Escuela Politécnica Superior

TravelSum

Diseño y desarrollo de una aplicación de generación automática de resúmenes abstractivos multigénero

Grado en Ingeniería Informática

Trabajo Fin de Grado

Autor:
Alberto Esteban García

Tutor/es:
Elena Lloret Pastor

Universitat d'Alacant
Universidad de Alicante

Julio 2016

TravelSum: diseño y desarrollo de una aplicación de generación automática de resúmenes abstractivos multigénero

La línea de investigación en generación automática de resúmenes es una línea de creciente interés dentro de las Tecnologías del Lenguaje Humano, debido a que la cantidad de información disponible aumenta a un ritmo vertiginoso. Esto hace que el ritmo de crecimiento de la información es mucho más rápido que el ritmo de desarrollo de herramientas inteligentes que ayuden a procesarla de manera eficiente. Además, con el nacimiento de la Web 2.0, se han creado una serie de nuevos géneros textuales (blogs, redes sociales, reseñas, microblogs, etc.) que han dificultado la utilización de las aplicaciones existentes en el contexto de las Tecnologías del Lenguaje Humano. Una de estas aplicaciones es la Generación Automática de Resúmenes, que despierta gran interés en la comunidad investigadora, gracias al potencial que tiene para la extracción de la información más relevante y su presentación de manera concisa y breve.

En este trabajo se propone el estudio y análisis de técnicas adecuadas para el diseño y desarrollo de un sistema de generación de resúmenes multigénero, tomando como partida distintas fuentes de datos pertenecientes a distintos géneros textuales (noticias de prensa, blogs, redes sociales, reseñas, etc). El objetivo final será combinar todos estos géneros y producir un nuevo texto coherente que capte las ideas fundamentales sobre un tema recogidas en las fuentes de datos originales.

Autor

Alberto Esteban García

Tutora

Elena Lloret Pastor

Departamento de Lenguajes y Sistemas Informáticos

Julio, 2016

Índice general

Índice de figuras

1.- Agradecimientos

He de reconocer que la realización del Trabajo de Fin de Grado me daba un poco de vértigo, así que acudí a Elena Lloret en busca de consejo. Durante casi toda mi estancia en la universidad me ha ayudado y aconsejado en varios aspectos y este trabajo hubiera sido imposible de realizar sin su ayuda. El campo en el que se enmarca el proyecto era desconocido para mí y por ello suponía un gran reto, muchas preguntas y muchas dudas en las que Elena me ayudó demostrando verdadera pasión por lo que hace y mostrándome lo interesante que puede ser la generación del leguaje natural.

Durante estos cuatro años de carrera he aprendido muchísimo, he conocido a fantásticas personas y es por ello que quiero agradecer a todas aquellas personas que me han ayudado a mejorar, permitiendo el desarrollo de este proyecto, en especial, a mi tan querido *Team Rocket* del que me llevo grandes recuerdos y la esperanza de trabajar en un futuro juntos, sois grandes.

Por supuesto no me puedo olvidar de agradecer a familiares y amigos que han estado ahí en los buenos y malos momentos, en especial a mis padres, los cuales han trabajado día a día para poder brindarme la oportunidad de estar en la Universidad.

Finalmente quiero agradecer a todos aquellos que habéis respondido el cuestionario de evaluación de la aplicación, sé que era un poco extenso, pero vuestras respuestas han sido de gran utilidad.

2.- Preámbulo

2.1.- Resumen

El presente Trabajo de Fin de Grado consiste en el análisis, desarrollo e implementación de una herramienta capaz de generar de forma automática resúmenes abstractivos multigénero.

El auge de la web 2.0 y la creación de nuevas formas de interaccionar en la web (blogs, redes sociales, reseñas, etc.) hacen que la búsqueda de cierta información implique un gran tiempo para su lectura y posterior comprensión.

Para abordar esta problemática se utilizarán técnicas del Procesamiento del Lenguaje Natural (PLN) así como diferentes técnicas informáticas que se explicarán en el cuerpo del trabajo. Además se procedió con el desarrollo de varios *crawlers* con el objetivo de la extracción de la información de las fuentes deseadas.

El primer paso del proyecto será la evaluación de las fuentes de información disponibles para su posterior elección como uso de fuentes primarias, las cuales serán de distinto tipo con el fin de que el resumen sea multigénero.

Las fuentes seleccionadas fueron *TripAdvisor* y *Twitter* y si bien la información recopilada tiene un gran valor, esta debe ser filtrada, dado que podemos encontrarnos con información muy heterogénea, en especial, la información procedente de la red social *Twitter*.

Una vez filtrada la información deberemos proceder con la creación los resúmenes utilizando técnicas de post-procesado con el objetivo de dotar de una mayor coherencia y claridad a los resúmenes generados.

Adicionalmente, con el objetivo de que la aplicación pudiera ser utilizada por cualquier persona, en cualquier lugar y sin necesidad de altos conocimientos informáticos, se decidió desarrollar y desplegar una aplicación web donde los usuarios podrían consultar los resúmenes generados por la herramienta y así poder conocer las características más relevantes para bien y para mal de cada establecimiento de la mano de las reseñas y *tweets* de los clientes.

Se han conseguido unos resultados notables dado que en el proceso de evaluación de la herramienta hubo respuestas donde los usuarios no distinguían si el proceso de realización de los resúmenes había sido obra de una persona o de una máquina, esto es el conocido test de *Turing*.

3.- Introducción

El gran volumen de datos con el que contamos hoy en día, es evidente en cualquiera de nuestras acciones en Internet, tanto si buscamos información sobre el último dispositivo móvil puesto a la venta como si la búsqueda es para encontrar una receta de cocina nos encontraremos con miles de resultados donde elegir.

Concretamente, en el dominio turístico, el auge de Internet y las Tecnologías de la Información han hecho que la forma de buscar un hotel o un restaurante cambie de manera drástica. El método tradicional era consultar en una agencia de viajes o a nuestros familiares si conocían un buen hotel o restaurante en el destino de nuestras próximas vacaciones.

En la actualidad esta tarea se ha simplificado mucho, podemos visitar varias páginas especializadas para tener toda la información a nuestro alcance.

Sin embargo, en estos momentos nos enfrentamos a un problema diferente al de antaño, y es que si antes teníamos menos información de la deseada, ahora es justo lo contrario, existe una ingente cantidad de información la cual necesita de mucho tiempo para ser analizada.

A continuación se muestra un ejemplo, se ha consultado una página especializada para conocer qué hotel podríamos escoger para hospedarnos en Alicante:

Hotel	Nº Opiniones
Meliá Alicante	2.590
NH Alicante	416
AC Alicante	579
Bonalba Alicante	922

Figura 3.1: Número de opiniones de varios hoteles en https://www.tripadvisor.es/, Alberto Esteban García

En la planificación de unas vacaciones o una estancia fuera de casa siempre queremos tomar una buena decisión en la elección del hotel donde nos hospedaremos y/o del restaurante que visitaremos. En la práctica esto conlleva la lectura de las opiniones de los antiguos clientes y como podemos apreciar el volumen de opiniones hace que no podamos leerlas todas, dado que es inviable en términos de tiempo y coste la lectura de todas y cada una de las opiniones.

En muchas ocasiones cuando visitamos un portal de reservas de hoteles o restaurantes podremos ver un breve resumen de las características de cada establecimiento. En la

mayoría de las ocasiones este resumen habrá sido realizado por fuentes cercanas al establecimiento, por lo que el resumen presentará una gran subjetividad y previsiblemente se enfatizarán las mejores características.

Por tanto la creación de una aplicación capaz de generar resúmenes que muestren tanto las bondades como las debilidades de los restaurantes es de una gran utilidad para los usuarios, los cuales en un corto espacio de tiempo conocerán los mejores y peores aspectos de cada establecimiento pudiendo así tomar la mejor decisión de donde alojarse o que restaurante visitar de una forma fácil y sencilla.

4.- Objetivos

4.1.- Generales

El objetivo principal del proyecto es el desarrollo de una aplicación de generación automática de resúmenes abstractivos multigénero que utilice como fuentes de información la web de reseñas especializada *TripAdvisor* y la red social de microblogging *Twitter*.

Mediante esta aplicación los usuarios podrán conocer de una manera rápida y sencilla lo mejor y lo peor de cada establecimiento en base a los comentarios realizados por los propios clientes, evitando así la falta de objetividad que podría tener leer un resumen realizado por el propio establecimiento.

Para la generación de los resúmenes se utilizarán herramientas de PLN que darán como resultado un conjunto de resúmenes abstractivos, esto quiere decir que parte del contenido del resumen no se encuentra en la información utilizada como fuente.

Se entregará una primera versión **terminada y disponible** para que los usuarios puedan realizar una valoración de la aplicación desarrollada y comprobar si se han cumplido los objetivos propuestos.

4.2.- Específicos

1. Analizar la información así como las herramientas necesarias para el correcto desarrollo del proyecto.

2. Planificar el desarrollo del proyecto, buscando herramientas que ayuden a la calendarización del trabajo a realizar.

3. Analizar los API de *TripAdvisor* y *Twitter* para comprobar la viabilidad de su utilización.

4. Analizar las páginas web de *TripAdvisor* y *Twitter* para descubrir posibles diferencias con los API que proporcionan.

5. Diseñar e implementar una base de datos capaz de albergar la información necesaria para el correcto desarrollo del proyecto.

6. Estudiar e implementar técnicas para la evaluación de los comentarios y *tweets* almacenados en la base de datos.

7. Estudiar diferentes herramientas para la división de los comentarios en frases así como emplear utilidades para el análisis del sentimiento de las frases.

8. Estudiar y proponer un método de identificación de información relevante que permita puntuar cada frase en base a su importancia para el futuro resumen.

9. Crear los resúmenes y almacenarlos en la base de datos.

10. Estudiar y aplicar las diferentes tecnologías web para el desarrollo de la interfaz del proyecto.

11. Analizar los API de *Flickr*, *Google Maps* y *Chart.js* para su utilización en la interfaz gráfica.

12. Investigar una posible utilidad que permita a los usuarios elegir el establecimiento en forma de sugerencia.

13. Crear un formulario de evaluación donde los usuarios puedan evaluar la aplicación desarrollada y su interfaz.

14. Desplegar la aplicación en un servidor web.

15. Permitir a cualquier usuario el acceso a la aplicación para que pueda ser utilizada y evaluada.

5.- Marco Teórico

En la siguiente sección vamos a explicar un conjunto de conceptos los cuales están relacionados con el proyecto desarrollado y muestran que el problema abordado está en auge actualmente.

5.1.- Procesamiento del Lenguaje Natural (PLN)

El campo en el que está englobado este proyecto es en el del Procesamiento del Lenguaje Natural (PLN). Para entender este concepto vamos a empezar con conocer qué es el lenguaje, el lenguaje natural y así terminar por comprender el PLN.

El término lenguaje desde el punto de vista funcional es una forma de expresar nuestros pensamientos y poder comunicarnos con los demás, desde el punto de vista formal es un conjunto de frases que se forma con combinaciones de elementos tomados de un conjunto, generalmente denominado alfabeto (*Marimón-Llorca, 2006*).

Podemos distinguir entre dos tipos de lenguajes:

- Lenguajes naturales: Español, inglés, etc.
- Lenguajes formales: Matemático, lógico, etc.

Por lo tanto el lenguaje natural es un idioma hablado o escrito que se utiliza con el fin de comunicarse.

Finalmente ya podemos definir el PLN como un campo de investigación de la Inteligencia Artificial muy amplio, el cual se centra en la creación de mecanismos eficaces que permitan la comunicación entre personas y máquinas a través del propio lenguaje natural sin la necesidad de utilizar lenguajes formales (*Moreno Boronat, 1999*).

Algunas de las aplicaciones más conocidas de este campo de investigación son:

- Generación automática de resúmenes
- Traducción automática
- Extracción de información
- Recuperación o búsqueda de información
- Minería de opiniones

Dentro del ámbito del PLN, el proyecto realizado se ha enfocado a la generación automática de resúmenes mediante el uso de tecnologías actuales como los *crawlers* o la minería de opiniones.

Realizar un resumen consiste en tratar la información para conseguir extraer de una forma breve y precisa la información esencial, esto implica conocer qué es lo más importante y expresarlo de forma abreviada. Por tanto, la generación automática de resúmenes es la producción de un resumen de manera automática, sin que nosotros nos preocupemos de qué aspectos son los más importante ni de cómo los queremos presentar (*Lloret, 2011*).

El proceso de generación de resúmenes automática plantea muchas dificultades, entre ellas:

- El proceso tiene que ser capaz de entender el texto.
- A su vez, el proceso debe poder determinar la información más relevante.
- Finalmente, el proceso debe poder generar un nuevo texto coherente con la información más relevante.

Es importante señalar que no existe un único tipo de resumen y es que según las entradas, el idioma, el medio, la finalidad o las salidas nos encontramos con una gran diversidad de resúmenes, a continuación se muestra este amplio abanico de posibilidades:

Figura 5.1: Tipos de resúmenes, Alberto Esteban García

En el caso del proyecto en desarrollo los resúmenes generados serán multigénero debido a que utilizaremos diferentes géneros textuales como información de entrada, en el caso de *TripAdvisor*, la información son reseñas y en el caso de *Twitter*, la información corresponde al género textual de los microblogs.

La salida del proceso producirá un resumen abstractivo dado que el resumen generado no estará formado solo con la información de entrada, además se utilizará información externa en forma de frases predefinidas que aportarán mayor coherencia al resumen generado.

El proceso para la generación del resumen ha comportado el uso de herramientas capaces de identificar el sentimiento de palabras, sentencias o documentos. Este hecho ha producido que nuestro resumen también se pueda catalogar como un resumen de opinión, dado que este aportará una visión positiva, negativa o neutral sobre el establecimiento consultado.

5.2.- Big Data

Otro tema al que está muy ligado el proyecto es el Big Data. Actualmente contamos con muchísima información, la cual nos es imposible procesar de manera tradicional ya que además del problema del volumen de información, nos enfrentamos al problema de la variedad de los datos así como las formas de representación de estos.

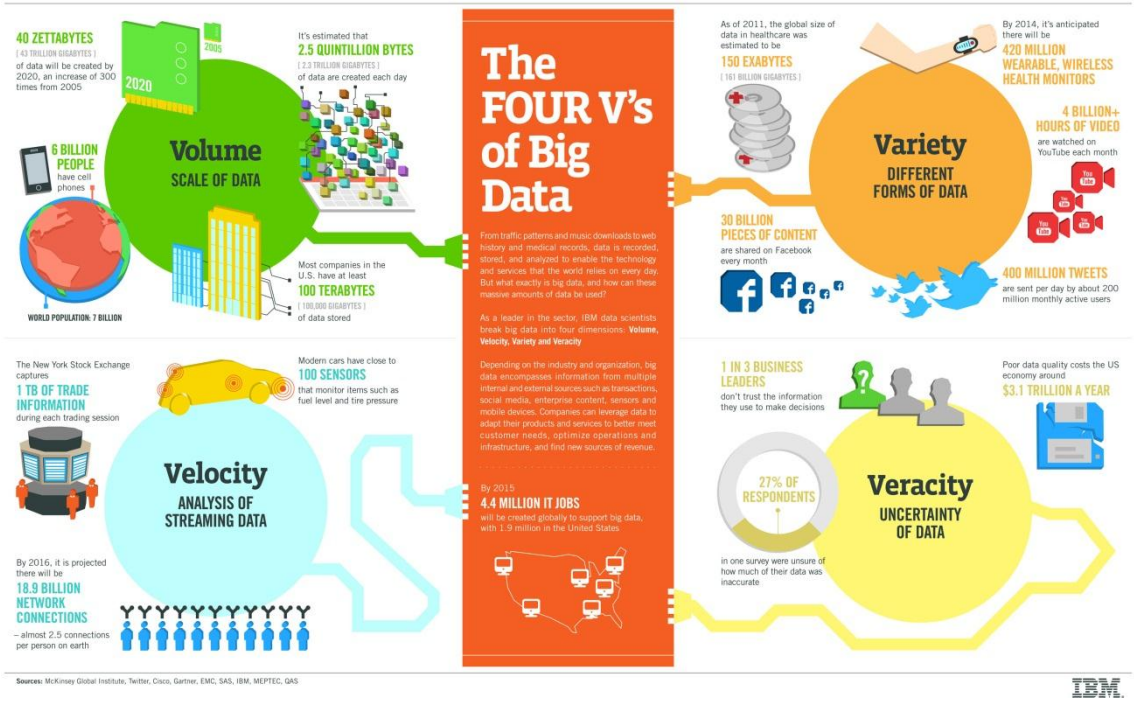


Figura 5.2: Las 4 V del Big Data (*The Four V's of Big Data*), http://www.ibmbigdatahub.com/

La imagen anterior corresponde a las 4 V del Big Data (*IBM, 2013*):

- **Volumen**: El volumen de datos actual es inimaginable, se estima que se generan alrededor de dos trillones de gigabytes cada día.

- **Variedad**: La diversidad de información es una de las características en la actualidad: *Smart cities*, *wearables*, redes sociales, etc.

- **Veracidad**: La calidad de la información en ciertos casos puede no ser la deseada.

- **Velocidad**: La velocidad de generación de información es mayor a la de tratamiento

Por tanto debemos proveer de mecanismos que automaticen y aporten valor a esta ingente cantidad de información que generamos los seres humanos (industrias, administraciones públicas, etc.)

La información por sí sola tiene poco valor, pero si somos capaces de desarrollar un proceso que utilice esta información y aporte valor, la información y el conocimiento resultante tendrán un gran valor para empresas, instituciones, etc.

Según el tipo de datos podemos realizar una clasificación de Big Data:

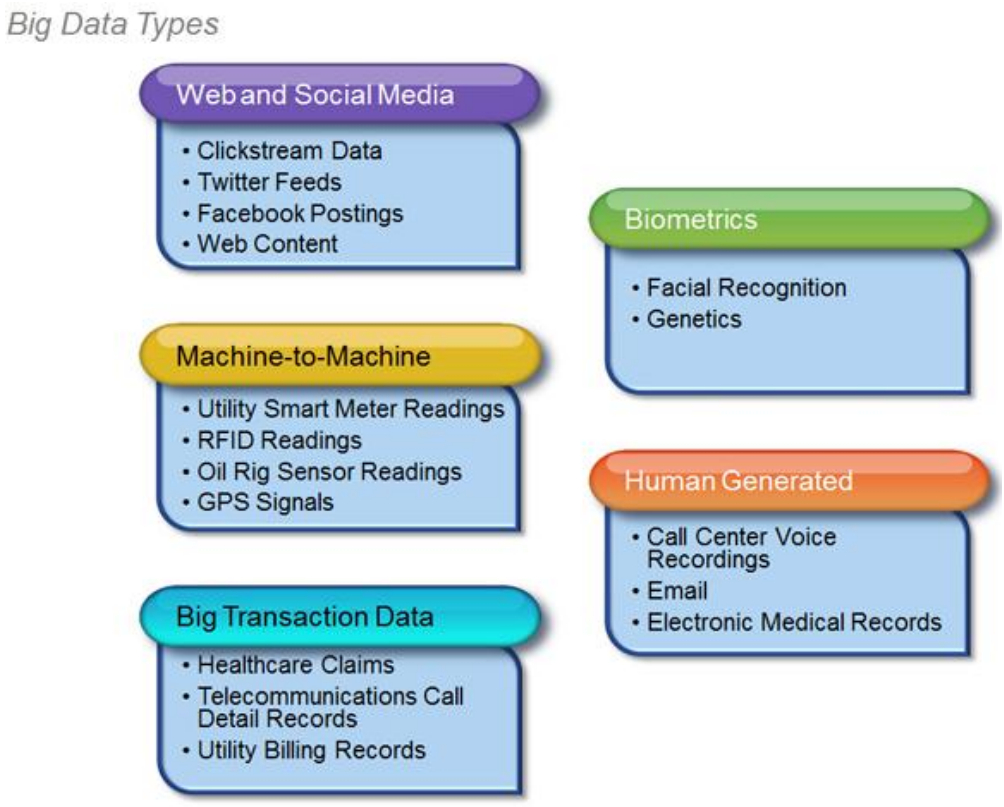


Figura 5.3: Tipos de Big Data, https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/

En nuestro caso hemos utilizado contenido web e información obtenida de la red social *Twitter* y la página especializada *TripAdvisor* por lo que el tipo de Big Data utilizado es el de *Web and Social Media.*

Los otros tipos de Big Data se relacionan con:

- **Biometrics**: Información biométrica de las personas como huellas dactilares, escáneres de retina, etc.

- **Machine-to-Machine**: Comunicación que permite enviar datos desde un sensor a una aplicación con el objetivo de mostrar información significativa

- **Big Data Transaction**: Registros de facturación, registros de llamadas, etc.

- **Human Generated**: Información generada por nosotros al establecer una llamada, enviar correos electrónicos, notas de voz, etc.

5.3.- Web 2.0

Además del PLN y el Big Data el proyecto también aborda el tema de la web 2.0, esto es el uso de webs que facilitan la compartición de información mediante la interoperabilidad y el diseño centrado en el usuario (*Wikipedia, 2016*).

Figura 5.4: Imagen Web 2.0, http://web2-0tye.blogspot.com.es/

La evolución más importante desde la web inicial es que los usuarios dejan de ser usuarios pasivos para adoptar un papel activo en el que muestran sus opiniones, experiencias, etc.

