



UNIVERSIDAD DE JAÉN
EPSJ

Trabajo Fin de Grado

DETECCIÓN DE PERFILES DE USUARIO EN EL DOMINIO TURÍSTICO

Alumno: Susana Torres Moreno

Tutor: Prof. Doña M^a Teresa Martin Valdivia
Doña Salud M^a Jiménez Zafra

Dpto: Informática

Junio, 2017



Universidad de Jaén
Escuela Politécnica Superior de Jaén
Departamento de Informática

Doña M^a Teresa Martin Valdivia y Doña Salud M^a Jiménez Zafra, tutoras del Proyecto Fin de Carrera titulado: Detección de perfiles de usuario en el dominio turístico, que presenta Susana Torres Moreno, autoriza su presentación para defensa y evaluación en la Escuela Politécnica Superior de Jaén.

Jaén, Junio de 2017

El alumno:

Los tutores:

Fdo: Susana Torres Moreno

Fdo: Doña M^a Teresa Martin Valdivia
Fdo: Doña Salud M^a Jiménez Zafra

ÍNDICE

Capítulo 1 :Introducción

1. Introducción al proyecto.....	1
2. Motivación.....	2
3. Propósito.....	4
4. Objetivos	5
5. Resultados esperados.	5
6. Planificación	6
6.1. Estimación de costes	7
6.1.1. Costes Hardware	7
6.1.2. Costes Software.....	8
6.1.3. Costes de personal.....	8
6.1.4. Costes Adicionales.....	10
6.1.5. Costes Totales	10

Capítulo 2: Análisis de perfiles de viajeros

1. Introducción.....	13
2. Estudio de sitios web turísticos.....	15
2.1. Booking.....	15
2.2. Wimdu	17
2.3. Airbnb	18
2.4. El tenedor	20
2.5. TripAdvisor.....	22
3. Portal web seleccionado.....	24
4. Extracción de información.....	26
5. Análisis de la información	29
6. Detección automática del tipo de viajero	30

Capítulo 3 : Desarrollo del proyecto

1. Descripción del problema.....	41
2. Objetivos del sistema.....	41
3. Metodología para el desarrollo de software y método a seguir	42
3.1. Metodología ágil elegida	43
4. Historias de usuario.....	45

5. Propuesta de solución	52
6. Justificación de la solución	52
7. Descripción de la solución	53
7.1. Iteración 1.....	55
7.2. Iteración 2.....	58
7.4. Iteración 4.....	67
7.5. Iteración 5.....	71
7.6. Iteración 6.....	73
7.7. Iteración 7.....	75
8. Herramientas para la implementación del sistema	76
8.1. Lenguajes de programación.....	76
8.2. Herramientas	76
8.3. Desarrollo página web	79
Capítulo 4: Conclusiones y trabajo futuro	
1. Conclusiones	83
2. Trabajo futuro	85
Bibliografía.....	86
Anexo A: Manual de instalación	89
Anexo B: Manual de usuario	93
Anexo C:Índice de ilustraciones	102

Capítulo 1 :

Introducción

1. Introducción al proyecto

La decisión de muchas personas a la hora de elegir destino o alojamiento es influenciada por la reputación en línea de estos (Mir et al.,2015). Una de las primeras menciones a este concepto de reputación online es la de Chun y Davies (2001) que la definen como “la reputación generada a partir de las Tecnologías de la Comunicación y de la Información”.

La gran cantidad de información existente en la Web acerca de los alojamientos gracias a las opiniones de los huéspedes, también ha generado una preocupación por parte de los alojamientos al no ser capaces de controlar la imagen que se ofrece de ellos en Internet. Ya que se ha demostrado que la presencia de éstos y su buen uso de las redes sociales es determinante en cuanto a su éxito y puntuación. (Leung, Bai and Stahura, 2013; Mellinas, Martínez and Bernal, 2016)

Por todo esto el objetivo de las empresas es generar conocimiento a partir de toda esta información. Conocimiento que les permita tomar decisiones más acertadas partiendo no solo de sus propios datos sobre los clientes sino también de las opiniones y valoraciones de sus clientes en la web.

A la hora de interpretar esta información surge un problema, los sistemas de información tienen la capacidad de procesar una gran cantidad de información siempre y cuando esté estructurada y tenga un único sentido. Sin embargo, el conflicto empieza cuando se intentan procesar opiniones, es decir, el lenguaje natural, ya que una máquina por sí sola es incapaz de procesar el lenguaje dándole un sentido y un contexto. Para ello, cada vez más se está estudiando sobre un campo de la informática denominado Procesamiento del Lenguaje Natural (PLN). El PLN es una disciplina de la Inteligencia Artificial cuyo objetivo es hacer posible la comprensión y procesamiento de información expresada en lenguaje humano por un ordenador.

Este proyecto se centrará en el uso de técnicas de PLN aplicadas al sector del turismo. En concreto se pretende hacer una aplicación que muestre

estadísticas basadas en las opiniones de los viajeros de un sitio web turístico con el objetivo de que estas estadísticas ayuden a las personas en la toma de decisiones.

Se ha elegido este tema por la relevancia que tiene este sector en el desarrollo de la economía de un país.

2. Motivación

El turismo es un sector clave tanto en el desarrollo económico de un país como en la generación de empleo, de hecho genera 1 de cada 10 empleos y aumentó el PIB global en un 3,1% en 2016 según el informe del World Travel & Tourism Council¹ desarrollado este año.

De entre todos los países, España se ha convertido en un país con una gran potencia turística. En 2013 estaba situado en cuarta posición en cuanto a países con una economía mejor preparada para el turismo, y su evolución ha sido tal que en 2015 ya estaba situado en primera posición, siendo también el sexto en cuando a países que le dan mayor prioridad a los viajes y al turismo según el ranking mundial de competitividad turística que desarrolló el Foro Económico Mundial en 2015².

En España, el PIB turístico ha acelerado su crecimiento hasta el 4,9% en 2016³, lo que supone un aumento de 6.414 millones de euros de actividad hasta alcanzar los 125 mil millones, que permite que la participación del turismo en la economía española ascienda al 11,2% del PIB.

Dentro de España, este proyecto se centrará en una de sus comunidades autónomas, Andalucía, que fue la tercera comunidad más visitada por los turistas en 2015 según el INE⁴.

¹ <https://www.wttc.org/-/media/files/reports/economic-impact-research/2017-documents/global-economic-impact-and-issues-2017.pdf>

² <http://ep00.epimg.net/descargables/2015/05/06/f2c1e517ce56a6453171d3a9d1b1da05.pdf>

³ <http://www.exceltur.org/wp-content/uploads/2017/01/Informe-Perspectivas-N59-Balance-del-a%C3%B1o-2016-y-previsiones-para-2017-web.pdf>

⁴ <http://www.ine.es/daco/daco42/frontur/frontur1215.pdf>

El auge de las Tecnologías de la Información y la Comunicación (TIC) ha cambiado totalmente la forma de interactuar entre las personas y la sociedad. Las agencias de viaje online (OTAs) están empezando a sustituir a las agencias de viajes tradicionales. En ellas no sólo disponemos de información acerca del hotel, sino que además contamos con una enorme cantidad de opiniones acerca del alojamiento y de sus diferentes características. Esto es debido a la llegada de la Web 2.0, cuyo propósito principal fue el intercambio ágil de información entre los usuarios. O'Reilly (2005), precursor del concepto web 2.0, la definió como el conjunto de aplicaciones que tratan de abarcar la red, entendiendo cómo funcionan los efectos de la red y aprovechándolos en todo lo que se hace. Su especificación dentro del mundo del turismo ha dado lugar a un nuevo fenómeno, el travel o turismo 2.0. El travel 2.0 es un término que representa la adaptación del concepto de Web 2.0 a la industria de los viajes y el turismo, el cual ha supuesto un cambio de enfoque para este sector. Los clientes se han vuelto más exigentes, quieren servicios personalizados, flexibles y además cuentan con la posibilidad de compartir su experiencia con otros viajeros por internet. Los consumidores se imitan unos a otros, y estas opiniones que comparten dirigen o alejan a otros usuarios hacia un servicio. Este efecto tiene una gran influencia en la toma de decisiones de un usuario (Aguilar, Gutiérrez y Hernanz, 2014; Reza, Shekarchizadeh y Samiei, 2010). Se denominó por (Dellarocas, 2003), como "Boca a Boca" (Word of Mouth) adaptado al dominio de internet. Podemos definir este fenómeno como una comunicación virtual entre consumidores que nunca se han conocido en la vida real (Gruen et al., 2006; Park and Lee, 2009).

Según el tripBarometer sobre tendencias de viaje en 2016⁵, "El 93 % de los hoteleros afirma que las opiniones online de los viajeros son importantes para el futuro de sus negocios". Es más, según este mismo informe, el área de mayor inversión para los hoteleros en 2016 fue la gestión de la reputación online como se puede ver en la Ilustración 1.1.

⁵https://d2bxpc4ajzxry0.cloudfront.net/TripAdvisorInsights/sites/default/files/downloads/2670/globalreport_spanish_dec2015_0.pdf

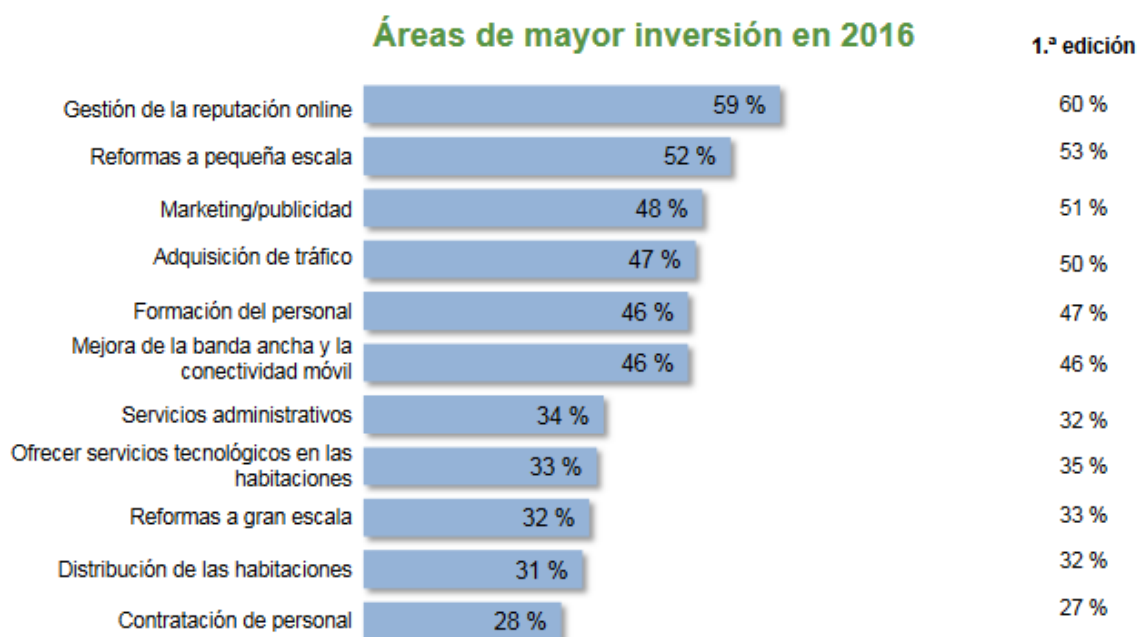


Ilustración 1.1. Inversiones en 2016

Por tanto, es creciente la preocupación de los propios alojamientos por conocer a sus clientes y atraerlos con un servicio más personalizado. Para ello les podría interesar conocer qué tipos de viajeros se alojan en cada alojamiento, en qué época del año, y sobre qué temas se preocupan más con el objetivo de organizar actividades orientadas hacia ese sector en concreto.

Debido a todo lo mencionado anteriormente, el proyecto se centrará en analizar las preferencias turísticas y los diferentes perfiles de viajeros de Andalucía en las distintas épocas del año.

3. Propósito

El propósito de este proyecto es aplicar técnicas de Recuperación de Información y de Procesamiento del Lenguaje Natural con el fin de extraer, analizar y sacar conclusiones de las opiniones de los viajeros de un portal web turístico.

Se pretende crear una herramienta que identifique automáticamente el tipo de viajero que escribe las opiniones, además de una aplicación que

muestre información útil relacionada con el tipo de viajero y la época del año en la que viaja, la polaridad de las valoraciones de los comentarios y otros datos de interés que puedan ayudar a la toma de decisiones de una entidad a ofrecer un servicio más personalizado a sus clientes.

4. Objetivos

Los objetivos del trabajo son:

- Realizar un estudio sobre distintos portales web relacionados con el dominio turístico.
- Seleccionar la fuente de información a tratar y extraer opiniones de la misma.
- Desarrollar una herramienta que identifique de forma automática el tipo de viajero que escribe las opiniones y que utilice esta información para mejorar la experiencia de usuario de otros viajeros.
- Desarrollar una aplicación que muestre los datos extraídos con gráficos que ayuden a sacar conclusiones sobre estos datos.
- Redactar una memoria que recoja todo el trabajo desarrollado así como los manuales de instalación y de usuario.

5. Resultados esperados.

Tras finalizar el proyecto, se deben obtener como resultados:

- Estudio de los diferentes portales web turísticos.
- Conjunto de datos formado por las opiniones de un portal web turístico.
- Detector automático del tipo de viajero dada una opinión.
- Aplicación web que muestre gráficos y estadísticas según los datos extraídos.
- Memoria del proyecto.

6. Planificación

La planificación temporal del proyecto se representará mediante un diagrama de Gantt que reflejará el tiempo estimado para cada una de las tareas que habrá que desempeñar en este proyecto.

El diagrama representará las tareas en el período de tiempo comprendido entre el 31 de Enero de 2017 y el 31 de mayo de 2017, es decir, se estima que el desarrollo del proyecto se completará en cuatro meses (Ilustración 1.2. y 1.3)



Ilustración 1. 2. Diagrama de Gantt

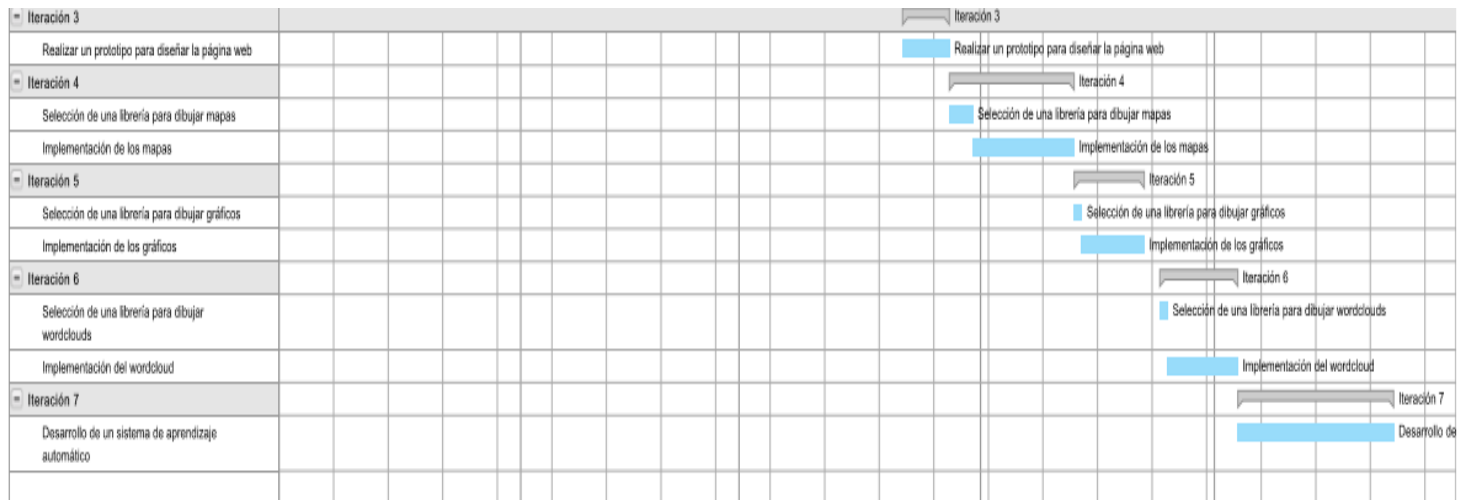


Ilustración 1.3. Diagrama de Gantt

6.1. Estimación de costes

La estimación de costes de un proyecto software es fundamental antes de su puesta en marcha. Esta estimación consiste en identificar los recursos que serán necesarios para llevar a cabo el proyecto en el tiempo deseado. Para ello se calcularán los costes hardware, software y de personal.

6.1.1. Costes Hardware

El único coste hardware será un equipo, en este caso pondré las especificaciones del que he usado pero podría usarse cualquier otro modelo para el desempeño del proyecto.

- Portátil ASUS-F552C 609€
- Procesador: Intel core i7-3537u
- Memoria Ram : 6GB
- HDD : 1TB
- Tarjeta gráfica: NVIDIA GeForce 710M
- SSOO : Windows 10

Se estima una vida útil al ordenador de cinco años. Esto supone un coste de 121,8€ al año, que serían 10,15€ al mes. Como el proyecto tendrá una duración de cuatro meses el coste del equipo sería de 40,6€. Como no se ha necesitado ningún componente hardware más, los costes hardware del proyecto serían 40,6€.

6.1..2. Costes Software

Los costes software en este caso serán las licencias de los programas que se han necesitado para el desarrollo del proyecto.

- Sistema Operativo Windows 10 0 € (licencia Universidad de Jaén)
- Java SE Development Kit 7 0 €
- JetBrains PyCharm Community Edition 0 €
- MySQL workbench 6.3 CE 0€

Paquete Microsoft Office 365 10€/mes, lo cual supondría un gasto de 40€ ya que la duración del proyecto es de cuatro meses.

6.1..3. Costes de personal

Los costes de personal de un proyecto software son los salarios del personal necesario para llevar a cabo el proyecto. En este caso el proyecto será desarrollado por una única persona pero que realizará diferentes perfiles.

Los perfiles necesarios para este proyecto serán:

- Programador : Pertenece al grupo II, técnico con un salario base de 20.734€
- Analista de sistemas: Pertenece al grupo II, analista con un salario base de 25.830€

La descripción del salario de estos perfiles aparece en el BOE en la resolución del 3 de junio de 2013, de la Dirección General de Empleo, del VIII Convenio colectivo de Telefónica Soluciones de Informática y Comunicaciones de España. Según este documento la jornada laboral con carácter general para

todos los empleados será de 39 horas semanales, lo cual supondría un total de 7,8h diarias.

Como en el desarrollo de este proyecto se han empleado 5h diarias en lugar de 7,8h se ha realizado una regla de tres para calcular el sueldo que equivaldría a trabajar 25h semanales. Que sería 16557,69€ al año en el caso del analista y 13291,02€ al año en el caso del programador.

Para calcular el coste por hora de cada tarea se ha dividido el salario anual de cada perfil entre 12 meses. El coste resultante se ha dividido entre 4 semanas que tiene cada mes y este coste entre 5 días laborables que tiene cada semana. Para conseguir el coste por día se ha dividido a su vez este resultado entre 5h trabajadas por día.

Teniendo en cuenta esto y la estimación en horas de cada tarea, obtendremos los siguientes resultados (Ilustración 1.4.):

Tarea	Perfil	Días	h	Salario/h			
Estudio portales web turísticos	Analista	4	20	176,898397			
Análisis de las ventajas de relacionar una opinión con el tipo de viajero	Analista	3	10	88,4491987			
Selección del portal web	Analista	2	10	88,4491987			
Selección del periodo de tiempo y de las ciudades a estudiar	Analista	2	10	88,4491987			
Estudio de la estructura de la página seleccionada para implementar un crawler	Programador	2	10	70,9990385			
Implementación del crawler	Programador	9	45	319,495673			
Realizar un prototipo para diseñar la página web	Analista	4	20	176,898397			
Definir la estructura de la base de datos	Analista	1	5	44,2245994			
Carga de información a la base de datos	Programador	1	5	35,4995192		Salarios Reales	
Selección de una API para dibujar mapas	Programador	3	15	106,498558		25.830	
Implementación de los mapas	Programador	9	45	319,495673		20.734	
Selección de una API para dibujar gráficos	Programador	1	5	35,4995192			
Implementación de los gráficos	Programador	6	30	212,997115			
Selección de una API para dibujar un wordcloud	Programador	1	5	35,4995192			
Implementación del wordcloud	Programador	7	35	248,496635			
Desarrollo de un sistema de aprendizaje automático	Programador	10	50	354,995192			
Redacción de la memoria	Analista	durante todo el proyecto	-	-			
Salario Analista al año	16557,69	Salario al mes/semana/día/h	1379,8075	344,951875	68,990375	8,844919872	
Salario Programador al año	13291,02	Salario al mes/semana/día/h	1107,585	276,89625	55,37925	7,099903846	
h al día	5						

Ilustración 1.4. Planificación de tareas y costes asignados

Sumando los costes de cada tarea obtendríamos un coste final de 2447,0€ ,obviando la tarea "Redacción de la memoria" ya que esta se realizará durante el desarrollo del proyecto.

6.1..4. Costes Adicionales

Además de los costes mencionados anteriormente, necesitaríamos contratar conexión a Internet. En este caso el coste de este servicio sería de 14,90€/mes por 50MB de fibra de ONO, lo cual supondría un total de 59,6€ ya que la duración del proyecto es de cuatro meses.

6.1..5. Costes Totales

Los costes totales del proyecto serían la suma de los costes hardware, software, de personal y los adicionales, que serían 2546,6€ según las estimaciones realizadas. Más el 10% de beneficio lo que supondría una cantidad de 2801,26€.

7. Metodología a desarrollar.

En esta memoria se desarrollará el siguiente contenido:

En el capítulo dos se estudiarán los diferentes sitios web turísticos que se han contemplado como posibles fuentes para extraer información. Se justificará la elección del portal web seleccionado así como el tiempo y las ciudades en las cuales se va a centrar el proyecto. Se explicarán las técnicas de extracción de información y las técnicas de análisis de información utilizadas, junto al sistema de detección automática del tipo de viajero.

En el capítulo tres se describirá la metodología que se ha seguido durante el desarrollo del proyecto, la descripción del problema a resolver, los objetivos del proyecto, las historias de usuario, la solución propuesta junto a su justificación y descripción. Además de la descripción de las herramientas utilizadas para desarrollar el proyecto.

En el capítulo cuatro se explicarán las conclusiones que se han sacado tras desarrollar el proyecto y las posibles vías de trabajo futuro que se contemplan tras la finalización de este trabajo.

Capítulo 2:

Análisis de perfiles de viajeros

1. Introducción

Según el I Barómetro Turístico de Brain Trust⁶ "El 59% de los viajeros está dispuesto a ceder información a cambio de servicios personalizados". En este mismo estudio también aparece el dato de que aproximadamente el 70% de los viajeros creen que el uso del Big Data por parte de las empresas ayudará a personalizar los servicios turísticos.

Estos datos arrojan la idea de que los viajeros desean un trato personalizado. Una manera de ofrecer este servicio podría ser organizando servicios para cada perfil de viajero. Hay diferentes perfiles de viajeros con diferentes intereses entre sí. Se podrían dividir en diversas categorías pero en este proyecto sólo se contemplarán los perfiles: en pareja, con la familia, con los amigos, solo y de negocios.

