha, en otro, (“test”), con el objetivo de realizar predicciones y validar los modelos. La estructura de sus campos es la siguiente:

1) service_request_id: Identificador único de la queja
2) status: Situación en la tramitación de la misma (“open”, “closed”)
3) service_code: identifica al departamento competente
4) service_name: nombre del departamento en cuestión
5) title: resumen del contenido de la reclamación
6) description: detalle de dicho contenido
7) requested_datetime: fecha de la queja/sugerencia
8) updated_datetime: fecha de la última edición del registro por parte del Ayto
9) address_string: Localización introducida por el usuario
10) long: longitud en el mapa de la localización introducida
11) lat: su latitud
12) address_id: identificador único de la localización

Como primer acercamiento al análisis de los datos, podemos observar la evolución del número de quejas desde el 1 de Enero de 2017 a 31 de Agosto de 2023 :




| Año | Número: |
|---|---|
| 2017 | 3.743 |
| 2018 | 4.061 |
| 2019 | 3.580 |
| 2020 | 4.703 |
| 2021 | 8.109 |
| 2022 | 7.939 |
| 2023 | 8.459 |

**3.1 Reclasificación datos:**

La primera tarea que voy a realizar es reorganizar el dataset en función del campo "service_code": en el conjunto descargado de la API, observo que en dicho campo se almacenan tanto las categorías como las subcategorías que se ofrecen al usuario en la web (además de otros códigos que no aparecen en la misma y que pueden corresponderse con registros no recodificados a lo largo de los años después de alguna actualización de la base de datos o bien alguna reclasificación realizada en la gestión diaria del departamento). Para realizar esta actualización me serviré de Python y en especial de su potente librería Pandas. Analizar sobre esos datos iniciales puede llevar a resultados engañosos como podemos observar si cuantificamos las reclamaciones totales realizadas a los departamentos:

**Antes de la actualización:**

| service_name | número |
|---|---|
| Policia Local | 3.398 |
| Limpieza viaria | 2.435 |
| Movilidad Urbana: Señalización | 2.372 |
| Calzada | 1.971 |
| Urbanismo | 1.681 |
| Contenedores | 1.667 |
| Parques y Jardines | 1.604 |
| Césped, arbustos, flor | 1.598 |
| Alumbrado Publico | 1.512 |
| Arbolado | 1.429 |

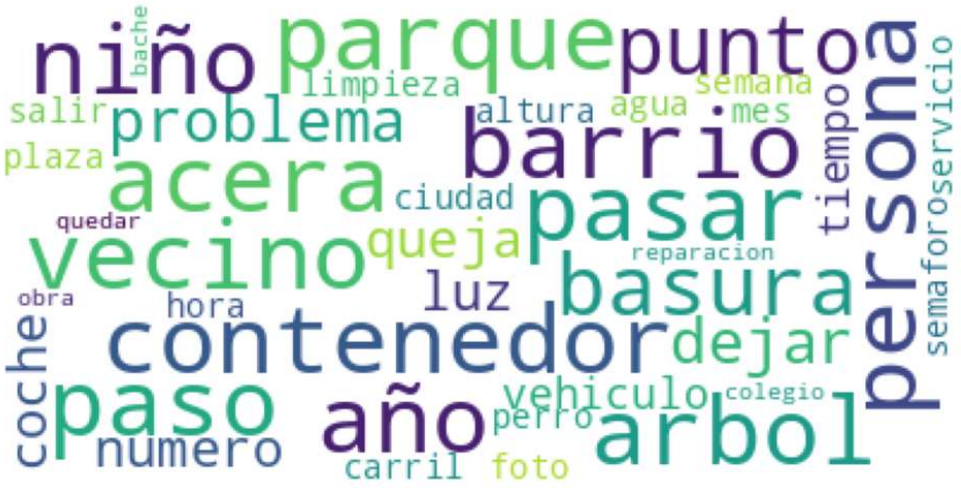
**Después de la actualización:**

| service_name | número |
|---|---|
| Parques y Jardines | 9.285 |
| Limpieza Pública | 5.850 |
| Infraestructuras:Conservacion | 4.747 |
| Policia Local | 3.398 |
| Alumbrado Publico | 2.785 |
| Movilidad Urbana: Señalización | 2.372 |
| Urbanismo | 1.682 |
| Movilidad Urbana. Transporte. | 1.278 |
| Instituto Mpal. Salud Publica | 847 |
| Cultura | 816 |