Los ejemplos más claros de esta evolución son:

- **Incremento del número de blogs**: Muchas personas participan activamente mediante la creación de un blog publicando sus opiniones sobre moda, tecnología, etc.

- **Auge de las redes sociales**: La irrupción de *Twitter*, *Facebook*, *Linkedin* han cambiado la forma de relacionarnos e incluso de buscar trabajo.

- **Facilidad de creación de webs**: Para crear una página web ya no hace falta tener unos sólidos conocimientos de diseño web si queremos tener una página web sencilla.

Por lo tanto la web 2.0 genera un volumen muy grande de información útil que debe ser tratada para aportar un valor único y diferenciador con el que el usuario final se sienta cómodo. Por ejemplo, leer cientos de opiniones sobre un producto puede hacer que al final no tengamos una idea clara de lo mejor y lo peor o que hayamos invertido una cantidad de tiempo muy grande.

Para que nos hagamos una idea estas son las estadísticas de generación de contenido de las fuentes de información utilizadas en el proyecto, es decir *Twitter* y *TripAdvisor*:

- Número de *tweets* generados cada día: 339.984.000 (*Onesecond, 2016*)
- Número de usuarios en *Twitter*: 310.000.000 (*Twitter, 2016*)
- Número de nuevas reseñas en *TripAdvisor* por día: 332.000 (*TripAdvisor, 2016*)
- Número de usuarios en *TripAdvisor*: 103.000.000 (*TripAdvisor, 2016*)

6.- Cuerpo del Trabajo

En esta sección se detallarán las fases planificadas para el desarrollo del proyecto.

6.1.- Fase de planificación

En esta etapa inicial vamos a definir las fases de las que va a constar el proyecto con el objetivo de acotar el trabajo que se ha de realizar y hacer una estimación del tiempo que nos va a llevar cada fase.

Es muy importante intentar realizar una estimación de tiempo para cumplir con el plazo de entrega previsto, el cual es el mes de junio de 2016.

La primera etapa que identificamos es la fase de extracción de datos, en ella investigaremos que fuentes podemos utilizar, con que herramientas contamos y cómo vamos a almacenar los datos una vez extraídos.

A continuación plantearemos un proceso de filtrado de esta información. Según la fuente de información nos podemos encontrar con información heterogénea la cual no tenga relación con el campo que deseamos.

Una vez tengamos los datos filtrados deberemos pensar cómo vamos a crear los resúmenes con los datos que tenemos, cómo podemos crear conocimiento y poner en valor ese gran volumen de datos.

Una vez tengamos los resúmenes creados deberemos analizar el resultado e incorporar técnicas de post-procesado con el objetivo de dotar al resumen de una mayor expresividad y naturalidad que permita al usuario leer un resumen coherente.

Finalmente cuando tengamos todos los resúmenes creados, lo que deberemos hacer es implementar una interfaz centrada en el usuario y en la temática para que el usuario final tenga todos los datos de interés disponibles.

Para juzgar si hemos logrado un buen resultado, la interfaz se desplegará en un servidor para hacer accesible la aplicación a cualquier persona. De este modo con un sencillo formulario las personas podrán valorar la aplicación desarrollada según diferentes aspectos y con ello ver que se puede mejorar.

Por lo tanto las fases que se han desarrollado son las siguientes:

- Fase de extracción de datos
- Fase de filtrado de datos
- Fase de elaboración de resúmenes
- Fase de post-procesado

- Fase de elaboración de la interfaz
- Fase de despliegue y evaluación

En la etapa de planificación comenzamos con la definición de los *mockups*[1] de la interfaz gráfica de la aplicación para enfatizar que la aplicación podría ser accesible por cualquier persona.

A continuación se muestran los *mockups* iniciales propuestos:

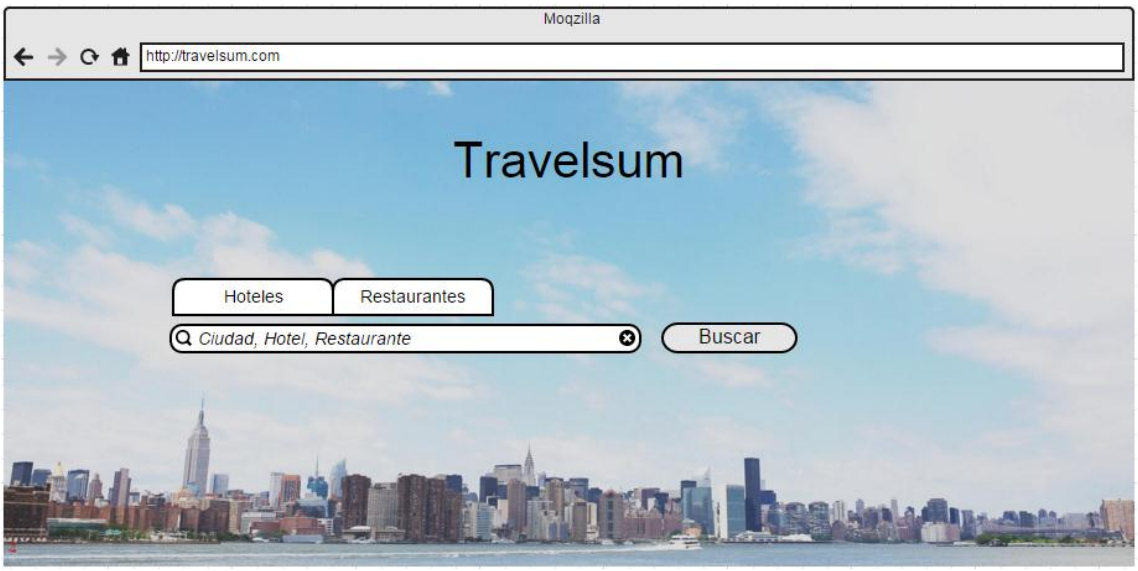


Figura 6.1: *Mockup* página de inicio, Alberto Esteban García

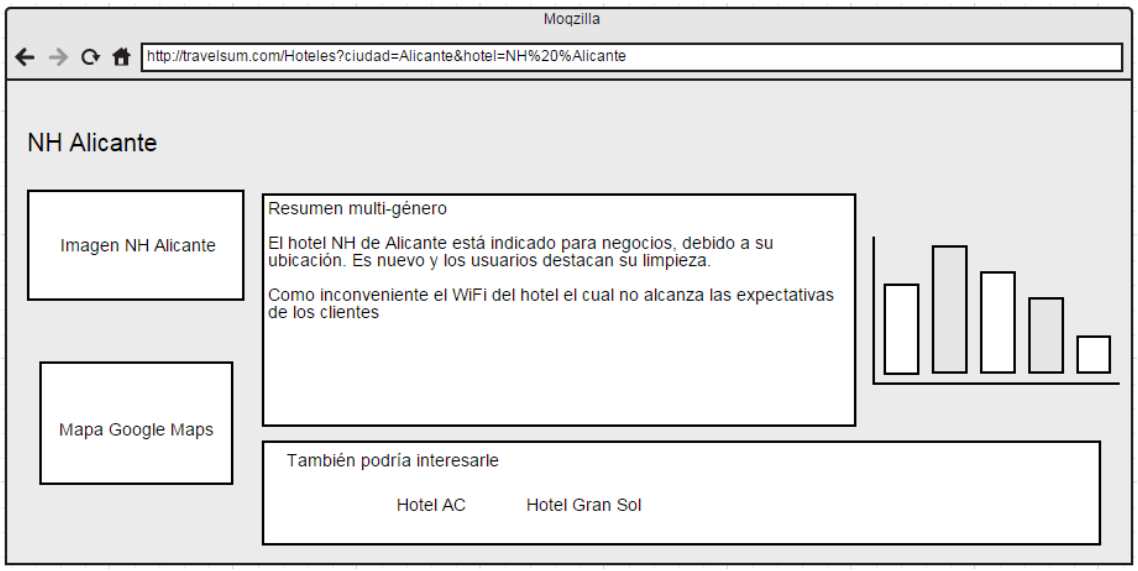


Figura 6.2: *Mockup* página del hotel, Alberto Esteban García

[1] *Mockup*: Prototipo que intenta mostrar el resultado final del producto o servicio a desarrollar.

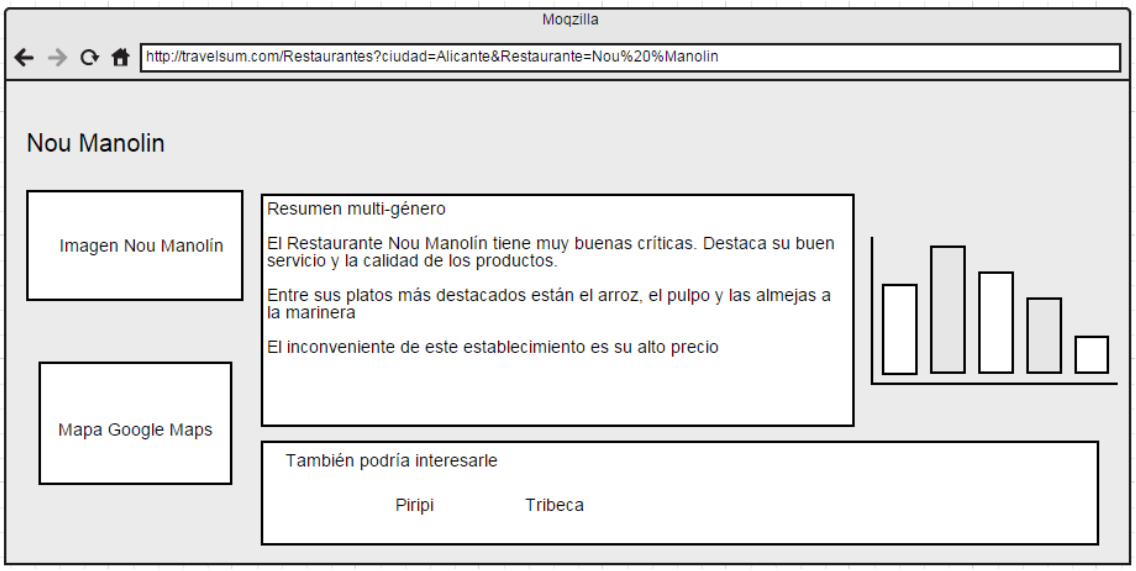


Figura 6.3: *Mockup* página del restaurante, Alberto Esteban García

6.2.- Fase de extracción de datos

El objetivo de esta fase es investigar qué fuentes de información podemos utilizar en nuestro proyecto para la generación de resúmenes y que permitan al cliente tener una opinión clara sobre el establecimiento que está consultando.

Puesto que queremos aplicar el proyecto al dominio turístico, buscamos páginas especializadas de turismo, foros, redes sociales, periódicos, etc.

Finalmente elegimos dos fuentes de información: la página especializada *TripAdvisor* y la red social *Twitter*, debido a dos razones:

- El gran volumen de datos que podíamos obtener
- El valor de la información obtenida para el dominio turístico, sector para el que se orienta la aplicación

6.2.1.- Sistema Gestor de Base de Datos (SGBD)

Dado que ya tenemos las fuentes de información se identificó el requisito de crear una base de datos para almacenar la información, para ello elegimos el Sistema Gestor de Base de Datos (SGBD) MySQL por las siguientes razones:

- Sencillo
- Intuitivo
- Gratuito
- Experiencia en el uso del SGBD escogido mediante un proyecto realizado en los estudios del Grado

El esquema de base de datos utilizado en el proyecto es el siguiente:

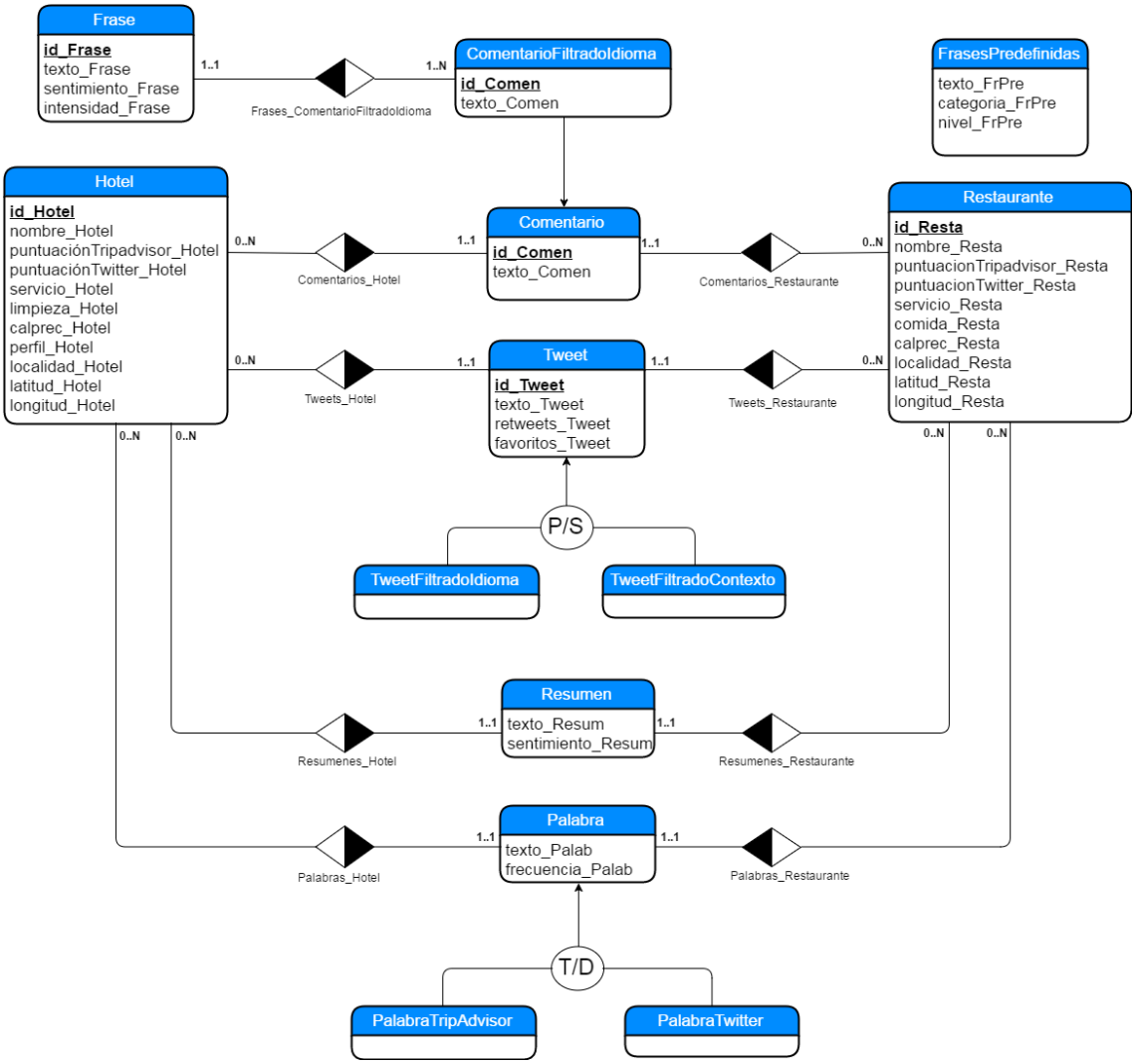


Figura 6.4: Esquema de bases de datos, Alberto Esteban García

A continuación se muestra la finalidad de cada tabla:

- ***Hotel***: Almacenará la información de los hoteles.
- ***Restaurante***: Guardará la información de los restaurantes
- ***Comentario***: Será la tabla encargada de almacenar la información de los comentarios originales extraídos de *TripAdvisor*
- ***Tweet***: Se encargará de almacenar la información de los *tweets*
- ***Resumen***: Almacenará la información de los resúmenes

- ***ComentarioFiltradoIdioma***: Guardará la información de los comentarios filtrados por idioma

- ***Frase***: Se encargará de almacenar las frases, las cuales son el resultado de la división de los comentarios filtrados por idioma

- ***PalabraTripAdvisor***: Se encargará de almacenar los cien lemas más relevantes de cada establecimiento utilizando como fuente de información las frases de *TripAdvisor*

- ***Tweet***: Será la tabla encargada de almacenar la información de los comentarios originales extraídos de *Twitter*

- ***TweetFiltradoPorIdioma***: Contendrá los *tweets* que han pasado el filtro de idioma partiendo como fuente de información la tabla *Tweet*

- ***TweetFiltradoPorContexto***: Contiene los *tweets* que han pasado el filtro de contexto partiendo como fuente de información la tabla *TweetFiltradoPorIdioma*

- ***PalabrasTwitter***: Se encargará de almacenar los cien lemas más relevantes de cada establecimiento utilizando como fuente de información los *tweets*

- ***FrasesPredefinidas***: Almacenará las frases predefinidas para la elaboración de los resúmenes

6.2.2.- Fuente de datos de reseñas: TripAdvisor

En este momento que tenemos decididas las fuentes de información, es hora de ver cómo obtener los datos. En primer lugar comenzamos con *TripAdvisor*, y comprobamos que existe un API[2] el cual permite la consulta de los datos de los establecimientos y las opiniones de los usuarios.

Sin embargo las condiciones del uso del API dejan muy claro que no es un API abierto, se ha de solicitar su utilización explicando el uso que se va a hacer y solo se permitirá

[2] https://developer-tripadvisor.com/content-api/request-api-access/

su utilización en aplicaciones *B2C*[3], en ningún caso se podrá utilizar para análisis de datos o investigaciones académicas.

Por tanto para la recopilación de datos de *TripAdvisor* se decidió desarrollar un *crawler*[4] específico. El *crawler* ha sido implementado en *Java* y permite almacenar de una forma rápida y sencilla los hoteles, restaurantes y comentarios.

El modo de utilización se ha hecho de una forma muy simple, se introduce una ciudad y se elige si se quiere recopilar la información de hoteles o restaurantes, tras ello automáticamente el *crawler* almacenará los treinta primeros hoteles o restaurantes y un máximo de quinientas reseñas por cada establecimiento en la base de datos explicada anteriormente (en ocasiones no existen tantas reseñas sobre un establecimiento).

Por cada hotel el *crawler* almacena en la base de datos:

- Nombre del hotel
- Nota del hotel: Esta nota no ha sido extraída directamente desde *TripAdvisor,* se realiza basándose en la puntuación otorgada por los clientes con las siguientes equivalencias

Puntuación	Correspondencia en base 10
Excelente	10
Muy bueno	8
Normal	6
Malo	4
Pésimo	2

Figura 6.5: Tabla de equivalencias entre *TripAdvisor* y la aplicación, Alberto Esteban García

$$Nota_{TripAdvisor} = \frac{10 \cdot n_{ex} + 8 \cdot n_{mu} + 6 \cdot n_{no} + 4 \cdot n_{ma} + 2 \cdot n_{pe}}{n_{ex} + n_{mu} + n_{no} + n_{ma} + n_{pe}}$$

Siendo:

- n_{ex} : Número de opiniones excelentes

[3] B2C: Abreviatura de *Business-to-Consumer* la cual se refiere a la estrategia que desarrollan las empresas comerciales para llegar directamente al cliente o consumidor final.

[4] *Crawler*: Programa que inspecciona las páginas del World Wide Web de forma metódica y automatizada.

- n_{mu} : Número de opiniones muy buenas
- n_{no} : Número de opiniones normales
- n_{ma} : Número de opiniones malas
- n_{pe} : Número de opiniones pésimas

- Valoración del servicio del hotel
- Valoración de la limpieza del hotel
- Valoración de la calidad-precio del hotel
- Perfil más aconsejado para visitar el hotel
  - En solitario, en pareja, familiar o de negocios
- Localidad del hotel
- Coordenadas del hotel

En el caso de los restaurantes la información almacenada en la base de datos es la siguiente:

- Nombre del restaurante
- Nota del restaurante: El procedimiento es análogo al visto para los hoteles, el *crawler* rastrea el número de opiniones excelentes, muy buenas, normales, malas y pésimas y mediante la siguiente fórmula asigna la puntuación:

$$Nota_{TripAdvisor} = \frac{10 \cdot n_{ex} + 8 \cdot n_{mu} + 6 \cdot n_{no} + 4 \cdot n_{ma} + 2 \cdot n_{pe}}{n_{ex} + n_{mu} + n_{no} + n_{ma} + n_{pe}}$$

- Valoración del servicio del restaurante
- Valoración de la comida del restaurante
- Valoración de la calidad-precio del restaurante
- Localidad del restaurante
- Coordenadas del restaurante

Tras la extracción de datos de *TripAdvisor*, ya tenemos almacenado un gran volumen de datos (*ver anexo*) de los que partir por lo que ahora podemos comenzar con la recopilación de datos de *Twitter*.

6.2.3.- Fuente de datos de microblogs: Twitter

En este caso *Twitter* tiene un API abierto, al contrario que *TripAdvisor*, como ya hemos comentado anteriormente.

Por lo tanto la extracción de la información se realizará mediante el uso del API a través del lenguaje de programación *Javascript*.

Dado que esta red social de microblogging es mucho menos especializada que la web de *TripAdvisor* en este caso vamos a almacenar en torno a mil quinientos *tweets* por cada establecimiento, dado que el posterior filtrado hará que muchos de los *tweets* no sean válidos.

