Trabajo de Fin de Grado

Procesamiento de Lenguaje Natural y su aplicación en servicios de hostelería

Natural Language Processing and its uses in hostelry services

José Gregorio Mesa Reyes

La Laguna, 5 de julio de 2016

Dña. **Isabel Sánchez Berriel**, con N.I.F. [42.885.838-S] profesor Titular de Universidad adscrito al Departamento de Nombre del Departamento de la Universidad de La Laguna, como tutor

C E R T I F I C A

Que la presente memoria titulada:

"Procesamiento del Lenguaje Natural y su aplicación en servicios de hostelería"

ha sido realizada bajo su dirección por D. **José Gregorio Mesa Reyes**,

con N.I.F. 78,633,637-H.

Y para que así conste, en cumplimiento de la legislación vigente y a los efectos oportunos firman la presente en La Laguna a 5 de julio de 2016

Agradecimientos

A mis padres, que siempre han estado ahí para apoyarme en todo cuanto he necesitado, y que ahora no están pasando por un buen momento por diversos motivos. Va por ellos antes que nadie.

A mis profesores, que de una manera u otra me han enseñado tanto los conocimientos necesarios para sacar las soluciones en todo momento en que se presenten problemas, y el aplomo para afrontar el futuro laboral.

A mi tutora Isabel, que "ha tenido que soportarme" durante todo el año tanto en las prácticas externas como en este trabajo, le agradezco enormemente por toda su ayuda, su predisposición tanto en horarios como en todos los problemas que hayan podido surgir y sus ánimos y energía para que todo salga adelante.

Y por último y no menos importante, a mis amigos... Sin todos esos momentos de risa y de nervios, y también los no tan buenos, sin todas esas noches en vela tratando de apurar al máximo posible para entregar las prácticas que nos quedan, incluso hasta esas varias ausencias no precisamente justificadas a clase... Sin todo esto y muchas cosas más, no habría existido el momento de descansar, de desconectar, de disfrutar de la vida, que muchas veces es más importante para uno mismo que cualquier otra cosa... De coger aliento para lo que venga al día siguiente, sea lo que sea.

Licencia

Resumen

El objetivo de este trabajo ha sido la elaboración de un sistema capaz de interpretar el lenguaje natural según la información extraída de una web de hoteles, para posteriormente llevar a cabo una clasificación de servicios ofrecidos y de satisfacción de cada uno de ellos por parte de los huéspedes.

En el caso de las características, serán obtenidas de manera objetiva y sin ningún patrón a seguir, debiendo desarrollar un algoritmo de reconocimiento a partir de la propia información y sin referencias, creando mecanismos que conduzcan a la obtención de las mismas de la manera más exacta posible.

En cuanto a las valoraciones, se analizará cada una de ellas de acuerdo a los valores contenidos en dos lexicones para determinar su índole y posteriormente determinar de qué características de las obtenidas se habla de forma positiva y de cuáles de forma negativa.

Todos los hoteles extraídos serán almacenados en una base de datos georreferenciada, con intención de un uso posterior en un mapa de zona que pueda mostrar los puntos únicos y servicios mejor y peor valorados de los mismos, y será capaz de procesar la información en dos idiomas diferentes, Español e Inglés.

A modo de resultado, se extraerá diversa información estadística en cuanto a características comunes de los hoteles, servicios que los distinguen entre el resto de la oferta hotelera, y respectivas valoraciones de los clientes de cada uno de ellas.

Palabras clave: Procesamiento del Lenguaje Natural, Análisis de hostelería, Georreferenciación de servicios hoteleros.

Abstract

The aim of this work is the development of a system capable of interpreting natural language according to the information extracted from hotel webpages and later on carry out a classification of services and guest satisfaction with each of them.

The characteristics will be obtained objectively and without any pattern followed, so I must develop a recognition algorithm from the information without any references, creating mechanisms that can lead to the most accurate solution.

The ratings will be analyzed according to the values contained in two lexicon so we can determine its nature and so which features have positive and negative ratings.

All extracted hotel information will be stored in a georreferenced database, with the intention of a later use in a zone map than could display single points features and ratings, and also will be able to process information in two different languages, Spanish and English.

Finally, as a result different statistical information will be extracted, such as common hotel features, unique services and respective customer ratings of each of them.