Algunos de los intereses de estos viajeros según el informe "Applying Consumer Insights To Attract North American Visitors"⁷ y un reportaje sobre la importancia de conocer al cliente en el turismo⁸ son los siguientes:

- Las parejas maduras prefieren socializar con otros viajeros o gente de la ciudad, uno de sus principales objetivos es conocer gente nueva, prefieren hacer turismo y no están interesados en escapar de su vida y vivir nuevas experiencias.
- El principal interés de las familias es conectar unos con otros y crear recuerdos duraderos. Buscan seguridad, comodidad y organización de una manera más estricta que otros perfiles ya que el viaje involucra a la familia entera y por tanto a sus hijos.

⁶ <http://braintrust-cs.eu/blog-turismo-ocio/>

⁷

https://www.tourismpartners.com/publications/Research/reas_segrpts_ConsumerInsightResearchSegmentationOverview.pdf

⁸ <http://www.uhu.es/uhutur/documentos/npturisticos1/5%20Turismo%20especializado.pdf>

- Los viajeros de negocios se preocupan primordialmente por la capacidad de desarrollar su trabajo de forma cómoda y con eficacia. No es tan importante el alojamiento sino la tranquilidad, situación y no ser molestados por otros viajeros. Además suelen ser personas que pasan demasiado tiempo fuera de casa por tanto tienen necesidades específicas como pueden ser el interés por que haya gimnasio en el hotel, comida sana y un baño bien equipado.
- las personas que viajan con amigos prefieren viajar con amigos cercanos, son extrovertidos y enérgicos durante las vacaciones y comparten sus viajes y experiencias en redes sociales.

Esto es sólo un ejemplo de la diferencia de intereses que hay de un colectivo a otro, por todo esto sería conveniente diferenciar entre perfiles de viajeros a la hora de realizar actividades para cada grupo.

Con el objetivo de sacar información que ayude a conocer los perfiles más comunes en cada ciudad y en cada época del año se extraerá información de un portal web de turismo y se sacarán conclusiones y estadísticas que ayuden a la toma de decisiones con información de cada perfil.

En este capítulo se hablará de los diferentes portales web que se han contemplado como posibilidades para extraer información para este proyecto, y se justificarán los motivos por los cuales se ha elegido uno en concreto. Además, también se hablará de las técnicas de recuperación de información y las técnicas de análisis de información que se han utilizado durante el desarrollo del proyecto. Por último se explicará el motivo de desarrollar un sistema que detecte automáticamente el tipo de viajero y cómo se ha realizado.

2. Estudio de sitios web turísticos

2.1. Booking

Booking es un sitio web para reservar hoteles o alquileres vacacionales. Por cada uno de ellos incluye una descripción del alojamiento, los servicios que incluye, ubicación, precio, lugares de interés más cercanos, lugares de interés más populares, hora de entrada y salida y forma de pago.

Inicialmente te permite hacer una búsqueda por fecha e incluir los siguientes datos: Destino, viaje de trabajo (Si/No), Fecha de check-in, Fecha de check-out, número de habitaciones, número de adultos y número de niños. Una vez realizada la búsqueda inicial te permite filtrar los resultados según los siguientes criterios:

- Presupuesto por noche
- Servicios
- Servicios de la habitación
- Disponibilidad
- Puntuación de los comentarios
- Categoría (estrellas) del hotel
- Recepción 24h
- Ofertas
- Reserva con facilidad
- Comidas
- Lugares de interés
- Tipo de alojamiento
- Preferencia de cama
- Zona
- Cadena de hoteles

Una vez introducidos los criterios de búsqueda nos aparece una selección con los alojamientos que cumplen con las especificaciones, pudiendo ordenar esta selección por precio, puntuación, categoría del hotel y distancia desde el centro de la ciudad.

Cada alojamiento está valorado con una calificación entre 1 y 10 basada en la media de las puntuaciones de todos los usuarios. Además de la nota global, Booking muestra el desglose según las características del lugar, y el número de usuarios que ha puntuado cada característica.

En la Ilustración 2.1. se muestra un ejemplo de la puntuación de un alojamiento y en la Ilustración 2.2 se puede observar la estructura de un comentario.



Ilustración 2. 1. Ejemplo puntuación alojamiento

Jonathan

España

2 comentarios

20 de diciembre de 2016

4,2

“Lo mejor la ubicacion”

• Viaje de ocio

• Habitación Individual con cama doble

• Enviado por móvil

• Pareja

• Estancia de 4 noches

– La limpieza, debajo de la cama estaba sucio y habia un calcetín de a saber cuando porque nuestro no era

+ La ubicación que esta cerca andando a todo los sitios que ibamos a visitar

Ilustración 2. 2. Ejemplo de comentario

Los comentarios se pueden filtrar por idioma y tipo de viajero, distinguiendo entre los siguientes: pareja, familia, grupo de amigos, personas que viajan solas y personas en viaje de negocios. Además se dividen en dos partes: parte positiva de la estancia y parte negativa. La estructura podría ser adecuada para nuestro proyecto ya que incluye el tipo de viajero y la puntuación en el comentario, pero a la hora de hacer el sistema de aprendizaje automático podríamos tener dificultades para analizar el texto por estar dividido en parte negativa y parte positiva.

2.2. Wimdu

Wimdu es un sitio web de alquileres vacacionales, en el cual existen dos tipos de usuarios: huéspedes y anfitriones. Los anfitriones son aquellas personas que desean alquilar su vivienda o alguna habitación dentro de ella, y los huéspedes son los usuarios interesados en alquilar un alojamiento. Inicialmente se pide hacer una búsqueda incluyendo los siguientes datos: fecha de check-in, fecha de check-out, y número de huéspedes. Una vez realizada la primera búsqueda se podrán filtrar los resultados por :

- Tipo de alojamiento, incluyendo entre las opciones: apartamento, casa o habitación privada
- Precio por noche
- Servicios incluidos
- Número de habitaciones

Cada alojamiento tendrá una puntuación entre 0 y 10 basándose en la media de las valoraciones de todos los comentarios. La estructura de la puntuación y de los comentarios se puede ver en las ilustraciones 2.3. y 2.4.

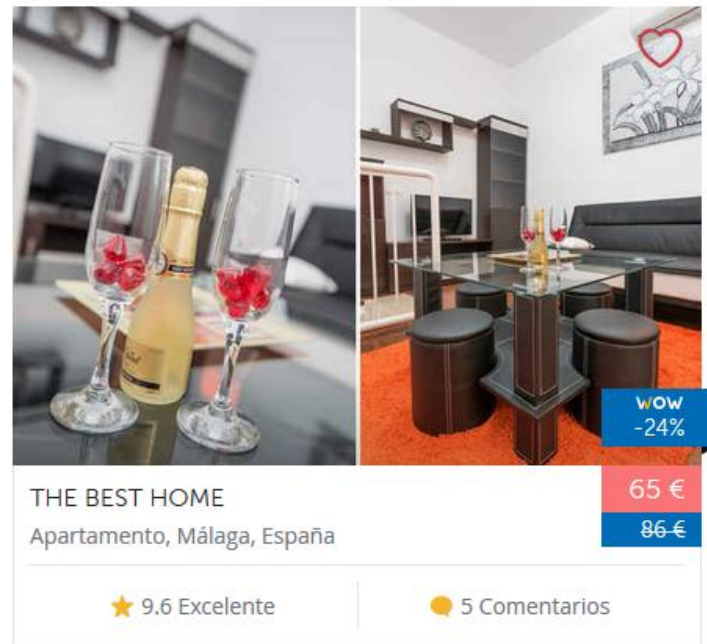


Ilustración 2. 3. Ejemplo puntuación alojamiento



BOBBY
10 Excelente
Mayo 2016

Apartamento perfecto. Limpio y con todo lo necesario para disfrutar. Seguro que volveremos.

Ilustración 2. 4. Ejemplo de comentario

La estructura de este comentario sería más útil para nuestra aplicación porque no viene dividido en parte negativa y positiva, lo que facilitaría su interpretación, pero no es tan completa como la de Booking ya que faltaría un dato fundamental para este proyecto como es el tipo de viajero. Además cuenta con menos número de comentarios por alojamiento.

2.3. Airbnb

Airbnb es un sitio web de alquileres vacacionales en el cual existen los mismos tipos de usuario que en Wimdu y funciona prácticamente de la misma forma. Para la búsqueda inicial hay que introducir fecha de check-in y check-out y número de huéspedes adultos, niños y bebés.

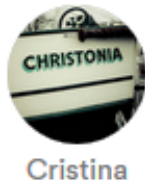
Una vez realizada la búsqueda podemos filtrar los alojamientos según los siguientes criterios:

- Tipo de habitación: casa entera, habitación privada y habitación compartida
- Precio
- Número de dormitorios, baños y camas
- Servicios

Cada alojamiento viene puntuado por estrellas entre 1 y 5, puntuando también de la misma manera las categorías: Veracidad, comunicación, limpieza, ubicación, llegada y calidad. En este sentido Airbnb tendría una puntuación más completa que la de Wimdu aunque seguiría teniendo un menor número de comentarios que Booking. La estructura de la valoración de un alojamiento y de un comentario se puede ver en las ilustraciones 2.5. y 2.6.



Ilustración 2. 5. Ejemplo puntuación alojamiento



La persona encargada de la gestión del apartamento, Paula, es encantadora. Su respuesta fue muy rápida y su trato muy atento. Nos hemos encontrado muy agusto y la cercanía con la estación de trenes lo hace un sitio ideal si viajas en tren.

Diciembre de 2016 · 

 Útil

Ilustración 2. 6. Ejemplo de comentario

El problema para tratar esta información estaría en que la puntuación está relacionada en global con el alojamiento y no viene especificada comentario a comentario. Además faltaría el filtro "Tipo de viajero" necesario para este proyecto.

2.4. El tenedor

El tenedor es un sitio web para reservar mesa y comparar restaurantes. En este sentido tiene un enfoque diferente a los mencionados anteriormente. Para realizar la búsqueda tenemos que introducir los siguientes datos: destino, cocina, fecha, hora y número de comensales

Una vez realizada la búsqueda podemos filtrar los resultados según los siguientes criterios: radio de búsqueda, tipo de cocina, "Best of ", promociones, precio por persona, punto de interés, nota, guía y programa yums.

Cada restaurante estaría valorado con una puntuación entre 1 y 10. Además podríamos visualizar cuántos usuarios han dado cada valoración, ver la puntuación según las categorías (cocina, servicio y ambiente), la valoración global calidad-precio, nivel de ruido y tiempo de espera.

Un ejemplo de como aparece la valoración de un restaurante y la estructura de un comentario son las ilustraciones 2.7., 2.8 y 2.9.



Dos Cielos Madrid by Hermanos... 

Cuesta Santo Domingo, 5 28013 Madrid 0.26Km

MEDITERRÁNEO BEST OF GRANDES CHEFS

BEST OF GRUPOS BEST OF VIDA NOCTURNA

Precio medio 50 €

-30% en cuenta sólo en ElTenedor *

9,8 /10

21 Opiniones

14:30
-30%

15:00
-30%

15:30
-30%

Ilustración 2. 7. Ejemplo valoración restaurante



Ilustración 2. 8.Ejemplo valoración restaurante



PEDRO H.

Conquistador/
(7 opiniones)

VALORACIÓN MEDIA

9,5 /10

COCINA 10

SERVICIO 10

AMBIENTE 8

Fecha de la comida : 28/1/2017

Los platos sorprendentes con sabores muy especiales...el local resulta inicialmente un tanto frio pero q luego se minimiza gracias a la atencion del personal. Volveremos.

 ¿ÚTIL?

Ilustración 2. 9.Ejemplo comentario restaurante

Como podemos ver, el sistema de puntuaciones es muy completo y cada valoración ofrece toda la información que necesitamos conocer a la hora de elegir restaurante, pero al estar enfocado a restaurantes y no a alojamientos no contamos con información relativa a los viajeros y épocas en las que viajan.

2.5. TripAdvisor

Sitio web que permite reservar y comparar hoteles, vuelos, alquileres vacacionales y restaurantes. En este sentido, TripAdvisor sería la página más completa de todas las vistas anteriormente.

En una primera búsqueda tendríamos que seleccionar que queremos reservar y en qué fecha. En el caso de los hoteles habría que indicar también cantidad de habitaciones y huéspedes. Se nos mostraría un listado de aquellos alojamientos que cumplan con los datos indicados, pudiendo ordenarlos por valoración de los viajeros, popularidad, precio y distancia. Además, por cada alojamiento nos aparecerá una comparación de precios con otras páginas de reserva de hoteles.

Cada alojamiento está valorado con una puntuación entre 1 y 5 y además, podemos visualizar la cantidad de comentarios totales del lugar.

Un ejemplo de como se visualiza la puntuación de un hotel y un comentario en TripAdvisor son los mostrados en las ilustraciones 2.10, 2.11 y 2.12.



Ilustración 2. 10. Ejemplo puntuación hotel

Lee lo que dicen los viajeros:

Puntuación de viajeros	Tipo de viajero	Época del año	Idioma
☐ Excelente 506	☐ Familias (96)	☐ Mar-may (117)	☐ Todos los idiomas
☐ Muy bueno 57	☐ En pareja (382)	☐ Jun-ago (163)	☒ Español (576)
☐ Normal 7	☐ En solitario (15)	☐ Sep-nov (179)	☐ Inglés (560)
☐ Malo 2	☐ De negocios (5)	☐ Dic-feb (117)	☐ Francés (80)
☐ Pésimo 4	☐ Amigos (40)		Más

Ilustración 2. 11. Ejemplo puntuación hotel

Podríamos filtrar las opiniones según la puntuación, el tipo de viajero, la época del año, el idioma y conocer qué número de comentarios hay de cada tipo.

Ejemplo de comentario:



juan s

Colaborador de nivel **2**

 4 opiniones

 8 votos útiles

“El mejor hotel de Sevilla”

 Opinión escrita hace 4 semanas  mediante dispositivo móvil

Arquitectonicamente maravilloso Excelentemente decorado Muy bien cuidados todos detalles Desayuno buenísimo Un patio de lujo con una vegetación exuberante Tranquilidad absoluta Se duerme en la gloria, no se oye ni el mas mínimo ruido




¿Útil?  **1** Gracias, **juan s**

 Denunciar

Ilustración 2. 12. Ejemplo de comentario

Cada comentario estaría formado por una valoración y el texto explicativo de la experiencia del usuario, del cual podríamos conocer el número de opiniones emitidas.

La gran ventaja de Tripadvisor es que contamos con dos filtros muy interesantes y necesarios para este proyecto, como son el tipo de viajero y la época del año. Además, contiene más número de comentarios por alojamiento que el resto de páginas contempladas en este estudio.

3. Portal web seleccionado

El portal web elegido para el desarrollo de este proyecto es "TripAdvisor". Como ya se ha mencionado en el estudio realizado previamente, es el sitio web con más comentarios por alojamiento/restaurantes. Además, es el único que reúne opiniones de hoteles, hostales, alquileres vacacionales y restaurantes. En cuanto al formato de los comentarios, son adecuados porque están constituidos por un texto sin dividir y una puntuación entre 1 y 5. La característica más importante es la existencia de sus filtros "tipo de viajero" y "época del año", ambos son fundamentales a la hora de desarrollar este proyecto.

TripAdvisor es la mayor web de viajes del mundo, cuenta con 500 millones de opiniones y comentarios sobre 7 millones de alojamientos, restaurantes y atracciones. Según el tripBarometer de 2016⁹, "El 73 % de los viajeros citan TripAdvisor como un importante factor de influencia a la hora de planificar sus viajes". Además, a la hora de reservar un alojamiento, un 86% de los encuestados en este estudio afirman que las opiniones de TripAdvisor son un factor decisivo (Ilustración 2.13).

⁹https://d2bxpc4ajzxry0.cloudfront.net/TripAdvisorInsights/sites/default/files/downloads/2670/globalreport_spanish_dec2015_0.pdf

Importancia para reservar un alojamiento (las 10 principales)**Ilustración 2. 13. Importancia para reservar alojamiento**

Por todo esto se ha elegido este portal web como fuente de información para realizar este proyecto. Además, se ha decidido que se extraerán únicamente los datos referentes a las capitales de Andalucía y durante el año 2016. El motivo de esta selección no es otro que la imposibilidad de incluir más cantidad de datos porque ocasionarían problemas de eficiencia y almacenamiento. A pesar de que se han elegido estos parámetros, no son más que un ejemplo y la aplicación podría extenderse hasta incluir todas las ciudades y durante los años que se eligieran.

4. Extracción de información

En este proyecto se han utilizado dos técnicas de extracción de información denominadas web crawling y web scrapping. Un web crawler , o araña de la Web es un programa automatizado que busca en la Web de una forma metódica y automatizada. Su objetivo es extraer una copia de una página web en formato html para su posterior procesamiento. Utilizaremos esta técnica para descargar las páginas de los comentarios de los hoteles de las capitales de Andalucía. Para ello, primero tenemos que hacer un estudio de la estructura de TripAdvisor para ver de qué manera se construyen las url. Cada ciudad está representada con el identificador que aparece posterior a "g" como se muestra en la siguiente imagen perteneciente a la url de la página principal de los hoteles de Málaga (Ilustración 2.14.)

https://www.tripadvisor.es/Hotels-g187438-Malaga_Costa_del_Sol_Province_of_Malaga_Andalu

Ilustración 2. 14. URL de la página principal de Málaga

En esta página nos aparecen todos los hoteles de la ciudad. Tenemos que tener en cuenta que la información puede aparecer paginada y hay que observar de que manera cambia la url al cambiar de página. En este caso los hoteles aparecen de treinta en treinta.

Cada hotel también tiene un identificador que aparecerá posterior al identificador de la ciudad y detrás de la letra "d" como aparece en la siguiente imagen (Ilustración 2.15.)

https://www.tripadvisor.es/Hotel_Review-g315916-d777672-Reviews-Hotel_Infanta_Cris

Ilustración 2. 15. URL del hotel infanta cristina de Jaén

Por cada hotel se descargarán todos los comentarios que pertenezcan al periodo de tiempo comprendido entre 2016 y 2017. Cada uno de ellos también tiene un identificador único que aparecerá detrás del identificador del hotel detrás de la letra "r" como aparece en la siguiente imagen (Ilustración 2.16.):



https://www.tripadvisor.es/ShowUserReviews-g315916-d777672-r481814490-Hotel_Infanta_Cristina_de_Jaén

Ilustración 2. 16. URL de un comentario del hotel infanta cristina de Jaén

Los comentarios también aparecen paginados y hay que tener en cuenta que aparecen de diez en diez.

Una vez observado esto se han implementado las funciones que extraerán estas páginas. Esta tarea se ha realizado en el lenguaje de programación Python y haciendo uso de la librería "urllib".

Una vez extraídas las páginas, la estructura de ficheros de la aplicación ha quedado como aparece en las ilustraciones 2.17 y 2.18.

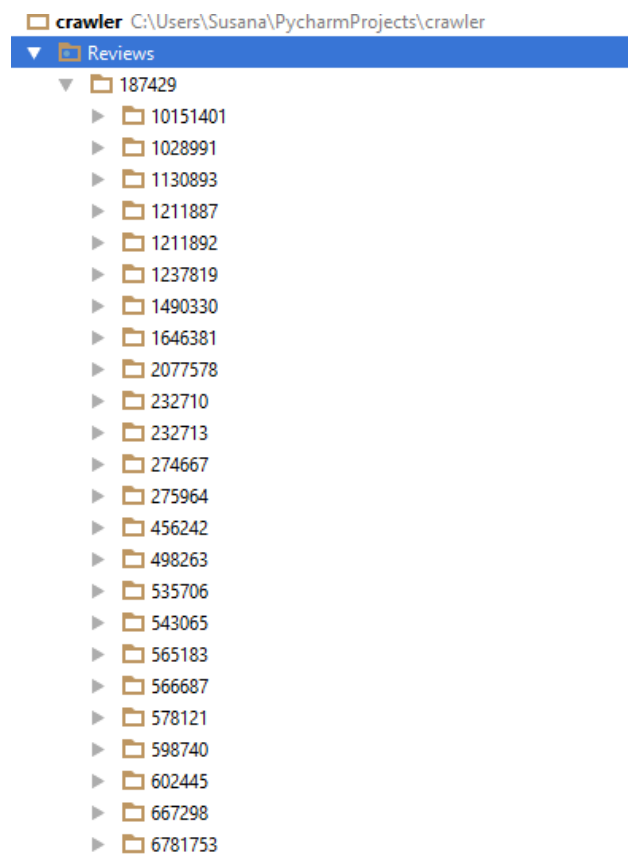


Ilustración 2. 17. Estructura de ficheros

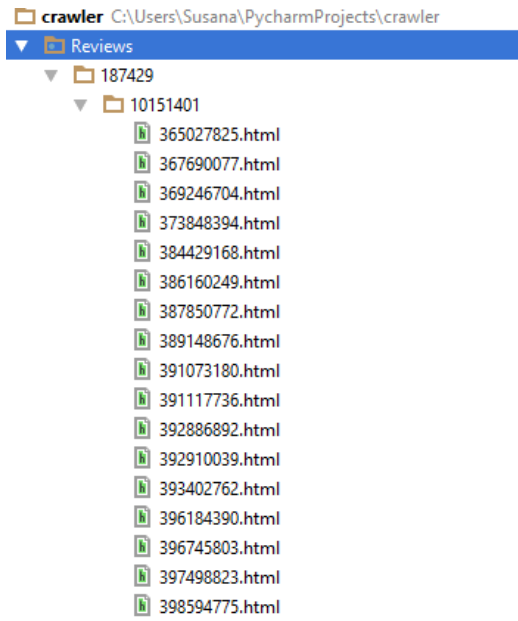


Ilustración 2. 18. Estructura de ficheros

Se ha creado una carpeta por cada ciudad, dentro de cada ciudad una carpeta para cada hotel, y dentro de cada hotel están las páginas descargadas de cada uno de los comentarios.

Una vez hecho esto el siguiente paso es extraer, mediante Web scraping, la información de interés para este proyecto. Web scraping es el proceso de recopilar información de forma automática de la Web, en este caso, mediante el uso de expresiones regulares. Una expresión regular es una secuencia de caracteres que forma un patrón que se utiliza para buscar coincidencias en un texto.

Se ha utilizado esta técnica para extraer los siguientes datos: título del comentario, nombre del hotel, contenido del comentario, valoración en estrellas, fecha del comentario, tipo de viajero que ha escrito el comentario y puntuación de las siguientes categorías: calidad del sueño, puntuación de la ubicación, puntuación de la relación calidad-precio, puntuación de las habitaciones, puntuación del servicio y puntuación de la limpieza. Para ello se ha utilizado Python, expresiones regulares y la librería BeautifulSoup. BeautifulSoup es un módulo para parsear páginas web.

Esta librería construye un árbol de etiquetas de la página HTML de tal manera que podemos navegar por ella de forma rápida y fácil.