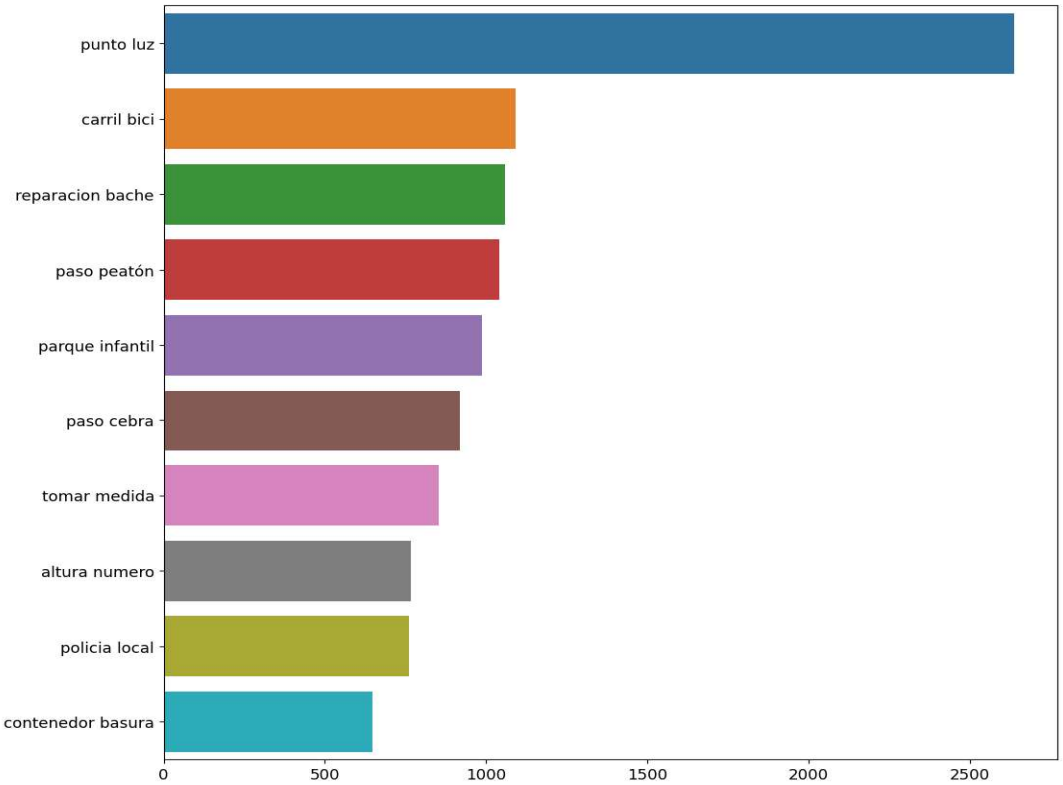
Como podemos observar, tras la correcta clasificación y agrupación de los códigos el ranking de departamentos con más reclamaciones cambia drásticamente y resulta evidente que las quejas se aglutinan en torno a tres departamentos concretos.