Keywords: Natural Language Processing, Hostelry analysis, Hotel services georeferencing.

Índice general

Capítulo 1
Introducción

Cuando se presenta una toma de decisiones, existen muchas variables que condicionan un posible resultado. En la gran mayoría de los casos, las personas solemos tener en cuenta a más información mejor, pudiendo ser ésta de índole totalmente dispar, pero todas igualmente útiles.

A la hora de trasponer todo ésto a una estancia hotelera de un viaje, prestamos especial atención a la oferta disponible en nuestro destino, la cual está compuesta por los diferentes servicios ofrecidos, el lugar de residencia, el clima y época del año, la facilidad para el acceso, etc.

Además, ponemos especial atención a los comentarios de aquellos que han estado antes que nosotros, y por los que podemos valorar en qué medida se cumplen cada una de dichas características, si son acordes a lo ofrecido y si las personas han quedado satisfechas con los servicios recibidos, o por el contrario no ha sido de su agrado.

Dado el inmenso volúmen de datos a los que somos capaces de acceder hoy en día, resulta una tarea titánica para una persona o grupo de personas el hecho de revisar y analizar cada una de estas características, así como las miles de valoraciones y opiniones que podemos recopilar.

Mediante el procesamiento del lenguaje natural y otros cómputos sobre información más estructurada somos capaces de determinar de manera bastante exacta todos estos aspectos, desarrollando algoritmos que nos permiten la automatización y procesamiento de toda esta información.

Este capítulo se orientará a la descripción general del problema, definición de los objetivos y alcance de los mismos, así como los antecedentes que han determinado la solución final del trabajo.

1.1 Objetivos

El objetivo principal de este proyecto se basa en la obtención de la información característica de hoteles y sus servicios llevando a cabo un procesamiento de las descripciones de los mismos, así como un análisis de las opiniones de los usuarios de cada uno de ellos, obteniendo qué aspectos de los mismos son valorados positivamente y cuáles de forma negativa.

Es importante mencionar que, pese a que de forma general los servicios ofrecidos por cada hotel tienen su mención aparte en algún punto de la web, el trabajo se centra en la extracción de las mismas a partir únicamente del análisis de sus descripciones, lo cual se ha intentado llevar a cabo a través de algoritmos de elaboración propia.

Para lograr este objetivo, es necesario un estudio y un trabajo previo de las diferentes técnicas, herramientas, módulos y fórmulas de procesado de texto que nos permitan alcanzar las metas propuestas. Cabe destacar que en el caso de las descripciones, en general las encontraremos escritas de manera correcta y objetiva, mientras que en las opiniones existe una complicación añadida dado el lenguaje coloquial utilizado por cada usuario.

Como desarrollo final, se pretende emplear todo lo estudiado y desarrollado en la creación de un sistema que permita la extracción y comparación de los servicios ofertados de cada hotel, así como un análisis de sentimiento sobre las opiniones que sus huéspedes merecen sobre los mismos.

El sistema actuará tanto en inglés como en español.

1.2 Alcance

Se pretende desarrollar un sistema que permita la extracción de un vocabulario clave de servicios de un hotel, así como un análisis de las valoraciones de los mismos. Con los resultados obtenidos se creará un sistema que permita la extracción de los servicios ofrecidos por cada hotel, así como la comparación de cada uno de ellos con el resto de hoteles analizados, de manera que podamos establecer qué características son comunes a una serie de hoteles y cuáles son únicas y, por tanto que diferencian a un hotel del resto de sus competidores.

Así mismo, también se hará un análisis de sentimiento sobre los comentarios de los usuarios, extrayendo de los mismos qué servicios son del agrado de los clientes y con cuáles están en desacuerdo. Se debe llevar a cabo:

- La creación de un sistema que permita la extracción de los servicios de cada hotel a través de su descripción.
- Comparación de los elementos del vocabulario clave extraído de la descripción de un hotel con el obtenido para el resto de hoteles, sintetizando qué características son comunes y cuáles no.
- Análisis de los comentarios de los usuarios, determinando qué valoraciones son positivas y cuáles negativas respecto a los servicios extraídos de ese hotel en cuestión.
- Extensión de lo expuesto en los puntos anteriores para dos idiomas, tanto inglés como español.
- Mostrar al usuario los resultados obtenidos, así como estadísticas y porcentajes relativos a los mismos.