Una vez finalizado esto se guardarán todos los comentarios en un único archivo en formato csv en el cual cada fila tendrá el contenido de un comentario. Queda estructurado de la siguiente forma (Ilustración 2.19.):

codigo	ciudad	nombreHotel	Fecha	Título	valoracion	GL	calidad	Sueño	Ubicacion	Habitaciones	Servicio	Calidad	Precio	Limpieza	Opinion	TipoViajero
228564-3369	Sevilla	Catalonia Gir	03/01/2016	"Para derribar"	1	None	None	None	None	None	None	None	None	None	Cuando te h	en pareja
228564-3375	Sevilla	Catalonia Gir	05/01/2016	"Pudiera est	2	None	None	None	None	None	None	None	None	None	Siento que e	None
228564-3380	Sevilla	Catalonia Gir	06/01/2016	"Vacaciones	4	None	4	None	4	None	None	None	None	None	3 El Hotel en G	en pareja
228564-3404	Sevilla	Catalonia Gir	16/01/2016	"Excelente p	5	None	None	None	None	None	None	None	None	None	Excelente pé	en pareja
228564-3419	Sevilla	Catalonia Gir	22/01/2016	"Muy buen h	5	None	None	None	None	None	None	None	None	None	Estuvimos cu	en pareja
228564-3443	Sevilla	Catalonia Gir	01/02/2016	"Buenas vist	3	None	None	None	None	None	None	None	None	None	Hotel norma	de negocios
228564-3476	Sevilla	Catalonia Gir	14/02/2016	"Muy buen h	4	None	5	None	None	None	None	4	None	None	5 Por estar cor	en pareja
228564-3481	Sevilla	Catalonia Gir	16/02/2016	"Buen hotel,	4	None	None	None	None	None	None	None	None	None	Si vas de turi	en pareja
228564-3496	Sevilla	Catalonia Gir	21/02/2016	"Excelente h	5	None	None	None	None	None	None	None	None	None	He estado er	en pareja
228564-3545	Sevilla	Catalonia Gir	11/03/2016	"Categoría d	1	None	None	None	None	None	None	None	None	None	No recomen	con los amig
228564-3546	Sevilla	Catalonia Gir	11/03/2016	"Hay mejore	3	None	None	None	None	None	None	None	None	None	Es un hotel c	con los amig
228564-3552	Sevilla	Catalonia Gir	14/03/2016	"Elefantes e	3	None	2	None	None	None	None	None	4	None	2 La ubicación	con los amig
228564-3557	Sevilla	Catalonia Gir	15/03/2016	"Excelente!"	5	None	None	None	None	None	None	None	None	None	El hotel está	con los amig
228564-3564	Sevilla	Catalonia Gir	17/03/2016	"Acogedor"	4	None	None	None	None	None	None	None	None	None	Es un hotel a	None
228564-3600	Sevilla	Catalonia Gir	30/03/2016	"Peor de lo e	3	None	3	None	None	None	None	3	None	None	2 Reservamos	con la famili
228564-3600	Sevilla	Catalonia Gir	30/03/2016	"Bastante bi	4	None	5	None	None	3	None	None	None	None	5 Lo primero q	en pareja
228564-3603	Sevilla	Catalonia Gir	31/03/2016	"No vendría	3	None	4	None	None	None	None	None	4	None	2 El hotel está	con los amig
228564-3631	Sevilla	Catalonia Gir	10/04/2016	"Bastante m	3	None	3	None	None	None	None	None	3	None	2 Esperaba un	con los amig
228564-3633	Sevilla	Catalonia Gir	11/04/2016	"Excelente H	5	None	None	None	None	None	None	None	None	None	Estuve en es	con la famili
228564-3652	Sevilla	Catalonia Gir	18/04/2016	"Feria"	1	None	None	None	None	None	None	None	None	None	Si este hotel	en pareja
228564-3654	Sevilla	Catalonia Gir	19/04/2016	"un cuatro e	1	None	3	None	2	2	None	1	3	None	2 Habitaciones:	con los amig
228564-3654	Sevilla	Catalonia Gir	19/04/2016	"Necesita re	3	None	3	None	3	4	None	2	3	None	2 Muy malos a	con los amig
228564-3655	Sevilla	Catalonia Gir	19/04/2016	"Es un buen	4	None	3	None	3	None	None	None	None	None	3 Fue una esta	con la famili
228564-3666	Sevilla	Catalonia Gir	23/04/2016	"No está mal	3	None	5	None	None	None	None	None	5	None	3 Antes que n	en pareja

Ilustración 2. 19. Estructura de los comentarios

5. Análisis de la información

El análisis de la información ha incluido todas las tareas necesarias para tratar los textos y sacar conclusiones de los comentarios de los viajeros. Para ello se han utilizado técnicas de Procesamiento del Lenguaje Natural (PLN). El PLN es una disciplina que nace en 1960, como un subárea de la Inteligencia Artificial y la Lingüística, con el objeto de estudiar los problemas derivados de la comprensión del lenguaje natural por una máquina.

Analizar los comentarios o críticas de los usuarios en plataformas de turismo como tripAdvisor nos puede ofrecer una imagen general de las características de hotel, como por ejemplo el tamaño de las habitaciones,

calidad o existencia de desayuno, trato del personal...etc (Fondevila-Gascón et al,2016).

En el caso de este proyecto, el objetivo del análisis es mostrar las palabras más utilizadas por los viajeros para describir su viaje, para ello se ha utilizado la librería NLTK (Natural Language Toolkit). NLTK es un conjunto de bibliotecas y programas para procesamiento del lenguaje natural en el lenguaje de programación Python. Contiene diversas herramientas para hacer cualquier tipo de operación con textos de forma simple.

Primero se ha aplicado una operación que convierte el texto a minúscula, con el objetivo de facilitar el reconocimiento de palabras iguales. Después se ha aplicado tokenización. Tokenizar es el proceso de dividir un texto en tokens, siendo un token una secuencia de caracteres entendidos como una unidad semántica útil para procesamiento. Podemos aplicar la tokenización para dividir el texto en sentencias o en palabras, en el caso de este proyecto se han dividido en palabras.

Después se ha aplicado un stopwords, un stopword es el proceso de eliminar palabras vacías, es decir, aquellas que no tienen ningún significado y por tanto no aportan nada al contenido del texto, como por ejemplo, artículos. Además, a las palabras se les ha tenido que pasar un POS tagger. POS tagging (part-of-speech tagging) es el proceso de etiquetar a cada palabra con su categoría gramatical. NLTK tiene herramientas para realizar esta tarea pero sólo para textos en inglés, por ello se ha utilizado el pos tagger de Stanford¹⁰ que tiene una extensión para textos en español. Una vez realizado esto, se han elegido únicamente las palabras con categoría de nombre y se han almacenado las veinte más comunes para posteriormente mostrarlas en una nube de palabras.

6. Detección automática del tipo de viajero

Como hemos visto en el estudio que se ha realizado sobre los portales web turísticos, la mayoría de los contemplados no contaban con el filtro "tipo

¹⁰ <https://nlp.stanford.edu/software/tagger.shtml>

de viajero". Por ello, otro de los objetivos de este proyecto es desarrollar una herramienta que, dada una opinión, determine el tipo de viajero mediante un proceso de clasificación basado en aprendizaje automático.

En primer lugar, definiremos lo que es el aprendizaje automático. El aprendizaje automático o machine learning es un campo de la inteligencia artificial que consiste en el diseño y estudio de algoritmos que utilizan la experiencia pasada para tomar decisiones futuras. En 1959, Arthur Samuel, pionero en el estudio de la inteligencia artificial, lo definió como "El estudio que da a las computadoras la capacidad de aprender sin ser programados explícitamente".

Dentro del aprendizaje automático se comprenden dos tipos, aprendizaje supervisado y no supervisado. El aprendizaje supervisado es aquel en el que a partir de un conjunto de ejemplos clasificados (train set o conjunto de entrenamiento) se pretende clasificar un segundo conjunto no clasificado (conjunto test o test set). El train set contiene la experiencia que el algoritmo utiliza para aprender. Está formado por una serie de características y su respectiva etiqueta o categoría asociada. El test set es una colección similar al train set que se utiliza para evaluar el rendimiento del modelo una vez que ya está entrenado.

El aprendizaje no supervisado sería aquel en el cual no contamos con un conjunto ya clasificado previamente como punto de partida para que la máquina aprenda, si no que el propio algoritmo por si mismo tiene que observar las propiedades de los ejemplos y agruparlos según su criterio a base de reconocer patrones.

En nuestro caso nos centraremos en el aprendizaje supervisado y en la tarea de clasificar textos. La clasificación consiste en intentar predecir la clasificación de un texto sobre un conjunto de etiquetas ya conocidas. Hay dos tipos de clasificación: binaria, si sólo se pueden clasificar en dos categorías, y multiclase si las opciones de clasificación contemplan más de dos categorías. En este proyecto estamos en un caso de clasificación multiclase ya que el objetivo es clasificar un texto según el tipo de viajero que lo ha escrito. Así que

tendremos cinco posibles categorías: pareja, familia, negocios, solo y amigos. Como conclusión, la idea principal de esta clasificación es suponer que podemos diferenciar significativamente textos, según las palabras comunes que tengan entre sí (Venegas,2007).

Para realizar este proceso se ha utilizado la herramienta Scikit-Learn (Hackeling, 2014), que contiene un conjunto de librerías para realizar tareas de aprendizaje automático, minería de datos y análisis de datos. Fue lanzada en 2017 y se ha convertido en una de las herramientas open source más populares para tareas de machine learning. Está construida sobre NumPy, SciPy y Matplotlib, y permite experimentar con una amplia cantidad de algoritmos de una manera muy sencilla.

Scikit-learn nos permite hacer diversas tareas como: clasificación, regresión, clustering, reducción de dimensionalidad, selección del modelo y preprocesamiento. Aunque nos centraremos como hemos mencionado anteriormente en sus algoritmos de clasificación.

Probaremos con los siguientes algoritmos:

- Máquinas de soporte vectorial (SVM) : Pertenecen a la categoría de clasificadores lineales o hiperplanos. Mapea los puntos de entrada a un espacio de características de una dimensión mayor y encuentra un hiperplano que los separe y maximice el margen entre las clases en este espacio (Betancourt,2005). Utilizaremos SVM utilizando el método .SVC() de la librería svm de Scikit-learn.
- K- vecinos más cercanos : La clasificación basada en vecinos almacena instancias de los datos de entrenamiento, es decir, los vectores con los ítem y las etiquetas de cada ítem. En la fase de clasificación, la evaluación del ítem del cual no conocemos su etiqueta es representada por un vector. Se calcula la distancia entre los vectores almacenados y el nuevo vector, y se seleccionan los k ejemplos más cercanos. Utilizaremos para esto el método . KNeighborsClassifier() de la librería neighbors de Scikit-learn.

- Naive Bayes: Está basado en la aplicación del teorema de Bayes suponiendo la independencia entre cada par de características. Funciona de la siguiente forma, dado un ejemplo representado por k valores, el clasificador debe encontrar la función clasificadora o hipótesis más probable que describa a ese ejemplo. Podemos aplicar Naive Bayes con los siguientes algoritmos: Gaussian Naive Bayes, Multinomial Naive Bayes y Bernoulli Naive Bayes. Utilizaremos la función `MultinomialNB()`.
- Árboles de decisión: Es uno de los algoritmos de clasificación más utilizados, su objetivo es crear un modelo que prediga el valor de la variable test deduciendo reglas de las características de los datos. En este caso utilizaremos el algoritmo Random Forest, que mejora la precisión en la clasificación gracias a su aleatoriedad en la construcción de cada árbol. Utilizaremos la función `RandomForestClassifier()`.

El primer paso para iniciar un proceso de clasificación es construir el modelo, para ello se entrena el conjunto de entrenamiento, y una vez entrenado se usará este modelo para clasificar nuevas instancias. Para evaluar la validez del modelo tendremos que utilizar como hemos mencionado anteriormente, un segundo conjunto test para medir la cantidad de acierto que tendría el modelo al clasificar este segundo conjunto. Una manera de validar modelos se denomina validación cruzada o Cross-Validation.

La técnica Cross-Validation (Ilustración 2.20.) consiste en dividir el conjunto de entrenamiento en k conjuntos más pequeños. Por cada uno de los conjuntos se repite el siguiente procedimiento: el modelo se entrena usando $k-1$ conjuntos como datos de entrenamiento, y una vez entrenado se valida en la parte restante de los datos. El resultado final sería una media de todos los resultados obtenidos.

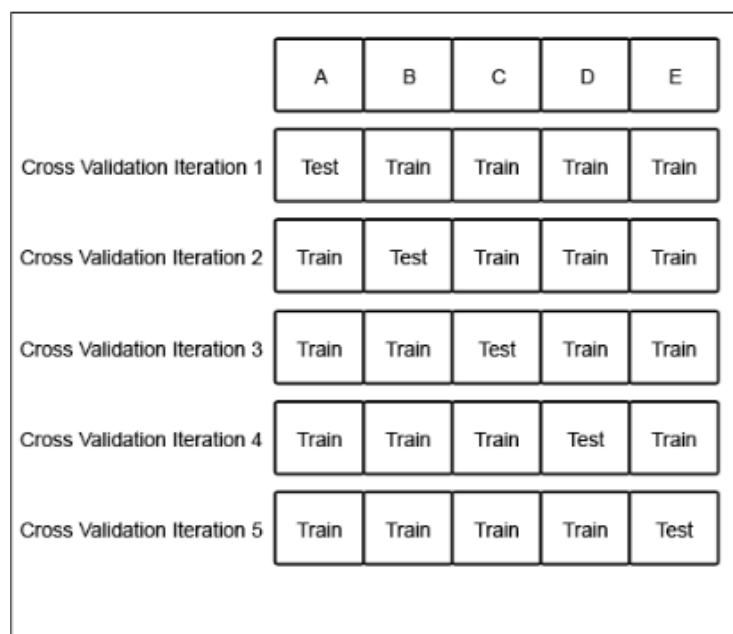


Ilustración 2. 20. Esquema Cross-Validation.

Además tenemos que tener en cuenta que los datos que vamos a tratar son textos. Para clasificar textos, debemos representarlos de manera que se codifique su sentido en un vector de características. Una de las representaciones más comunes utilizadas es el modelo llamado bolsa de palabras o bag-of-words. Esta representación consiste en codificar las palabras que aparecen en el texto, sin tener en cuenta la sintaxis, orden y gramática del mismo. Una colección de documentos es lo que se denomina como corpus. El conjunto de palabras únicas del corpus se denomina vocabulario. Cada documento será representado por un vector binario con tantos elementos como palabras haya en el vocabulario, tomando el valor 1 si la palabra aparece en el texto, y el valor 0 en caso contrario. La idea principal es suponer que los textos contienen palabras similares que nos permitirán establecer patrones.

La clase `CountVectorizer` de Scikit-learn puede crear representaciones de bolsas de palabras a partir de textos. Para ello, convierte las palabras a minúscula y tokeniza el documento. Una estrategia para mejorar el resultado que obtendríamos haciendo únicamente eso, es eliminar las palabras que más se repiten en todos los textos. Como ya se ha explicado en un apartado

anterior, este proceso se consigue eliminando las palabras vacías (stopwords) del texto. La clase `CountVecorizer` permite pasarle el argumento `stopwords=list`, siendo `list` la lista de palabras vacías en español.

Otra posible mejora es utilizar lo que se conoce como Stemmer. Un Stemmer es un método que se utiliza para reducir una palabra a su raíz. De esta manera aquellas palabras que se refieran a cosas similares se contarán como la misma. Es importante tener en cuenta estas similitudes a la hora de buscar patrones en los textos. En este caso se utilizará el algoritmo Snowball ya que es un Stemmer que contempla el idioma español.

También es importante tener en cuenta la cantidad de veces que se repite una palabra en un texto. Que se repita una vez no tiene porque significar que se esté hablando sobre ese tema en concreto, pero si se repite varias veces sí puede indicar que es un concepto relevante en ese texto.

Con la representación planteada hasta este momento no se tenía en cuenta este dato. Para solucionar esto, en vez de utilizar un valor binario en los vectores de características indicando si la palabra aparece o no en el texto, podemos crear un vector cuyo contenido sea la cantidad de veces que se repite cada palabra en el texto. Para ello se utiliza la función `TfidfTransformer()`, que se encarga de construir una matriz de frecuencias de palabras.

Una vez construido el modelo teniendo en cuenta todo lo mencionado anteriormente, el siguiente paso es evaluarlo. Las medidas que se tendrán en cuenta para evaluar la calidad del modelo son: precisión, recall y f1- score.

Precisión: $P = TP / TP + FP$

Recall: $R = TP / (TP + FN)$

F1-Score: se considera una media ponderada entre los dos anteriores.

$F1 = 2 * (precision * recall) / (precision + recall)$

Entendiendo como TP (True Positives) de una clase dada "x" a aquellas instancias que pertenecen a "x" y son clasificadas correctamente como pertenecientes a "x".

True Negatives (TN) de una clase "x" serán aquellas instancias que no pertenecen a "x" y no se clasifican como pertenecientes a "x".

False Positives (FP) de una clase "x" serán aquellas instancias no pertenecientes a la clase "x" pero que se clasifican como clase "x".

False Negatives (FN) de una clase "x" serán aquellas instancias pertenecientes a la clase "x" pero que no se clasifican como clase "x".

Los valores que podrá tomar x son: "Pareja", "Amigos","Solo","Negocios" y "Familia".

El conjunto de datos estará formado por 22.777 comentarios, de los cuales: 5520 son comentarios de familias, 2266 son comentarios de viajeros que viajan con amigos, 3448 son de viajeros que viajan con motivo de negocios, 10810 son parejas y 733 son personas que viajan en solitario.

Los resultados obtenidos para este conjunto de datos con el algoritmo SVM utilizando validación cruzada con un tamaño de 5 son los que se muestran en la ilustración 2.21.

	precision	recall	f1-score	support
pareja	0.53	0.94	0.68	10810
amigos	0.58	0.08	0.13	2266
familia	0.76	0.24	0.36	5520
negocios	0.59	0.30	0.39	3448
solo	0.50	0.00	0.00	733
avg / total	0.60	0.56	0.48	22777

Ilustración 2. 21. Resultados SVM

Los resultados obtenidos para este conjunto de datos con el algoritmo K-vecinos más cercanos utilizando validación cruzada con un tamaño de 5 son los que se muestran en la ilustración 2.22.

	precision	recall	f1-score	support
pareja	0.50	0.89	0.64	10810
amigos	0.20	0.02	0.03	2266
familia	0.39	0.17	0.24	5520
negocios	0.47	0.12	0.20	3448
solo	0.00	0.00	0.00	733
avg / total	0.42	0.48	0.39	22777

Ilustración 2. 22. Resultados K-vecinos más cercanos

Los resultados obtenidos usando el algoritmo Random Forest utilizando validación cruzada con un tamaño de 5 son los que se muestran en la ilustración 2.23.

	precision	recall	f1-score	support
pareja	0.51	0.87	0.64	10810
amigos	0.23	0.03	0.06	2266
familia	0.44	0.22	0.29	5520
negocios	0.46	0.14	0.22	3448
solo	0.20	0.00	0.00	733
avg / total	0.44	0.49	0.41	22777

Ilustración 2. 23. Resultados Random Forest

Los resultados obtenidos usando el algoritmo Naive Bayes utilizando validación cruzada con un tamaño de 5 son los que se muestran en la ilustración 2.24.

	precision	recall	f1-score	support
pareja	0.48	1.00	0.65	10810
amigos	0.00	0.00	0.00	2266
familia	0.92	0.02	0.03	5520
negocios	1.00	0.00	0.00	3448
solo	0.00	0.00	0.00	733
avg / total	0.60	0.48	0.31	22777

Ilustración 2. 24. Resultados Naive Bayes

Los resultados obtenidos usando el algoritmo SGD Classifier utilizando validación cruzada y un tamaño de cinco son los que se muestran en la ilustración 2.25.

	precision	recall	f1-score	support
pareja	0.51	0.97	0.66	10810
amigos	0.51	0.01	0.02	2266
familia	0.68	0.12	0.20	5520
negocios	0.66	0.18	0.29	3448
solo	0.00	0.00	0.00	733
avg / total	0.56	0.52	0.41	22777

Ilustración 2. 25. Resultados SGD Classifier

Estos resultados serían una aproximación inicial para resolver este problema, pero tal y como se observa aún son muy preliminares y tendrían que mejorarse. Para ello habría que hacer un estudio más exhaustivo de las características a utilizar para representar la información y probar técnicas más personalizadas a estos datos. Aun así, el algoritmo que parece dar mejores resultados son las máquinas de soporte vectorial o SVM.

Capítulo 3 :

Desarrollo del proyecto

1. Descripción del problema

Con el auge de las Tecnologías de la Información y la Comunicación y la llegada de la Web 2.0, la cantidad de información en Internet ha aumentado de forma exponencial en los últimos años. Cada vez más aumenta la preocupación por parte de las empresas por conocer la opinión de sus clientes, y por parte de los usuarios por compartir sus experiencias y basarse en las de otros usuarios a la hora de tomar decisiones. Este fenómeno aplicado al campo del turismo es lo que se ha denominado como turismo 2.0., que ha supuesto una revolución a la hora de concebir el turismo y la forma de reservar y organizar los viajes para las personas. El principal problema de todo esto, es que actualmente la información que encontramos en la web no está estructurada, lo que supone una gran pérdida de información que podría ser aprovechada para obtener conocimiento útil de las preferencias de los usuarios. Por todo ello surge la necesidad de crear un sistema que sea capaz de extraer, analizar, estructurar y sacar conclusiones de esta información con el objetivo de ayudar a las personas interesadas en este campo a la toma de decisiones teniendo en cuenta las opiniones de los turistas en la web.

2. Objetivos del sistema

El objetivo principal del sistema es mostrar el tipo de viajero más común en cada una de las capitales de Andalucía diferenciando por época del año. Se tendrán en cuenta como épocas: verano, Navidad, Semana Santa y el resto de días que no estén comprendidos en ninguna de las mencionadas anteriormente. Se pretende que la información se visualice de forma clara con el objetivo de que permita tomar decisiones rápidas. Además, por cada ciudad y época se mostrarán la cantidad de comentarios, la polaridad y las palabras mas utilizadas por cada tipo de viajero.

Se podrán visualizar los datos desde dos puntos de vista: diferenciando los gráficos por época de manera que en cada ciudad se marcará el viajero más común, y al revés, es decir, se agruparán los datos por tipo de viajero, y se

visualizará por ciudad, la época vacacional en la que más viaja cada tipo de viajero. Esta doble visualización es simplemente con el objetivo de que los datos se puedan ver con total claridad y desde los dos puntos de vista.

3. Metodología para el desarrollo de software y método a seguir

La metodología para el desarrollo de software se define como un marco de trabajo usado para estructurar, planificar y controlar el proceso de desarrollo en sistemas de información. Hay dos metodologías principales: la metodología tradicional y la ágil.

La metodología tradicional se basa en un ciclo de vida en cascada¹¹, esto quiere decir que el proyecto se organiza en etapas que se ejecutan secuencialmente. Cada etapa se realiza una única vez y hasta que no se ha finalizado no se pasa a la siguiente. Este tipo de metodología es apropiada para proyectos con equipos grandes, proyectos con requisitos estables e inmutables en el tiempo, aplicaciones críticas o equipos de desarrollo distribuidos geográficamente. En esta metodología, si un requisito cambia o hay que implementar algún cambio, los costes y el tiempo del proyecto aumentan, ya que generalmente no ofrecen buenas soluciones para proyectos con requisitos cambiantes.