Para ejemplificar este hecho a continuación se muestra un *tweet* extraído para el hotel *Meliá Alicante*:

"Simulacro contra incendios #lovemyjob #MhiSpain @meliaalicante"

Como podemos apreciar, el *tweet* no aporta ninguna información relevante sobre las características del hotel.

Para la búsqueda de la información de cada establecimiento partiremos de los nombres de los establecimientos que tenemos de *TripAdvisor,* se tendrá en cuenta si el nombre del establecimiento contiene la ciudad origen, por ejemplo, en el caso del hotel "*Meliá Alicante*" se hará la búsqueda del término "*Meliá Alicante*", pero en el caso de "Piripi" introduciremos su localidad para acotar la búsqueda y evitar información no concerniente al tema tratado, por lo que buscaremos "Piripi Alicante".

En definitiva el término buscado en *Twitter* será de la siguiente forma:

- Si el nombre del establecimiento contiene la localidad, buscaremos *"Nombre_Establecimiento"*
- Si por el contrario el nombre del establecimiento no contiene la localidad, buscaremos *"Nombre_Establecimiento"* + localidad

Una vez definidas las formas de búsqueda, lanzamos el proceso de recopilación de datos, pero nos topamos con un problema importante.

El API de *Twitter* tiene una restricción de solo recuperar los *tweets* de hasta dos semanas antes de la búsqueda. Por lo que los *tweets* más antiguos a esta fecha no son devueltos por el API.

Como alternativa se utilizó *Twitter4J*[5]*,* librería *Java* para el uso del API de *Twitter* desde *Java*, es sencilla e intuitiva pero al utilizar el API de *Twitter* el problema persiste, no podemos recuperar *tweets* con más de dos semanas de antigüedad.

En relación a los detalles técnicos de implementación y puesto que la utilización del API de *Twitter* no cumplía nuestros requisitos, se decidió realizar un *crawler* para recuperar todos los *tweets* existentes.

[5] http://twitter4j.org/en/index.html

La principal diferencia con el crawler de *TripAdvisor*, es que la página de *Twitter* tiene un estilo de página infinita donde al llegar al final, carga los *tweets* más antiguos a los ya mostrados.

Finalmente se utilizó *Selenium Web Driver*[6], una aplicación que permite emular al navegador para poder realizar las acciones que haría un usuario convencional al visitar la página de resultados de los términos buscados.

Mediante este método pudimos almacenar toda la información de los *tweets* que queríamos, la cual es:

- Texto del tweet
- Número de *retweets* del tweet
- Número de favoritos del tweet

6.2.4.- Resultados de la extracción de datos

Para este TFG, se decidió realizar la recopilación de datos sobre seis ciudades que incluyeran tanto ciudades nacionales como internacionales, las seleccionadas fueron:

- Alicante
- Valencia
- Madrid
- Barcelona
- Londres
- Roma

A continuación se muestra muy brevemente algunos datos de interés:

Número de Hoteles	180
Número de Restaurantes	180
Número de reseñas *TripAdvisor*	91.505
Número de *Tweets*	78.713

Figura 6.6: Estadísticas de la información extraída, Alberto Esteban García

Los resultados de la recopilación de información de forma desagregada de las dos fuentes de información se pueden ver a continuación:

[6] http://www.seleniumhq.org/projects/webdriver/

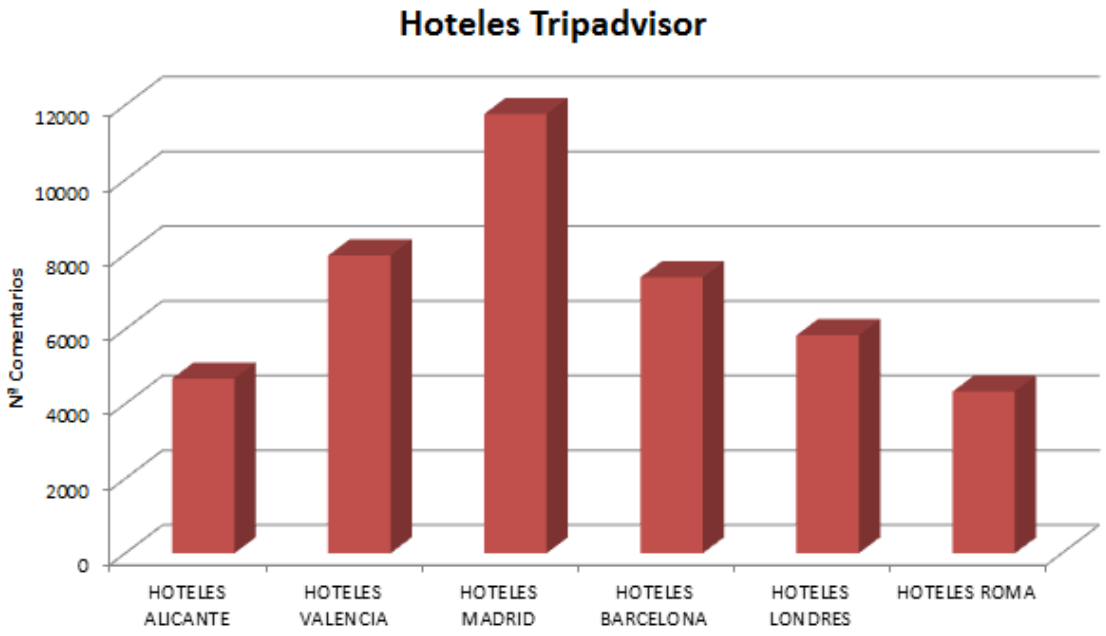


Figura 6.7: Gráfico sobre el número de reseñas de hoteles, Alberto Esteban García

Los hoteles de Madrid son los que más comentarios tienen en *TripAdvisor* con diferencia, seguidos de los de Valencia y Barcelona.

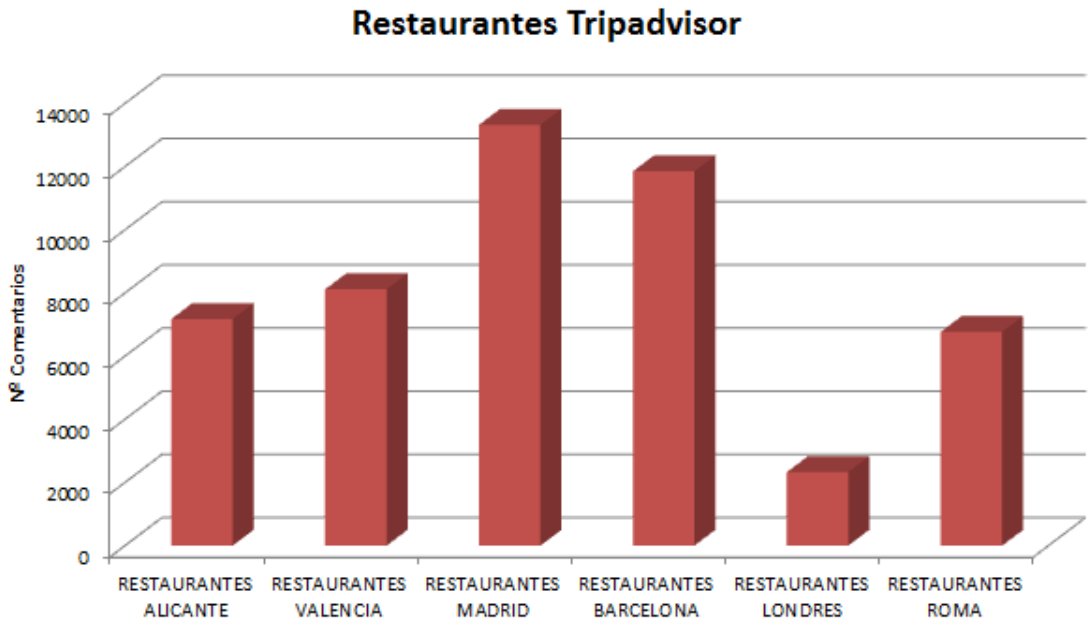


Figura 6.8: Gráfico sobre el número de reseñas de restaurantes, Alberto Esteban García

En el caso de los restaurantes, Madrid sigue siendo el que más comentarios tiene, seguido muy de cerca por Barcelona. Los comentarios de los restaurantes londinenses son bastante escasos como podemos apreciar.

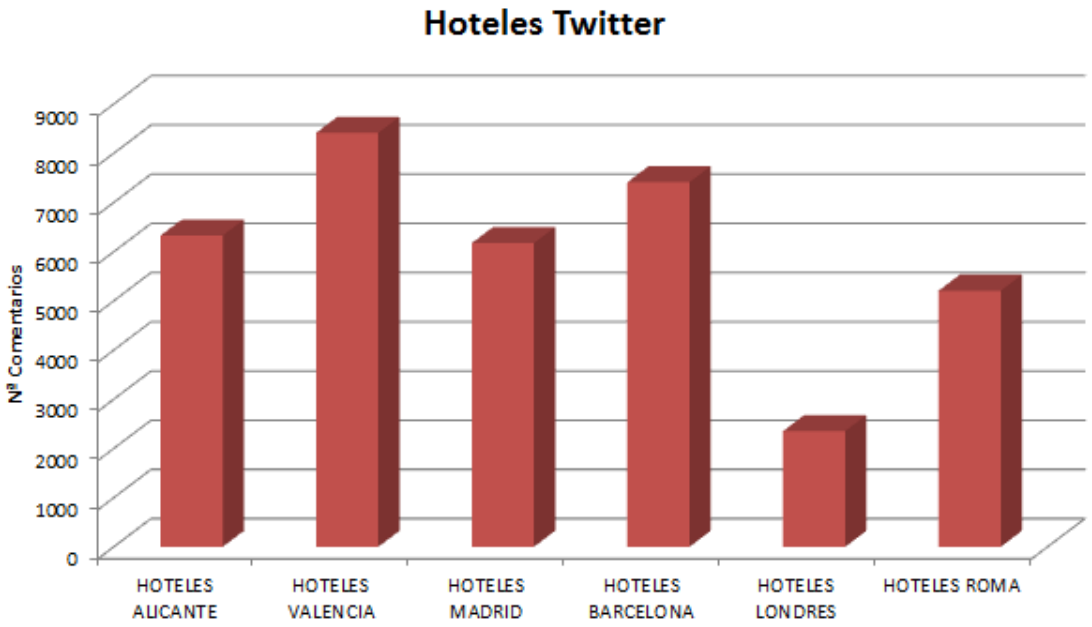


Figura 6.9: Gráfico sobre el número de *tweets* de hoteles, Alberto Esteban García

Si hablamos de *Twitter*, Valencia reina en el apartado de los hoteles seguido de cerca por Barcelona.

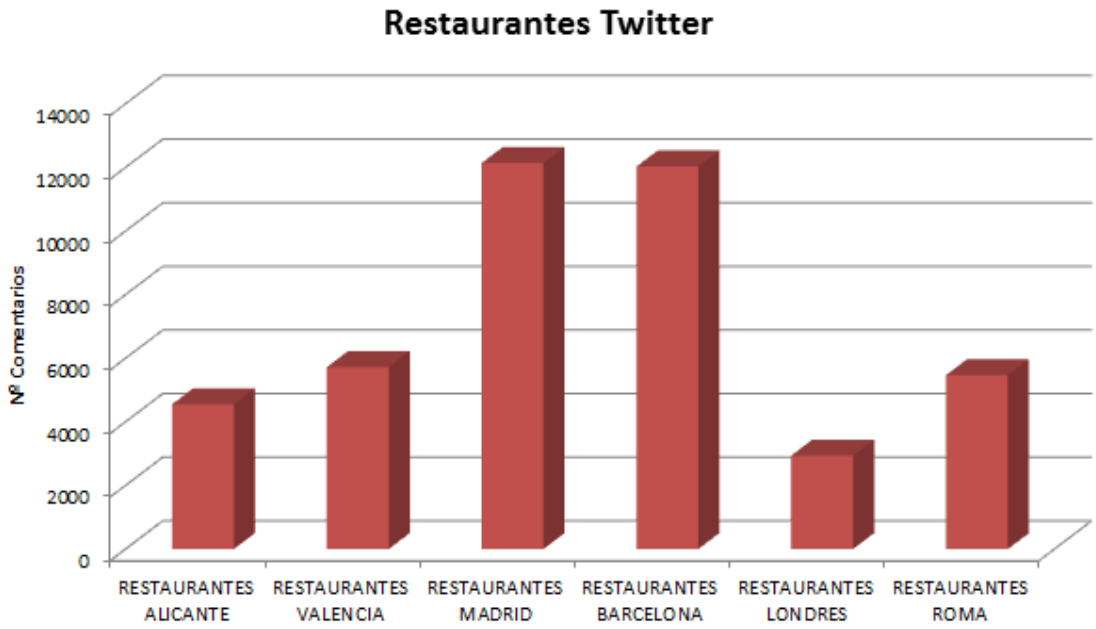


Figura 6.10: Gráfico sobre el número de *tweets* de restaurantes, Alberto Esteban García

En el caso de los *tweets* recabados sobre restaurantes, Madrid y Barcelona lideran con diferencia sobre los demás.

A modo de resumen los gráficos siguientes muestran el peso que tiene cada ciudad sobre los datos recopilados:

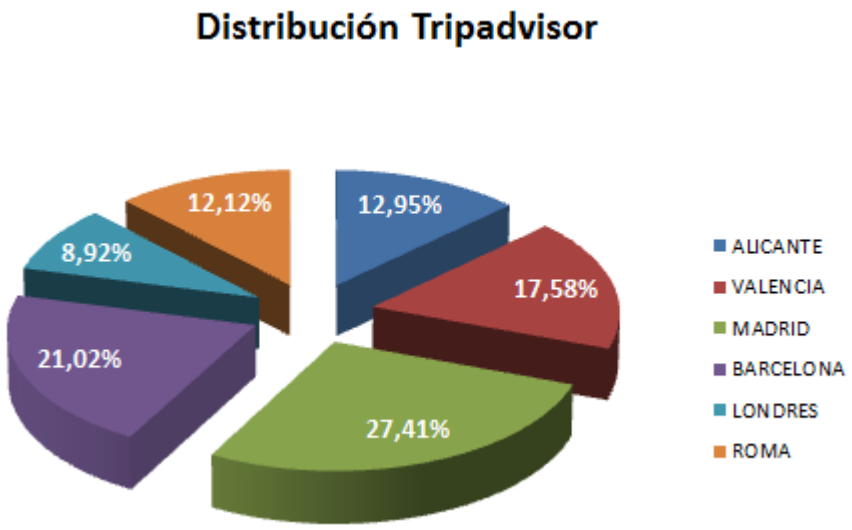


Figura 6.11: Gráfico sobre la distribución de reseñas por ciudad, Alberto Esteban García

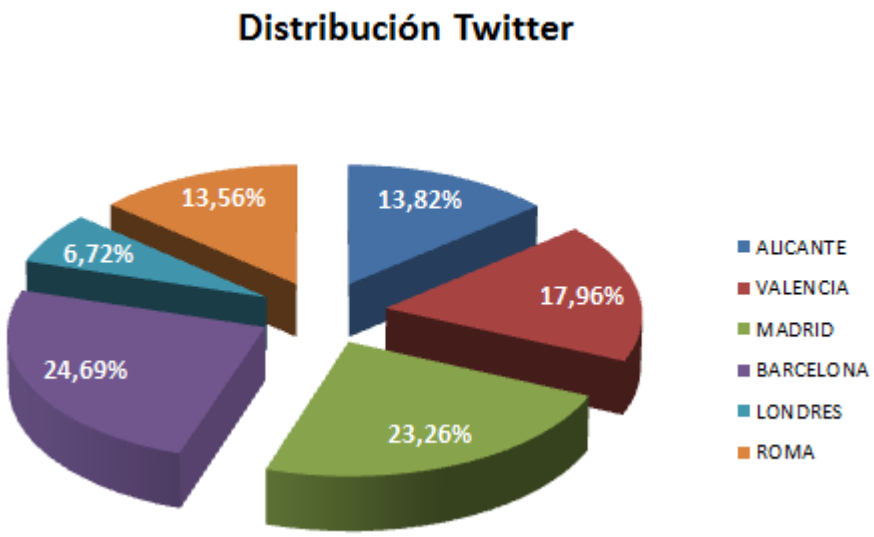


Figura 6.12: Gráfico sobre la distribución de *tweets* por ciudad, Alberto Esteban García

Podemos apreciar que el mayor número de datos es de Madrid y Barcelona, seguidas por Valencia y Alicante quedando las ciudades extranjeras al final.

Esto es lógico dado que en la búsqueda de *Twitter* hemos buscado el nombre del establecimiento junto con el tipo de establecimiento (hotel o restaurante) en castellano. Mediante esta técnica acotamos los *tweets,* los cuales en su mayoría serán en castellano lo cual nos interesa mucho, dado que queremos tener solo los datos en castellano.

6.3.- Fase de filtrado de datos

El propósito de esta fase es la de filtrar los datos obtenidos en la etapa anterior para ser utilizados posteriormente en la generación de los resúmenes.

En este punto ya tenemos toda la información guardada en la base de datos, esta información es muy valiosa por lo que procedemos a realizar una copia de seguridad utilizando la utilidad de *mysqldump*:

```
mysqldump –uusername –ppassword –hhost databaseName >
nombreArchivo.sql
```

Ahora debemos especificar las heurísticas necesarias para realizar un buen filtrado de la información.

Se ha de tener en cuenta que el tipo de información recopilado es diferente en cada fuente de información. En el caso de *TripAdvisor* contamos con reseñas las cuales están muy enfocadas a realizar un comentario de opinión sobre el establecimiento, mientras que en *Twitter* la información extraída puede ser de otra índole.

6.3.1.- Filtrado de la información obtenida de TripAdvisor

En el caso de *TripAdvisor* el filtrado se ha realizado para eliminar los comentarios que no están en castellano, dado que utilizar un software de traducción podría afectar a los resultados posteriores.

Para realizar este filtrado hemos utilizado una herramienta de detección automática de idioma del GPLSI [7]de la Universidad de Alicante.

La herramienta es muy sencilla de utilizar, mediante una petición de tipo POST[8] le proporcionamos el texto a la herramienta y esta nos devuelve el código del idioma que del texto. En nuestro caso, almacenaremos los comentarios que tengan como código de respuesta ***"es" (español)***.

[7] http://gplsi.dlsi.ua.es/services/pln/doc/index.html#!/lang/detect
[8] http://gplsi.dlsi.ua.es:80/services/pln/rest/v1/lang/detect

A continuación se muestra un ejemplo utilizando la interfaz gráfica de la web:

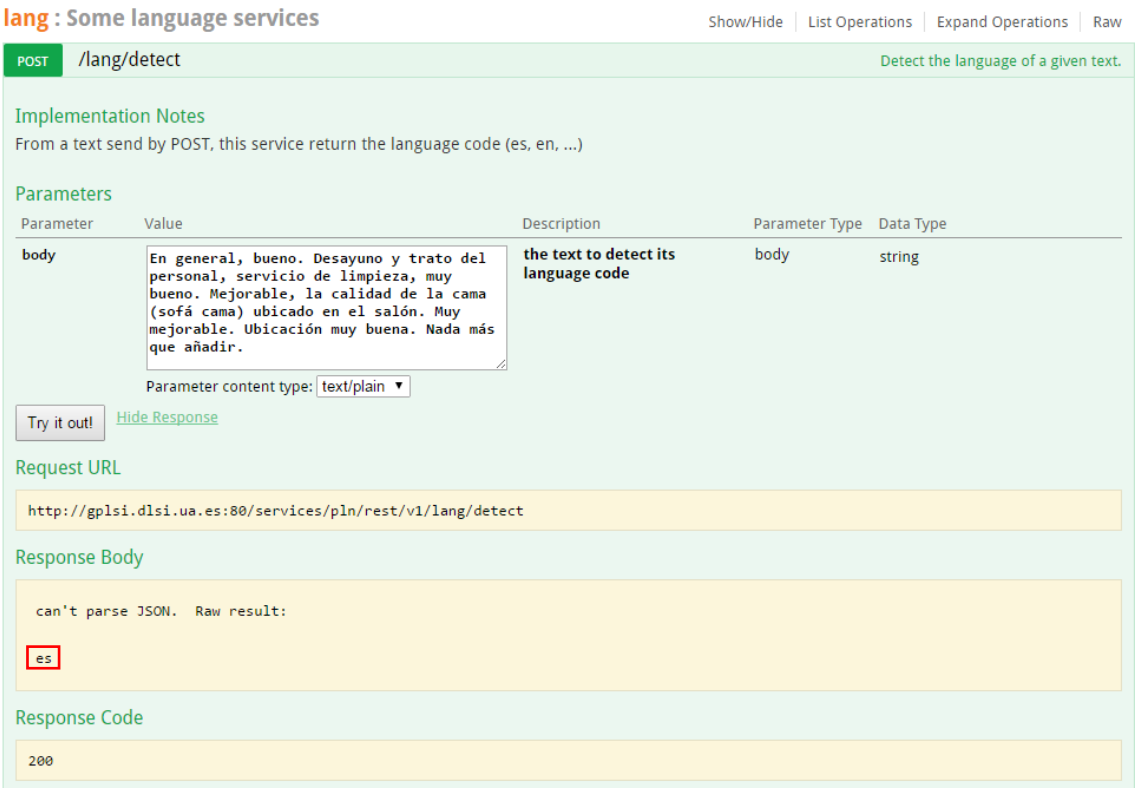


Figura 6.13: Interfaz gráfica de la herramienta de detección automática de idioma, Alberto Esteban García

A continuación se muestran los resultados del filtrado por idioma de los comentarios:

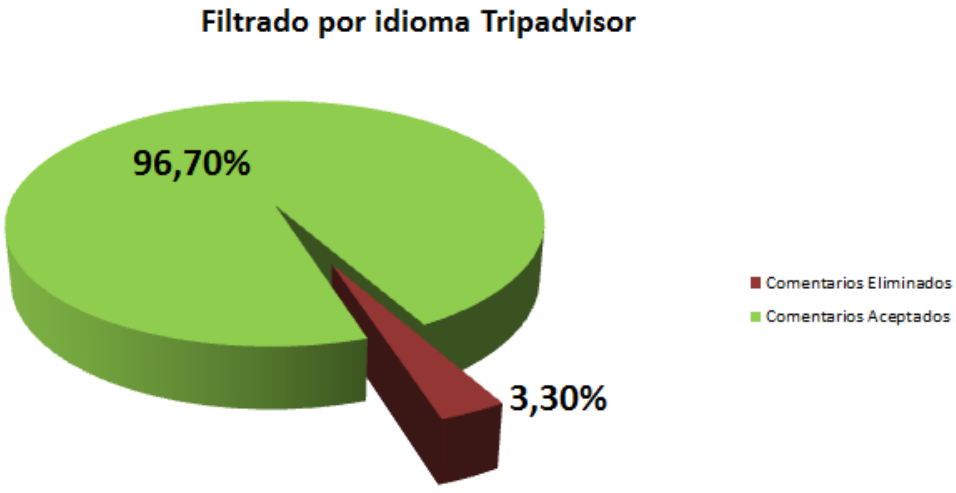


Figura 6.14: Gráfico sobre el filtrado de idioma de las reseñas, Alberto Esteban García

Como podemos apreciar la inmensa mayoría de los comentarios estaban en castellano debido al mínimo impacto que el filtrado ha causado sobre las reseñas originales.