Esta aplicación almacenará la información georreferenciada en una base de datos, de manera que pueda ser accedida una vez finalizada la extracción facilitando su uso.

Se ha implementado tanto para el inglés como para el español. Pese a que las herramientas y recursos de éste último no están tan extendidas ni alcanzan la misma calidad, los resultados han sido bastante satisfactorios, con lo que finalmente se han desarrollado ambos.

1.3 Antecedentes

En primer lugar, en cuanto al análisis de las descripciones y extracción de servicios se refiere, no se ha encontrado ninguna aplicación ya definida que lo lleve a cabo.

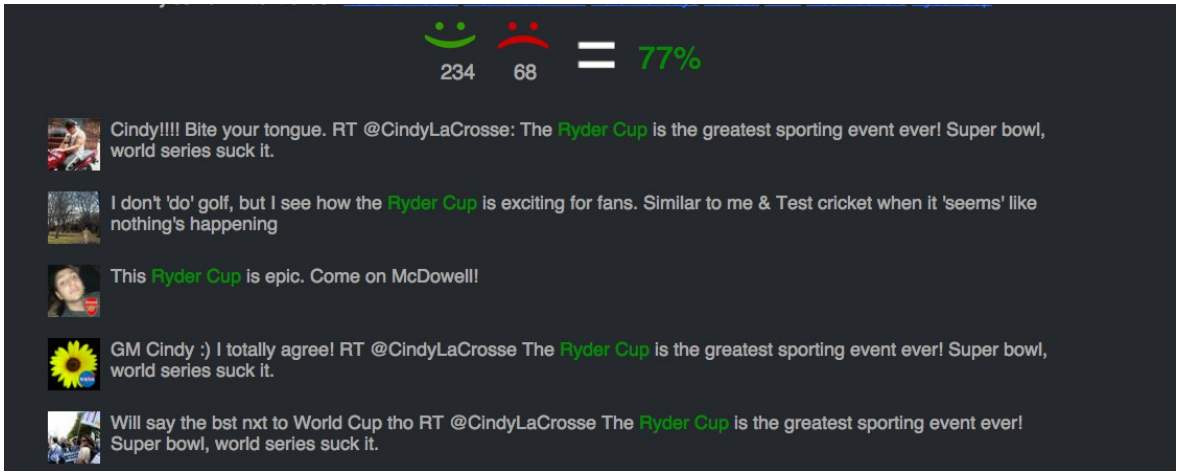
Esto es un hecho normal, dado que como citamos anteriormente, de manera general siempre aparecen los servicios más generales como: *piscina, caja fuerte, terraza, solarium*, etc. enlazados en algún punto de la web del hotel, con lo que extraerlos a partir de la descripción no suele ser nunca necesario.

Sin embargo, en las descripciones pueden aparecer otros servicios o especificaciones más concretas de estos, por ejemplo:

"Piscina al aire libre del hotel incluye cascada y zona infantil"

Este proyecto está más centrado en la obtención de dichas características que en el análisis de sentimiento propio de los comentarios, y constituyendo la tarea más complicada debido a la necesidad de tratar el bloque de información de manera completa y objetiva, debiendo determinar qué datos implican un servicio, cuáles un grado de dicho servicio, y qué parte de la descripción no aporta nada y por tanto será desechada.

En segundo lugar, en cuanto al análisis de sentimiento, existen diversas webs que podemos citar que cumplen con lo que se busca, como por ejemplo http://www.tweetfeel.com, las cuales permiten realizar búsquedas de texto y calificar los comentarios obtenidos como positivos o negativos. Por ejemplo, para el término *"Ryder Cup"*:



En nuestro caso, más que en la índole del comentario en su totalidad, nuestro objetivo es buscar qué valoración tienen para los usuarios las características obtenidas en el primer punto.

1.4 Destinatarios

En primer lugar, y otorgándole la mayor importancia, la parte fundamental de este proyecto ha sido la investigación de técnicas y herramientas para la implementación del sistema de análisis del lenguaje aplicados al ámbito turístico. El trabajo que conlleva la exploración de los diferentes módulos para cada uno de los lenguajes, su integración y prueba en los diferentes entornos y el análisis de los resultados obtenidos con cada uno de ellos es muy amplio, y ha conllevado el grueso de este proyecto. Es por eso que la investigación se lleva el papel principal de todo el desarrollo, sin que haya una finalidad expresa para el mismo.