En cambio, las metodologías ágiles son un enfoque basado en el desarrollo y en la entrega de "incrementos" de funcionalidad limitada. Suelen ser metodologías iterativas e incrementales que tienen como idea principal la mejora constante del código y la implicación del usuario en el equipo de trabajo. Al contrario que la metodología tradicional, se centran en fijar un coste y tiempo para el proyecto y que los requisitos se vayan adecuando a los cambios durante el desarrollo del mismo. Se basan en la idea de que es difícil saber con exactitud que los requisitos propuestos en un primer momento por el cliente se mantendrán hasta el final, por este motivo hay que crear una metodología que se adapte ante los posibles cambios de forma eficaz. La

¹¹ <https://inventtatte.com/metodologia-tradicional-vs-agil/>

comparativa de este ciclo de desarrollo se representa en la siguiente ilustración (Ilustración 3.1.):



Ilustración 3. 1. Comparativa ágil y tradicional

Se ha elegido seguir una metodología ágil debido a las características del proyecto. Es un equipo de trabajo pequeño, cuya aplicación no es crítica, además los requisitos pueden cambiar al ser un proyecto del cual se desconocen los resultados que se van a obtener una vez analizados los datos. Además, se pretende ir realizando entregas parciales del proyecto e iterativamente ir modificando los requisitos según las reuniones que se tengan con las tutoras y los posibles inconvenientes que puedan ir surgiendo durante el desarrollo del proyecto, lo cual encaja con el enfoque de las metodologías ágiles.

3.1. Metodología ágil elegida

Hay una gran variedad de metodologías ágiles. Cada una de ellas tiene una serie de características diferentes que serán más ventajosas o menos dependiendo del tipo de proyecto que queramos realizar. La metodología que he elegido es Kanban y explicaré sus características a continuación.

Kanban tiene como propósito gestionar la forma de realizar las tareas en un proyecto. Se basa en dividir el flujo de trabajo en fases o bloques, en los cuales para pasar al siguiente bloque se debe haber completado el anterior. Pone especial énfasis en la idea de que el equipo pueda visualizar el estado del proyecto, las tareas que se tienen que hacer, en qué está trabajando cada persona y la prioridad de cada una de las tareas. Esto se consigue con un tablero. Las columnas del tablero representan los posibles estados por los que puede pasar una tarea hasta completarse. Las tareas se escriben en tarjetas y se colocan en el tablero. Su posición sobre el tablero refleja el estado en el que se encuentran. Los estados que se suelen representar en un tablero Kanban son “pendiente”, “en curso” y “terminado”.¹² (Ilustración 3.2.)



Ilustración 3. 2. Tablero Kanban

Es útil establecer el WIP (Work In Progress), que indica el máximo de tareas que pueden llegar a acumularse en una fase o columna del tablero. Este término se fundamenta en la idea de no comenzar una tarea hasta no haber terminado la anterior.

¹² <http://www.scrummanager.net/blog/2013/09/kanban-las-4-cosas/>

Se determina un WIP por cada fase del tablero. Además se debe medir el tiempo que se tarda en terminar cada tarea. Este tiempo se denomina "lead time" y se mide contando desde que se hace una petición de una tarea hasta que se realiza la entrega. Hay otra métrica también importante que se denomina "cycle time", y que mide el tiempo de trabajo sobre una tarea desde que comienza hasta que termina. Este concepto se encarga de medir el rendimiento del proceso.

He elegido la metodología Kanban porque tiene numerosas ventajas que encajan con el enfoque de este proyecto. Su énfasis en mostrar el flujo de trabajo ayuda a mejorar la productividad del proyecto y a tener claras las tareas a realizar. Su flexibilidad a la hora de planificar entregas y reuniones es adecuada ya que las reuniones con las tutoras se organizarán de acuerdo a conforme se vayan finalizando las tareas preestablecidas, y no de una manera fija. El carácter del proyecto es cambiante ya que consiste en mostrar gráficos y conclusiones sobre unos datos que no sabemos qué información arrojarán, por tanto es probable que los requisitos cambien y que haya que realizar cambios en el planteamiento del proyecto. Además, Kanban es una metodología que funciona bien en entornos cambiantes y por tanto sería adecuada. Kanban no impone restricciones en cuanto a los roles, y esto también sería ventajoso ya que al tratarse de un trabajo fin de grado los únicos roles que se podrían contemplar son el del desarrollador y el del cliente, que en este caso será un rol asumido por las tutoras del proyecto.

En definitiva, la flexibilidad de esta metodología es la ventaja que ha supuesto que Kanban sea la elegida entre las demás. Otras metodologías ágiles, como puede ser Scrum, proponen más rigidez a la hora de establecer reuniones con el cliente o en las iteraciones.

4. Historias de usuario

Las historias de usuario son una herramienta utilizada en las metodologías ágiles para obtener los requisitos de la aplicación a partir de la información obtenida de los usuarios. Intentan combinar las fortalezas de la

comunicación oral y escrita. En definitiva, son una descripción escrita de parte de la funcionalidad del software, expresada de forma valiosa para las partes involucradas y útil para la gestión de proyectos. Su ventaja es que son una manera más simple de describir una tarea. Cuentan con el entendimiento del cliente para especificar el problema y su posible solución. Al ser tareas concretas y pequeñas permiten centrarse en su desarrollo, aplazando detalles posteriores que podrían ofuscar el desempeño de esa tarea.

Las historias de usuario se representan en forma de tarjetas con la siguiente estructura :

- Id: Identificador de la historia de usuario.
- Título: Título descriptivo de la historia de usuario.
- Descripción: Descripción breve de la historia de usuario usando la estructura "Como... Quiero... Para...".
- Estimación: Estimación del coste de implementación en unidades de desarrollo, en Kanban se expresa en semanas.
- Prioridad: Prioridad en la implementación de la historia de usuario respecto al resto de las historias de usuario. La prioridad se puede expresar a través de un número, que cuanto mayor sea, mayor será la prioridad de la historia.
- Dependencias: Se indican los ids de las tareas de las que depende esta tarea. Idealmente una historia de usuario no debería ser dependiente de otra historia, pero a veces es inevitable.
- Criterios de aceptación: Descripción de las pruebas que se llevarán a cabo para decir que la tarea se ha completado con éxito. Los criterios de aceptación incluyen: funcionalidad que el sistema debe realizar, aspecto y funcionamiento de la interfaz y la documentación necesaria.

La estructura de las tarjetas de las historias de usuario tendrán el formato que se muestra en la ilustración 3.3.

ID	TITULO
DESCRIPCIÓN	
ESTIMACIÓN	DEPENDENCIA
PRIORIDAD	
CRITERIOS DE ACEPTACIÓN	

Ilustración 3. 3. Estructura historia de usuario

El proyecto tendrá dos aplicaciones, una que contendrá la página web en la que se visualizarán los datos extraídos de forma gráfica, y otra anterior que se encargará de la extracción y filtrado de los datos de la página web TripAdvisor. En este sentido habrá dos tipos de usuarios a tener en cuenta, los de la primera aplicación y los de la segunda. Las historias de usuario del proyecto serán la que se muestran en las ilustraciones 3.4, 3.5, 3.6, 3.7, 3.8, 3.9, 3.10, 3.11 y 3.12.:

1	EXTRACCIÓN DE DATOS
Como usuario quiero extraer los datos de la página web TripAdvisor para obtener las opiniones de los turistas.	
Estimación: 2 Semanas	Dependencia: Ninguna
Prioridad: 10	

Ilustración 3. 4.Historia de usuario 1

2 FILTRADO DE DATOS	
Como usuario quiero filtrar los datos extraídos para obtener los comentarios y valoraciones de los turistas.	
Estimación: 2 Semanas	Dependencia: 1
Prioridad: 10	

Ilustración 3. 5. Historia de usuario 2

3 VISUALIZACIÓN DE MAPAS	
Como usuario quiero poder visualizar los viajeros más comunes de cada ciudad y época del año.	
Estimación: 2 Semanas	Dependencia: 1,2
Prioridad: 10	
Comprobar que los mapas se visualizan de forma clara y descriptiva y con los datos correctos.	

Ilustración 3. 6. Historia de usuario 3

4 VER EN DETALLE	
Como usuario quiero poder visualizar la cantidad de comentarios de cada tipo de viajero en cada ciudad y época del año.	
Estimación: 1 Semana	Dependencia:1,2
Prioridad: 10	
Comprobar que los gráficos representan los datos de forma clara y fácil de interpretar.	

Ilustración 3. 7. Historia de usuario 4

5	VISUALIZACIÓN DE POLARIDAD
Como usuario quiero poder visualizar la cantidad de comentarios positivos y negativos de los turistas distinguiendo por tipo de viajero, ciudad y época del año.	
Estimación: 1 Semanas	Dependencia:1,2
Prioridad: 10	
Comprobar que los gráficos se visualizan de forma clara y con los datos correctos.	

Ilustración 3. 8. Historia de usuario 5

6 VISUALIZACIÓN DE LAS PALABRAS MÁS USADAS POR LOS VIAJEROS

Como usuario quiero poder visualizar las palabras que más utilizan los viajeros en los comentarios distinguiendo por tipo de viajero, ciudad y época del año.

Estimación: 1 Semana

Dependencia:1,2

Prioridad: 10

Comprobar que la nube de palabras se visualiza correctamente y que las palabras que utiliza son las veinte más comunes de cada ciudad, época del año y tipo de viajero.

Ilustración 3. 9. Historia de usuario 6

7 INTERFAZ WEB

Como usuario quiero poder visualizar los datos en una página web con una interfaz amigable.

Estimación: 2 Semanas

Dependencia: Ninguna

Prioridad: 7

Comprobar que los colores y disposición de la página son adecuados. El diseño debe ser usable para cualquier usuario.

Ilustración 3. 10. Historia de usuario 7

8 DISPONIBLE EN TODOS LOS NAVEGADORES	
Como usuario quiero poder visualizar la página web en cualquier navegador.	
Estimación: no estimada	Dependencia: Ninguna
Prioridad: 7	
Comprobar que la página web se visualiza correctamente en los navegadores chrome, firefox, opera y edge.	

Ilustración 3. 11. Historia de usuario 8

9 DISPONIBILIDAD EN CUALQUIER MOMENTO	
Como usuario quiero poder visualizar la página web en cualquier momento.	
Estimación: no estimada	Dependencia: Ninguna
Prioridad: 10	
Comprobar que la página web se visualiza correctamente en un porcentaje mayor al 90% de las veces que accedemos.	

Ilustración 3. 12. Historia de usuario 9

5. Propuesta de solución

La solución propuesta para llevar a cabo las historias de usuario y resolver el problema propuesto es el desarrollo de una aplicación web que permita visualizar los datos obtenidos de TripAdvisor en forma de gráficos que representen de manera visual los contenidos. El sistema tendrá los siguientes componentes:

- Sistema de extracción de opiniones, que permita filtrar los datos de interés con el fin de almacenarlos de manera estructurada en una base de datos.
- Base de datos de comentarios y valoraciones.
- Plataforma web que muestre los datos extraídos de forma visual.

Los gráficos de la plataforma web se mostrarán distinguiendo por ciudad, época del año y tipo de viajero. Con el fin de que se pudiera utilizar como fuente de información otra plataforma web distinta a TripAdvisor que pudiera no tener explícito el tipo de viajero que ha escrito el comentario, como algo adicional se pretende desarrollar un sistema que detecte el tipo de viajero dada una opinión.

6. Justificación de la solución

Se ha seleccionado esta estructura porque como primera aproximación , se ha decidido utilizar únicamente los datos de las capitales de Andalucía. Por ello en un primer proceso se ha extraído la información de TripAdvisor, se ha volcado en la base de datos este contenido, y la aplicación web simplemente hace consultas a la base de datos. Debido a esto se ha considerado que la mejor manera de llevar a cabo el proyecto es considerar ambas aplicaciones independientes y que únicamente se comuniquen mediante una base de datos.

7. Descripción de la solución

En este apartado se abordará la solución propuesta en el apartado seis. La solución se irá desarrollando en forma de iteraciones que irán completando las historias de usuario ya mencionadas. El tablero Kanban que se utilizará será virtual y se usará la página web KanbanFlow¹³, una herramienta para gestión de proyectos Kanban que nos permite crear y administrar tareas.

Las "tarjetas" del tablero Kanban serán todas las tareas que hay que realizar en el proyecto. Cada iteración se compondrá de la realización de un conjunto de ellas. Inicialmente tendremos las siguientes tareas:

Iteración 1:

- Estudio portales web turísticos
- Análisis de las ventajas de relacionar una opinión con el tipo de viajero
- Selección del portal web
- Selección del periodo de tiempo y de las ciudades a estudiar

Iteración 2:

- Estudio de la estructura de la página seleccionada para implementar un crawler
- Implementación del crawler
- Definir la estructura de la base de datos
- Carga de información a la base de datos

Iteración 3:

- Realizar un prototipo para diseñar la página web

Iteración 4:

- Selección de una librería para dibujar mapas
- Implementación de los mapas

¹³ <https://kanbanflow.com/>

Iteración 5:

- Selección de una librería para dibujar gráficos
- Implementación de los gráficos

Iteración 6:

- Selección de una librería para dibujar un wordcloud
- Implementación del wordcloud

Iteración 7:

- Desarrollo de un sistema de aprendizaje automático

La situación del tablero Kanban inicialmente sería la que se muestra en la Ilustración 3.13.

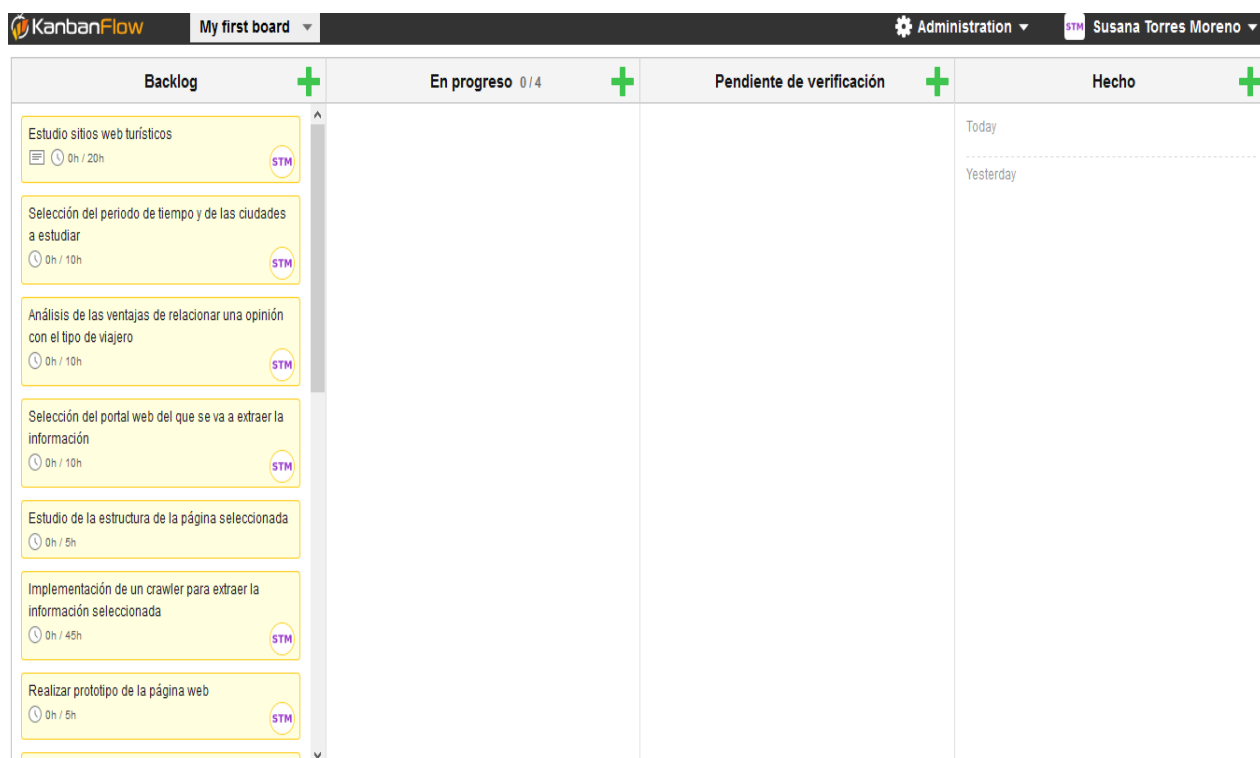


Ilustración 3. 13. Situación inicial tablero Kanban

7.1. Iteración 1

En esta iteración se realizarán las tareas relacionadas con los estudios necesarios para contextualizar el proyecto. La situación del tablero Kanban al comenzar esta iteración es la que se visualiza en la Ilustración 3.14.

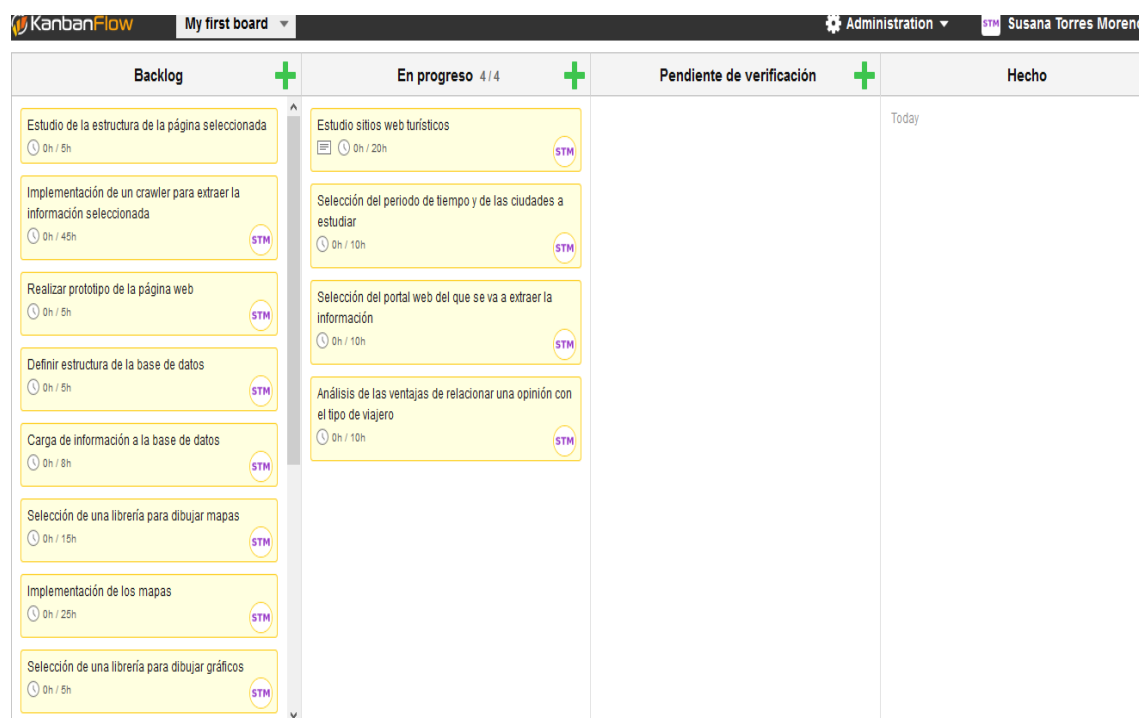


Ilustración 3. 14. Situación tablero Kanban al inicio de la iteración 1

En esta iteración se realizó un estudio sobre los sitios web turísticos que se podían utilizar como posible fuente de información para desarrollar la plataforma web. Se contemplaron como posibilidades Booking, Wimdu, Airbnb, El tenedor y TripAdvisor. Tras estudiar la estructura de cada una de las páginas y la forma en la que los usuarios podían expresar sus valoraciones y opiniones se decidió elegir TripAdvisor porque era la que tenía una estructura más adecuada y además, la que tenía mas cantidad de opiniones por hotel o apartamento.

Una vez que teníamos clara la fuente de información, se realizó un segundo estudio para decidir el periodo de tiempo y las ciudades en las que se iba a centrar el proyecto. Debido a la gran cantidad de información que maneja TripAdvisor se decidió centrarse únicamente en España debido a su gran

potencia turística, y dentro de España en Andalucía, por su gran importancia turística dentro del país. Se decidió extraer únicamente las capitales y no las comunidades autónomas al completo por motivos de eficiencia al tratar con tanta cantidad de datos. Como una primera versión nos pareció suficiente hacer una aproximación inicial únicamente con las capitales de provincia. También por este mismo motivo se decidió que únicamente se utilizarían los datos pertenecientes al periodo de tiempo 2016 - 2017.

En esta iteración también se realizó un estudio sobre las posibles ventajas que podría tener relacionar el tipo de viajero con la opinión de un viajero, y las posibles mejoras en la personalización de servicios que podría desencadenar tener en cuenta este dato. La industria del turismo está cambiando, cada vez más se enfoca hacia la personalización de servicios¹⁴. Es importante para el éxito de las empresas aprender a utilizar la inteligencia de mercado basada en los datos para dirigir ofertas¹⁵. En definitiva, la información que los viajeros comparten en Internet sobre sus viajes permiten a las empresas turísticas distinguir perfiles y tendencias que pueden ayudar a la toma de decisiones a la hora de decidir una estrategia de marketing.

Los datos sobre los perfiles de los turistas se pueden utilizar para ayudar a predecir el comportamiento del turista, su elección de vacaciones o destino y sus formas de experimentar¹⁶. A pesar de que esta predicción puede verse influenciada también por factores externos ajenos al perfil del turista, puede ser útil a la hora de ofrecer servicios más personalizados de los hoteles, conocer los perfiles de turistas que más visitan cada alojamiento, ciudad y en qué época lo hacen. Relacionarlo con la opinión puede ayudarnos a sacar conclusiones sobre las cuestiones que preocupan a cada perfil en sus comentarios.

¹⁴ <http://www.puromarketing.com/14/27470/reservas-turisticas-sera-online.html>

¹⁵ <http://www.ciospain.es/movilidad/el-sector-turistico-evolucion-a-hacia-estrategias-digitales>

¹⁶ <http://www.tourismtheories.org/?p=999>

Estos estudios tienen como fin contextualizar el proyecto y decidir los parámetros iniciales que se tomarán para su desarrollo. Una vez realizadas todas las tareas correspondientes a la iteración 1, pasarán al estado pendiente de verificación en el tablero Kanban hasta tener la reunión con las tutoras que revisarán los resultados obtenidos como se muestra en la Ilustración 3.15.