6.3.2.- División comentarios de TripAdvisor

Una vez tenemos los comentarios filtrados por idioma el paso siguiente ha sido la división de los comentarios en frases.

Para ello se han probado dos herramientas

- La clase *Java*, *BreakIterator.class*
- Analizador sintáctico desarrollado por la Universidad de Stanford[9]

Tras realizar varias pruebas con ambas herramientas se optó por la utilización del software de Stanford dado que es más preciso en la división de las frases.

Como ejemplo, la clase *BreakIterator* interpretaba el siguiente fragmento como una única sentencia mientras que el software de Stanford distinguía dos sentencias:

El hotel me pareció maravilloso. está muy bien ubicado

Si observamos el ejemplo, el único defecto apreciable a primera vista es que la letra posterior al punto no comienza en mayúscula, el software de Stanford determina que son dos frases pero la clase de *Java* lo interpreta como una única sentencia.

La tarea de la división automática de sentencias es muy complicada dado que las reseñas han sido escritas por todo tipo de usuarios por lo que la informalidad o las faltas de ortografía son muy comunes, lo que afecta a un proceso automatizado.

6.3.3.- Recuento lemas más relevantes de TripAdvisor

Una vez divididos los comentarios en frases, el paso siguiente que debemos hacer es realizar un proceso para almacenar los términos más comentados de cada establecimiento, que sería equiparable a la detección de temas tratados en cada comentario.

En este proceso se involucrarán dos técnicas de PLN, estas son:

- Eliminación de *stopwords*
- Lematización de las palabras

Las *stopwords* son palabras sin significado las cuales enlazan las palabras con el objetivo de crear un texto coherente, es decir, conjunciones, preposiciones, artículos, pronombres, etc. (*Wikipedia, 2016*).

[9] http://stanfordnlp.github.io/CoreNLP/

Las *stopwords* son muy frecuentes en cualquier texto, pero estas no aportan ningún significado, por lo que será necesario eliminarlas con el fin de obtener las palabras más frecuentes y que aporten valor al resumen.

Por otro lado, la lematización de las palabras consiste en conocer el lema de cada palabra. Como ejemplo, las palabras "habitación" y "habitaciones" no son la misma palabra pero si tienen el mismo lema, "habitación".

El software utilizado para realizar la lematización ha sido *TreeTagger*[10].

Es un software muy sencillo de utilizar donde como entrada le pasamos un fichero de texto con las palabras que queremos lematizar, y el proceso generará otro fichero de texto donde por cada palabra en el fichero original, obtendremos:

- La categoría de la palabra (*Post-Tagger*) indicando si es un sustantivo, un verbo, un adjetivo, etc.
- El lema correspondiente a la palabra

Así que de esta forma hemos almacenado en la base de datos los cien lemas más repetidos del tipo sustantivo de cada establecimiento (la etiqueta asignada para un sustantivo es "NC").

6.3.4.- Filtrado de la información obtenida de Twitter

En el caso de *Twitter* el filtrado debe ser más exhaustivo dada la heterogeneidad de la información por ello se definieron varias las reglas de filtrado:

- Al igual que hicimos en el filtrado de las reseñas de *TripAdvisor,* los *tweets* escritos en un idioma distinto del castellano serán eliminados.

- Las menciones (p. ej: @perfil) y los enlaces (p. ej. http://www.melialicante.com) contenidos en cada *tweet* serán eliminados.

- Se filtrarán los *tweets* por contexto, esto quiere decir, que deberemos buscar la manera de eliminar los *tweets* que no estén relacionados con opiniones o comentarios sobre los establecimientos.

El filtrado por idioma es idéntico al explicado anteriormente, ha sido realizado mediante la herramienta del GPLSI[11].

[10] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[11] http://gplsi.dlsi.ua.es/services/pln/doc/index.html#!/lang/detect

El resultado del filtrado por idioma de los *tweets* es el siguiente:

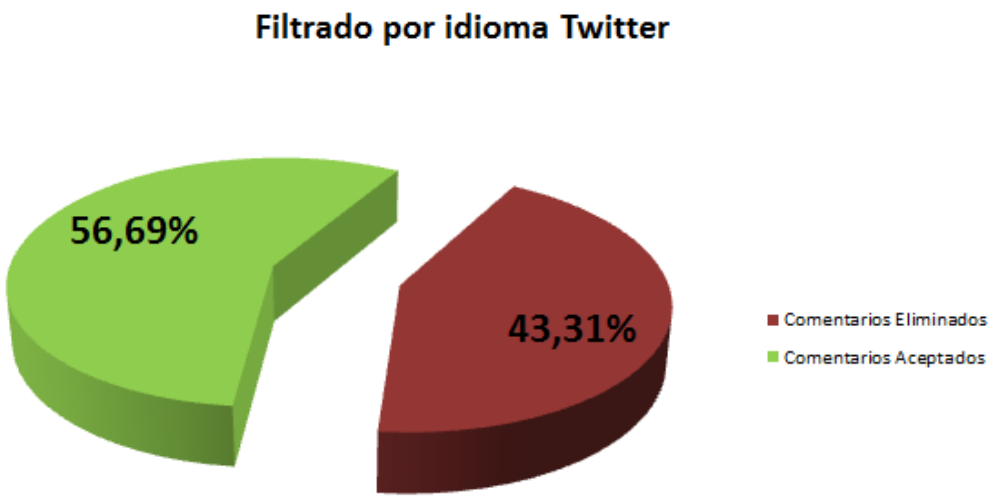


Figura 6.15: Gráfico sobre el filtrado de idioma de los *tweets*, Alberto Esteban García

Podemos ver que en este caso, buena parte de los *tweets* han sido eliminados dado que no estaban en castellano.

Sin embargo, la heterogeneidad de la red social *Twitter* hace que además de filtrar los *tweets* por idioma, necesitemos filtrarlos por contexto.

El filtrado por contexto se refiere a intentar acotar los *tweets* que verdaderamente están mostrando alguna opinión sobre el establecimiento, es decir, que no solo citen al establecimiento sin añadir un comentario sobre él.

Para ello nos hemos hecho servir de las palabras más comentadas por cada establecimiento almacenadas anteriormente en la base de datos que tienen como fuente los comentarios de *TripAdvisor*, con el objetivo de afinar un poco más los *tweets* que vamos a emplear para la creación del resumen.

El proceso que se ha realizado es la división del *tweet* en palabras, proceso que en PLN se conoce como tokenización. Acto seguido se ha realizado el proceso de eliminación de *stopwords* y el de lematización de las palabras que da como resultado un conjunto de lemas.

En este momento comparamos los lemas más frecuentes del establecimiento con los del *tweet.* El umbral fijado es que existan tres o más lemas coincidentes en el *tweet* para que este pase el filtro, sino este será descartado.

Para ver la evolución del número de *tweets* durante el proceso de filtrado, a continuación se muestra la siguiente gráfica:

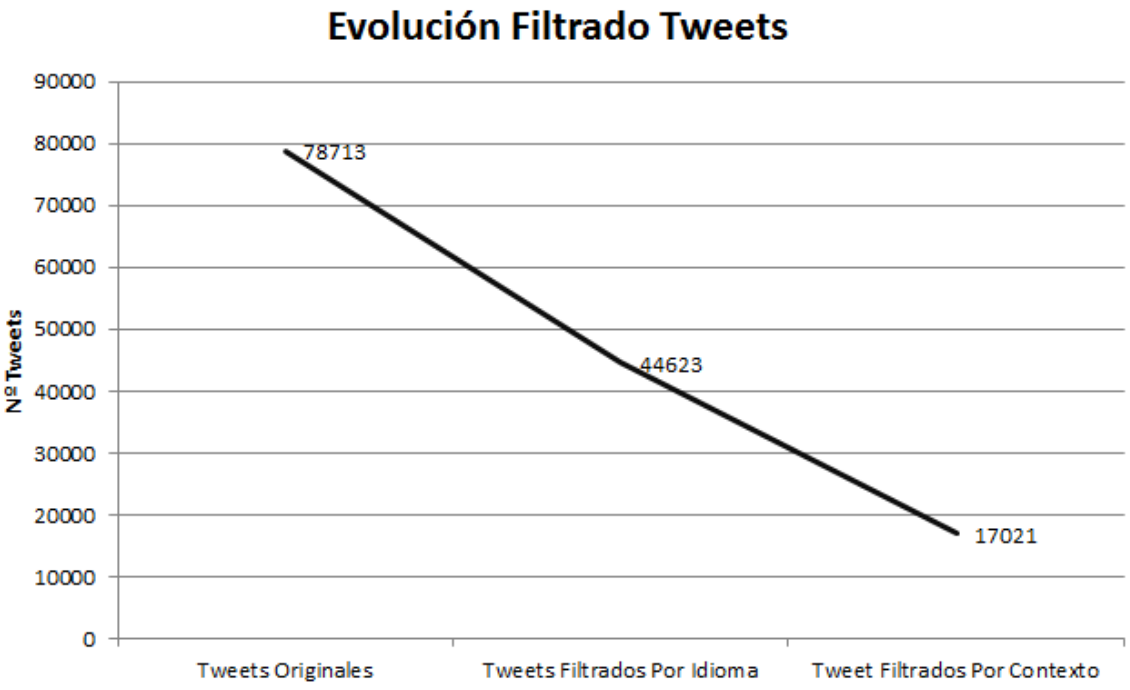


Figura 6.16: Gráfico sobre la evolución del filtrado sobre los *tweet,* Alberto Esteban García

Como podemos ver sólo vamos a emplear el 21,62% de los *tweets* que originalmente teníamos, este dato prueba la heterogeneidad de *Twitter*.

6.4.- Fase de elaboración de resúmenes

El objetivo de esta fase es la generación de los primeros resúmenes con el objetivo de evaluarlos y poder depurarlos en la fase de post-procesado.

6.4.1.- Asignación puntuación de Twitter

En este momento ya tenemos toda la información filtrada y lista para ser utilizada.

Sin embargo, antes de la creación de los resúmenes vamos a asignar la puntuación de cada establecimiento en base a *Twitter*.

Para ello nos hemos hecho servir de la API Rest del GPLSI[12] sobre análisis de sentimientos.

Su utilización es sencilla, como entrada recibe un texto y la aplicación devuelve la siguiente información:

- **Intensidad**: Valor discreto que informa de la intensidad que tiene el texto.
- **Sentimiento**: Sentimiento asociado al texto (positivo, negativo o neutral)

[12] http://gplsi.dlsi.ua.es/services/socialobserver3/doc/

A continuación se muestra un ejemplo de su utilización mediante la interfaz web:

Figura 6.17: Interfaz gráfica de la herramienta de análisis de sentimiento, Alberto Esteban García

El resultado que nos devuelve es el siguiente:


```
{
  "semanticAttributes": [
    [
      {
        "subject": "OVERALL",
        "intensity": 0.44444445,
        "emotionLabels": null,
        "sentimentCategory": "positive"
      }
    ]
  ]
}
```


Figura 6.18: Respuesta de la herramienta de análisis de sentimiento, Alberto Esteban García

Además de la herramienta del GPLSI hemos utilizado la información almacenada de cada *tweet* (*retweets* y favoritos) para la puntuación del propio *tweet*.

Por ello la nota de cada *tweet* dependerá de cuatro factores:

- Sentimiento
- Intensidad
- Nº de *retweets*
- Nº de favoritos

Se aplica una única fórmula donde el signo dependerá de si el sentimiento es positivo o neutral (signo positivo) o si por el contrario el sentimiento es negativo (signo negativo).

Se ha considerado que la importancia de un *retweet* es el doble que la de un favorito.

La fórmula de puntuar cada *tweet* se muestra a continuación

- Si el *tweet* es positivo o neutral:

$$Nota_{Tweet} = 10 \cdot intensidad_{Tweet} + 2 \cdot retweets_{Tweet} + favoritos_{Tweet}$$

- Si el *tweet* es negativo:

$$Nota_{Tweet} = (-1) \cdot 10 \cdot intensidad_{Tweet} + 2 \cdot retweets_{Tweet} + favoritos_{Tweet}$$

Para calcular la nota de cada establecimiento hemos de saber cuál hubiera sido la máxima nota posible atendiendo al número de *tweets* de cada establecimiento, por lo que la máxima nota posible viene determinada por la siguiente fórmula:

$$Nota_{Máxima} = 10 \cdot n_{Tweets} + 2 \cdot n_{TOTALretweets} + n_{TOTALfavoritos}$$

Finalmente la nota de cada establecimiento se calcula con la siguiente fórmula:

$$Nota_{Establecimiento} = \frac{\sum_{i=0}^{n} Nota_{Tweet}}{Nota_{Máxima}} + n_{Tweets} \cdot 0{,}01$$

La última parte de la fórmula, se utiliza para aplicar un factor de popularidad, con el objetivo de que tener un alto número de *tweets*, incremente la puntuación del establecimiento.

6.4.2.- Recuento lemas más relevantes de Twitter

Como hicimos previamente con las frases resultantes de los comentarios de *TripAdvisor*, vamos a contabilizar los cien lemas más relevantes de cada establecimiento pero en este caso basándonos en los ***tweets* filtrados por idioma**, con el objetivo de no sesgar la información obtenida en *Twitter*.

En este caso no se ha hecho uso del software para la división del *tweet* en sentencias debido a que como el tamaño de un *tweet* es como máximo de 140 caracteres, esto comporta que en la mayoría de las ocasiones solo exista una única sentencia.

El proceso es el explicado anteriormente, cada *tweet* es dividido a nivel de palabra, se eliminan las *stopwords* y se realiza la lematización de las palabras obteniendo como resultado un conjunto de lemas.

Finalmente almacenaremos los cien sustantivos que más frecuencia tengan en los *tweets* de cada establecimiento.

Por ejemplo, tras el proceso explicado, los cinco lemas más utilizados del hotel "*Meliá Alicante*" son:

Lema	Frecuencia
Habitación	416
Vista	226
Playa	157
Desayuno	156
Mar	139

Figura 6.19: Tabla con los lemas más relevantes del hotel "*Meliá Alicante*", Alberto Esteban García

6.4.3.- Agrupación de las frases de TripAdvisor

A modo de recapitulación vamos a enumerar la información de la que disponemos por cada establecimiento para elaboración de los resúmenes:

- Comentarios filtrados por idioma y divididos en frases
- *Tweets* filtrados por idioma y contexto
- Cien lemas más utilizados por cada establecimiento en *Tripadvisor*
- Cien lemas más utilizados por cada establecimiento en *Twitter*

Tras examinar la información recopilada de *Twitter* comprobamos que incluso mediante el filtrado de la información, los *tweets* filtrados en ocasiones eran bastante distantes del tema que estamos tratando, por ello se decidió utilizar como base de los resúmenes las frases de *TripAdvisor* y la información de *Twitter* serviría como refuerzo.

El primer paso para la elaboración del resumen es la agrupación de las frases por sentimiento, esto lo tenemos dado que cada frase está catalogada gracias a la herramienta del GPLSI[13] de detección de sentimientos.

El objetivo es agrupar las frases con el fin de conocer las semejanzas que hay entre ellas y poder mostrar la frase más relevante y frecuente que pueda contribuir al resumen.

Para explicar esta idea vamos a ver un ejemplo, tenemos las dos frases siguientes:

La ubicación del hotel es muy buena

El hotel está muy bien ubicado

[13] http://gplsi.dlsi.ua.es/services/socialobserver3/doc/

Nosotros podemos identificar que las dos frases anteriores quieren decir lo mismo, sin embargo esta tarea no es sencilla para una máquina. Por ello vamos a utilizar la librería *simmetrics*[14].

Utilizaremos la función del coseno, el cual informará del grado de similitud entre las frases. Para determinar el umbral de similitud se han realizado varias pruebas con el objetivo de conocer el grado de similitud en el cual debemos agrupar las diferentes frases.

Si introducimos las dos frases anteriores, el grado de similitud que da como resultado es:

0.30860671401023865

Por tanto, una similitud superior a ese factor es correcta. Vamos a realizar una prueba utilizando las frases que están en nuestra base de datos, a continuación se muestran las frases agrupadas con una similitud del 0.3 o superior:

```
FRASE ORIGINAL: En general  bueno
Frase coincidente: Buenos días  Mi opinión  en lo esencial y en general  es buena
Frase coincidente: Vacaciones en familia
Frase coincidente: Estuvimos dos noches en fin de semana y la experiencia general ha sido buena
Frase coincidente: Excelente trato y en general el personal de diez
Frase coincidente: El desayuno muy bueno en un salon precioso
Frase coincidente: El desayuno muy bueno y servido en un comedor muy luminoso  Más
Frase coincidente: El servicio  bueno
Frases coincidentes: 7
```

Figura 6.20: Imagen con las frases coincidentes sobre la frase original mostrada con similitud mayor a 0.3, Alberto Esteban García

Podemos ver que la similitud entre las frases existe, si somos un poco más exigentes y aumentamos el factor a 0.4, obtenemos lo siguiente:

```
FRASE ORIGINAL: En general  bueno
Frase coincidente: El desayuno muy bueno en un salon precioso
Frases coincidentes: 1
```

Figura 6.21: Imagen con las frases coincidentes sobre la frase original mostrada con similitud mayor a 0.4, Alberto Esteban García

Como podemos ver la agrupación es menor y nuestro objetivo no es realizar una agrupación demasiado estricta, por lo que un factor de similitud de 0.3 es correcto para cumplir con nuestro objetivo de agrupación de frases.

Hemos de decir que los grupos de frases son conjuntos disjuntos, es decir, una frase no puede pertenecer a dos conjuntos de frases, esta decisión fue tomada dado que si no la complejidad de las operaciones sería mucho mayor y el tiempo de procesamiento también.

[14] http://mvnrepository.com/artifact/com.github.mpkorstanje/simmetrics-core/3.0.1

6.4.4.- Asignación de puntuación a cada frase

Una vez tenemos el conjunto de frases agrupadas por similitud, debemos escoger la frase más relevante de ese grupo.

Para ello se puntúa cada frase teniendo en cuenta los siguientes factores:

- Relevancia en *TripAdvisor*
- Relevancia en *Twitter*
- Complementos del sustantivo: Se valorará positivamente que las frases contengan adjetivos calificando a los sustantivos con el fin de aportar mayor expresividad al resumen

La relevancia en *TripAdvisor* es el resultado de realizar un proceso de eliminación de *stopwords* y lematización de la frase que da como resultado los lemas de la frase.

Una vez tenemos los lemas de la frase los comparamos con los lemas más utilizados del establecimiento en *TripAdvisor* y cuando coincide un lema de la frase con un lema de la lista de lemas del establecimiento, la puntuación aumenta con el valor de la frecuencia de este lema.

Pongamos un ejemplo para entender el proceso explicado, si tenemos una frase con los siguientes lemas:

- Habitación
- Vista
- Canal

Y tenemos la siguiente lista de lemas de *TripAdvisor* con su frecuencia:

Nº	Lema	Frecuencia
1	Habitación	325
….	…	…
100	Vista	5

Figura 6.22: Tabla para la explicación del ejemplo propuesto, Alberto Esteban García

El resultado de la relevancia de *TripAdvisor* es:

$$RelevanciaTripAdvisor_{Frase} = 325 + 5 = 330$$

La relevancia de *Twitter* se realiza de forma análoga, comparamos los lemas de cada frase con los lemas de *Twitter* del establecimiento y aumentamos la puntuación según la coincidencia de las dos listas de lemas.