A continuación, se intenta aportar al usuario unos resultados lo más exactos y objetivos posibles acerca de cada una de las características y las respectivas opiniones, de manera que se pueda comparar por medio de datos procesados y estadísticas la veracidad y el grado de satisfacción sobre los servicios buscados.

La herramienta puede interesar en aquellos casos que se quiera:

- Medir la satisfacción de los clientes respecto a los servicios ofertados por el hotel.
- Determinar qué características del hotel lo hacen único dentro del conjunto completo, y por tanto establecer distinciones.

Capítulo 2
Análisis del problema

2.1 Definición del problema

Como se ha introducido en el capítulo anterior, en nuestro caso contamos con dos aspectos del tratamiento del lenguaje; el más importante, orientado a la obtención de un vocabulario clave de servicios ofrecidos por un hotel a través de la información provista del mismo, y en segundo lugar el estudio de opiniones o sentimientos expresados por los clientes sobre cada uno de ellos.

En el primer caso, los textos a tratar contendrán casi en su totalidad una descripción objetiva, esperando una tasa de errores tipográficos bastante baja y con diversos calificadores para cada uno de los servicios ofertados. Un ejemplo de análisis de las mismas, resaltando característica y calificador, sería como sigue:

"*Todas las habitaciones incluyen baño de lujo de mármol, TV interactiva y conexión a internet.*"

En el segundo caso, la información obtenida será subjetiva, de manera que se analizará el contenido en busca de las claves que queremos valorar y el grado de satisfacción de cada una, sea positiva o negativa. Como ejemplo encontramos:

"*El baño de la habitación era precioso.*"

Destacar que en general, las personas sólo citamos la entidad en concreto, con lo que el procesamiento de las referencias se hará sólo con las claves, en este caso *baño*, y no *baño de lujo de mármol*.

2.2 Análisis de los textos tratados

Podemos discernir los dos análisis desarrollados y tratar cada uno de forma independiente:

2.2.1 Análisis de descripción

El primer apartado a tratar constituye las descripciones de los hoteles. Encontraremos en cada uno de ellos una sección dedicada a la presentación del hotel, indicando su ubicación, zonas atractivas de las cercanías, y exponiendo los servicios más relevantes y las características más destacadas del mismo, de manera que se atraiga la mayor cantidad de clientes posible.

Se ha decidido no tomar en cuenta ningún tipo de recurso, ya sea de servicios comúnmente ofertados, palabras clave de un vocabulario hotelero, ni grado de valoración alguna.

Para cada descripción, se recorrerá buscando todas aquellas palabras que puedan suponer un identificador (sustantivos), y a raíz de las referencias que pueda haber sobre ella (adjetivos, complementos, referencias...) decidir si es una palabra clave para nuestro objetivo.

Una vez obtenidas las claves se llevará a cabo una síntesis de las mismas, de manera que tengamos además como resultado final una comparación entre los hoteles analizados, separando los servicios afines a un conjunto de ellos de aquellas características que puedan hacer único a un hotel en concreto (o a un conjunto muy limitado).

Esto es así debido a la intención de crear un análisis puro de la información, sintetizando la misma a partir de identificadores de palabras, referencias de unas con otras e intentar obtener de la manera más exacta posible los datos que nos interesan a partir de una fuente completamente objetiva, lo cual añade un grado de dificultad importante al proceso.

2.2.2 Análisis de comentarios

Como segundo punto analizaremos los comentarios de los clientes, buscando las palabras clave previamente obtenidas en las valoraciones que hayan podido aportar.

De forma general, las valoraciones en páginas dedicadas no suelen tener opiniones dispares en una misma frase, como sí suele ocurrir por ejemplo en las redes sociales. Por tanto se ha desarrollado un análisis a nivel de frase, en las que valoraremos si se está hablando de forma positiva o negativa atendiendo al sentimiento asociado, para a continuación sintetizar las palabras clave en cada una de ellas y asociar qué servicios ha mencionado el cliente, así como establecer su grado de satisfacción o decepción con cada uno de ellos.