Backlog	En progreso 0 / 4	Pendiente de verificación	Hecho
Estudio de la estructura de la página seleccionada 0h / 5h Implementación de un crawler para extraer la información seleccionada 0h / 45h Realizar storyboard de la página web 0h / 5h Definir estructura de la base de datos 0h / 5h Carga de información a la base de datos 0h / 8h Selección de una API para dibujar mapas 0h / 15h Implementación de los mapas 0h / 25h Selección de una API para dibujar gráficos 0h / 5h		Estudio sitios web turísticos 0h / 20h Selección del portal web del que se va a extraer la información 0h / 10h Selección del periodo de tiempo y de las ciudades a estudiar 0h / 10h Análisis de las ventajas de relacionar una opinión con el tipo de viajero 0h / 10h	Today

Ilustración 3. 15. Situación tablero Kanban al finalizar la iteración 1

Una vez hecho esto, se tuvo la primera reunión con las tutoras para hablar sobre las conclusiones que se habían sacado tras la búsqueda de información. Y una vez decidido el sitio web del cual extraer la información, el periodo de tiempo y las ciudades en las cuales centrar el proyecto, estas tareas pasaron al estado "Hecho" del tablero Kanban.

Se han mostrado los pasos que se siguieron en el tablero para esta primera iteración a modo de ejemplo, Para el resto de tareas e iteraciones del proyecto se llevó a cabo el mismo proceso.

7.2. Iteración 2

La iteración 2 tendrá como objetivo la descarga del contenido deseado de tripAdvisor para satisfacer las historias de usuario 1 y 2 "Extracción de datos" y "Filtrado de datos", cuya descripción es " Como usuario quiero extraer los datos de la página web TripAdvisor para obtener las opiniones de los turistas" y " Como usuario quiero filtrar los datos extraídos para obtener los comentarios y valoraciones de los turistas" , respectivamente.

Para ello se tuvo que realizar un estudio de la estructura de la página y de la forma en la que iban cambiando las urls para poder descargar los comentarios, como ya se ha detallado en el apartado cuatro del capítulo dos. El resultado de utilizar las técnicas de web crawling y web scrapping será un fichero con todos los comentarios estructurados con sus respectivos atributos. Una vez obtenido este fichero, se importarán los datos a la base de datos. Para el manejo de la base de datos se ha utilizado el programa MySQL Workbench¹⁷ que es una herramienta visual para el manejo de MySQL. El esquema conceptual de la base de datos es el que se puede visualizar en la ilustración 3.16.

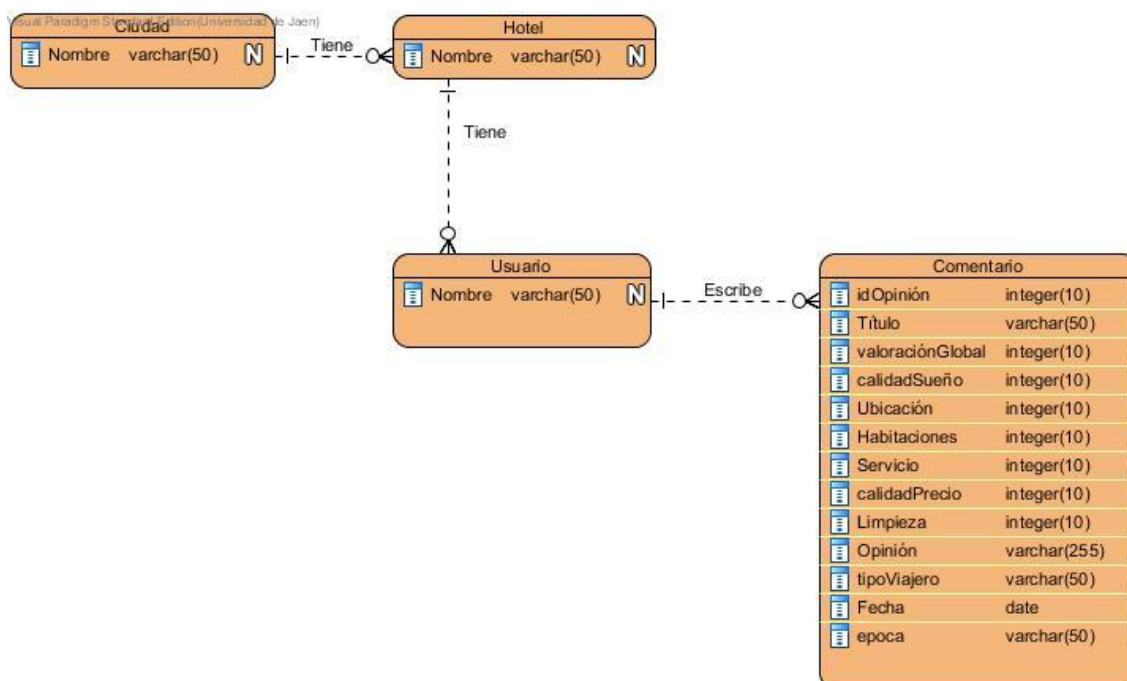


Ilustración 3. 16. Esquema conceptual BBDD

¹⁷ <https://www.mysql.com/products/workbench/>

A partir del esquema conceptual, se obtiene el esquema conceptual modificado. Para tratar las relaciones de uno a muchos tenemos que agregar en la tabla del lado de la cardinalidad N de la relación, la clave de la otra tabla y las columnas que correspondan a atributos de la relación. En este caso al realizar ese proceso desaparecerían las tablas hotel, usuario y ciudad ya que sólo tienen un atributo.

De esta manera, la base de datos tiene una única tabla en la que se han importado todos los comentarios. La estructura de la base de datos es la que se muestra en la Ilustración 3.17. e Ilustración 3.18.

código	ciudad	nombreHotel	Título	valoracionGlobal	calidadSueño	Ubicación	Habitaciones	Servicio	CalidadPrecio	Limpieza	Opinión
228564-336983701	Sevilla	Catalonia Giralda Hotel	Para derribarlo v reconstruirlo entero	1	-1	-1	-1	-1	-1	-1	Cuando te hospedas en un hotel de 4* esperas
228564-337519324	Sevilla	Catalonia Giralda Hotel	Pudiera estar mejor	2	-1	-1	-1	-1	-1	-1	Siento que este hotel no merece cuatro estrellas.
228564-338076593	Sevilla	Catalonia Giralda Hotel	Vacaciones bonitas	4	4	4	-1	-1	-1	3	El Hotel en General está bien situado. Está a 10.
228564-34047847	Sevilla	Catalonia Giralda Hotel	Excelente personal v confort	5	-1	-1	-1	-1	-1	-1	Excelente personal de recepción. muy amables.
228564-341933436	Sevilla	Catalonia Giralda Hotel	Muy buen hotel	5	-1	-1	-1	-1	-1	-1	Estuvimos cuatro noches en este gran hotel en
228564-344351549	Sevilla	Catalonia Giralda Hotel	Buenas vistas	3	-1	-1	-1	-1	-1	-1	Hotel normalito. pero si te dan una habitación d.
228564-347695442	Sevilla	Catalonia Giralda Hotel	Muy buen hotel	4	5	-1	-1	4	-1	5	Por estar completos. nos pasaron de una habita.
228564-348109175	Sevilla	Catalonia Giralda Hotel	Buen hotel. excelente trato.	4	-1	-1	-1	-1	-1	-1	Si vas de turista a Sevilla es un hotel que está
228564-349678487	Sevilla	Catalonia Giralda Hotel	Excelente hotel v magnifica acogida	5	-1	-1	-1	-1	-1	-1	He estado en muchas ocasiones en este hotel o.
228564-354507279	Sevilla	Catalonia Giralda Hotel	Categoría de pensionucha	1	-1	-1	-1	-1	-1	-1	No recomendamos para nada este hotel!! Cuan.
228564-354643006	Sevilla	Catalonia Giralda Hotel	Hav mejores opciones en Sevilla!	3	-1	-1	-1	-1	-1	-1	Es un hotel que se auto proclama de 4* pero ao.
228564-355224370	Sevilla	Catalonia Giralda Hotel	Elefantes en una cacharrería	3	2	-1	-1	-1	4	2	La ubicación está muy bien v personal de receo.
228564-355718713	Sevilla	Catalonia Giralda Hotel	Excelente!	5	-1	-1	-1	-1	-1	-1	El hotel está ubicado en una zona inmejorable d.
228564-356405393	Sevilla	Catalonia Giralda Hotel	Acogedor	4	-1	-1	-1	-1	-1	-1	Es un hotel acogedor v tiene una buena ubicad.
228564-360007001	Sevilla	Catalonia Giralda Hotel	Peor de lo esperado	3	3	-1	-1	3	-1	2	Reservamos una tr. suite v no tenía nada que v.
228564-360038122	Sevilla	Catalonia Giralda Hotel	Bastante bien!!!	4	5	-1	3	-1	-1	5	Lo primero que quiero comentar es la atención d.
228564-360374843	Sevilla	Catalonia Giralda Hotel	No vendría mal una remodelación.	3	4	-1	-1	-1	4	2	El hotel está bien situado a unos 10 min del cent.
228564-363155040	Sevilla	Catalonia Giralda Hotel	Bastante mediocre	3	3	-1	-1	-1	3	2	Esperaba un hotel acorde al número de estrella.
228564-363388265	Sevilla	Catalonia Giralda Hotel	Excelente Hotel	5	-1	-1	-1	-1	-1	-1	Estuve en este hotel con mi familia v la verdad ..
228564-365236329	Sevilla	Catalonia Giralda Hotel	Feria	1	-1	-1	-1	-1	-1	-1	Si este hotel.no mejora la acústica.aislacion .No.
228564-365409740	Sevilla	Catalonia Giralda Hotel	un cuatro estrellas que es una pensión	1	3	2	2	1	3	2	Habitaciones pequeñas v calurosas. servicio de
228564-365414727	Sevilla	Catalonia Giralda Hotel	Necesita rehabilitación v mal acceso	3	3	3	4	2	3	2	Muy malos accesos. inadmisibles para minusválid.
228564-365512344	Sevilla	Catalonia Giralda Hotel	Es un buen hotel de 3 estrellas	4	3	3	-1	-1	-1	3	Fue una estancia agradable. Estuve alojado en
228564-366696472	Sevilla	Catalonia Giralda Hotel	No está mal...	3	5	-1	-1	-1	5	3	Antes que nada, destacó el excelente personal
228564-367875802	Sevilla	Catalonia Giralda Hotel	LA CAMA ES ESTUPENDA	4	3	4	5	3	4	4	La mejor cama que he dormido en todos mis viaj.
228564-368366556	Sevilla	Catalonia Giralda Hotel	Para ser un cuatro estrellas. deia m....	2	3	3	3	1	4	2	Hotel anticuado (baños con desaoberfectos. sin al.
228564-370946833	Sevilla	Catalonia Giralda Hotel	Un aniversario maravilloso	5	5	-1	-1	5	3	-1	Me aloje aquí en diciembre de 2015 v tenía pen.
228564-371387224	Sevilla	Catalonia Giralda Hotel	Decepcionante	2	2	2	3	2	3	2	Reservamos una habitación triple que en la web.
228564-371732756	Sevilla	Catalonia Giralda Hotel	Falta de confort	3	3	4	2	2	4	4	Falto aire , no buen sueño habitación. Caliente
228564-372064960	Sevilla	Catalonia Giralda Hotel	Buena opción para alojarse en Sevilla	4	4	-1	-1	-1	4	3	Estuvimos 3 días alojados en este hotel. No est.
228564-373843952	Sevilla	Catalonia Giralda Hotel	mejorable se le nota antiguo	3	3	-1	-1	2	4	-1	Hotel algo anticuado v precio 100 noche SIN DE.
228564-373893261	Sevilla	Catalonia Giralda Hotel	NO LO RECOMIENDO	1	2	2	3	2	2	2	El hotel está bastante escondido. en un callejón.
228564-373960062	Sevilla	Catalonia Giralda Hotel	Estancia de 3 noches	3	5	4	3	3	5	3	La ubicación v atención del personal que allí tra.
228564-374186376	Sevilla	Catalonia Giralda Hotel	Buena ubicación cerca del centro v d...	3	5	4	-1	-1	-1	3	Cerca del centro. se puede ir andando a casi to.

Ilustración 3. 17. Estructura base de datos

Opinión	TipoViajero	Fecha	epoca
Cuando te hospedas en un hotel de 4* esperas ...	en pareja	2016-01-03	Navidad
Siento que este hotel no merece cuatro estrella...	-1	2016-01-05	Navidad
El Hotel en General está bien situado. Está a 10...	en pareja	2016-01-06	Navidad
Excelente personal de recepción. muy amables ...	en pareja	2016-01-16	Resto
Estuvimos cuatro noches en este gran hotel en ...	en pareja	2016-01-22	Resto
Hotel normalito, pero si te dan una habitación d...	de negocios	2016-02-01	Resto
Por estar completos, nos pasaron de una habita...	en pareja	2016-02-14	Resto
Si vas de turismo a Sevilla es un hotel que está ...	en pareja	2016-02-16	Resto
He estado en muchas ocasiones en este hotel p...	en pareja	2016-02-21	Resto
No recomendamos para nada este hotel!! Cuan...	con los amigos	2016-03-11	Resto
Es un hotel que se auto proclama de 4* pero ap...	con los amigos	2016-03-11	Resto
La ubicación está muy bien y personal de recep...	con los amigos	2016-03-14	Resto
El hotel está ubicado en una zona inmejorable d...	con los amigos	2016-03-15	Resto
Es un hotel acogedor y tiene una buena ubicad...	-1	2016-03-17	Resto
Reservamos una ir. suite y no tenía nada que v...	con la familia	2016-03-30	Seman...
Lo primero que quiero comentar es la atención d...	en pareja	2016-03-30	Seman...
El hotel está bien situado a unos 10 min del cent...	con los amigos	2016-03-31	Seman...
Esperaba un hotel acorde al número de estrella...	con los amigos	2016-04-10	Resto
Estuve en este hotel con mi familia y la verdad ...	con la familia	2016-04-11	Resto
Si este hotel, no mejora la acústica, aislación .No...	en pareja	2016-04-18	Resto
Habitaciones pequeñas y calurosas, servicio de ...	con los amigos	2016-04-19	Resto
Muy malos accesos, inadmisibles para minusválid...	con los amigos	2016-04-19	Resto
Fue una estancia agradable. Estuve alojada en ...	con la familia	2016-04-19	Resto
Antes que nada, destaco el excelente personal ...	en pareja	2016-04-23	Resto
La mejor cama que he dormido en todos mis via...	en pareja	2016-04-26	Resto
Hotel anticuado (baños con desperfectos, sin ai...	de negocios	2016-04-28	Resto
Me alojé aquí en diciembre de 2015 y tenía pen...	en pareja	2016-05-07	Resto
Reservamos una habitación triple que en la web...	con la familia	2016-05-09	Resto
Falto aire , no buen sueño habitación. Caliente ...	en pareja	2016-05-10	Resto
Estuvimos 3 días alojados en este hotel. No est...	con la familia	2016-05-11	Resto
Hotel algo anticuado y precio 100 noche SIN DE...	con los amigos	2016-05-17	Resto
El hotel está bastante escondido, en un callejón...	de negocios	2016-05-17	Resto
La ubicación y atención del personal que allí tra...	en pareja	2016-05-17	Resto
Cerca del centro, se puede ir andando a casi to...	en pareja	2016-05-18	Resto

Ilustración 3. 18. Estructura base de datos

La cual sigue el esquema que se muestra en la ilustración 3.19.

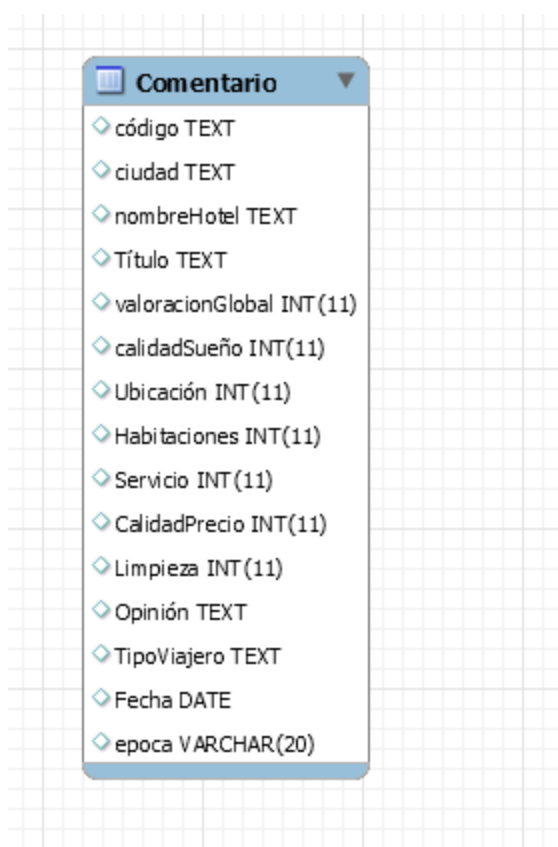


Ilustración 3. 19. Esquema base de datos

El atributo "época" referido a la época del año en la cual ha hecho un comentario un determinado viajero, se ha añadido posteriormente a la descarga de los datos. Su valor se ha establecido según los siguientes criterios:

Se ha considerado que el periodo de tiempo comprendido entre el 23-12-2016 y el 31-12-2016 , y el comprendido entre el 01-01-2016 y el 10-01-2016 pertenecerá a la época "Navidad". Esto se ha escogido así para únicamente tener en cuenta los comentarios pertenecientes a 2016.

Se ha considerado que el periodo comprendido entre el 01-06-2016 y el 31-08-2016 pertenecerá a la época "verano".

Se ha considerado que el periodo comprendido entre el 20-03-2016 y el 03-04-2016 pertenecerá a la época "Semana Santa", a pesar de que la Semana Santa de 2016 realmente fue la semana del 20 al 27 de marzo. Esto se ha decidido así debido a que muchas personas dejan sus comentarios algún tiempo después de su viaje. Para intentar agrupar también esos comentarios se ha dejado una semana más de margen, a pesar de que haciendo esto puede que se incluyan comentarios que no pertenezcan a esta época del año.

Se considerará como época "resto" a todos aquellos comentarios que se hayan realizado algún día fuera de los periodos mencionados anteriormente.

Una vez cargados los datos a la base de datos, se realizó una segunda reunión en la cual se comprobó que el contenido descargado era correcto, y en la cual se habló de lo que se realizaría durante la siguiente iteración.

7.3. Iteración 3

Durante la iteración 3 se diseñó la estructura de la página web, con el fin de satisfacer la historia de usuario seis referente a la interfaz web " Como usuario quiero poder visualizar los datos en una página web con una interfaz amigable". Se ha prestado especial atención a su diseño.

El diseño de una interfaz puede definir el éxito de cualquier proyecto, ya que si una interfaz está bien diseñada, el usuario encontrará fácilmente la respuesta que espera a su acción. Para ello, primero se realizó un storyboard con el fin de establecer la apariencia general de la interfaz. El storyboard se ha realizado usando la herramienta boords¹⁸, que permite dibujar storyboards online. El resultado de este storyBoard se muestra en la Ilustración 3.20.

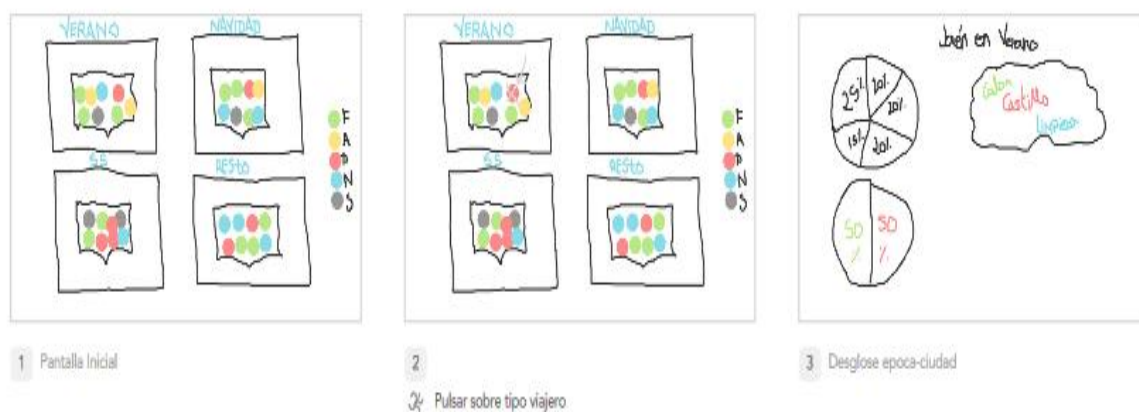


Ilustración 3. 20. StoryBoard de la aplicación

Se pretende crear una pantalla inicial con cuatro mapas, uno por cada época del año, en los cuales se debería visualizar por colores el tipo de viajero que más viaja en cada ciudad. Se puede pulsar sobre cualquier ciudad para obtener una visión más detallada de los datos de esa ciudad en esa época del año. En el caso del ejemplo se ha pulsado sobre Jaén en verano. En este caso debería mostrarse la cantidad de comentarios del resto de viajeros en un diagrama de sectores, una nube de palabras con las veinte más comunes escritas por los viajeros de Jaén en verano, y un diagrama con la cantidad de comentarios positivos y negativos de los viajeros. Esta misma interacción sería posible pulsando sobre cualquier otra ciudad en cualquier otra época del año. Además, pulsando sobre un tipo de viajero del gráfico de sectores, se puede obtener la nube de palabras específica de ese tipo de viajero y el gráfico con la polaridad de los comentarios de ese tipo. Se ha decidido colocar los mapas en la misma pantalla con el objetivo de que permita hacer comparativas entre épocas con más facilidad.

¹⁸ <https://app.boords.com/>

Una vez realizado el storyboard, se hizo un prototipo con la herramienta Axure RP 8¹⁹, con el objetivo de tener una imagen más clara de las posibles interacciones y estructura de la página.

A continuación, se mostrarán algunas imágenes del prototipo simulando la interacción de un usuario que pulsa sobre Jaén en Verano (Ilustración 3.21 e Ilustración 3.22.), y posteriormente sobre el sector del diagrama referente al tipo de viajero "en pareja" (Ilustración 3.23.).