La puntuación de los complementos del sustantivo se realiza mediante un análisis morfológico donde intentamos puntuar la aparición de adjetivos seguidos por sustantivos o viceversa con el objetivo de seleccionar las frases que sean más descriptivas.

De esta manera si identificamos que la frase contiene un adjetivo y está precedido o seguido de un sustantivo la puntuación relativa a los complementos del sustantivo de la frase (que hemos denominado *concordancia*) aumenta en diez puntos. En el caso de que se detecte el adjetivo aislado, la puntuación para este factor aumenta en dos puntos.

Para comprenderlo vamos a poner un ejemplo, si tenemos la siguiente frase:

Buenas vistas y trato exquisito

Su concordancia asociada es:

$$Concordancia_{Frase} = 10 + 10 = 20$$

Finalmente la relevancia total se calcula mediante la siguiente fórmula:

$$RelevanciaTotal_{Frase} = 0{,}7 \cdot RelevanciaTripAdvisor_{Frase} + 0{,}3 \cdot RelevanciaTwitter_{Frase}$$

Esta ponderación es debido a que consideramos que la información de *TripAdvisor* es más fidedigna que la de *Twitter* y por ello le aportamos un mayor peso.

Finalmente la puntuación final de la frase viene determinada por la siguiente fórmula:

$$Nota_{Frase} = 10 \cdot intensidad_{Frase} + RelevanciaTotal_{Frase} + Concordancia_{Frase} + n_{Frases} + n_{Tweets}$$

Donde n_{Frases} se corresponde con el número de frases del conjunto, con el fin de utilizarse como factor de popularidad sobre la información de *TripAdvisor*, n_{Tweets} es el número de *tweets* coincidentes con el grupo de frases, el cual se utiliza de forma análoga al anterior, es decir, como factor de popularidad sobre la información de *Twitter*.

En este momento tenemos un conjunto de frases positivas, negativas y neutrales, las cuales son diferentes entre ellas y estamos en disposición de crear los resúmenes.

6.4.5.- Elaboración del resumen

En un principio la generación de resúmenes iba a consistir en generar un resumen por establecimiento. Sin embargo, debido a la gran cantidad de información y como previamente realizamos una clasificación según el sentimiento, decidimos generar tres resúmenes por cada establecimiento:

- **Resumen mixto**: Mostrará los aspectos más importantes para bien y para mal del establecimiento.

- **Resumen positivo**: Mostrará las características más relevantes que han gustado más a los clientes.

- **Resumen negativo**: Mostrará los aspectos más importantes que no han gustado a los clientes.

Para realizar cada tipo de resumen tenemos un conjunto de frases las cuales no tienen relación entre sí (previamente nos hemos asegurado de escoger la más significativa de su grupo) y el proceso de generación de resumen diferirá en base al sentimiento.

Para generar el resumen mixto, utilizaremos:

1. La frase neutral con mayor puntuación
2. La frase positiva con mayor puntuación
3. La frase negativa con mayor puntuación

En el caso de la generación del resumen positivo utilizaremos el conjunto de frases positivas ordenadas por puntuación, donde escogeremos las tres frases con mayor puntuación.

La generación del resumen negativo, se realiza de forma análoga a la del resumen positivo, es decir, se escogen las tres frases negativas con mayor puntuación.

6.4.6.- Generación de resúmenes inicial

En este momento estamos en disposición de crear nuestro primer resumen, en este caso generamos el resumen mixto del hotel "*NH Valencia*":

"Hotel ubicado en la zona de la ciudad de las artes & ciencias de valencia, lobby pequeño y habitación individual chica pero con un plus inmejorable, una terraza enorme con vista a la ciudad de valencia donde vale la pena apreciar el atardecer y amanecer"

Desde la llegada trato e informacion humana muy acogedora las habitaciones modernas, espaciosas y luminosas con hermosas vistas, baños confortables tiene instalaciones para ejercicios, piscina y solárium el bufet para desayuno variado y rico.

Las habitaciones un poco pequeñas pero cuidadas, pedid mejor las habitaciones que dan a la plaza interior porque sino son algo ruidosas las exteriores."

El primer párrafo es la frase positiva con más puntuación, el segundo se corresponde con la frase neutral más puntuada y el último párrafo es la frase más puntuada dentro del sentimiento negativo.

Como podemos ver el resultado es interesante pero se necesitan corregir varias cosas:

- La similitud entre las frases es apreciable en ocasiones a simple vista, así que habrá que realizar algún proceso para evitar que las frases resultantes tengan similitud entre ellas.

- Se ha de crear un conjunto de frases de enlace que doten al resumen de una mayor coherencia.

Además en otros resúmenes se detectaron otros aspectos que deben ser mejorados:

- Se deben eliminar posibles preposiciones, conjunciones o artículos al final de cada frase debido al proceso automatizado de división de frases.

- Para aportar mayor legibilidad al texto se aplicarán reglas para intentar convertir la frase a tercera persona.

6.5.- Fase de post-procesado

En esta fase el objetivo es mejorar los resultados obtenidos en la creación automática de resúmenes aplicando varias técnicas.

6.5.1.- Evitar similitud entre frases

Para evitar la similitud entre frases crearemos un nuevo proceso de selección el cual no se basará únicamente en la puntuación de la frase.

En el caso de la generación del resumen mixto utilizaremos tres conjuntos de frases:

- El conjunto de frases neutrales
- El conjunto de frases positivas
- El conjunto de frases negativas

Para la creación del resumen mixto seguiremos el siguiente procedimiento:

1. Escogemos la frase neutral con mayor puntuación
2. La siguiente frase a escoger será la mayor puntuada dentro del grupo de las positivas donde su similitud con la frase anterior no supere el umbral de 0.2
3. Si no existiera ninguna frase positiva con similitud menor a 0.2, se escogería la frase positiva con menor similitud dentro del conjunto de las posibles.

4. Finalmente escogeremos la frase negativa con mayor puntuación que tenga una similitud menor a 0.2 con la frase neutral y con la positiva.
5. Si se diera el caso de que no existe una frase negativa que supere este requisito, se ordenará el conjunto de frases negativas y se escogerá la que menor similitud tenga con la frase positiva, esto es debido, a que la frase negativa estará tras la positiva.

En el caso de la generación del resumen positivo utilizaremos el conjunto de frases positivas ordenadas por puntuación, donde escogeremos tres frases atendiendo a los siguientes criterios:

1. Escogemos la frase positiva con mayor puntuación
2. La siguiente frase positiva escogida será la de mayor puntuación donde su similitud con la anterior sea menor a 0.2.
3. En el caso de que no exista, es decir no existe ninguna frase con similitud menor a 0.2, se escogerá la frase con menor similitud entre las posibles.
4. Finalmente, escogeremos la última frase positiva atendiendo a su puntuación siempre y cuando su similitud con las anteriores sea menor a 0.2
5. Si dentro del conjunto de frases positivas no existe ninguna que cumpla el requisito anterior, se ordenará el conjunto de frases positivas y se escogerá la que menor similitud tenga con la segunda frase positiva seleccionada.

Este proceso se realiza para asegurar que la información aportada en el resumen no es repetitiva. Previamente hicimos una agrupación de frases con una similitud mayor a 0.3 y ahora exigimos que la similitud entre frases sea menor a 0.2 para intentar crear un resumen rico en contenido.

La generación del resumen negativo, se realiza de forma análoga a la del resumen positivo pero escogiendo las frases del conjunto negativo.

6.5.2.- Verificación de final de frase

Anteriormente hemos visto que en el proceso de división de frases, algunas de estas frases se quedaron con un artículo, preposición o conjunción al final.

Para evitar incoherencias en el resumen se ha creado una función recursiva para que elimine este tipo de palabras, en el caso de que se encontraran al final de la frase, si por ejemplo tenemos la siguiente frase:

"El hotel está ubicado en una cala o en"

Tras pasar por la función, la frase queda como a continuación:

"El hotel está ubicado en una cala"

6.5.3.- Creación de las frases de enlace

Vamos a estudiar cómo podemos realizar un conjunto de frases que aporten información relevante y hagan que el resumen gane en coherencia.

De cada hotel conocemos las siguientes características:

- Nota de servicio
- Nota de limpieza
- Nota de calidad-precio
- Nota de perfil del cliente
- Nota *TripAdvisor*
- Nota *Twitter*

Por ello se ha optado por la creación de un conjunto de frases predefinidas que aporten coherencia a los resúmenes y enmascaren que los resúmenes se han generado de forma automática.

Esta técnica se ha basado en la propuesta de (*Gerani et al., 2014*).

El servicio, la limpieza y el servicio de cada hotel tienen una valoración de cero a diez, por lo que vamos a definir cinco tramos:

Puntuación Decimal	Nivel
8-10	5
6-8	4
4-6	3
2-4	2
0-2	1

Figura 6.23: Tabla de equivalencias entre la puntuación decimal y el nivel asignado para las características de servicio, comida, calidad-precio y limpieza, Alberto Esteban García

Para la elaboración de los resúmenes positivos y mixtos, por cada nivel se han definido un conjunto de cuatro frases, las cuales serán elegidas aleatoriamente con el objetivo de que no se repita la misma frase predefinida en multitud de resúmenes.

Así, por tanto, por ejemplo, para elaborar la información del servicio del hotel en un resumen positivo o mixto contaremos con un total de veinte frases predefinidas que podemos utilizar.

En el caso de que queramos generar un resumen negativo, se han creado un conjunto de frases siguiendo la misma técnica, pero las cuales aportan un tono de negatividad o "soberbia" con el objetivo de dotar al resumen de una mayor intensidad.

Al igual que para los resúmenes positivos o mixtos, se definió un conjunto de veinte frases con un tono negativo para informar del servicio del hotel.

Este mismo planteamiento se aplica a la limpieza y a la relación calidad-precio del hotel.

Para la evaluación de la nota de *TripAdvisor* y *Twitter* se realizó una ponderación donde la nota global del hotel se correspondía con la siguiente fórmula

$$Nota\ Global_{Hotel} = 0{,}7 \cdot nota_{TripAdvisor} + 0{,}3\ \cdot nota_{Twitter}$$

Esta ponderación ya fue utilizada anteriormente (*sección 6.4.4*) debido a que consideramos que la información de *TripAdvisor* es más fidedigna que la de *Twitter*.

Al igual que antes se utilizó la tabla de equivalencias anterior, generando un total de veinte frases relacionadas con la nota del hotel para los resúmenes positivos y mixtos, y otro conjunto de veinte frases para los resúmenes negativos.

Finalmente, se generaron un conjunto de frases, para informar del tipo de cliente al que está recomendado el hotel, un total de cuatro frases por cada categoría (En solitario, en pareja, familias, de negocios), la asignación del nivel con el perfil del cliente viene determinada por la siguiente tabla:

Perfil Cliente	Nivel
En solitario	1
En pareja	2
Familias	3
De negocios	4

Figura 6.24: Tabla de equivalencias entre la puntuación decimal y el nivel asignado para la característica de perfil del cliente, Alberto Esteban García

Por tanto contamos con ochenta frases para generar los resúmenes positivos o los resúmenes mixtos así como otras ochenta para generar los resúmenes negativos, junto con dieciséis frases en común para evaluar el perfil del cliente. Esto hace un total de 176 frases predefinidas que posibilitarán una mayor expresividad en los resúmenes así como evitar la detección de la repetición de frases, a continuación se muestra una tabla resumen para mayor claridad:

Característica	Resumen positivo/mixto	Resumen negativo
Servicio	20 frases	20 frases
Limpieza	20 frases	20 frases
Calidad-Precio	20 frases	20 frases
Nota Global	20 frases	20 frases
Perfil del cliente	16 frases	

Figura 6.25: Tabla resumen de las frases predefinidas existentes para los hoteles, Alberto Esteban García

Algunos ejemplos de estas frases predefinidas se muestran a continuación:

- **Categoría: Servicio, Nivel: 5, Sentimiento: Positivo:**

  "La gran mayoría de las opiniones destacan su excelente servicio"

- **Categoría: Servicio, Nivel: 5, Sentimiento: Negativo:**

  "Las opiniones indican que el servicio es correcto"

- **Categoría: Perfil del cliente, Nivel: 4 (De Negocios):**

  "Los clientes lo utilizan para hospedarse en viaje de negocios"

- **Categoría: Limpieza, Nivel: 2, Sentimiento: Positivo:**

  "Los clientes opinan que se debería prestar mucha más atención a la limpieza"

El mismo planteamiento fue realizado para los restaurantes, de los que conocemos:

- Nota de servicio
- Nota de comida
- Nota de calidad-precio
- Nota *TripAdvisor*
- Nota *Twitter*

Aplicando el procedimiento anterior contamos con ochenta frases para la elaboración de los resúmenes positivos y mixtos de los restaurantes así como otras ochenta frases para la generación de los resúmenes negativos contando con un total de 160 frases.

Para ejemplificar lo explicado se muestra la siguiente tabla:

Característica	Resumen positivo/mixto	Resumen negativo
Servicio	20 frases	20 frases
Comida	20 frases	20 frases
Calidad-Precio	20 frases	20 frases
Nota Global	20 frases	20 frases

Figura 6.26: Tabla resumen de las frases predefinidas para los restaurantes, Alberto Esteban García

Algunos ejemplos de las frases creadas se muestran a continuación:

- **Categoría: Comida, Nivel: 5, Sentimiento: Positivo:**

  "Los clientes recomiendan este restaurante por su magnífica gastronomía"

- **Categoría: Comida, Nivel: 5, Sentimiento: Negativo:**

  "La oferta gastronómica de este restaurante es aceptable"

- **Categoría: Nota Global, Nivel: 5, Sentimiento: Positivo:**

  "Según la valoración de los clientes es uno de los mejores de la ciudad"

- **Categoría: Calidad-Precio, Nivel: 5, Sentimiento: Negativo:**

  "Las opiniones de los clientes indican que la relación calidad-precio es suficiente"

Además se crearon oraciones predefinidas con el objetivo de introducir las frases elegidas que formarán parte del resumen. A continuación se muestran un par de ejemplos:

"De lo más comentado por los clientes es que"

"Las opiniones de los clientes enfatizan que"

Por tanto, para la generación de los resúmenes de los hoteles seguiremos el siguiente criterio:

Resumen Positivo y Mixto

1. Introducir frase predefinida con la valoración de la nota global del hotel con sentimiento positivo.
2. Introducir dos frases predefinidas sobre el servicio, la limpieza, la relación calidad-precio o el perfil de cliente recomendado con sentimiento positivo.
3. Introducir frase predefinida de introducción de los comentarios de las reseñas de los clientes.

4. Añadir las frases de las reseñas atendiendo al criterio explicado anteriormente (*sección 6.5.1*).

Resumen Negativo

1. Introducir frase predefinida con la valoración de la nota global del hotel con sentimiento negativo.
2. Introducir dos frases predefinidas sobre el servicio, la limpieza, la relación calidad-precio o el perfil de cliente recomendado con sentimiento negativo.
3. Introducir frase predefinida de introducción de los comentarios de las reseñas de los clientes.
4. Añadir las frases de las reseñas atendiendo al criterio explicado anteriormente (*sección 6.5.1*).

En el caso de la generación de los resúmenes de los restaurantes el procedimiento es el mismo salvo que deberemos elegir frases predefinidas sobre el servicio, la comida o la relación calidad-precio en formato positivo o negativo según el tipo de resumen que se vaya a generar.

6.5.4.- Reglas para pasar a tercera persona la frase

Con el objetivo de generar un resumen en tercera persona, se han definido un conjunto de reglas para convertir sentencias en primera persona del plural a tercera persona del plural.

El proceso ha consistido en realizar un análisis morfológico de las palabras utilizando la herramienta *TreeTagger* que ya utilizamos anteriormente para conocer los lemas de cada frase y su tipo.

En este caso nos interesa conocer si la frase contiene verbos y si es así convertirlos a tercera persona del plural, para dotar al resumen de un tono más impersonal, y por tanto, más “objetivo”.

Por ello utilizamos el lema que nos da la herramienta y la palabra original para conocer el tiempo verbal utilizado, pongamos un ejemplo para ver este proceso de forma sencilla:

Palabra	Lema	Tagger
Compramos	Comprar	VLfin

Figura 6.27: Tabla con palabra, lema y *tagger*, Alberto Esteban García

Mediante el *tagger* identificamos que la palabra se corresponde con un verbo, además el *TreeTagger* nos informará del lema, lo que podemos utilizar para conocer la raíz de la palabra:

Verbo	Lema	Raíz
Compramos	Comprar	Compr

Figura 6.28: Tabla con verbo, lema y raíz, Alberto Esteban García

Una vez tenemos la raíz, podemos construir la forma verbal en tercera persona:

Raíz	Verbo en Tercera Persona
Compr	Compraron

Figura 6.29: Tabla con raíz y verbo en tercera persona, Alberto Esteban García

Como ejemplo, a continuación se muestra el paso a tercera persona de la siguiente frase:

"Nos fuimos a esperar a los amigos, hemos comprado una bebida y esperamos a que vinieran"

Cuando aplicamos la función obtenemos lo siguiente:

"Se fueron a esperar a los amigos, han comprado una bebida y esperaron a que vinieran"

6.5.5.- Generación de resúmenes final

A continuación se muestra el resultado obtenido, una vez explicado el proceso de generación del resumen, podemos detectar la composición del resumen y distinguir las frases predefinidas de las reseñas de los clientes:

Resumen Mixto

"El hotel NH Valencia Las Artes según la valoración de los clientes es un buen hotel para hospedarse. La mayoría de las opiniones recomienda el hotel por la buena relación calidad-precio, además los clientes destacan que es excelente para venir en pareja. Entre lo más comentado por los clientes, se encuentra que hotel ubicado en la zona de la ciudad de las artes & ciencias de valencia, lobby pequeño y habitación individual chica pero con un plus inmejorable, una terraza enorme con vista a la ciudad de valencia donde vale la pena apreciar el atardecer y amanecer. desde la llegada trato e informacion humana muy acogedora las habitaciones modernas, espaciosas y luminosas con hermosas vistas, baños confortables tiene instalaciones para ejercicios, piscina y solárium el bufet para desayuno variado y rico. Por el contrario las habitaciones un poco pequeñas pero cuidadas, pedid mejor las habitaciones que dan a la plaza interior porque sino son algo ruidosas las exteriores."

Resumen Positivo

"El hotel NH Valencia Las Artes según la valoración de los clientes es un buen hotel para hospedarse. El hotel es recomendable para venir en pareja, además el hotel destaca por tener una gran relación calidad-precio. Las opiniones destacan que habitación normal, hotel cerca de la ciudad de las artes y la playa de la malvarrosa La limpieza de la habitación era inexistente, polvo acumulado por toda la habitación, sin embargo el baño estaba limpio. Otro aspecto importante es que hotel muy céntrico y con muy buen servicio, las habitaciones están muy bien y con buenas vistas, el buffet del desayuno muy completo y todo muy bueno, el personal de servicio muy agradable y se puede descansar muy bien en las habitaciones, muy tranquilo para estar en el centro de valencia. Algo a destacar es que a nuestro parecer este hotel fue realmente perfecto, su ubicación es inmejorable, al lado de la ciudad de las ciencias, al lado de un centro comercial grande, al lado de unos restaurantes, se gustó la taberna de maria, tampoco tan alejado del centro."

Resumen Negativo

"El hotel NH Valencia Las Artes según la valoración de los clientes es un hotel correcto para pasar unos días. El hotel cuenta con un servicio del que no debemos preocuparnos según la mayoría de las opiniones, además la limpieza del hotel es aceptable. Las opiniones destacan que la ubicación no puede ser mejor, en la puerta de la ciudad de las artes y las ciencias, cerca del parque, dos centros comerciales y. Adicionalmente las habitaciones y el desayuno muy bien pero es caro la atenxion a la hora dde habrir[15] las habitaciones comunicadas entre si mal. Además algo lejos del centro pero de fácil acceso desde allí, lo peor la limpieza de la habitación, que dejaba algo que desear."

6.6.- Fase de elaboración de la interfaz

El objetivo de esta etapa es la creación de una interfaz gráfica sencilla y vistosa, que permita a los usuarios ver los resultados de la aplicación desarrollada, es decir, los resúmenes generados automáticamente.

Las tecnologías utilizadas en el desarrollo de la aplicación fueron *HTML*, *CSS* y *JavaScript* para capa de presentación (*Front-end*) y para la capa de acceso a datos (*Back-End*) se hizo uso de los *Servlets* de *Java*.

Con el objetivo de minimizar los costes de diseño, se utilizó el framework *Bootstrap* dado que es uno de los más utilizados y el cual permitió un rápido desarrollo de la interfaz debido a la gran comunidad de usuarios que lo utilizan.

En el desarrollo de la página principal, desde un principio se deseaba que fuese un sencillo buscador. Para su implementación se ha hecho uso de la herramienta

[15] Como se observa, debido a que las opiniones han sido extraídas de usuarios reales, a veces podemos encontrar faltas de ortografía (como ocurre en el caso de *habrir*). Este es uno de los aspectos que se plantean como mejora para el futuro.

"TextComplete"[16] la cual permite sugerir los establecimientos al tiempo que el usuario introduce el hotel o restaurante que desea consultar.

Por ello tras añadir los logos de los correspondientes departamentos colaboradores y el apartado para informar al usuario el fin de la aplicación, ya tenemos desarrollada la página principal.