2.2.3 Implementación bilingüe

Por último, cabe destacar el desarrollo llevado a cabo en ambos idiomas, puesto que también ha conllevado problemas adicionales debido a que en términos de procesamiento de lenguaje no se trata de una simple traducción.

El inglés cuenta con una construcción en general más “cerrada” que el español. Además de las diferencias propias entre ambos idiomas tanto morfológica como sintácticamente, las frases tienen un orden más establecido y más restrictivo, con lo que la implementación de ambos módulos ha sido totalmente independiente la una de la otra.

Así mismo, dada la gran escala del inglés y por tanto la mayor disponibilidad de herramientas, módulos y técnicas de procesamiento de datos para el mismo, el procesamiento del español resulta una tarea algo más ardua de llevar a cabo.

2.3 Problemas abordados

Nuevamente debemos abordar cada problema por separado, dado que funcionan de forma independiente incluso dentro del tratamiento de la información:

2.3.1 Extracción de características

En primer lugar, los problemas con el análisis de la descripción han sido los más complicados. Contando sólo con la información objetiva que podemos sacar del análisis de los datos (morfología, sintaxis y correferencias), el problema se nos presenta a la hora de identificar qué aspectos entre los procesados nos hablan realmente de una característica o servicio del hotel y cuáles no.

Por tanto, el problema principal radica en encontrar las claves válidas. Entre los muchos sustantivos que encontraremos en cada descripción, debemos encontrar la manera de clasificarlos lo más exacto posible para que podamos distinguirlos. Del mismo modo, debemos poder identificar de la mejor forma todos aquellos modificadores y complementos que les hacen referencia, como adjetivos o bloques preposicionales, que serán necesarios para ubicar cada una de las claves y calificarlas.

Existen además muchos más puntos a tener en cuenta, como las conjunciones que suelen enlazar objetos (pudiendo ser éstos útiles para nosotros o no), o los propios signos de puntuación. Por ejemplo, en la siguiente frase:

"Todas incluyen baño de lujo de mármol y TV interactiva con conexión a internet. Hay un gran spa con salas de tratamientos, circuito de agua y gimnasio."

Vemos que nos podemos encontrar diversos complementos habilitando a un sustantivo, pudiendo cambiar indistintamente la posición (ir delante o detrás del nombre al que hace referencia) y varias características contenidas en una misma frase separadas por signos de puntuación y conjunciones.

2.3.2 Valoraciones de los usuarios

En segundo lugar, el problema a la hora de evaluar los comentarios se basa en determinar si una opinión es positiva o negativa. Para poder clasificarla, debemos contar con una base previa que nos permita conocer si una opinión es positiva o negativa, para lo que realizaremos un análisis a nivel de frase.

Una vez obtenida y con las características previamente extraídas, podremos buscar de cuáles de ella se habla y en qué sentido lo hace, determinando si es para bien o para mal:

Positivo: "La playa acogedora y las habitaciones muy bonitas."

Negativo: "El aislamiento de las habitaciones respecto al pasillo es mejorable."

A modo de ejemplo, vemos cómo los complementos utilizados en las oraciones nos indican su categoría, así como la evaluación de los servicios *"playa"* y *"habitación"* en cada uno de ellos.

Capítulo 3
Estudio previo

Constituye la parte más extensa del proyecto llevado a cabo. El estudio que se ha tenido que realizar de manera previa ha cubierto la mayor parte del tiempo y del trabajo realizado, siendo los resultados muy positivos en todos los aspectos.

Para poder hacer frente a los objetivos propuestos, se ha llevado a cabo un análisis bastante extenso acerca de muchos puntos, tanto a nivel de lenguaje de programación como en el ámbito del tratamiento del lenguaje natural, como expondremos más adelante; conceptos, formas de trabajo, posibles herramientas y módulos a utilizar, capacidades y filtros de las distintas librerías y salidas correspondientes, etc.

Así mismo, se tuvo especial cuidado en la valoración de cada una de ellas, puesto que el hecho de querer desarrollar el proyecto en dos idiomas diferentes provoca que la evaluación sea aún mayor (para el inglés hay mucha más variedad de opciones a tratar que para el español).

Como idea inicial, se planteó intentar extraer un “vocabulario común” a todos los hoteles a través de sus descripciones, para posteriormente abordar en profundidad cada una de ellas. Para el análisis de comentarios, sí deberíamos tomar una fuente que nos permita indicar si un comentario es positivo o negativo, para poder luego procesar un resultado a partir del mismo.