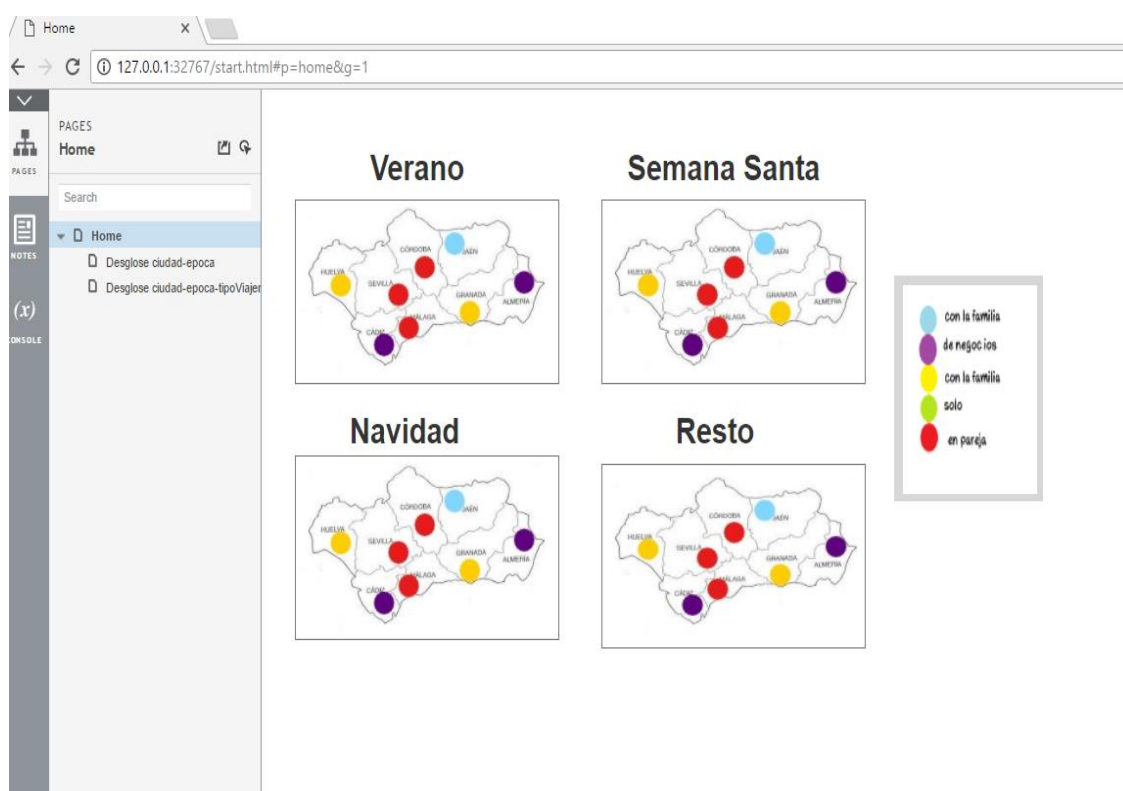


Ilustración 3. 21. Página principal prototipo

¹⁹ <https://www.axure.com/>



Ilustración 3. 22. Ver detalle Jaén en Verano

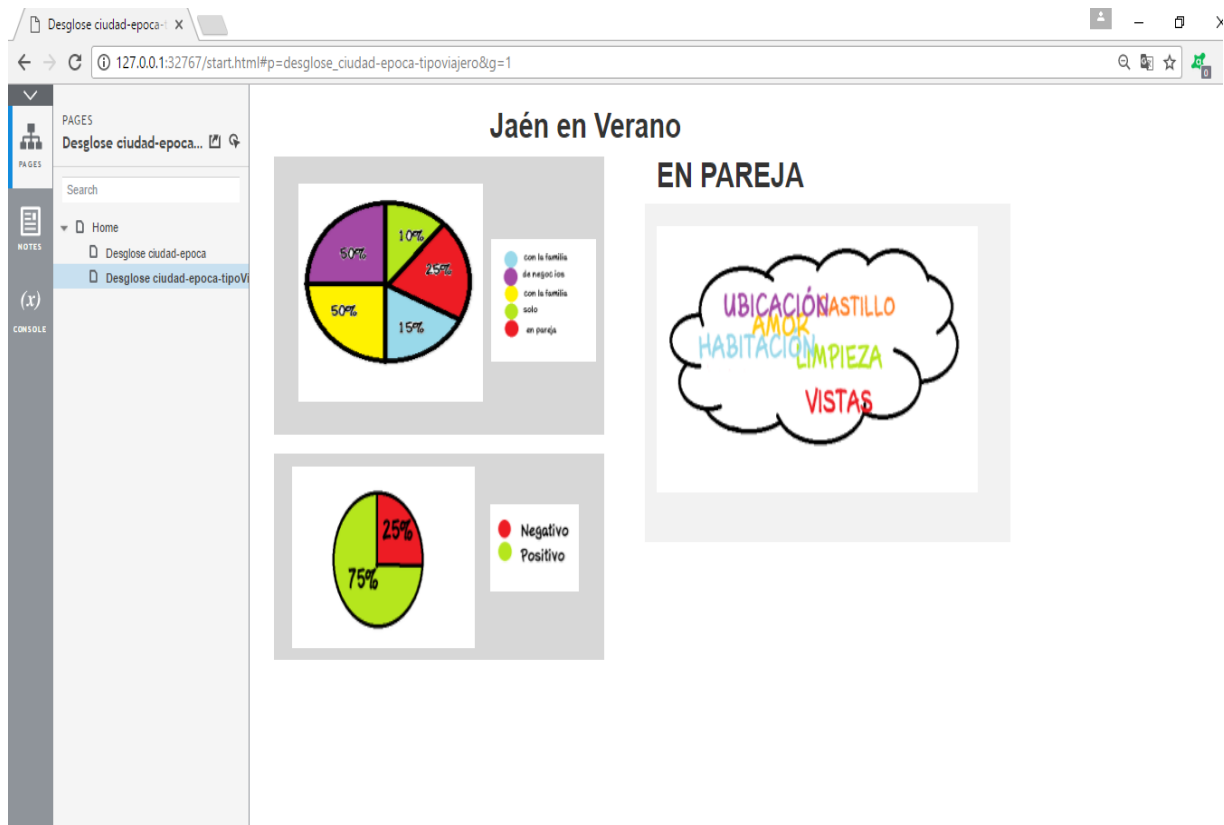


Ilustración 3. 23. Ver detalle Jaén en verano en pareja

Una vez clara la estructura que iba a tener la aplicación web, se tuvo que decidir cómo se iba a implementar y qué patrones se iban a utilizar durante su desarrollo. Se ha decidido utilizar el patrón arquitectónico "Modelo-Vista-Controlador" (MVC). Este patrón se basa en la idea de separar los datos y la lógica de negocio de una aplicación de la interfaz de usuario y del módulo encargado de gestionar los eventos y las comunicaciones. Está formado por tres componentes principales que son los siguientes y cuya interacción se puede visualizar en la Ilustración 3.24.

- Modelo: Aporta la funcionalidad central de la aplicación.
- Vista: Corresponde a un estilo y formato particular de presentación de información al usuario.
- Controlador: Acepta entradas del usuario en forma de eventos que disparan la ejecución de operaciones dentro del modelo.

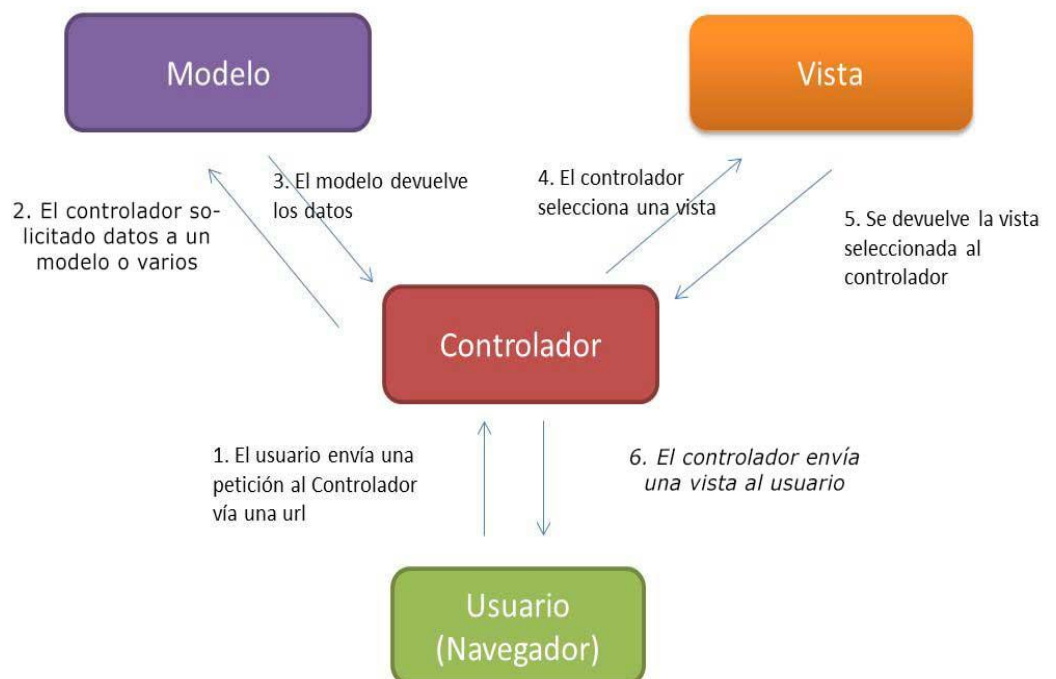


Ilustración 3. 24. Esquema patrón MVC

Las interacciones del usuario con la aplicación quedan representadas en el siguiente diagrama de secuencia (Ilustración 3.25), para el caso de uso de un usuario que quiera ver en detalle los gráficos correspondientes a una ciudad y una época del año concreta.

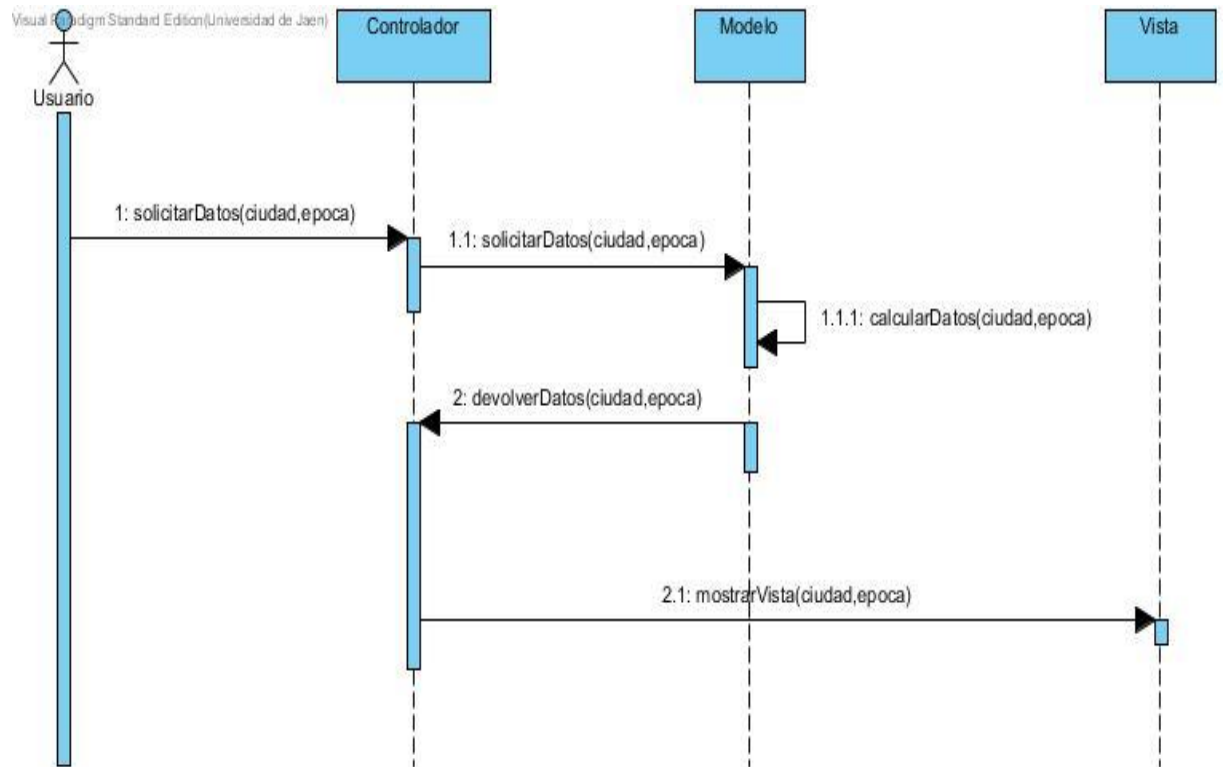


Ilustración 3. 25. Diagrama de secuencia

La función `calcularDatos()` englobaría todas las funciones que hay en el modelo para calcular los datos necesarios para mostrar los gráficos de las vistas correspondientes.

Tras esta iteración se tuvo una reunión con las tutoras y una vez aprobado el diseño de la página esta tarea pasó al estado "Hecho".

7.4. Iteración 4

Durante la iteración 4 se trabajó en la historia de usuario 3 "Visualización de mapas", cuya descripción era " Como usuario quiero poder visualizar los viajeros más comunes de cada ciudad y época vacacional". Para ello se investigó sobre algunas librerías para dibujar mapas. Al final se decidió utilizar MapBox²⁰ (Ilustración 3.26).



Ilustración 3. 26. Logo MapBox

Mapbox es una plataforma de mapeo para desarrolladores. Construye mapas y aplicaciones mediante APIs sencillas y potentes, utilizando para ello bibliotecas de código abierto. Dentro de Mapbox se ha utilizado, Mapbox GL JS, que es una biblioteca de JavaScript que utiliza WebGL para procesar mapas interactivos. MapBox tiene una amplia gama de mapas, pero como ninguno se adaptaba a las necesidades del proyecto, se utilizó MapBox Studio, una herramienta dentro de MapBox que permite crear tus propios estilos a partir de sus mapas.

El estilo generado con mapBox Studio es el que se muestra en la Ilustración 3.27.

²⁰ <https://www.mapbox.com/mapbox-gl-js/api/>



Ilustración 3. 27. Mapa de Andalucía MapBox Studio

Sobre los mapas se pueden dibujar distintas capas, representar puntos y cargar datos sobre ellos. Se pueden cargar los datos en diferentes formatos. Uno de ellos y el que se ha elegido para realizar el proyecto es el formato geojson.

Geojson es un formato estándar abierto diseñado para representar elementos geográficos sencillos, junto con sus atributos no espaciales, basado en JavaScript Object Notation.

De esta forma los mapas dibujarán el tipo de viajero más común de acuerdo a un fichero en formato geoJson. Habrá un fichero de este tipo por cada mapa, con la estructura que se visualiza en la Ilustración 3.28.

```

{"type": "FeatureCollection",
  "features":
    [{"type": "Feature", "properties": {"porcent": "49.64%", "epoca": "Verano", "viajero": " en pareja", "ciudad": "Cádiz"},
      "geometry": {"type": "Point", "coordinates": [-6.2684345022, 36.5008762379097]}},

    {"type": "Feature", "properties": {"porcent": "43.119%", "epoca": "Verano", "viajero": " en pareja", "ciudad": "Jaén"},
      "geometry": {"type": "Point", "coordinates": [-3.7331593542, 37.8258451226024]}},

    {"type": "Feature", "properties": {"porcent": "49.859%", "epoca": "Verano", "viajero": " en pareja", "ciudad": "Granada"},
      "geometry": {"type": "Point", "coordinates": [-3.5907775816, 37.1886273389356]}},

    {"type": "Feature", "properties": {"porcent": "46.565%", "epoca": "Verano", "viajero": " en pareja", "ciudad": "Almería"},
      "geometry": {"type": "Point", "coordinates": [-2.45, 36.83333]}},

    {"type": "Feature", "properties": {"porcent": "42.857%", "epoca": "Verano", "viajero": " en pareja", "ciudad": "Huelva"},
      "geometry": {"type": "Point", "coordinates": [-6.929323, 37.2582017]}},

    {"type": "Feature", "properties": {"porcent": "43.106%", "epoca": "Verano", "viajero": " en pareja", "ciudad": "Sevilla"},
      "geometry": {"type": "Point", "coordinates": [-5.9591776906, 37.3914105564361]}},

    {"type": "Feature", "properties": {"porcent": "47.289%", "epoca": "Verano", "viajero": " en pareja", "ciudad": "Córdoba"},
      "geometry": {"type": "Point", "coordinates": [-4.7086738547, 37.8550964348286]}},

    {"type": "Feature", "properties": {"porcent": "44.918%", "epoca": "Verano", "viajero": " en pareja", "ciudad": "Málaga"},
      "geometry": {"type": "Point", "coordinates": [-4.3971722687, 36.7585406465564]}}}
  ]
}

```

Ilustración 3. 28. Ejemplo fichero GeoJson

Una vez dibujados los mapas y representados los datos sobre ellos se obtuvo el resultado que se visualiza en la Ilustración 3.29.

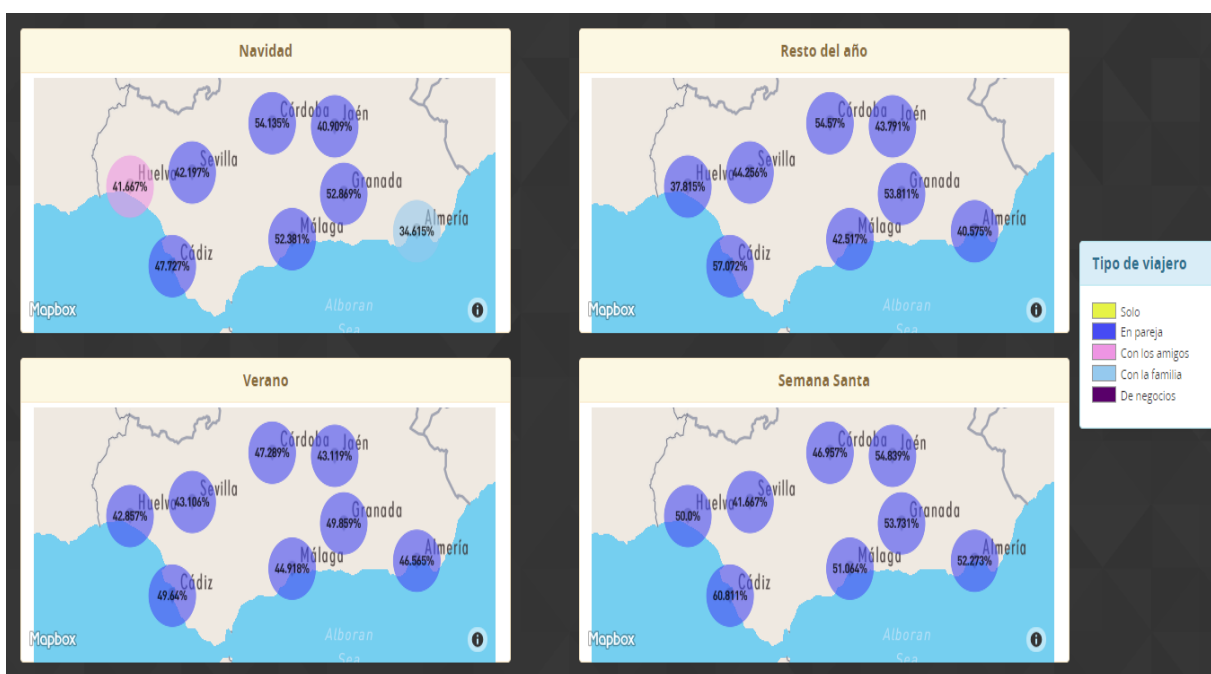


Ilustración 3. 29. Visualización de mapas por épocas

Durante la reunión con las tutoras se llegó a la conclusión de que el tipo de viajero más común en cualquier época del año, y prácticamente en todas las ciudades eran aquellos que viajaban en pareja. Como visualizando el mapa desde este punto de vista no quedaba una imagen demasiado clara, se decidió representar los datos desde el otro punto de vista, es decir, mostrar las épocas del año en las cuales viajaba más cada tipo de viajero. Esta segunda visualización quedó como se puede ver en la Ilustración 3.30.

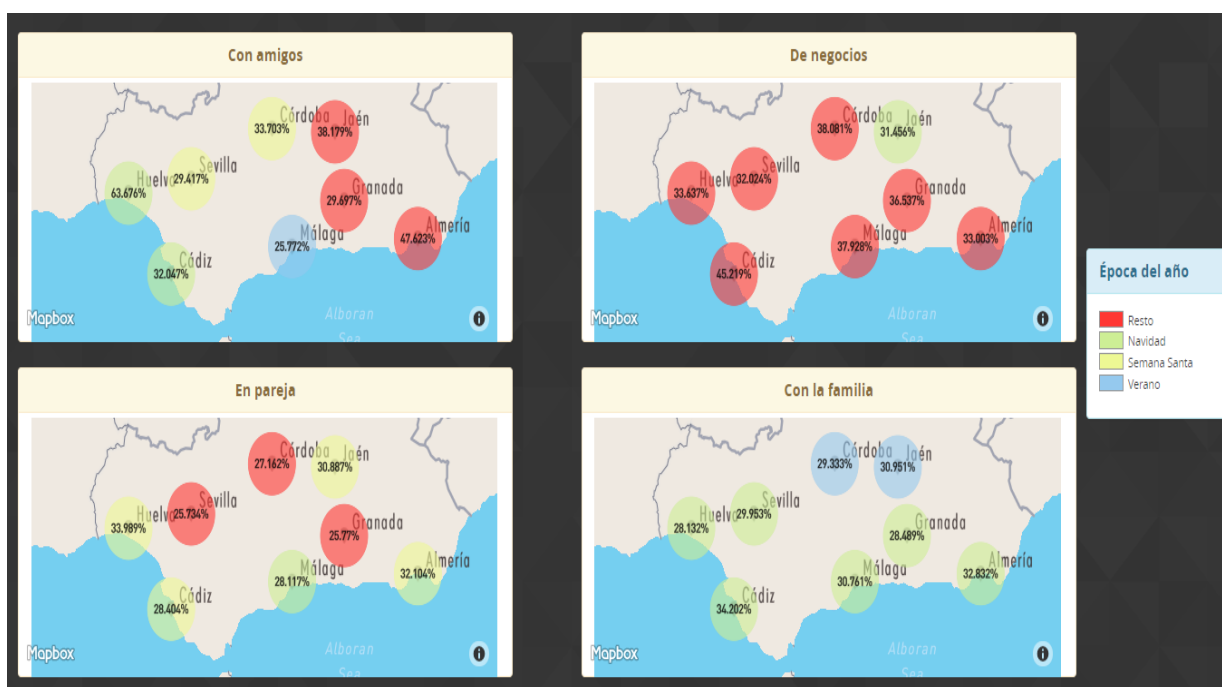


Ilustración 3. 30. Visualización de mapas por tipo de viajero

En esta segunda visualización no se muestra el tipo de viajero que viaja en solitario simplemente por el motivo de no recargar la página con cinco mapas en la misma pantalla. Se ha obviado este tipo porque son los viajeros con menos cantidad de comentarios.

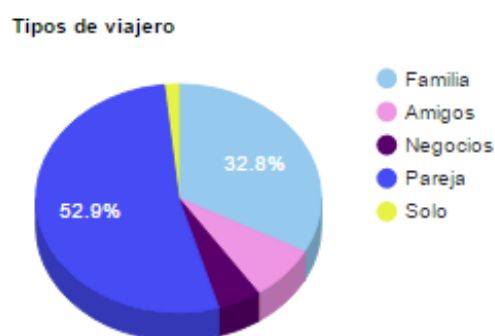
7.5. Iteración 5

Durante la iteración 5 se realizaron las tareas necesarias para satisfacer las historias de usuario 4 y 5, es decir, " Como usuario quiero poder visualizar la cantidad de comentarios de cada tipo de viajero en cada ciudad y época del año" y " Como usuario quiero poder visualizar la polaridad de los comentarios distinguiendo por tipo de viajero, ciudad y época del año" respectivamente.

Para la implementación de los dos gráficos se ha utilizado Google Chart. El primero será un gráfico de sectores, en google chart denominado "Pie chart", que representará la cantidad de comentarios de cada tipo de viajero en la época y ciudad seleccionadas.