| Antes de la actualización: | | | Después de la actualización: | | |
|---|---|---|---|---|---|
| Departamento | Año | Número | Departamento | Año | Número |
| Césped, arbustos, flor | 2023 | 764 | Parques y Jardines | 2023 | 2.213 |
| Movilidad Urbana. Transporte | 2023 | 667 | Limpieza Pública | 2023 | 1.399 |
| Policia Local | 2023 | 560 | Infraestructuras:Conservacion | 2023 | 724 |
| Parques y Jardines | 2022 | 879 | Parques y Jardines | 2022 | 1.970 |
| Policia Local | 2022 | 697 | Limpieza Pública | 2022 | 1.259 |
| Limpieza viaria | 2022 | 553 | Infraestructuras:Conservacion | 2022 | 814 |
| Policia Local | 2021 | 700 | Parques y Jardines | 2021 | 1.965 |
| Calzada | 2021 | 512 | Limpieza Pública | 2021 | 1.132 |
| Arbolado | 2021 | 502 | Infraestructuras:Conservacion | 2021 | 1.002 |
| Policia Local | 2020 | 527 | Parques y Jardines | 2020 | 855 |
| Calzada | 2020 | 302 | Limpieza Pública | 2020 | 688 |
| Movilidad Urbana: Señalización | 2020 | 288 | Infraestructuras:Conservacion | 2020 | 550 |
| Policia Local | 2019 | 323 | Parques y Jardines | 2019 | 781 |
| Movilidad Urbana: Señalización | 2019 | 303 | Limpieza Pública | 2019 | 495 |
| Limpieza viaria | 2019 | 246 | Infraestructuras:Conservacion | 2019 | 493 |
| Policia Local | 2018 | 302 | Parques y Jardines | 2018 | 789 |
| Movilidad Urbana: Señalización | 2018 | 272 | Infraestructuras:Conservacion | 2018 | 676 |
| Calzada | 2018 | 241 | Limpieza Pública | 2018 | 459 |

Seguidamente, procedo, mediante la librería Spacy a la limpieza, tokenización y lematización del texto de descripción de la reclamación. En la siguiente imagen, quedan reflejadas las palabras que aparecen más de quiinientas veces en el corpus de estudio. Podemos ir haciéndonos una idea de las preocupaciones de los zaragozanos en lo que al estado de la ciudad se refiere.




Si ampliamos la representación al estudio de los bigramas más repetidos, vemos en que se concretan con más nitidez las palabras más frecuentes del gráfico anterior. Los puntos de luz, ocupan con diferencia el primer puesto según este análisis, seguidos de otros que aparecerán posteriormente con más nitidez en los tópicos resultantes.



### 3.2: Topic Modelling:

El siguiente paso es obtener los Topics del dataset, es decir, los temas latentes en el corpus de documentos. Para ello he utilizado la librería Gensim, el modelo LDA y el modelo Kmeans de clasificación. Gensim es una herramienta que nos permite obtener un diccionario, una bolsa de palabras y un conjunto de vectores numéricos que podemos volcar sobre los otros dos modelos. Después del análisis de los resultados obtenidos, creo que el modelo más eficiente ha sido Kmeans, que nos suministra los siguientes ocho tópicos con sus términos más frecuentes (tras su optimización):

| "limpieza" | "baches" | "farolas" | "parque" | "semáforos" | "contenedores" | "señalización" | "árboles" |
|---|---|---|---|---|---|---|---|
| agua | bache | luz | Parque | semaforo | contenedor | carril | arbol |
| barrio | reparacion | punto | infantil | fundido | basura | paso | rama |
| año | paseo | farola | niño | reparacion | cubo | señal | poda |
| limpieza | plaza | apagado | perro | paseo | lleno | acera | podar |
| vecino | solicito | fundido | columpio | posicion | papel | peatón | tapar |
| pasar | puente | alumbrado | juego | correcto | recogida | vehiculo | plantar |
| persona | aragon | fundida | roto | conductor | bolsa | coche | situado |
| tarde | señora | noche | fuente | colocar | organico | baldosa | caer |
| servicio | ciudadania | funcionar | agua | roto | vidrio | bici | año |
| queja | rotonda | encender | jugar | plaza | reciclaje | vertical | acera |