3.1 Python

En primer lugar, se dio libertad para elegir el lenguaje de programación, siendo Java y Python las dos opciones consideradas. Tanto durante la carrera como en las Prácticas Externas he estado trabajando mucho sobre Java, y en su defecto lenguajes fuertemente tipados (C++, Pascal), teniendo bastante control sobre los mismos y soltura a la hora de desarrollar código.

A raíz de ésto, la decisión ha sido escoger Python, tomándolo como un reto tanto a nivel personal como de cara al mundo laboral, puesto que es un lenguaje que siempre he tenido curiosidad por utilizar y del que además obtendré conocimientos de cara al futuro.

A modo de definición, Python constituye un lenguaje de programación de alto nivel, interpretado y multipropósito. Su filosofía hace hincapié en una sintaxis que favorezca un código legible, siendo ésta clara y concisa, y cuya potencia y flexibilidad lo convierten en un lenguaje muy productivo. Soporta orientación a objetos, programación imperativa y funcional, es multiplataforma y utiliza un tipado dinámico (comúnmente denominado lenguaje débilmente tipado).

3.1.1 Base de Python

Ha sido necesario un estudio previo acerca de este lenguaje así como bastante práctica, tanto del uso de sus tipos de datos como de la manera de trabajar sobre él. Dado que no es un modelo de lenguaje con el que estoy acostumbrado a trabajar se ha hecho algo complicado en ocasiones; el tipeo de datos, las implementaciones genéricas y la utilización de módulos han conllevado los mayores problemas. Por ejemplo, en ocasiones y sobre todo al principio, no sabía distinguir qué tipo de resultado devolvía un determinado método, y por tanto tampoco cómo debía tratarse. En ocasiones aún conociendo el tipo devuelto tampoco estaba claro qué métodos se podían aplicar, ni el origen de los errores que se producían.

3.1.2 Entorno de trabajo Windows

Se ha decidido trabajar en un entorno de Windows, con la consecuencia de búsqueda de IDE y varios problemas asociados muy leves. A medida que se ha ido avanzando en el proyecto, surgieron algunos problemas más importantes, debido sobre todo a diversos errores sobre las propias librerías, rutas de acceso, variables del sistema y descarga y actualización de módulos, pero con mayor o menor dedicación se han conseguido resolver todos.

Entre los más importantes podemos destacar las incidencias con la biblioteca de Python *"vcvarshall.dll"*, un error recurrente del compilador de C++ al trabajar sobre este entorno y para el cual no existe una solución determinada, pasando las opciones desde

actualizar la herramienta *setuptools* de Python, como instalar un compilador *MS Visual C++* o instalar las librerías precompiladas.

Así mismo, se me presentó la necesidad de compilar librerías procedentes de diferentes lenguajes, e incluso de configurar algún API de integración con Python para su funcionamiento. Para ello, se llevó a cabo un estudio y posterior uso de SWIG, una herramienta de desarrollo software que nos permite conectar programas escritos en C o C++ con una variedad de lenguajes de alto nivel, entre los que se incluyen lenguajes de scripting como JavaScript, PHP o Python.

3.1.3 Threading

Por último, cabe destacar el estudio llevado a cabo sobre el uso de hilos en este lenguaje. De esta manera podemos segmentar la carga de trabajo en distintos hilos que trabajarán de forma asíncrona, y teniendo especial cuidado de no hacer uso de una información que no haya sido completada.

Esto no ha sido necesario para que el sistema cumpla con los objetivos propuestos, pero se trabajó sobre ello con la idea a priori de que optimizaría el procesamiento de la información sobre todo en cuestiones de rapidez, y una vez llevado a cabo se ha confirmado un aumento considerable en la velocidad de procesamiento del programa.