El segundo gráfico será un "Donut chart" con la cantidad de comentarios positivos y negativos de los viajeros. En caso de que se pulse sobre un tipo de viajero en el gráfico de sectores, el donut de la polaridad global se cambiará por el específico de ese tipo de viajero. Se considerarán como comentarios positivos aquellos comentarios con una valoración de tres o superior, y se considerarán como comentarios negativos los que tengan una valoración de uno o dos.

El gráfico de los tipos de viajeros quedó como se puede ver en la Ilustración 3.31.



**Ilustración 3. 31. Ejemplo gráfico de sectores
Granada en Navidad**

El gráfico de la polaridad quedó como se puede ver en la Ilustración 3.32.

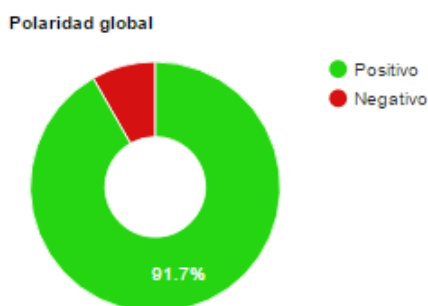


Ilustración 3. 32. Ejemplo de gráfico de polaridad Granada en Navidad

Como se ha explicado en el apartado anterior, también estaría la opción de ver el gráfico de sectores desde el punto de vista de las épocas del año en lugar de por tipo de viajero.

Si pulsamos por ejemplo sobre "Granada en pareja" ,el gráfico de sectores que se mostraría es el que se puede ver en la Ilustración 3.33.



Ilustración 3. 33. Ejemplo gráfico de sectores Granada en pareja

Tras esta iteración se comprobó que los datos mostrados por los gráficos eran correctos y que se visualizaban de forma clara y completa. Una vez hecho esto, esta tarea pasó al estado "Hecho".

7.6. Iteración 6

Durante esta iteración se completó la historia de usuario número 6 "Visualización de las palabras más usadas por los viajeros", que decía " Como usuario quiero poder visualizar las palabras que más utilizan los viajeros en los comentarios distinguiendo por tipo de viajero, ciudad y época del año". Para ello, se ha decidido representar las palabras mediante una nube de palabras.

Una nube de palabras o nube de etiquetas es una representación visual de las palabras que conforman un texto, en donde el tamaño es mayor para las palabras que aparecen con más frecuencia. En esta nube se representarán las veinte palabras más repetidas en los comentarios por cada ciudad, época vacacional y tipo de viajero. Para ello se utilizó la librería d3-cloud de Jason Davies²¹ escrita en JavaScript.

Como aparece detallado en el apartado 2.5. Análisis de la información, se han utilizado técnicas de Procesamiento del Lenguaje Natural y la librería NLTK para procesar los textos y obtener las palabras más comunes.

Una vez obtenidos los resultados, se observó que dos de las palabras más utilizadas por todos los viajeros eran "Hotel" y "Habitaciones". Con motivo de que no aportan ningún significado se han eliminado del resultado final. Una vez hecho esto por cada ciudad, época del año y viajero habrá un fichero con las palabras más comunes mencionadas en los comentarios y su frecuencia tal y como se muestra en la Ilustración 3.34.

²¹ <https://github.com/jasondavies/d3-cloud>

```
habitación 404
granada 270
centro 262
desayuno 245
parking 168
precio 153
ubicación 144
ciudad 142
zona 129
recepción 122
baño 119
personal 118
servicio 114
piscina 113
cama 108
alhambra 101
coche 94
calidad 89
calle 86
día 80
```

Ilustración 3. 34. Ejemplo de fichero de frecuencias de palabras Granada en verano en pareja

Habrán tres carpetas de ficheros de palabras, una para las palabras más repetidas por ciudad y época, otra para las palabras más repetidas por ciudad y tipo de viajero, y otra para las palabras más repetidas por ciudad, época y tipo de viajero.

El ejemplo de la nube de palabras para el fichero anterior es la que se muestra en la Ilustración 3.35.

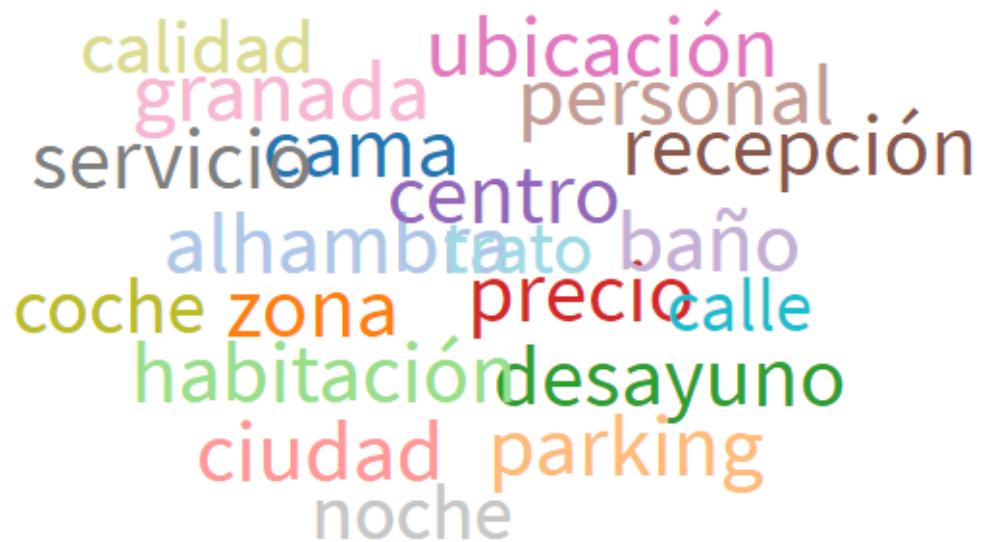


Ilustración 3. 35. Ejemplo nube de palabras Granada en verano en pareja

7.7. Iteración 7

Con el objetivo de satisfacer las historias de usuario siete y ocho, "Disponible en todos los navegadores" y "Disponibilidad en cualquier momento" cuyas descripciones son "Como usuario quiero poder visualizar la página web en cualquier navegador" y "Como usuario quiero poder visualizar la página web en cualquier momento" respectivamente. En esta iteración se comprobó que en todas las ocasiones en las cuales se intentó acceder a la página el acceso fue correcto. Y además se accedió desde Edge, Google Chrome, Firefox y Opera y su visualización fue correcta también.

Además en esta iteración se completó la tarea "Desarrollo de un sistema de detección del tipo de viajero". Las técnicas y herramientas utilizadas para el desarrollo de este sistema están detalladas en el apartado 2.6. Detección automática del tipo de viajero.

8. Herramientas para la implementación del sistema

8.1. *Lenguajes de programación*

Para el desarrollo del sistema se ha utilizado el lenguaje de programación Python²², en su versión 3.5. Python (Ilustración 3.36) es un lenguaje de programación interpretado cuya filosofía pone especial énfasis en que su sintaxis favorezca un código legible. Es un lenguaje de programación multiparadigma, soporta orientación a objetos, programación imperativa y, en menor medida, programación funcional.



Ilustración 3. 36. Logo Python

Además, se ha utilizado la distribución Python Anaconda (Ilustración 3.37.). Anaconda es un administrador de paquetes, administrador de entorno y contiene una colección de más de 720 paquetes de código abierto.



Ilustración 3. 37. Logo Anaconda

8.2. *Herramientas*

Durante todo el desarrollo del proyecto se ha utilizado PyCharm Community Edition 2016 (Ilustración 3.38.). Pycharm es un entorno de

²² <https://www.python.org/>

desarrollo integrado (IDE) multiplataforma utilizado para desarrollar en el lenguaje de programación Python. Es desarrollado por la compañía checa JetBrains. Proporciona análisis de código, un depurador gráfico, un probador de unidad integrado, integración con sistemas de control de versiones , y soporta el desarrollo web con Django .



Ilustración 3. 38. Logo PyCharm

La interfaz del programa tiene el aspecto que se muestra en la Ilustración 3.39.

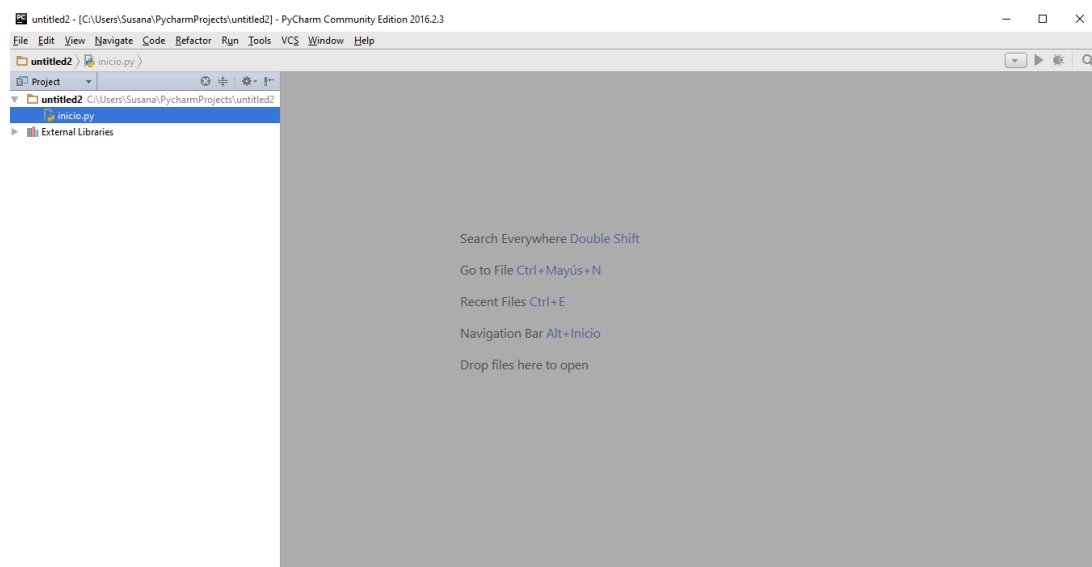
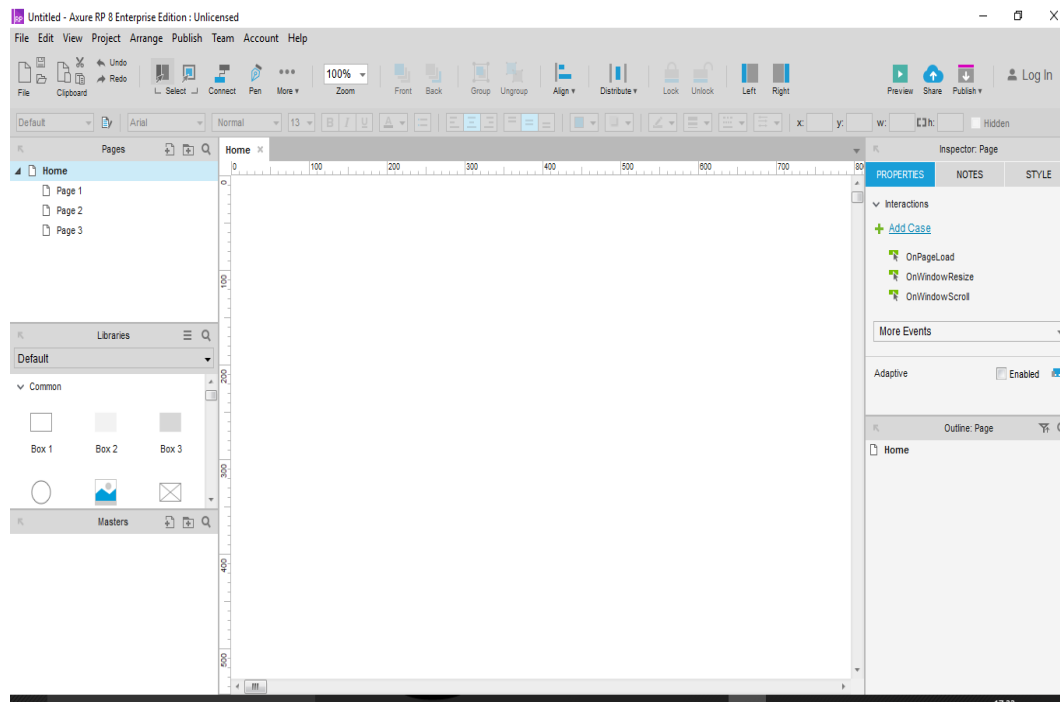
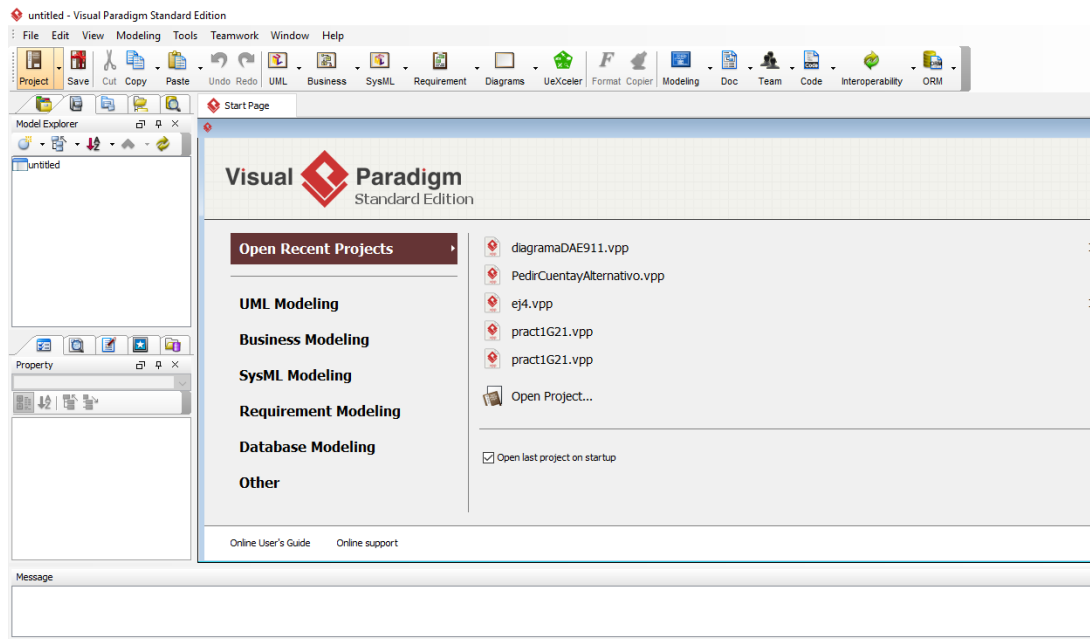


Ilustración 3. 39. IDE PyCharm Community Edition

Para el desarrollo del prototipo de la página web se ha utilizado el programa Axure RP 8 (Ilustración 3.40), una herramienta que permite realizar prototipos de aplicaciones web de una manera simple, permitiendo no solo el diseño de la interfaz sino la simulación de las interacciones.

**Ilustración 3. 40. Axure RP 8**

Para el desarrollo de los diagramas se ha utilizado el programa Visual Paradigm 12.2. (Ilustración 3.41). Visual Paradigm es una herramienta para desarrollo de aplicaciones utilizando modelado UML. Se ha elegido este programa por su familiaridad ya que es una herramienta que se ha utilizado durante todo el transcurso de la carrera.

**Ilustración 3. 41. Visual Paradigm**

8.3. Desarrollo página web

Para el desarrollo de la página web se ha utilizado el framework Flask (Ilustración 3.42.). Flask²³ es un framework escrito en Python. Su gran ventaja es su facilidad a la hora de construir una página web de forma rápida y con una cantidad mínima de líneas de código. Está basado en la especificación WSGI de Werkzeug, el motor de templates Jinja2 y tiene una licencia BSD.

**Ilustración 3. 42. Logo Flask**

Para el diseño de la interfaz web se ha utilizado Bootstrap²⁴. He elegido usar Bootstrap por su familiaridad ya que lo utilizamos en la asignatura

²³ <http://flask.pocoo.org/>

²⁴ <http://getbootstrap.com/>

"Desarrollo de aplicaciones web". Bootstrap es un framework desarrollado y liberado por Twitter que tiene como objetivo facilitar el diseño web.

Como sistema de gestión de base de datos se ha utilizado MySQL (Ilustración 3.43.). MySQL es un sistema de administración de bases de datos (*Database Management System, DBMS*) para bases de datos relacionales. MySQL²⁵ es la base de datos de código abierto más popular existente actualmente. Se ha elegido MySQL por su rendimiento, fiabilidad y facilidad de uso.



Ilustración 3. 43. Logo MySql

²⁵ <https://www.mysql.com/>

Capítulo 4:

Conclusiones y trabajo futuro

1. Conclusiones

La decisión de elegir este proyecto fue motivada por una asignatura optativa que curse en cuarto curso "Procesamiento del Lenguaje Natural". Decidí que quería aprender más sobre este tema ya que hoy en día es una rama de la informática en auge y con numerosas aplicaciones. Mis tutoras, Maite Martín Valdivia y Salud María Jiménez Zafra, me propusieron este tema en concreto porque yo quería indagar más en temas relacionados con perfiles de usuario y redes sociales. Se decidió elegir como ámbito de aplicación el turismo, por la cantidad de páginas web que ofrecen información sobre alojamientos y experiencias de los usuarios, cuya extracción de conocimiento podría ser mejor aprovechada. Además, es indudable el papel fundamental que toma el turismo en la economía de España.

Durante la primera fase del proyecto, en la cual busqué información, descubrí que la llegada de la Web 2.0. y su aplicación al ámbito del turismo, el turismo 2.0., supusieron un gran giro entorno a la planificación de viajes de los turistas. Generando todo esto una gran cantidad de información que permitía tanto a usuarios como a empresas tomar decisiones en base a ella. Este tema me ha resultado muy interesante ya que hasta ahora desconocía que la aplicación de la informática al marketing pudiera suponer una ayuda a la hora de establecer estrategias.

Además, en la última fase del proyecto, realicé algunas tareas relacionadas con el aprendizaje automático que es un tema que no se ve en profundidad en la carrera, y me ha parecido una rama muy interesante y de la cual me gustaría aprender más en un futuro. He aprendido mucho en este sentido sobre algoritmos y librerías que no conocía y que pueden ser muy útiles a la hora de desempeñar tareas en este ámbito.

En un sentido más práctico he aprendido una gran cantidad de tareas que desconocía hasta ahora como son:

- Qué es un web crawler y cómo se extrae información de la web.

- Herramientas y librerías para procesar textos. A pesar de que ya conocía NLTK, en este proyecto he aprendido nuevas funcionalidades.
- La librería Scikit-Learn. Me sorprendió la facilidad con la que se podía implementar un algoritmo de aprendizaje automático y la potencia y variedad de funcionalidades que tenía esta librería.
- He aprendido a programar en Python. Tenía unas nociones básicas que he mejorado durante el desarrollo del proyecto.
- Desarrollo web con un framework en Python, Flask.
- Uso de MapBox, una herramienta muy útil a la hora de incluir mapas interactivos en una aplicación.

En cuanto a los resultados obtenidos de las palabras más utilizadas por los viajeros para describir su viaje se ha llegado a la conclusión de que muchas de las preocupaciones de los viajeros son comunes entre sí. Las palabras más repetidas en prácticamente todos los casos son palabras relacionadas con el personal, la ubicación, y que haya desayuno. Además, otras preocupaciones de los viajeros son el baño, el precio, la existencia de parking, la limpieza, el trato y calidad del hotel. También dependiendo de la ciudad se habla sobre lugares de interés de esa ciudad, en el caso de Cádiz, Málaga y Almería se habla de mar y playa. En el caso de Jaén se habla del Castillo, la Catedral y el Parador, en Granada se habla de la Alhambra, en Córdoba de la Mezquita y en Sevilla de la Giralda. Distinguiendo por época del año las diferencias que encontramos son las apariciones de las palabras "año", "cena", "celebraciones" en Navidad. "Piscina" y "Aire" en Verano, y "semana", "paseo", "patio", "calle" y "familia" en Semana Santa. Distinguiendo por tipo de viajero podemos observar como las familias hablan de "familia" y "niño". Entre los comentarios de viajeros que viajan con amigos podemos encontrar palabras como "tapas", "fiesteros" y "salmorejo". Los viajeros de negocios hablan sobre "trabajo" y "negocios".

Como resumen podemos deducir que en general los viajeros se preocupan principalmente por los mismos temas aunque hay diferencias entre los temas de los que hablan según la época, ciudad y tipo de viajero.

Habría que hacer un estudio más exhaustivo eliminando todas aquellas palabras comunes entre todos para sacar más conclusiones de estos datos.

En cuanto a mi conclusión del proyecto, creo que la elección tanto el enfoque del proyecto como del ámbito de aplicación fueron acertadas y me han permitido aprender sobre diversos temas que no había profundizado durante la carrera y que, a mi parecer, tienen un futuro prometedor en la informática.

2. Trabajo futuro

Durante la implementación del detector del tipo de viajero se observó que la mayoría de opiniones recogidas eran de turistas que viajaban en pareja, lo cual ha supuesto que los datos estén desbalanceados. Una mejora para esto podría ser recopilar más datos de viajeros e igualar la cantidad de opiniones. Esto se podría hacer extrayendo opiniones de otros años o de otros portales web turísticos. También habría que hacer un estudio más exhaustivo sobre las palabras más relevantes para cada tipo de viajero con el objetivo de darles más peso y que el detector diera mejores resultados.

En cuanto a la aplicación web se debería aumentar el enfoque del proyecto para todas las ciudades de España. También se podrían extraer más conclusiones y estadísticas y aumentar la información que ofrece. Sería interesante mostrar gráficos sobre las valoraciones de cada tipo de viajero en las categorías limpieza, ubicación, calidad del sueño, relación calidad-precio, habitaciones y servicio.

Bibliografía

Arcos, V. A., Gutiérrez, S. S. M., & Hernanz, R. P. (2014). La aplicación empresarial del marketing viral y el efecto boca-oreja electrónico. Opiniones de las empresas/Business application of viral marketing and Electronic Word-of-Mouth. Firm opinions. *Cuadernos de Gestión*, 14(1), 15.

Betancourt, G. A. (2005). Las máquinas de soporte vectorial (svms). *Scientia et technica*, 1(27).

DELLAROCAS, C. (2003): «The digitization of word of mouth: promise and challenges of online feedback mechanisms». *Management Science*, vol. 49, nº. 10, pp. 1407-1424

Fondevila-Gascón, J., Mir-Bernal, P., Puiggròs-Román, E., Muñoz-González, M., Berbel-Giménez, G., Gutiérrez-Aragón, Ó., Feliu-Roé, L., Santana-López, E., Rom-Rodríguez, J., Sorribas-Morales, C., L. Crespo, J. and Marqués, J. (2016). Semantic Fields to Improve Business: the hotels case. *Fronteiras: Journal of Social, Technological and Environmental Science*, 5(2), p.47.

Gruen, T. W., T. Osmonbekov, and A. J. Czaplewski. 2006. eWOM: The impact of customer-to-customer online know-how exchange on customer value and loyalty. *Journal of Business Research* 59 (4): 449–56.

Hackeling, G. (2014) *Mastering machine learning with scikit-learn*. 1st ed.

Leung, X., Bai, B. and Stahura, K. (2013). The Marketing Effectiveness of Social Media in the Hotel Industry: A Comparison of Facebook and Twitter. *Journal of Hospitality & Tourism Research*, 39(2), pp.147-169.