La imagen[17] de fondo utilizada permite su reutilización, esta fue encontrada utilizando el buscador de *Google* con la opción de buscar las imágenes que pueden ser reutilizadas sin comportar por ello ninguna violación de *copyright*.

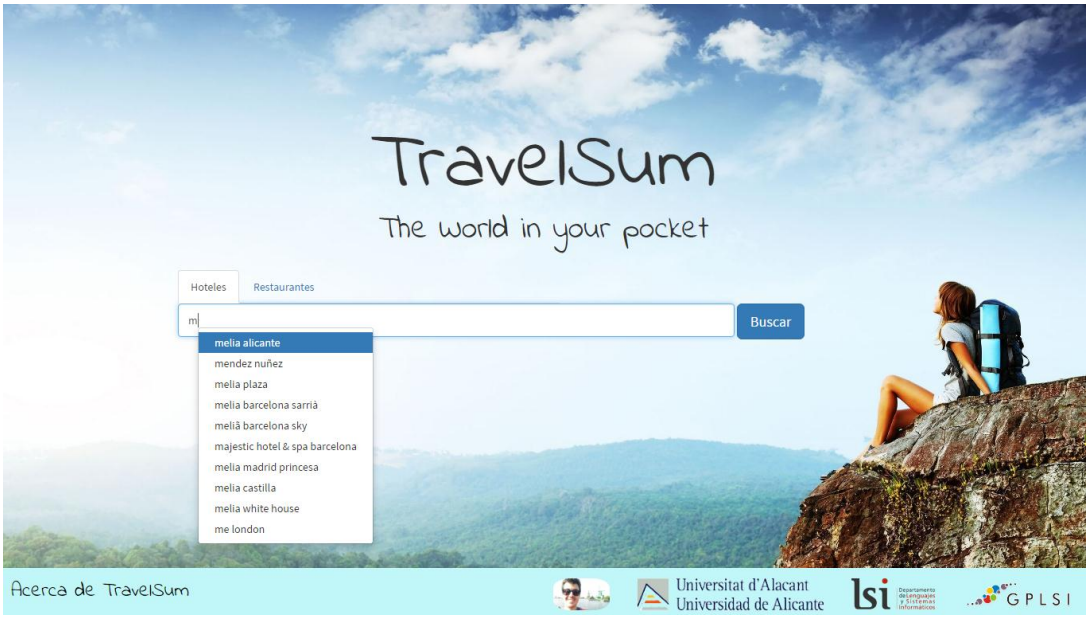


Figura 6.30: Página principal de la interfaz, Alberto Esteban García

El apartado para informar sobre los detalles de la aplicación desarrollada muestra la siguiente información:

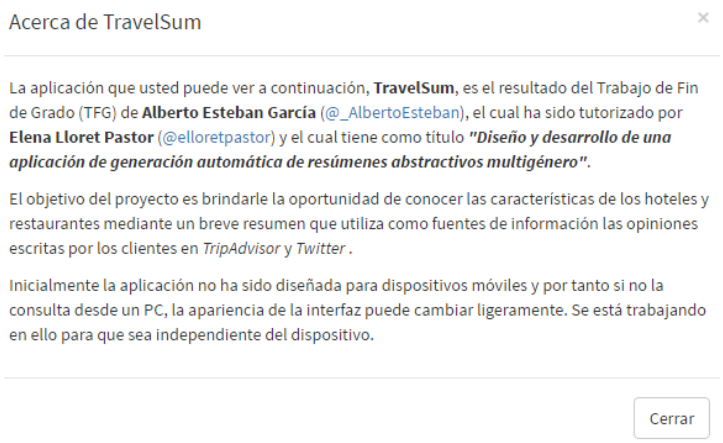


Figura 6.31: Información mostrada al consultar el apartado de *"Acerca de TravelSum"*

[16] https://github.com/yuku-t/jquery-textcomplete
[17] http://ejocurionline.net/data/wallpapers/16/DDW_927318.jpg

Una vez desarrollada la página principal, se implementó la interfaz para mostrar los resúmenes abstractivos multigénero generados automáticamente.

Con el objetivo de no limitarnos sólo a mostrar el resumen generado, se hicieron uso de varias herramientas que se citan a continuación:

- *Flickr*[18]: Las imágenes de los hoteles y restaurantes mostradas tienen como fuente *Flickr*, el cual proporciona un API abierto para utilizar sus fotografías.

- *Chart.js*[19]: Se trata de una librería de *JavaScript* para la realización de gráficos. Se utilizará para mostrar las características de los hoteles y restaurantes y poder realizar una comparación con la media de la zona del establecimiento.

- *Google Maps*[20]: Mediante el uso del API de *Google Maps*, mostraremos la localización de los establecimientos.

Finalmente al final de la página de los establecimientos añadiremos sugerencias de hoteles o restaurantes que le podrían interesar al usuario, atendiendo a la localización del establecimiento que está consultando, con el objetivo de facilitar la búsqueda de establecimientos cercanos.

El resultado de la interfaz se puede ver a continuación:

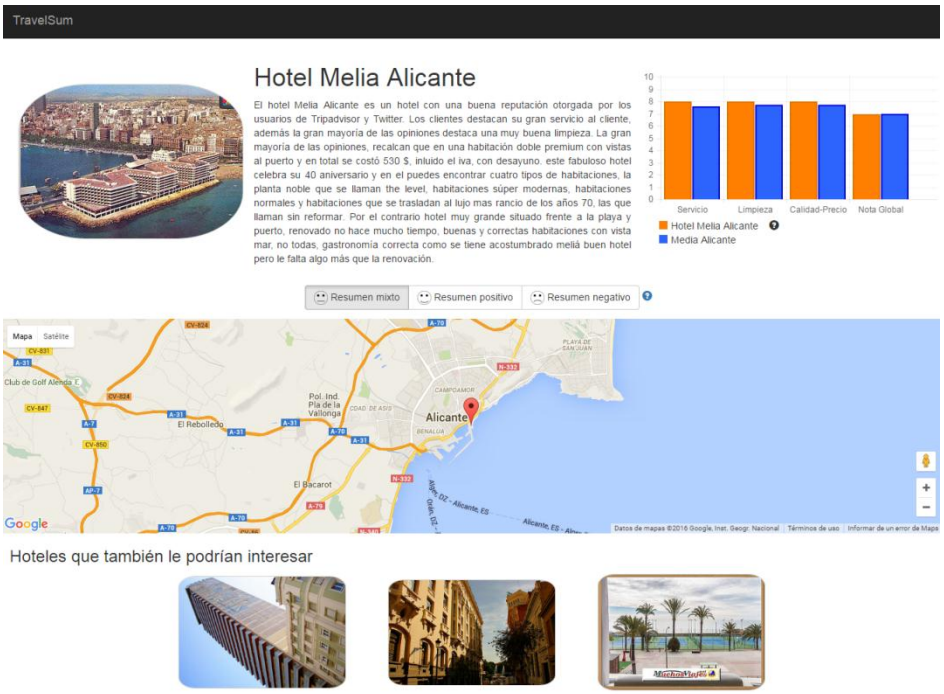


Figura 6.32: Interfaz para mostrar la información del hotel, Alberto Esteban García

[18] https://www.flickr.com/services/developer/
[19] http://www.chartjs.org/
[20] https://developers.google.com/maps/

Figura 6.33: Interfaz para mostrar la información del restaurante, Alberto Esteban García

Para facilitar la comprensión de los tipos de resúmenes generados se ha incluido una ayuda que el usuario puede consultar junto a la elección de cada tipo de resumen:

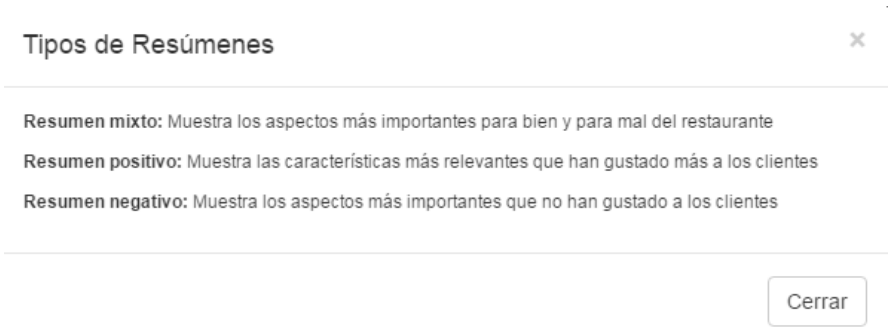


Figura 6.34: Información mostrada al consultar la ayuda de los tipos de resúmenes, Alberto Esteban García

Además, como la nota global de los establecimientos no es la media de las características mostradas se ha incluido otro mensaje de ayuda para informar de ello a los usuarios:

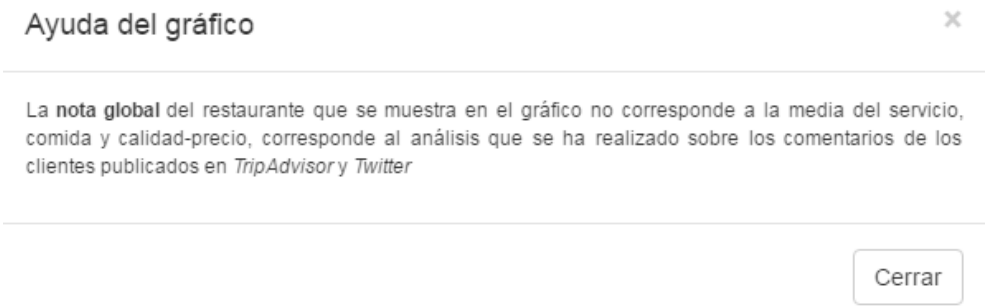


Figura 6.35: Información mostrada al consultar la ayuda del gráfico, Alberto Esteban García

La nota de cada establecimiento se calcula mediante la siguiente fórmula (*sección 6.4.4.*)

$$Nota_{Establecimiento} = 0{,}7 \cdot Nota_{TripAdvisor} + 0{,}3 \cdot Nota_{Twitter}$$

6.7.- Fase de despliegue

Una vez creada la aplicación, debíamos desplegarla en un servidor, para ello se recurrió al grupo GPLSI el cual nos brindó la posibilidad de tener una máquina virtual donde albergar el proyecto.

Los requisitos para el despliegue son:

- *Java* 7 o superior
- *Apache Tomcat* (versión 7.0.67)
- *MySQL* (versión 5.5)

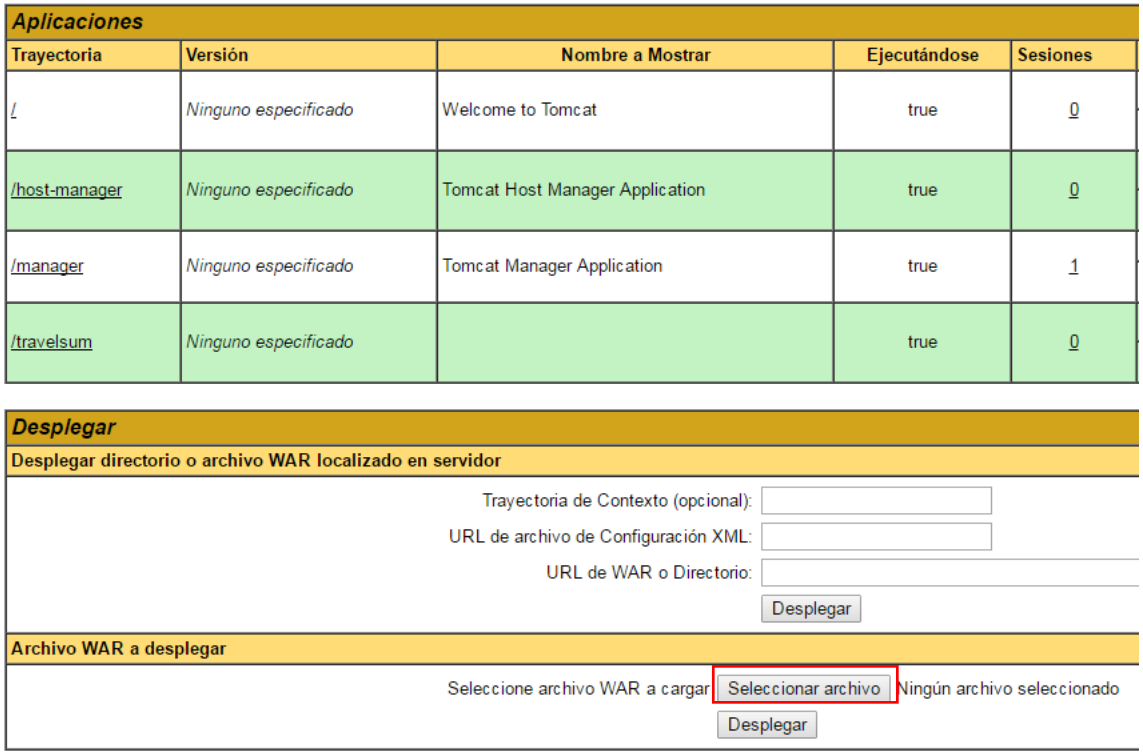
La interfaz solo hace uso de tres tablas de la base de datos (hotel, restaurante y resumen) por ello en la exportación de la base de datos, solo se exportaron estas tres tablas:

```
mysqldump –uuser –ppassword –hhost databaseName hotel restaurant
                    resumen > nombreArchivo.sql
```

Una vez en el servidor la importación de las tablas fue sencilla:

```
--user=username –password=password –host=localhost databaseName
                    < nombreArchivo.sql
```

Para el despliegue de la aplicación se generó un WAR y mediante la interfaz gráfica de *Apache Tomcat* se pudo subir el fichero empaquetado, con lo que el proceso fue realmente simple, como muestra la siguiente imagen:



Aplicaciones

Trayectoria	Versión	Nombre a Mostrar	Ejecutándose	Sesiones
/	*Ninguno especificado*	Welcome to Tomcat	true	0
/host-manager	*Ninguno especificado*	Tomcat Host Manager Application	true	0
/manager	*Ninguno especificado*	Tomcat Manager Application	true	1
/travelsum	*Ninguno especificado*		true	0

Desplegar

Desplegar directorio o archivo WAR localizado en servidor

Trayectoria de Contexto (opcional):
URL de archivo de Configuración XML:
URL de WAR o Directorio:
Desplegar

Archivo WAR a desplegar

Seleccione archivo WAR a cargar | Seleccionar archivo | Ningún archivo seleccionado
Desplegar

Figura 6.36: Interfaz gráfica de Apache Tomcat, Alberto Esteban García

Una vez desplegada se puede comprobar que la aplicación funciona igual que de manera local:

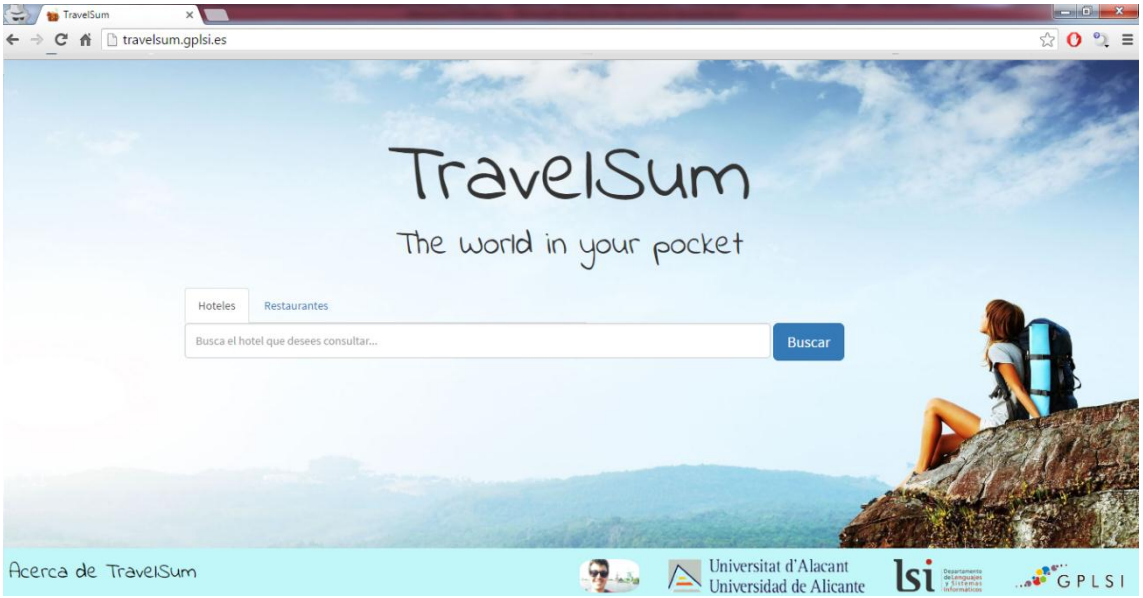


Figura 6.37: Imagen de la aplicación desplegada en http://travelsum.gplsi.es/, Alberto Esteban García

Por tanto la interfaz de la aplicación es accesible desde el siguiente enlace:

http://travelsum.gplsi.es/

7.- Evaluación y resultados

Una vez desplegada la aplicación, en esta fase determinaremos cómo vamos a evaluar la generación de los resúmenes automática así como la interfaz gráfica desarrollada.

7.1.- Formulario de evaluación

Se ha decidido crear un formulario para que los usuarios puedan evaluar la aplicación desarrollada, sobre todo en cuanto a la generación de los resúmenes.

El formulario[21] fue creado con *Google Forms* y su contenido puede ser consultado en el *anexo*.

Las preguntas planteadas intentan conocer los hábitos de los usuarios así como la opinión sobre diferentes aspectos de la aplicación:

- Conocer las formas de buscar información sobre establecimientos.
- Conocer la opinión sobre la utilidad del proyecto desarrollado.
- Determinar si el proceso de creación de los resúmenes está cerca de pasar el test de Turing.
- Conocer la opinión sobre la coherencia, utilidad y errores ortográficos en los resúmenes.
- Conocer la opinión sobre la usabilidad y utilidad de la interfaz desarrollada.

7.2.- Resultados de la evaluación

El cuestionario ha sido respondido por un total de 41 usuarios, en su mayoría residentes en las provincias de Alicante y Burgos.

A continuación se muestran los resultados obtenidos gracias a los usuarios que han respondido el formulario planteado,

[21] https://t.co/1b8BZaEm2s

Figura 7.1: Gráfico sobre el uso de Internet en la búsqueda de información turística, Alberto Esteban García

El resultado de este gráfico confirma que Internet se ha convertido en la mayor herramienta para la toma de decisiones como la búsqueda de estancias hoteleras o para consultar la oferta gastronómica de una zona.

Figura 7.2: Gráfico sobre la utilidad de las reseñas de *TripAdvisor*, Alberto Esteban García

Los resultados del gráfico ponen de manifiesto que, casi la totalidad de los usuarios conoce la página especializada por lo que es utilizada para consultas de hoteles y restaurantes con asiduidad.

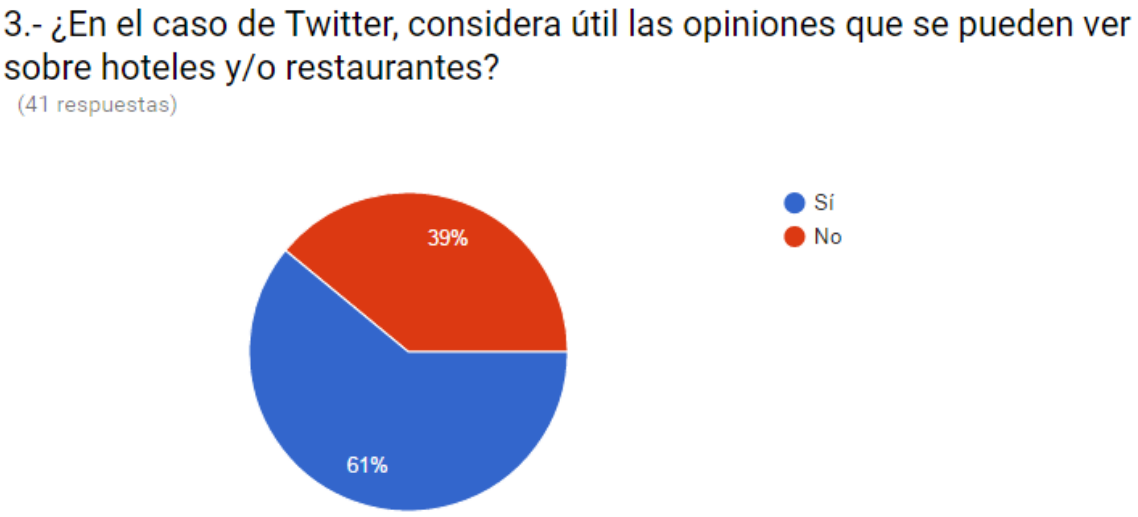


Figura 7.3: Gráfico sobre la utilidad de *Twitter* en cuanto a búsqueda de opiniones sobre establecimientos, Alberto Esteban García

El gráfico muestra la controversia que la red social *Twitter* genera ante la heterogeneidad de la información a la que podemos acceder.