Y estos serían los departamentos más representativos, y por lo tanto susceptibles de mejora en su actuación:

| Departamento | Clúster | Quejas |
|---|---|---|
| Parques y Jardines | árboles | 2.355 |
| Infraestructuras:Conservacion | señalizacion | 2.264 |
| Movilidad Urbana: Semáforos Fundidos | semaforos | 525 |
| Parques y Jardines | parques | 2.615 |
| Parques y Jardines | limpieza | 3.817 |
| Alumbrado Publico | farolas | 2.316 |
| Limpieza Pública | contenedores | 2.514 |
| Infraestructuras:Conservacion | baches | 1.043 |

Según lo anterior, Zaragoza tiene una asignatura pendiente con sus parques y jardines. La limpieza también es una de las grandes preocupaciones de los zaragozanos, aglutinando el mayor número de registros a través de los clústers. Hay otros topics, como por ejemplo, la señalización, que pueden tener un componente también estructural, pero no hay que dejar a un lado, problemas del día a día como por ejemplo la gestión de contenedores, el alumbrado y la conservación de la calzada. Estas conclusiones surgen del estudio de la totalidad de los datos, la evolución de la política municipal podría fundamentarse en el análisis de la evolución anual del contenido y número de estos topics, aplicando los procedimientos vistos a conjuntos de datos anuales.

**3.3 Predicción:**

Dado que el conjunto de datos, como he comentado anteriormente está compuesto de textos que surgen mayoritariamente y dada su naturaleza, de un sentimiento negativo salvo en contadas ocasiones, no resulta razonable realizar una predicción de dicho sentimiento. No obstante, podemos utilizar los modelos predictivos para tratar de clasificar los nuevos registros en función de modelos ajustados a los datos de años anteriores en lo que a su correcta asignación de departamento corresponde. Es decir, vamos a asignar un código de departamento en función del texto contenido en la nueva reclamación gracias a la aplicación del modelo predictivo construido en base al contenido de quejas de años anteriores y el departamento que se les asignó.

El objetivo va a ser el siguiente: vamos a comparar el código de servicio asignado, tras la reclasificación y reagrupamiento de los registros en base a la codificación presentada en la página web del ayuntamiento, con el código resultante mediante el modelo predictivo en función del texto contenido en cada uno de los documentos. Con ello obtenemos, por un lado una valoración de esa reclasificación original, y por otro, un modelo que nos permite mejorarla y aplicarla a la clasificación de nuevos registros.

Para ello, alimentamos varios modelos de predicción (Regresión Logística, MultinomialNB, DecisionTreeClassifier) con los vectores numéricos obtenidos mediante TfidfVectorizer. De todos ellos, el que mejor ajusta, con un 71% de accuracy es el primero, el modelo logístico.

Aplicando los mismos modelos a los conjuntos de datos en bruto, tal y como se han descargado de la API, se constata que la accuracy es mucho menor. Resulta lógico pues los modelos han de predecir entre un mayor número de clases con un menor número de registros asociadas a las mismas.

Dicho lo anterior, puede parecer que dicha agrupación en un número más reducido de clases, limite las posibilidades de extraer más y mejores conclusiones. Esto depende básicamente de la utilidad final y concreta que quiera darse al análisis de la información y a la predicción sobre la misma, tanto más, cuando dicha información, mediante las adecuadas herramientas de programación y bases de datos, no se pierde ni se desvirtúa, quedando más bien ampliada y preservada para futuros estudios.

**3.4 Extracción de información:**

Analizando el dataset pude constatar la escasa información en relación a la localización geográfica de los registros: queda a la voluntad del usuario el indicar o no la dirección donde se origina la queja que relata. Esta se almacena en tres campos concretos, address_string, lat y long, estas dos últimas se pueblan al elegir el usuario sobre un mapa interactivo la ubicación del problema.