María-Dolores, S., García, J. and Mellinas, J. (2017). *Análisis del nivel de presencia de los establecimientos hoteleros de la Región de Murcia en la Web 2.0*. [online] *Revistas.um.es*. Available at: <http://revistas.um.es/turismo/article/view/170861> [Accessed 16 May 2017].

Mellinas, J., Martínez María-Dolores, S. and Bernal García, J. (2016). El uso de redes sociales por los hoteles como indicativo de gestión eficiente. *Tourism & Management Studies*, 12(2), pp.78-83.

Mir, P Calderón, R, Recalde, M. and Fondevila Gascón J.F. (2015) Cómo Internet transforma la gestión de la reputación. *TELOS*, 101. Recuperado desde:

https://telos.fundaciontelefonica.com/DYC/TELOS/LTIMONMERO/DetalleArticulo_101TELOS_PERSPECT1/seccion=1288&idioma=es_ES&id=2015061812440001&activo=6.do

Mohammad Reza Jalilvand, Sharif Shekarchizadeh Esfahani, Neda Samiei, Electronic word-of-mouth: Challenges and opportunities, *Procedia Computer Science*, Volume 3, 2011, Pages 42-46, ISSN 1877-0509, <http://dx.doi.org/10.1016/j.procs.2010.12.008>.

(<http://www.sciencedirect.com/science/article/pii/S1877050910003832>)

Keywords: Word of mouth; Online consumer; Consumer behavior

O'Reilly, T.(2005): "What is Web 2.0". Disponible en:

<http://oreilly.com/web2/archive/what-is-web-20.html>

Park, C., and T. M. Lee. 2009. Information direction, website reputation and eWOM effect: A moderating role of product type. *Journal of Business Research* 62 (1): 61–67.

Venegas, R. (2007). Clasificación de textos académicos en función de su contenido léxico-semántico. *Revista signos*, 40(63), 239-271.

Anexo A:

Manual de instalación

En el siguiente manual se explicarán los pasos a seguir para poner en funcionamiento la aplicación.

1. Instalación Python

Para instalar Python, en concreto la versión 3.5., hay que acceder a <https://www.python.org/downloads/> e instalar el archivo correspondiente para el sistema operativo en cuestión.

2. Instalación Flask

Flask es el framework que se ha utilizado para desarrollar la aplicación web, para instalarlo hay que utilizar la siguiente instrucción:

Para Windows:

- *pip install virtualenv*

Para Linux o MacOSX:

- *sudo pip install virtualenv*

Virtualenv es una herramienta de desarrollo en Python que se usa para crear entornos aislados para Python, en los que es posible instalar paquetes sin interferir con otros paquetes de Python del sistema.

El siguiente paso es crear un directorio y ejecutar virtualenv en dicho directorio. Se puede hacer con las siguientes instrucciones:

- *mkdir myproject*
- *cd myproject*
- *virtualenv venv*
- *venv\Scripts\activate*

Una vez hecho esto ya podemos instalar Flask con la siguiente instrucción:

- `pip install Flask`

3. Instalación MySQL

Hay que acceder a la página <https://dev.mysql.com/downloads/installer/> e instalar la versión que corresponda para cada equipo.

Una vez hecho todo esto basta con poner en la consola la instrucción siguiente en la carpeta contenedora del proyecto:

- `Python control.py`

Anexo B:

Manual de usuario

En este apartado se explicará cómo funciona la aplicación web a modo de manual de usuario. Para usarla es necesario un ordenador con conexión a Internet.

Esta aplicación tiene como función mostrar cuatro vistas de un mapa de Andalucía con los viajeros más comunes de cada capital en cada época del año. Estos mapas pueden verse desde dos puntos de vista diferentes, "Por tipo de viajero" o por "época del año ". La vista por tipo de viajero nos muestra un mapa por cada tipo de viajero, y en cada capital se coloreará la ciudad con la época del año en la que más turistas de ese tipo viajan a esa ciudad. En la vista por época del año se mostrará un mapa por cada época y se coloreará la ciudad con el color del tipo de viajero más común de esa ciudad en esa época.

Podemos ver la vista por tipo de viajero en la siguiente ilustración (Ilustración B.1.):

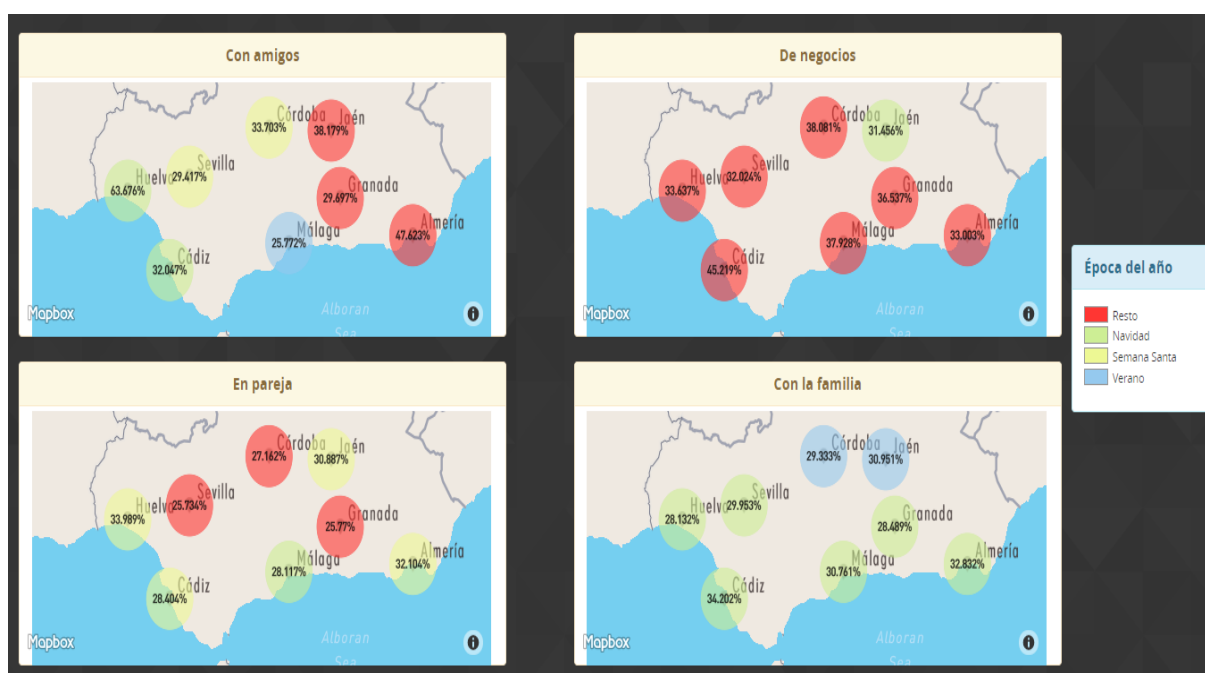


Ilustración B. 1. Vista por tipo de viajero

Podemos ver la vista por época del año en la siguiente ilustración (Ilustración B.2.):

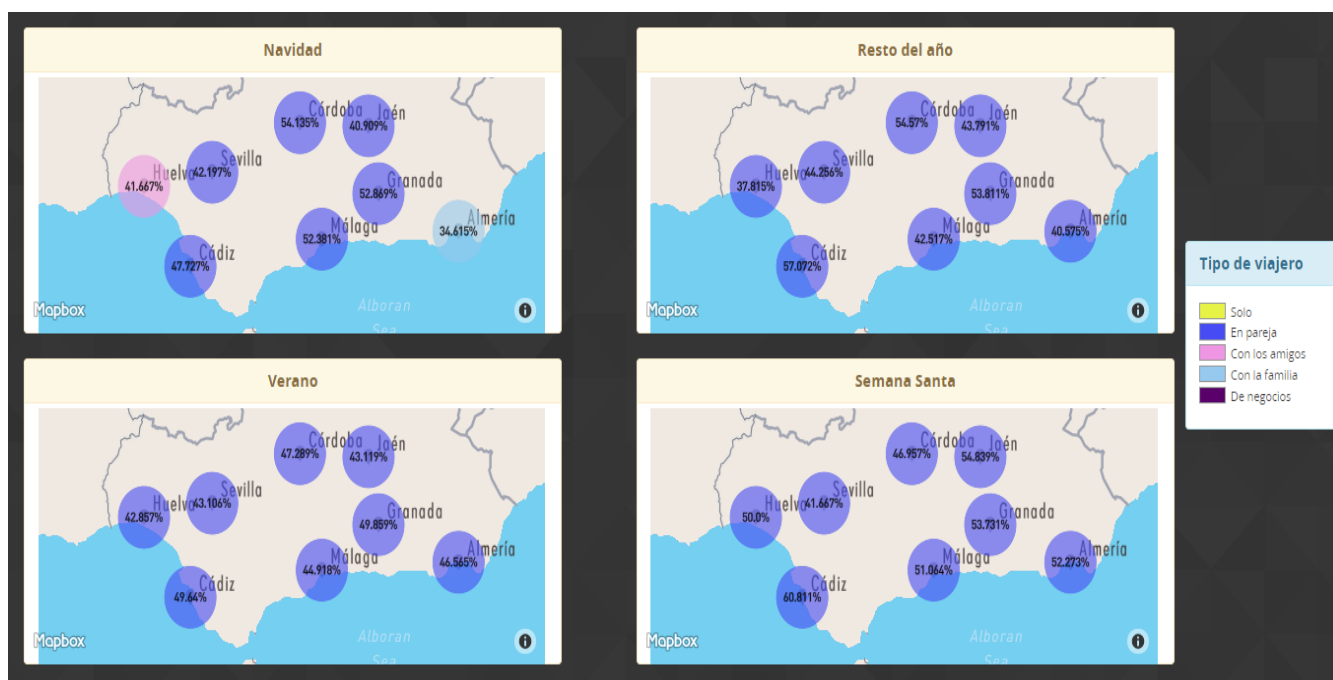


Ilustración B. 2. Vista por época del año

En cualquiera de las dos vistas se puede pulsar sobre una ciudad para ver en detalle los datos de esa ciudad.

Como ejemplo, si pulsamos sobre Málaga en Verano obtendremos el siguiente resultado (Ilustración B.3.):

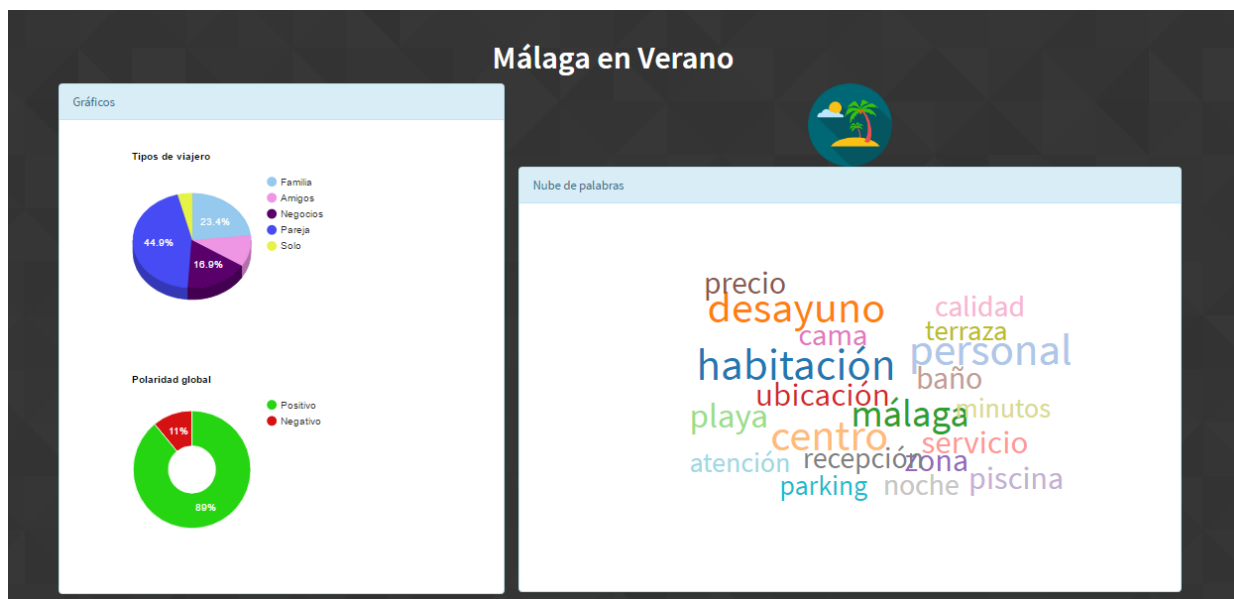


Ilustración B. 3. Vista en detalle de Málaga en verano

En esta ventana nos aparecen los porcentajes y cantidad de comentarios de cada tipo de viajero que ha comentado sobre Málaga en verano en forma de diagrama de sectores, la cantidad de comentarios positivos y negativos de los viajeros en forma de gráfico donut y las veinte palabras más comunes utilizadas por los viajeros para describir Málaga en verano.

Dentro de esta ventana tenemos la opción de pulsar sobre el gráfico de sectores sobre cualquier tipo de viajero. Una vez pulsado un tipo de viajero, el gráfico de polaridad y la nube de palabras cambiarán y se mostrará los específicos de ese tipo de viajero.

Por ejemplo, si pulsamos sobre "En pareja" se nos mostrará el siguiente resultado (Ilustración B.4.)



Ilustración B. 4. Vista en detalle de Málaga en verano en pareja

El tipo de viajero pulsado se representa en forma de icono sobre la nube de palabras.

La vista "Por época del año" funcionaría de la misma manera, mostraría en detalle el porcentaje de cada época del año en la que viaja un tipo de viajero en concreto en forma de gráfico de sectores, la polaridad de los comentarios en esa época y las palabras más usadas por los viajeros en esa época. Por ejemplo, si pulsáramos sobre Granada con la familia obtendríamos el siguiente resultado (Ilustración B.5.):



Ilustración B. 5. Granada con la familia

Para ver el detalle de una época habría que pulsar sobre ella en el gráfico de sectores. Si por ejemplo pulsamos sobre "Resto del año", obtendríamos los siguientes cambios (Ilustración B.6.):

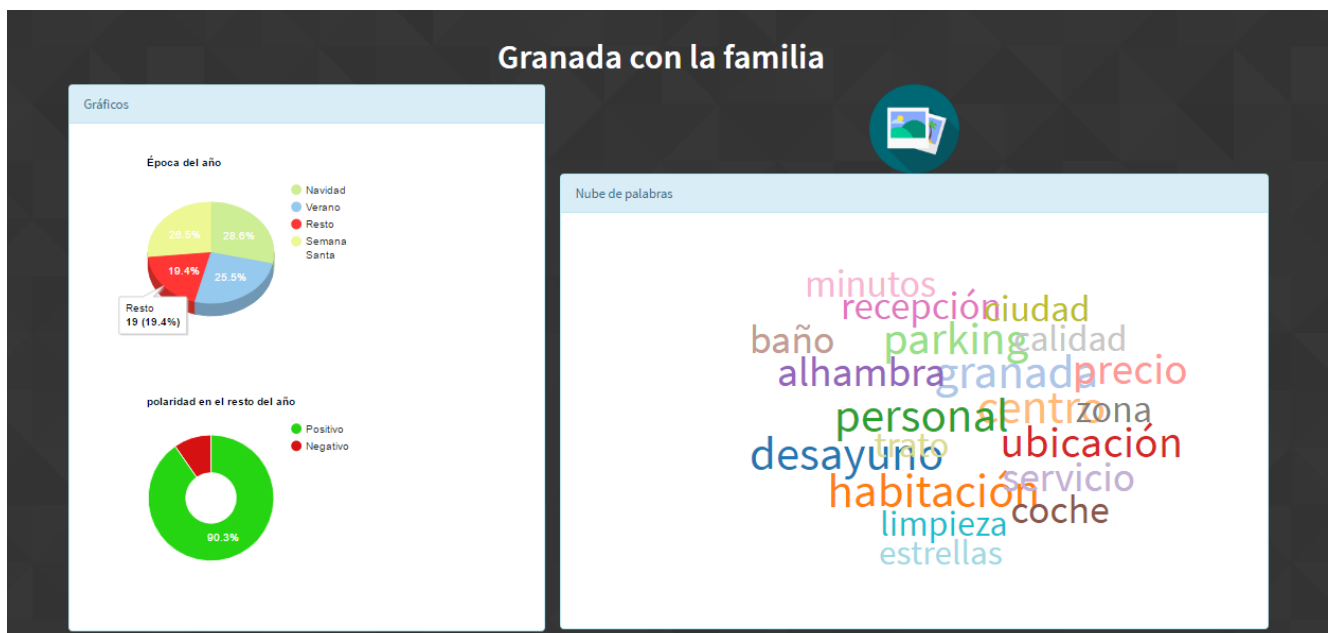


Ilustración B. 6. Granada con la familia en el resto del año

Las épocas del año aparecen indicadas como icono encima de la nube de palabras.

El significado de los iconos es el siguiente:

Épocas del año:

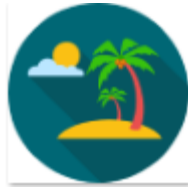


Ilustración B. 7. Icono Verano

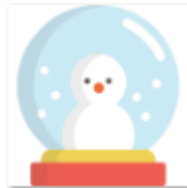


Ilustración B. 8. Icono Navidad



Ilustración B. 9. Icono Semana Santa

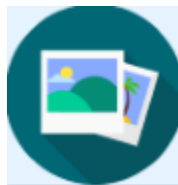


Ilustración B. 10. Icono Resto del año

Tipos de viajero:



Ilustración B. 11. Icono En pareja



Ilustración B. 12. Icono en familia



Ilustración B. 13. Icono de negocios



Ilustración B. 14. Icono en solitario



Ilustración B. 15. Icono con amigos

Anexo C:

Índice de ilustraciones

Capítulo 1:

Ilustración 1.1. Inversiones en 2016	4
Ilustración 1. 2. Diagrama de Gantt	6
Ilustración 1.3. Diagrama de Gantt	7
Ilustración 1.4. Planificación de tareas y costes asignados	9

Capítulo 2:

Ilustración 2. 1. Ejemplo puntuación alojamiento	16
Ilustración 2. 2. Ejemplo de comentario	16
Ilustración 2. 3. Ejemplo puntuación alojamiento	18
Ilustración 2. 4. Ejemplo de comentario	18
Ilustración 2. 5. Ejemplo puntuación alojamiento	19
Ilustración 2. 6. Ejemplo de comentario	20
Ilustración 2. 7. Ejemplo valoración restaurante	21
Ilustración 2. 8. Ejemplo valoración restaurante	21
Ilustración 2. 9. Ejemplo comentario restaurante	21
Ilustración 2. 10. Ejemplo puntuación hotel	22
Ilustración 2. 11. Ejemplo puntuación hotel	23
Ilustración 2. 12. Ejemplo de comentario	23
Ilustración 2. 13. Importancia para reservar alojamiento	25
Ilustración 2. 14. URL de la página principal de Málaga	26
Ilustración 2. 15. URL del hotel infanta cristina de Jaén	26
Ilustración 2. 16. URL de un comentario del hotel infanta cristina de Jaén	27
Ilustración 2. 17. Estructura de ficheros	27
Ilustración 2. 18. Estructura de ficheros	28
Ilustración 2. 19. Estructura de los comentarios	29
Ilustración 2. 20. Esquema Cross-Validation.	34
Ilustración 2. 21. Resultados SVM	36
Ilustración 2. 22. Resultados K-vecinos más cercanos	37
Ilustración 2. 23. Resultados Random Forest	37
Ilustración 2. 24. Resultados Naive Bayes	38
Ilustración 2. 25. Resultados SGD Classifier	38

Capítulo 3:

Ilustración 3. 1. Comparativa ágil y tradicional	43
Ilustración 3. 2. Tablero Kanban	44
Ilustración 3. 3. Estructura historia de usuario	47
Ilustración 3. 4. Historia de usuario 1	47
Ilustración 3. 5. Historia de usuario 2	48
Ilustración 3. 6. Historia de usuario 3	48
Ilustración 3. 7. Historia de usuario 4	49
Ilustración 3. 8. Historia de usuario 5	49
Ilustración 3. 9. Historia de usuario 9	50
Ilustración 3. 10. Historia de usuario 7	50
Ilustración 3. 11. Historia de usuario 8	51
Ilustración 3. 12. Historia de usuario 9	51
Ilustración 3. 13. Situación inicial tablero Kanban	54
Ilustración 3. 14. Situación tablero Kanban al inicio de la iteración 1	55
Ilustración 3. 15. Situación tablero Kanban al finalizar la iteración 1	57
Ilustración 3. 16. Esquema conceptual BBDD	58
Ilustración 3. 17. Estructura base de datos	59
Ilustración 3. 18. Estructura base de datos	60
Ilustración 3. 19. Esquema base de datos	60
Ilustración 3. 20. StoryBoard de la aplicación	62
Ilustración 3. 21. Página principal prototipo	63
Ilustración 3. 22. Ver detalle Jaén en Verano	64

Ilustración 3. 23. Ver detalle Jaén en verano en pareja	64
Ilustración 3. 24. Esquema patrón MVC	65
Ilustración 3. 25. Diagrama de secuencia	66
Ilustración 3. 26. Logo MapBox	67
Ilustración 3. 27. Mapa de Andalucía MapBox Studio	68
Ilustración 3. 28. Ejemplo fichero GeoJson	69
Ilustración 3. 29. Visualización de mapas por épocas	69
Ilustración 3. 30. Visualización de mapas por tipo de viajero	70
Ilustración 3. 31. Ejemplo gráfico de sectores Granada en Navidad	71
Ilustración 3. 32. Ejemplo de gráfico de polaridad Granada en Navidad	72
Ilustración 3. 33. Ejemplo gráfico de sectores Granada en pareja	72
Ilustración 3. 34. Ejemplo de fichero de frecuencias de palabras Granada en verano en pareja	74
Ilustración 3. 35. Ejemplo nube de palabras Granada en verano en pareja	75
Ilustración 3. 36. Logo Python	76
Ilustración 3. 37. Logo Anaconda	76
Ilustración 3. 38. Logo PyCharm	77
Ilustración 3. 39. IDE PyCharm Community Edition	77
Ilustración 3. 40. Axure RP 8	78
Ilustración 3. 41. Visual Paradigm	79
Ilustración 3. 42. Logo Flask	79
Ilustración 3. 43. Logo MySQL	80

Anexos:

Ilustración B. 1. Vista por tipo de viajero	95
Ilustración B. 2. Vista por época del año	96
Ilustración B. 3. Vista en detalle de Málaga en verano	97
Ilustración B. 4. Vista en detalle de Málaga en verano en pareja	98
Ilustración B. 5. Granada con la familia	99
Ilustración B. 6. Granada con la familia en el resto del año	99
Ilustración B. 7. Icono Verano	100
Ilustración B. 8. Icono Navidad	100
Ilustración B. 9. Icono Semana Santa	100
Ilustración B. 10. Icono Resto del año	100
Ilustración B. 11. Icono En pareja	100
Ilustración B. 12. Icono en familia	100
Ilustración B. 13. Icono de negocios	101
Ilustración B. 14. Icono en solitario	101
Ilustración B. 15. Icono con amigos	101