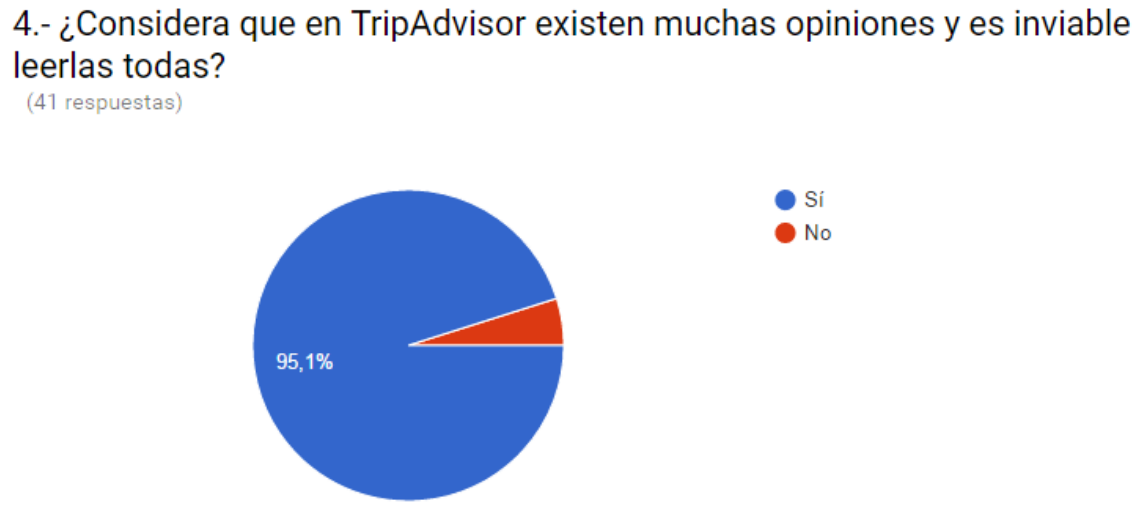


Figura 7.4: Gráfico sobre el volumen de información disponible en *TripAdvisor*, Alberto Esteban García

El gran volumen de datos de la página especializada es visible por todos los usuarios como se puede apreciar en el gráfico mostrado.

Figura 7.5: Gráfico sobre la heterogeneidad de *Twitter*, Alberto Esteban García

Los resultados muestran que casi todos los usuarios identifican una heterogeneidad de información muy grande en la red social de microblogging.

En el siguiente gráfico podemos ver que la utilidad de la aplicación desarrollada es por todos los usuarios reconocida:

Figura 7.6: Gráfico sobre la opinión de la utilidad de una aplicación que elabore resúmenes de forma automática y utilizando varios géneros textuales, Alberto Esteban García

La siguiente pregunta, fue una pregunta para determinar si la aplicación estaba próxima a pasar el test de Turing, podemos ver que la mayoría de los usuarios identificaron que el proceso era automático, aunque un significativo 30% no fue capaz de determinar si los resúmenes habían sido generados por una persona o una máquina, lo que significa que la aplicación va por buen camino:

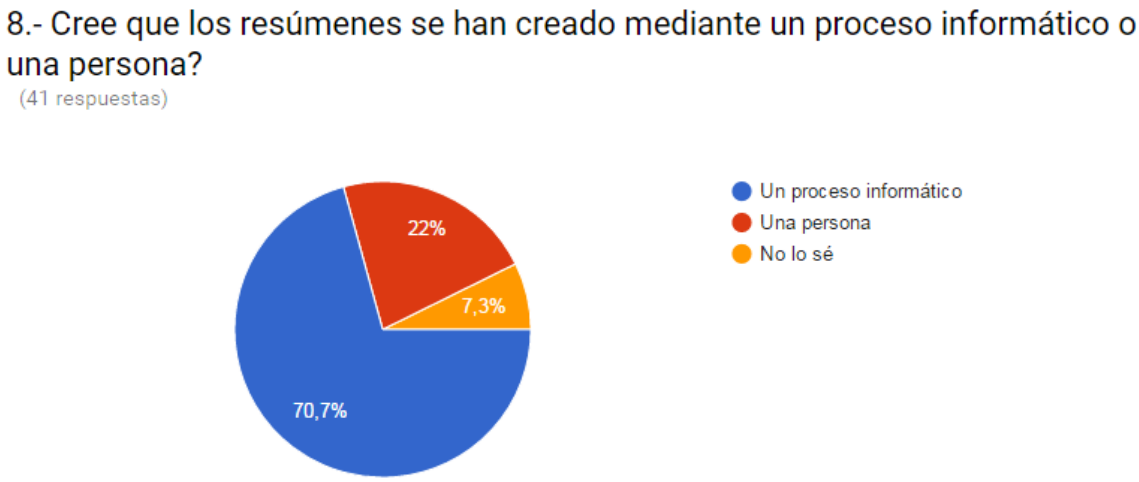


Figura 7.7: Gráfico para determinar si los usuarios distinguen el proceso automático de la generación de resúmenes, Alberto Esteban García

Los resultados del siguiente gráfico comportan que la aplicación tiene mucha utilidad y que tiene capacidad de mejora en varios aspectos, con ello el interés de su uso incrementaría.

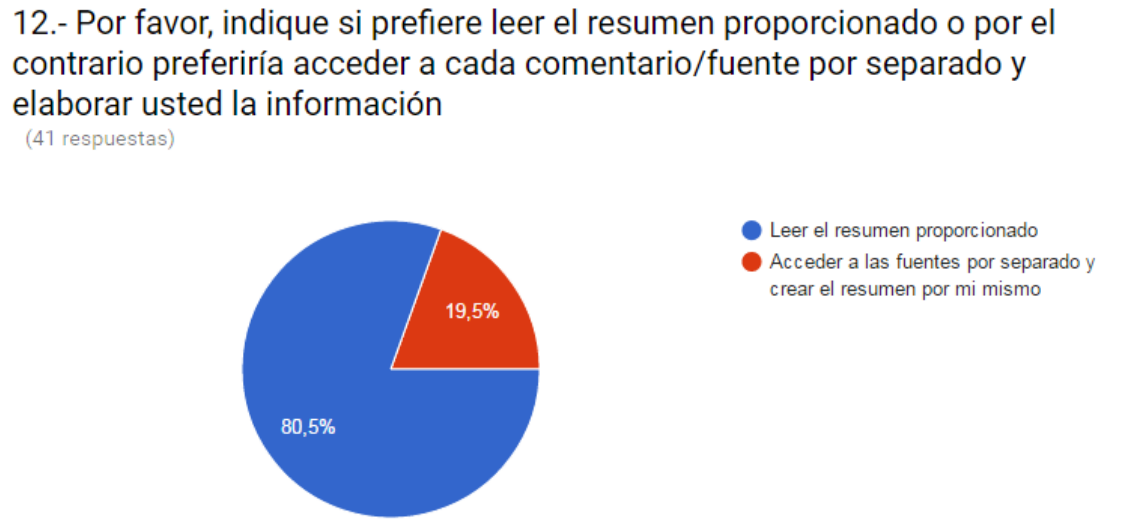


Figura 7.8: Gráfico sobre la preferencia de leer un resumen o elaborarlo manualmente, Alberto Esteban García

La evaluación de los resúmenes la podemos ver a continuación, el apartado más valorado fue la utilidad de los resúmenes alcanzando una nota media que supera el notable, sin embargo el punto más criticado es la existencia de errores ortográficos y gramaticales en los resúmenes.

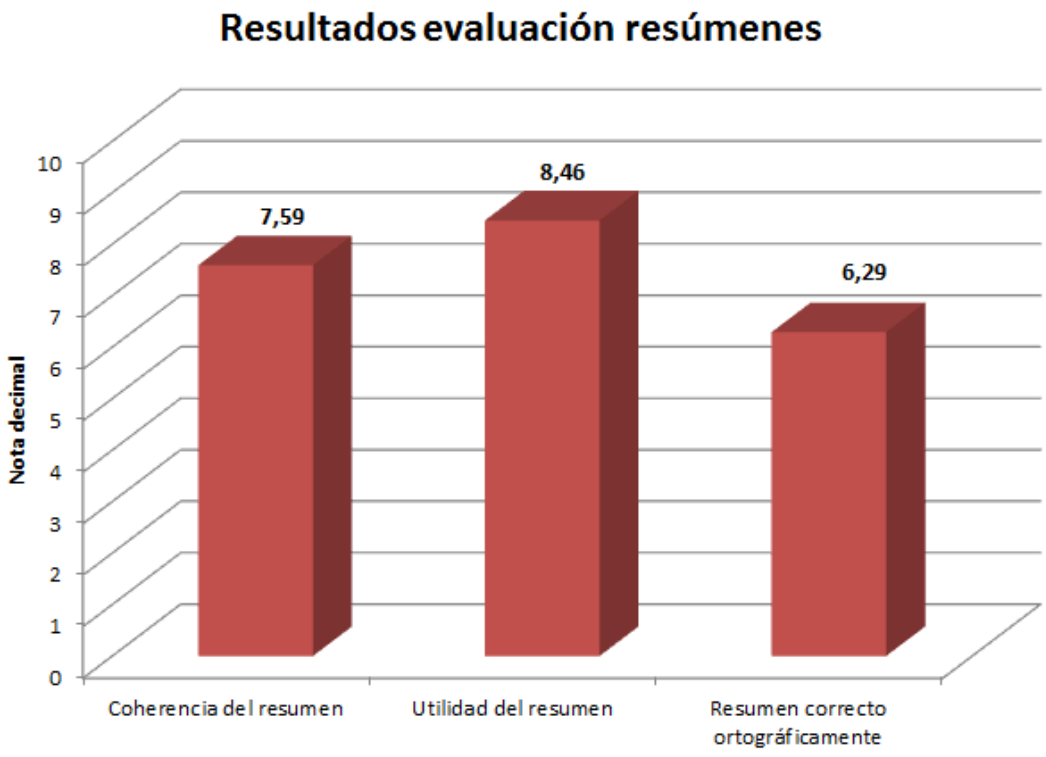


Figura 7.9: Gráfico sobre la evaluación de los resúmenes generados, Alberto Esteban García

En el caso de la evaluación de la interfaz, todos los apartados superan el notable, por lo que la mayoría de los usuarios están satisfechos con la interfaz desarrollada:

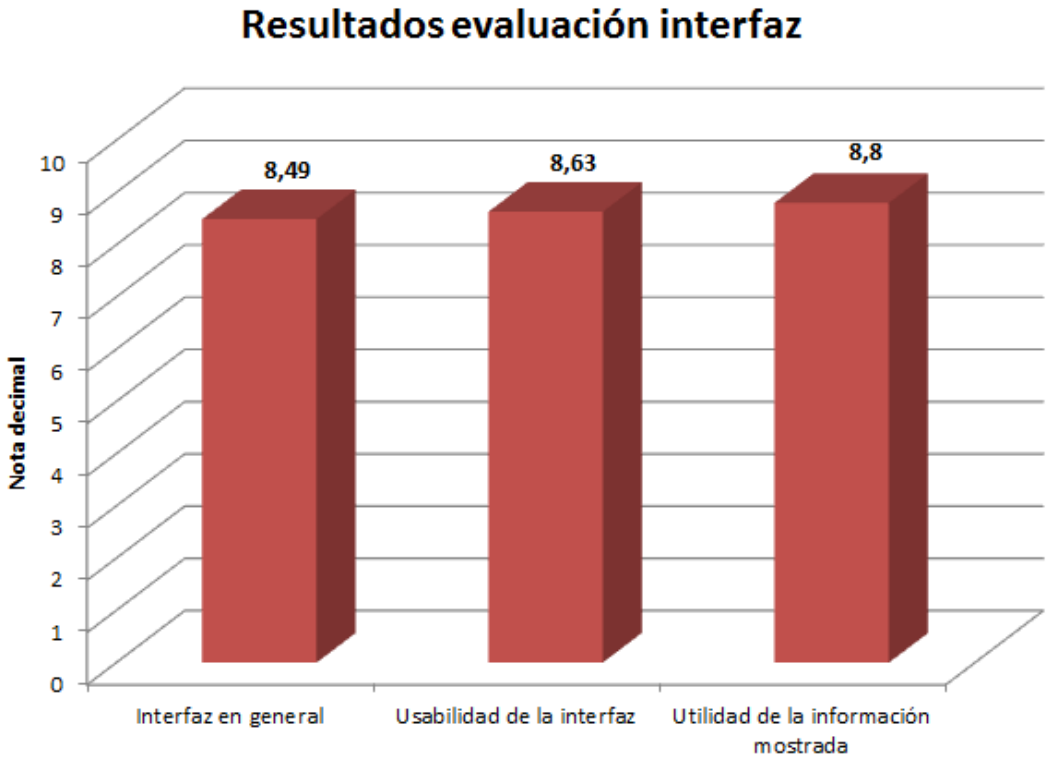


Figura 7.10: Gráfico sobre la evaluación de la interfaz, Alberto Esteban García

La siguiente gráfica muestra que el resumen mixto es el más consultado, seguido del positivo y finalmente el menos leído es el resumen negativo.

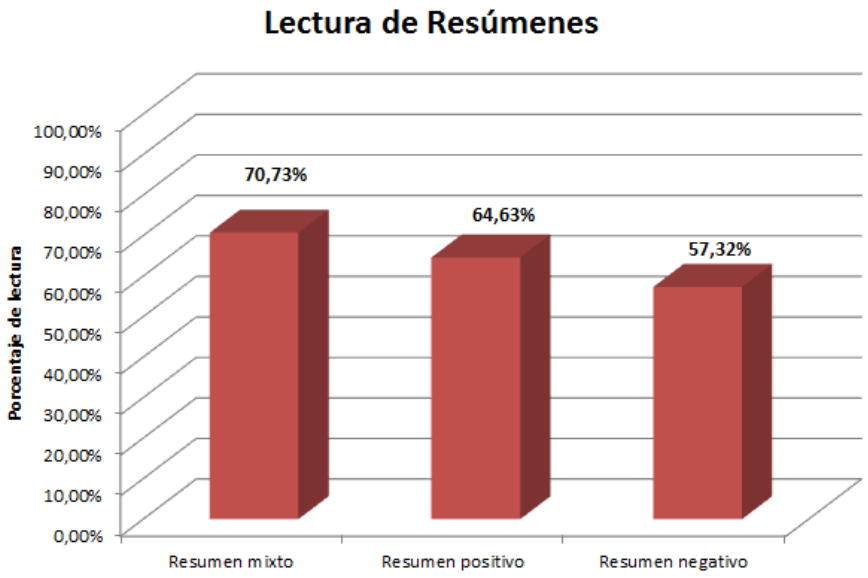


Figura 7.11: Gráfico sobre el porcentaje de lectura según el tipo de resumen, Alberto Esteban García

Como la evaluación fue realizada en su mayoría por usuarios residentes en Alicante, los hoteles y restaurantes más consultados se corresponden con establecimientos de la zona como se puede ver a continuación:

HOTELES	
Posición	**Nombre del hotel**
1	Meliá Alicante
2	Hotel Bonalba Alicante
3	NH Alicante

Figura 7.12: Tabla con los hoteles más consultados, Alberto Esteban García

RESTAURANTES	
Posición	**Nombre del restaurante**
1	La Taberna Del Gourmet
2	Katagorri
3	Yakitoro

Figura 7.13: Tabla con los restaurantes más consultados, Alberto Esteban García

8.- Análisis de errores de los resúmenes y posibles mejoras

En este apartado se van a exponer ciertos aspectos de la aplicación que podrían mejorarse atendiendo a la generación automática de los resúmenes así como a cambios en la interfaz desarrollada.

En líneas generales los resultados obtenidos son buenos, los usuarios opinan que la aplicación desarrollada es útil y que la interfaz implementada para mostrar los resúmenes generados de forma automática es fácil de utilizar.

En vista a los resultados obtenidos, en el apartado de la generación de resúmenes automática el punto más criticado ha sido la inclusión de faltas de ortografía provenientes de las fuentes de información, por ello se deberá hacer hincapié en mejorar este aspecto.

En el apartado de la interfaz gráfica, los resultados han sido buenos, pero los usuarios han comentado posibles mejoras que permitirán mejorar la experiencia de uso de la aplicación.

8.1.- Mejoras en la creación de los resúmenes

El apartado más criticado en la generación de los resúmenes son los errores ortográficos y gramaticales, es por ello que se deberían refinar tanto la extracción de los datos, intentando identificar informalidades y faltas de ortografía, así como la división en frases de las reseñas de *TripAdvisor* con el objetivo de minimizar los errores.

Otra propuesta podría ser añadir un factor de errores gramaticales en la ponderación de las frases para evitar que las frases con faltas de ortografía formaran parte de los resúmenes generados.

En el filtrado de la información extraída podemos evitar la duplicidad de información en la base de datos utilizando columnas adicionales para indicar la información que ha pasado el filtro en lugar de crear una nueva tabla con los mismos campos e introducir la información filtrada en ella.

El análisis de sentimiento de la herramienta del GPLSI utilizada para conocer el sentimiento de las frases en ocasiones falla, por lo que se podría hacer uso de otra herramienta adicional con el mismo cometido y si ante un mismo texto las dos herramientas obtienen el mismo resultado almacenar el sentimiento de la frase, de esta forma evitamos el uso exclusivo de una única herramienta.

En ocasiones el paso de algunos verbos a tercera persona ha comportado faltas de ortografía, en especial cuando el verbo es irregular, se debería utilizar una lista completa de los verbos irregulares conjugados, para que en el caso de detectarse la existencia de uno de ellos, se haga una sustitución simple, sin comportar ningún tipo de procesamiento que diera lugar a errores.

Una posible mejora en cuanto a espacio de almacenamiento podría ser la generación de cada resumen en el momento de su solicitud por parte del usuario, en lugar de estar almacenado en la base de datos, sin embargo esto implicaría un aumento considerable del tiempo de respuesta lo que repercutiría en un peor servicio al usuario.

Finalmente se podría realizar la inserción de todos los establecimientos y su información lanzando un proceso en un servidor.

8.2.- Mejoras en la interfaz gráfica

Gracias a los comentarios aportados por los usuarios al final del formulario contamos con varios cambios que podrían mejorar la interfaz:

- Introducir el buscador de establecimientos en la página del hotel o restaurante consultado para facilitar la búsqueda de otro establecimiento sin tener que volver a la página principal.

- Ampliar la información de cada establecimiento: horarios, año de creación, número de habitaciones en el caso de los hoteles, etc.

- Añadir la posibilidad de que en el mapa de *Google Maps* se pueda consultar como llegar a los establecimientos mostrados.

- Mejorar el apartado de fotografías de los establecimientos, por ejemplo, crear una galería de imágenes por cada establecimiento.

- Añadir el nombre de los establecimientos recomendados junto a sus respectivas imágenes.

- Añadir la búsqueda por ciudad.

- Mejorar el diseño *responsive*[22] de la interfaz gráfica.

[22] *Responsive: Filosofía de diseño y desarrollo cuyo objetivo es adaptar la apariencia de las páginas web al dispositivo que se esté utilizando para visualizarlas.*

9.- Conclusiones y trabajo futuro

Durante el desarrollo del proyecto hemos contado con numerosos retos los cuales hemos ido superando uno a uno, donde se han aplicado multitud de técnicas de PLN así como otras aprendidas durante la carrera.

El objetivo más importante que tenía en el desarrollo de mi Trabajo de Fin de Grado era el desarrollo de una aplicación que fuera útil y en la que evolucionara mis capacidades y siento que lo he conseguido.

En mi opinión la aplicación desarrollada tiene mucho futuro, en este caso nos hemos centrado en el sector turístico, pero existen mucho más ámbitos donde podría ser una utilidad sobresaliente, por ejemplo, enfocada en la generación de resúmenes de productos tecnológicos.

En multitud de portales de compras online (*Amazon*, *PCComponentes*, etc.) los clientes expresan su opinión de los productos adquiridos. Podemos utilizar esta información junto a la provista por el fabricante del producto y la recogida en páginas especializadas (en el caso de la tecnología: *Xataka*, *Android4All*, etc.) para elaborar un resumen que condense los mejores y peores aspectos y donde el usuario solo habrá invertido una pequeña parte de su tiempo en conocer el producto para decidir si es una compra acertada o no.

Como creo que este trabajo es fruto de un gran esfuerzo y puede ser de utilidad en un futuro, será presentado al certamen Arquímedes[23], convocado por el Ministerio de Educación, Cultura y Deporte (MECD) dado que pienso que debe ser propuesto para su reconocimiento. Posteriormente me gustaría exponerlo en un *workshop* o conferencia así como su envío en forma de artículo científico.