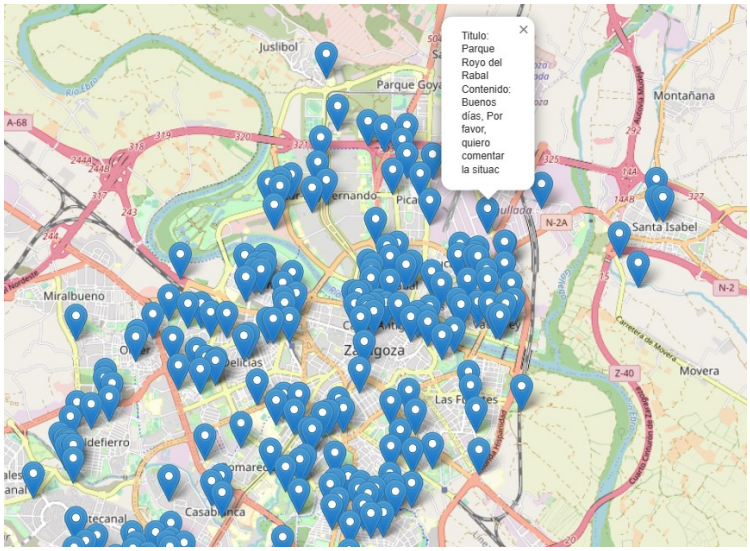
No pudiendo obligar al usuario a cumplimentar dicha información (no estaría de más, aunque es posible que le desalentara), entiendo que ésta resultaría de mucha utilidad al Ayuntamiento. Es por eso, que mediante la aplicación de paquetes como Stanza, Geocoder o simplente programación Python, se puede llegar a extraerla y tratarla, como explicaré a continuación:

- Stanza es una herramienta de Procesamiento de Lenguaje Natural, que nos interesa por su módulo NER, de reconocimiento de entidades con nombre, puesto que es capaz de extraer entidades de tipo “LOC”, es decir, localizaciones.
-Programación Python: basándonos en el callejero de Zaragoza, extraemos mediante código, las direcciones incluidas en el texto de los documentos.
-Posteriormente, introduzco las localizaciones en un procedimiento basado en la librería Geocoder, que recogiendo sus datos en buscadores de diverso tipo logra las coordenadas geográficas de las mismas, y otras informaciones, como por ejemplo el código postal. En algunos casos, el campo address_string, aparece vacío, pero no asi las coordenadas, con las cuales, mediante geolocalización inversa podemos concluir la dirección física.

Aplicando dichas técnicas, he llegado a la conclusión de que si bien Stanza es una valiosa herramienta, se pueden alcanzar mejores resultados con una implementación de Python. Dicho esto, ambas técnicas pueden ser complementarias y seguramente se pueden desarrollar mucho más allá de los objetivos de este trabajo, sin olvidarnos tampoco de algunas otras tecnologías como las redes neuronales, eso sí, sobre un número mucho mayor de registros, que pudieran arrojar incluso mejores resultados.

### 3.5 Visualización de los registros:

Hago uso de la librería Folium, que me permite a través de las coordenadas almacenadas en un GeoDataFrame, representar sobre un mapa la ubicación de cada queja, pudiéndose visualizar mediante un pop-up, el contenido de la misma. Es una tecnología potente que permitiría aplicando filtros sobre el GeoDataFrame (por fecha, por departamento, por topic…) una visualización general que podría servir al personal competente para, por ejemplo, la planificación de actividades. En la imagen he aplicado un filtro sobre la categoría Parques y Jardines.




## 4. Resultados y discusión:

Aunque quizás no sea uno de los fines últimos de este trabajo, porque parece que haya de centrarse más en la analítica de datos que en otros asuntos, comenzaré este apartado valorando la propia operativa del Ayuntamiento en lo que a su página web del registro se refiere (que puede afectar a la calidad de la base de datos subyacente). Y es que echo de menos, no ya una mejora del diseño y la apariencia, sino de la forma en que los datos se recogen y el contenido de los mismos.