[23] http://www.mecd.gob.es/educacion-mecd/areas-educacion/universidades/convocatorias/estudiantes/certamen-arquimedes.html

10.- Referencias

Marimón-Llorca, C. *Temas de Análisis y Redacción de Textos*, 2008, http://hdl.handle.net/10045/4023

Moreno Boronat, L. *Introducción al procesamiento del lenguaje natural*, 1999, Publicaciones de la Universidad de Alicante

Elena Lloret, *Text Summarisation based on Human Language Technologies and its Applications*, 2011, http://www.dlsi.ua.es/~elloret/publications/elloret_tesis_FINAL_WEB.pdf

Gerani S., Mehand, Y, Carenini, G. Raymond, T Ng., Njeat B. *Abstractive Summarization of Product Reviews Using Discourse StructureProceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP*), *pages 1602–1613, October 25-29, 2014, Doha, Qatar.*

Draw.io, *Página Web*, https://www.draw.io/ [Accedida 28 de abril de 2016]

IBM*, Página Web*, https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/ [Accedida 6 de junio de 2016]

Wikipedia, *Página Web,* https://es.wikipedia.org/wiki/Web_2.0 [Accedida 6 de junio de 2016]

Onesecond, *Página Web,* http://onesecond.designly.com/ [Accedida 22 de mayo de 2016]

Twitter, *Página Web*, https://about.twitter.com/company [Accedida 23 de mayo de 2016]

TripAdvisor, *Página Web*, https://www.tripadvisor.com/PressCenter-c4-Fact_Sheet.html [Accedida 30 de mayo de 2016]

Wikipedia, *Página* Web, https://en.wikipedia.org/wiki/Stop_words [Accedida 22 de Mayo de 2016]

CSS Tricks, *Página Web*, https://css-tricks.com/ [Accedida 15 de mayo de 2016]

Text 2 Mind Map, *Página Web*, https://www.text2mindmap.com/ [Accedida 6 de junio de 2016]

Lunapic, *Página Web*, http://www191.lunapic.com/editor/ [Accedida 14 de mayo de 2016]

Stackoverflow, *Página Web*, http://stackoverflow.com/ [Accedida 5 de febrero de 2016]

MoocTLH, *Página Web*, http://mooctlh.uaedf.ua.es/course [Accedida 1 de marzo de 2016)

Ejocurio, *Página Web*, http://ejocurionline.net/data/wallpapers/16/DDW_927318.jpg [Accedida 16 de mayo de 2016]

11.- Anexo

11.1.- Datos de la extracción de información

A continuación se muestran los resultados de la fase de extracción de datos:

Número de Hoteles	180
Número de hoteles por ciudad	30
Número de restaurantes	180
Número de restaurantes por ciudad	30
Número de reseñas de *TripAdvisor*	91.505
Número de reseñas de *TripAdvisor* por hotel	235
Número de reseñas de *TripAdvisor* por restaurante	275
Número de *tweets*	78.713
Número de *tweets* por hotel	200
Número de *tweets* por restaurante	240

Figura 11.1: Tabla sobre los resultados obtenidos en la fase de extracción de datos, Alberto Esteban García

La información desagregada por ciudad puede verse a continuación:

HOTELES ALICANTE		
Nombre Hotel	**Reseñas**	**Tweets**
AC Hotel Alicante by Marriott	150	130
Albergue Juvenil Alicante Villa Universitaria	19	0
Areca Hotel	85	121
Barcelo Asia Gardens Hotel & Thai Spa	499	668
Boutique Hotel Sierra de Alicante	69	58
Campanile Alicante	65	134
Chameleon Hostel Alicante	16	58
ESTUDIOTEL ALICANTE	36	87
Eurostars Lucentum	213	243
Exe Alicante Hills	122	26

Explanada Hotel Alicante	15	18
Goya Hotel de Alicante	54	2
Holiday Inn Elche	64	102
Hospes Amerigo	133	588
Hotel Albahía	126	511
Hotel Alicante Golf	146	395
Hotel Bonalba Alicante	487	539
Hotel Cervantes	26	35
Hotel Les Monges Palace Boutique	33	0
Hotel Maya Alicante	155	496
Hotel Spa Porta Maris & Suites del Mar	195	231
Ibis budget Alicante	52	65
Melia Alicante	498	460
Mendez Nuñez	36	513
NH Alicante	277	615
NH Rambla de Alicante	182	6
Old Centre Inn Alicante	3	0
Pueblo Acantilado Suites	500	67
Servigroup Montiboli	192	18
Tryp Gran Sol	229	140

Figura 11.2: Tabla sobre el número de reseñas y *tweets* de los hoteles de Alicante, Alberto Esteban García

HOTELES VALENCIA		
Nombre Hotel	**Reseñas**	**Tweets**
AC Hotel Valencia by Marriott	212	21
Ayre Hotel Astoria Palace	168	471
Barcelo Valencia	373	512
Embassy Suites by Hilton Valencia-Downtown	352	12
Expo Hotel Valencia	274	477
Hesperia WTC Valencia	428	300
Holiday Inn Valencia	212	415
Hospes Palau de la Mar Hotel	206	201
Hotel Benetusser	129	606
Hotel Dimar	153	136
Hotel Las Arenas Balneario Resort	251	154
Hotel Meliá Valencia	498	626
Hotel Olympia	153	564
Hotel Zenit Valencia	106	30
ILUNION Aqua 3	184	1
ILUNION Aqua 4	254	52
Melia Plaza	111	149

NH Ciudad de Valencia	349	157
NH Valencia Center	352	323
NH Valencia Las Artes	498	417
NH Valencia Las Ciencias	390	25
Primus Valencia	317	383
Senator Parque Central Hotel	359	68
Sercotel Sorolla Palace	314	197
Silken Puerta Valencia	240	700
The Westin Valencia	181	600
Tryp Oceanic	498	410
TRYP Valencia Feria	116	15
Vincci Lys	169	216
Vincci Palace Valencia	127	176

Figura 11.3: Tabla sobre el número de reseñas y *tweets* de los hoteles de Valencia, Alberto Esteban García

HOTELES MADRID		
Nombre Hotel	**Reseñas**	**Tweets**
Ayre Gran Hotel Colon	494	0
Emperador Hotel Madrid	379	580
Eurostars Madrid Tower	496	499
Gran Melia Fenix	359	219
Hesperia Madrid	325	0
Hospedaje Romero	398	1
Hotel Liabeny	370	428
Hotel Mayorazgo	492	512
Hotel Paseo del Arte	433	456
Hotel Preciados	313	0
Hotel Regina	478	721
Melia Castilla	499	616
Melia Madrid Princesa	500	474
NH Madrid Alberto Aguilera	451	0
NH Madrid Nacional	443	344
NH Madrid Príncipe de Vergara	455	0
NH Madrid Ribera del Manzanares	498	222
NH Madrid Suecia	246	63
NH Madrid Zurbano	499	354
Novotel Madrid Puente de la Paz	319	0
Praktik Metropol	299	0
Room Mate Óscar	497	600
Senator Gran Via 70 Spa Hotel	439	37
Sercotel Suites Viena	258	0
TRYP Madrid Atocha	427	0

TRYP Madrid Centro	356	0
Tryp Madrid Cibeles	314	46
Vincci Capitol Hotel	392	0
Vincci Soho	334	0

Figura 11.4: Tabla sobre el número de reseñas y *tweets* de los hoteles de Madrid, Alberto Esteban García

HOTELES BARCELONA		
Nombre Hotel	**Reseñas**	**Tweets**
Alexandra Barcelona A DoubleTree By Hilton	225	93
Alimara Barcelona Hotel	108	0
Alma Barcelona	129	520
Aparthotel Silver	99	0
Axel Hotel Barcelona	278	519
Ayre Hotel Gran Via	191	0
Barcelo Sants	371	735
Catalonia Barcelona Plaza	295	540
Catalonia Born	168	0
Condes De Barcelona	177	518
El Balcon del Born	58	0
Eric Vokel Boutique Apartments - Gran Via Suites	137	0
Expo Hotel Barcelona	451	697
Gran Hotel La Florida	158	0
Gran Hotel Princesa Sofia	317	358
Gran Hotel Torre Catalunya	323	126
Hotel Arts Barcelona	225	0
Hotel Constanza Barcelona	170	0
Majestic Hotel & Spa Barcelona	133	359
Melia Barcelona Sarrià	373	360
Meliã Barcelona Sky	370	277
NH Barcelona Centro	218	204
NH Collection Constanza	428	200
NH Sants Barcelona	373	95
Novotel Barcelona City	123	0
Senator Barcelona Spa Hotel	264	354
Tryp Barcelona Aeropuerto	254	136
Tryp Barcelona Apolo	303	800
Tryp Condal Mar	241	0
W Barcelona	425	520

Figura 11.5: Tabla sobre el número de reseñas y *tweets* de los hoteles de Barcelona, Alberto Esteban García

HOTELES LONDRES

Nombre Hotel	Reseñas	Tweets
DoubleTree by Hilton Hotel London -Tower of London	98	0
Doubletree by Hilton London - Westminster	147	0
easyHotel London Earls Court	122	0
H10 London Waterloo	372	25
Harrington Hall Hotel	231	10
Hotel de Londres y de Inglaterra	500	415
Hotel The Caesar	178	59
Ibis Budget London Whitechapel Hotel	126	0
Ibis London Blackfriars	139	12
Ibis London City	133	35
Ibis London Earls Court	195	13
Ibis London Shepherds Bush	65	1
Luna & Simone Hotel	73	638
ME London	185	556
Melia White House	500	249
NH London Kensington	231	6
Novotel London Waterloo	99	3
Park Plaza County Hall London	256	2
Park Plaza Victoria London	101	15
Park Plaza Westminster Bridge London	399	3
Peckham Lodge	90	13
Phoenix Hotel	95	6
Premier Inn London Blackfriars (Fleet Street) Hotel	107	0
Premier Inn London County Hall Hotel	148	1
Premier Inn London Victoria Hotel	93	0
President Hotel	228	57
Royal Eagle Hotel	94	13
Royal National Hotel	480	198
Tavistock Hotel	222	10
Tune Hotel Kings Cross	129	10

Figura 11.6: Tabla sobre el número de reseñas y *tweets* de los hoteles de Londres, Alberto Esteban García

HOTELES ROMA		
Nombre Hotel	**Reseñas**	**Tweets**
Accademia Hotel	159	35
Artemide Hotel	162	88
Augusta Lucilla Palace	157	33
Ergife Palace Hotel	130	314

Eurostars International Palace	329	42
Eurostars Roma Aeterna	225	394
Exe Domus Aurea	148	7
Gran Melia Rome	154	210
H10 Roma Citta	140	639
Hiberia Hotel	109	55
Hotel Anglo Americano	119	50
Hotel De Petris	165	31
Hotel Diocleziano	104	34
Hotel Giolli Nazionale	132	4
Hotel Golden	149	103
Hotel Selene Roma	109	178
Hotel Trevi	149	345
Hotel Trevi Collection	133	13
Hotel Virgilio	126	65
iQ Hotel Roma	211	842
Mercure Roma Piazza Bologna	49	213
NH Collection Roma Giustiniano	238	27
NH Collection Roma Vittorio Veneto	157	30
NH Roma Leonardo da Vinci	150	39
NH Roma Villa Carpegna	79	119
Quirinale Hotel	149	277
Smart Hotel Roma	84	240
Starhotels Metropole	100	93
Twentyone Hotel	95	56
UNA Hotel Roma	116	630

Figura 11.7: Tabla sobre el número de reseñas y *tweets* de los hoteles de Roma, Alberto Esteban García

RESTAURANTES ALICANTE		
Nombre Restaurante	**Reseñas**	**Tweets**
Cactus Alicante	141	297
Casa Mia Italia	199	23
Cerveceria Sento	550	43
Cervecería ESTIU Bar	342	3
Daikichi	185	73
Darsena	264	533
Don Carlos Alicante	237	59
El Canto	146	173
El Portal Taberna & Wines	547	341
El suquet de castaños 16	128	3
Enso Sushi	78	27
Heladería Artesana Felici e Contenti	120	0

Horchateria Azul	100	107
Irreverente	185	89
L arruzz Alicante	150	72
L Atelier	158	51
La Barra de Cesar Anca	230	72
La Mary Restaurant Alicante	220	120
La taberna de Tito	166	32
La Taberna del Gourmet	533	384
Liberty Kitchen	187	115
Livanti Gelato di Sicilia	328	55
Monastrell	206	420
Nou Manolin	481	391
Peccati di Gola	152	26
Piripi	334	327
Restaurante Katagorri	142	75
Sale & Pepe Pizzeria - Barrio	207	508
Sudeste	144	121
Tribeca Music Bar	311	10

Figura 11.8: Tabla sobre el número de reseñas y *tweets* de los restaurantes de Alicante, Alberto Esteban García

RESTAURANTES VALENCIA		
Nombre Restaurante	**Reseñas**	**Tweets**
Al Pomodoro	272	113
Alma del temple	227	193
Alqueria del Pou	208	32
Arribar	146	580
BURGER BEER	212	61
Café de Las Horas	86	281
Canalla Bistro	550	346
Canela	338	478
Casa Carmela	384	157
Casa Roberto	258	332
Don Salvatore	177	68
Horchateria Daniel	328	186
Kamon	265	204
Komori	127	344
L Alquimista	163	68
La Cantinella	228	30
La Pappardella	295	156
La Pepica	550	513
Los Toneles	165	189
Mediterranea de Hamburguesas	223	194

Navarro	548	153
Nozomi Sushi Bar	179	429
Rausell	138	380
Restaurand Marchica	186	16
Restaurant Alqueria del Brosquil	305	1
Restaurante Commo Fusion	283	4
Ricard Camarena Restaurant	281	85
San Tommaso	501	31
The Sushi Room	240	93
Trattoria Napoletana Da Carlo	251	3

Figura 11.9: Tabla sobre el número de reseñas y *tweets* de los restaurantes de Valencia, Alberto Esteban García

RESTAURANTES MADRID		
Nombre Restaurante	**Reseñas**	**Tweets**
Alfredo s Barbacoa	406	760
Asador Donostiarra	532	520
Bodega de la Ardosa	292	159
Cafe del Art	550	41
Casa Hortensia	298	743
Casa Labra	550	600
Chocolateria San Ginés	550	520
Docamar	257	492
El Brillante	550	27
El Neru	332	449
El Paraguas	550	600
Goiko Grill	374	457
Grazie Mille	287	611
La Mallorquina	481	506
La Terraza de Oscar	289	280
Mad Cafe	213	654
Malacatin	386	245
Mercado de la Reina	548	820
Mercado de San Miguel	550	434
Miyama Flor Baja	353	49
Museo del Jamon	550	530
Museo del Jamon	550	530
Naomi	308	531
New York Burger General Yagüe	420	115
Platea Madrid	547	527
Restaurante Filandon	550	130
Restaurante La Tragantúa	546	20
Ten Con Ten	550	156

Trattoria Malatesta	402	90
Yakitoro	550	539

Figura 11.10: Tabla sobre el número de reseñas y *tweets* de los restaurantes de Madrid, Alberto Esteban García

RESTAURANTES BARCELONA		
Nombre Restaurante	**Reseñas**	**Tweets**
4 gats	428	520
Bacoa Kiosko	540	37
Bar Tomas	327	840
Cera 23	545	241
Cerveceria Catalana	550	760
DelaCrem	208	205
Domino Bar	520	800
El Nacional Barcelona	550	500
Gusto	364	579
Igueldo	280	84
La Bella Napoli	322	285
La Esquinica	402	509
La Flauta	550	660
Martinez	316	649
Matsuri	307	460
Mirablau	155	680
Moritz	548	360
Paco Meralgo	458	553
Petit Bangkok	284	162
Piazze D Italia	403	54
Ramen-ya Hiro	335	491
Restaurante Arume	454	25
Restaurante Taktika Berri	270	17
Shunka	463	538
Sports Bar Italian Food	192	5
Tanta Barcelona	517	711
Teresa Carles	550	640
Thai Barcelona | Royal Cuisine	505	0
Tlaxcal	215	496
Vinitus Barcelona	291	164

Figura 11.11: Tabla sobre el número de reseñas y *tweets* de los restaurantes de Barcelona, Alberto Esteban García

RESTAURANTES LONDRES		
Nombre Restaurante	**Reseñas**	**Tweets**

Ametsa with Arzak Instruction	88	13
Angus Steakhouse	117	24
Aqua Shard	78	208
Baileys Fish and Chips	103	0
Barrafina	77	43
Bilbao Berria	48	67
Byron Westbourne Grove	38	0
Casa Brindisa	63	5
Centro Galego de Londres	14	110
Dickens Inn	96	32
Five Guys	95	213
Garfunkel s	79	32
Goodman	66	163
Hispania London	49	55
Iberica Marylebone	94	21
Jamon Jamon - Soho	80	1
Londres	24	529
Misato Japanese	70	0
Oxo Tower Restaurant, Bar & Brasserie	123	274
Pizza Hut	103	418
Pret a Manger	148	494
Rules Restaurant	91	33
The Five Fields	65	0
The Wolseley	78	120
Three O Two at H10 London Waterloo	48	0
Tokyo Diner	62	6
Vapiano	123	48
Vapiano Bankside	67	0
Veeraswamy	67	19
Wahaca Soho	70	10

Figura 11.12: Tabla sobre el número de reseñas y *tweets* de los restaurantes de Londres, Alberto Esteban García

RESTAURANTES ROMA		
Nombre Restaurante	**Reseñas**	**Tweets**
Bar Sant Eustacchio il Caffe	122	1
Cajo e Gajo	202	14
Cantina e Cucina	278	124
Carlo Menta Talevi Luigi e Luciano	550	1
Gelateria La Romana	198	38
Gelateria Valentino	449	43
Giolitti	379	520
Grazia & Graziella	383	0

Il Gelato di San Crispino	142	158
L Antica Birreria Peroni	158	157
La Fontana di Venere	207	22
La Gelateria Frigidarium	272	3
La Montecarlo	346	732
La Prosciutteria Trevi	181	14
La Taverna dei Fori Imperiali	124	40
Life	147	520
Likeat	152	31
Luzzi	303	0
Mercado Roma	251	491
Navona Notte	152	143
Old Bridge Gelateria	152	64
Palazzo del Freddo Giovanni Fassi	112	186
Piccolo Buco	208	35
Pinsere	214	129
Pompi	232	556
Ristorante Alessio	154	68
Roma Lleida	131	620
Roma Mia	193	460
Spaghetteria L Archetto	211	46
Trattoria Vecchia Roma	161	250

Figura 11.13: Tabla sobre el número de reseñas y *tweets* de los restaurantes de Roma, Alberto Esteban García

11.2.- Formulario de evaluación

A continuación se muestran las preguntas planteadas y sus posibles respuestas:

1. *¿De qué forma busca información acerca de un hotel o un restaurante?*
    - *Internet (TripAdvisor, Booking, Twitter, Foros, etc)*
    - *Pregunto a familiares, amigos, conocidos, etc.*
    - *Acudo a establecimientos especializados (agencia de viajes, etc)*

2. *¿En el caso de TripAdvisor, considera útil las opiniones de los clientes?*
    - *Sí*
    - *No*

3. *¿En el caso de Twitter, considera útil las opiniones que se pueden ver sobre hoteles y/o restaurantes?*
    - *Sí*
    - *No*

4. *¿Considera que en TripAdvisor existen muchas opiniones y es inviable leerlas todas?*

- *Sí*
- *No*

5. *¿Considera que en Twitter existe información de todo tipo: opiniones, recomendaciones, críticas, etc.?*
    - *Sí*
    - *No*

6. *¿Considera útil poder ver un resumen de las opiniones con el fin de poder reconocer fácilmente lo mejor y lo peor de cada establecimiento utilizando la información de varias fuentes, por ejemplo, TripAdvisor y Twitter?*
    - *Sí*
    - *No*

7. *Por favor, entre en la siguiente dirección web http://travelsum.gplsi.es y busque 2 hoteles y 2 restaurantes e indique qué hotel ha consultado y qué resúmenes ha leído de cada hotel*
    - *Nombre Hotel 1*
    - *¿Qué resumen ha leído del Hotel 1? (puede seleccionar más de una opción)*
        - *Resumen mixto*
        - *Resumen positivo*
        - *Resumen negativo*

    - *Nombre Hotel 2*
    - *¿Qué resumen ha leído del Hotel 2? (puede seleccionar más de una opción)*
        - *Resumen mixto*
        - *Resumen positivo*
        - *Resumen negativo*

    - *Nombre Restaurante 1*
    - *¿Qué resumen ha leído del Restaurante 1? (puede seleccionar más de una opción)*
        - *Resumen mixto*
        - *Resumen positivo*
        - *Resumen negativo*

    - *Nombre Restaurante 2*
    - *¿Qué resumen ha leído del Restaurante 2? (puede seleccionar más de una opción)*
        - *Resumen mixto*
        - *Resumen positivo*
        - *Resumen negativo*

8. *¿Cree que los resúmenes se han creado mediante un proceso informático o una persona?*
    - *Un proceso informático*
    - *Una persona*
    - *No lo sé*

9. Por favor, valore en general la coherencia de los resúmenes leídos
    - Escala de 0 a 10 donde 0 es nada coherente y 10 muy coherente

10. Por favor, valore en general la utilidad de los resúmenes leídos
    - Escala de 0 a 10 donde 0 es nada útil y 10 muy útil

11. Por favor, indique si ha encontrado errores ortográficos o gramaticales en los resúmenes leídos
    - Escala de 0 a 10 donde 0 es muchos errores y 10 ningún error

12. Por favor, indique si prefiere leer el resumen proporcionado o por el contrario preferiría acceder a cada comentario/fuente por separado y elaborar usted la información:
    - Leer el resumen proporcionado
    - Acceder a las fuentes por separado y crear el resumen por mí mismo

13. Por favor, valore la interfaz de manera general
    - Escala de 0 a 10 donde 0 es mala interfaz y 10 buena interfaz

14. Por favor, valore la usabilidad de la interfaz
    - Escala de 0 a 10 donde 0 es difícil de utilizar y 10 es fácil de utilizar

15. Por favor, valore si es útil la información mostrada en la interfaz
    - Escala de 0 a 10 donde 0 es nada útil y 10 muy útil

16. Si desea realizar alguna sugerencia o comentario, adelante :)