Inicialmente, quizás sería buena idea separar, en la misma página, quejas, sugerencias y agradecimientos en su caso. No supondría un cambio drástico en el código de la página, y solamente la adición de un campo en la base de datos.

Dicho lo anterior, creo que debería facilitarse al usuario la elección del departamento competente, porque aunque algunas de las categorías presentadas, como por ejemplo “Limpieza Pública” o “Policía Local”, no dejan lugar a interpretación, hay otras como “Urbanismo”, que son difícilmente elegibles sin una información adicional o un conocimiento avanzado del usuario. Existen otras, como “Bicicletas”, muy concreta y muy genérica a la vez e “Infraestructuras: Explotación” , “Infraestructuras Conservación” y “Servicio Técnico de Infraestructuras” que dan lugar a una ambigüedad patente.

Esta clasificación, evidentemente se ha hecho desde el Ayuntamiento, conocedor de la variedad de problemas, su frecuencia y los correspondientes departamentos que los atienden, pero pienso que su origen debería partir de los conocimientos del usuario, para posteriormente, poder reclasificar a nivel interno valiéndose de herramientas como las que he expuesto.

De seguir con esta misma estructura, sería beneficioso, de alguna manera, informar de las competencias de cada departamento de una forma más amplia. Otra posibilidad, sería por ejemplo, implementar un buscador, que basado en las palabras que componen los tópicos que hemos desarrollado, ofreciese al ciudadano, los departamentos que más se ajusten a sus necesidades, mostrando una casuística de ejemplos aplicables sobre quejas anteriores.

El resto de campos presentes en el formulario parecen los adecuados y necesarios para la operativa de gestión del servicio. Por otra parte y con el fin de facilitar el trabajo de quienes tienen que gestionar los registros recibidos, se podría implementar un procedimiento de búsqueda y eliminación de reclamaciones duplicadas, por un mismo usuario o por varios cuando detallan el mismo problema. Para ello, existen librerías en Python , tales como Levenshtein o PyEnchant, con las cuales podemos determinar la similitud entre cadenas de texto.

En relación al análisis de datos, a falta de objetivos más ambiciosos (como una ampliación en el apartado de sugerencias, al que se le podría cambiar el nombre por otro más amable si se quiere fomentar la participación ciudadana), este registro se convierte en un diario de problemas cotidianos y corrientes, que tras su estudio, como hemos visto, nos permite ver los problemas estructurales latentes en la gestión de la ciudad.

En cierta manera, y no pudiéndose aportar ideas o solicitar la creación de nuevos servicios, los ciudadanos tratan de que los existentes se mantengan adecuadamente. Es por ello, que los departamentos más demandados sean Parques y Jardines, Limpieza Pública y Conservación de Infraestructuras. Se echa en falta pues, la existencia de categorías como podrían ser “Solicitud apertura de Biblioteca”, “Nuevo centro de Salud” , “Mejora de Servicios Sociales”.

En lo que se refiere a la utilización de tecnologías de Big Data, es evidente que el uso de la vía telemática para la comunicación con cualquier entidad social se ha convertido en el principal medio de intercambio de información entre agentes de toda naturaleza, como ha ocurrido en muchas otras áreas de la sociedad. La masificación de ese intercambio obligará a dichas entidades a recurrir a herramientas como las que he detallado en este trabajo para asegurar su gestión y respuesta adecuada y oportuna, modificando drásticamente la composición de su estructura departamental y sus recursos humanos y materiales.

**Citas:**

1. Ayuntamiento de Zaragoza. Servicio de Quejas y Sugerencias. 'https://www.zaragoza.es/ciudad/ticketing/verNuevaQuejaAnonima_Ticketing'

2. API Quejas y Sugerencias. Datos Abiertos de Zaragoza. Ayuntamiento de Zaragoza. 'https://www.zaragoza.es/sede/portal/datos-abiertos/open311'