3.2 Web Scraping

Se trata de una técnica utilizada mediante programas de software para la extracción de información de los sitios web. Se enfoca sobre todo en la transformación de datos sin estructura en la web (tales como el formato HTML) en datos estructurados que pueden ser almacenados y analizados en una base de datos central, una hoja de cálculo o alguna otra fuente de almacenamiento. Los usos más comunes del web scraping son la comparación de diversos elementos (precios en tiendas, datos del clima de cierta región, etc.) y para obtener información relevante de un sitio a través de los *rich snippets* (referencia a los fragmentos enriquecidos extraídos de algo, específicamente de las páginas de resultados de búsqueda). Para analizar todas estas variables se han utilizado dos módulos diferentes, UrlLib y BeautifulSoup:

3.2.1 UrlLib

Se trata de una herramienta contenida en las bibliotecas estándar de Python para la gestión de direcciones web y su contenido, de forma que podamos por ejemplo aportarle parámetros con el fin de utilizarlos en una búsqueda en dicha web, o algún tipo de configuración sobre la url, u obtener el código html de la página en cuestión.

Por medio de Urllib hemos podido formular la búsqueda de hoteles y extraer el código de la página, para posteriormente pasar al tratamiento del texto y obtención de datos.

3.2.2 Beautiful Soup

Se trata de una librería creada por terceros para el tratamiento y extracción de datos de archivos html y xml, agilizando en gran medida las búsquedas, navegación, filtros, modificaciones y parseos de la información contenida en los mismos.

Tenemos disponibles una gran cantidad de métodos, permitiendo realizar búsquedas según diversos patrones, como pueden ser un id o class concreta, obtener el valor de alguna de las variables definidas, o el contenido de texto de cierto párrafo, con unos resultados excelentes y una gran velocidad de procesado.

3.3 Procesamiento del Lenguaje Natural

En segundo lugar, se ha llevado a cabo una investigación extensa acerca del procesamiento del lenguaje natural. Comprender bien los conceptos y cómo se desgrana la información obtenida una vez se analiza cualquier tipo de texto es básico para poder trabajar sobre ello.

3.3.1 Conceptos y técnicas

Es necesario comprender cada aspecto implicado en el procesamiento del lenguaje, de los que se hablará de forma continua en este documento.

- Lexicón

Es una serie abstracta no ordenada de palabras pertenecientes a un lenguaje, una persona o una región, así como las reglas que permiten combinar las mismas.

- Etiquetado gramatical (Part-Of-Speech Tagging)

Comúnmente denominado POS Tagging, es un etiquetado de las palabras de un texto según su categoría gramatical. Puede presentarse de dos maneras diferentes, según se realice en base a la definición propia de una palabra, o bien en base a la función que ésta desempeña en su contexto. Por ejemplo, en la oración:

"El guapo entró en la sala."

Observamos que el término *guapo* analizado según su definición propia constituye un adjetivo, mientras que se trata de un sustantivo si atendemos a su desempeño en la frase.

No existe una única manera de etiquetado ni hay un convenio establecido acerca del mismo, con lo que dependiendo de los módulos utilizados para llevar a cabo el POS Tagging podremos encontrar diferentes resultados. Así mismo, diferentes idiomas tienen diferentes códigos para etiquetar cada palabra, con lo que tendremos que prestar especial atención.

A modo de ejemplo, y por medio del etiquetado utilizado finalmente en este proyecto, se muestran dos tablas con los tagsets de ambos idiomas:

English Tagset: Penn Treebank Part-Of-Speech Tags			
Tag	**Description**	**Tag**	**Description**
CC	Coordinating conjunction	***PRP$***	Possessive pronoun
CD	Cardinal Number	***RB***	Adverb

DT	Determiner	***RBR***	Adverb, comparative
EX	Existential *there*	***RBS***	Adverb, superlative
FW	Foreign word	***RP***	Particle
IN	Preposition or subordinating conjunction	***SYM***	Symbol
JJ	Adjective	***TO***	*to*
JJR	Adjective, comparative	***UH***	Interjection
JJS	Adjective, superlative	***VB***	Verb, base form
LS	Last item marker	***VBD***	Verb, past tense
MD	Modal	***VBG***	Verb, gerund or present participle
NN	Noun, singular or mass	***VBN***	Verb, past participle
NNS	Noun, plural	***VBP***	Verb, non-3rd person singular present
NNP	Proper noun, singular	***VBZ***	Verb, 3rd person singular present
NNPS	Proper noun, plural	***WDT***	Wh-determiner
PDT	Predeterminer	***WP***	Wh-pronoun
POS	Possessive ending	***WP$***	Possessive wh-pronoun
PRP	Personal pronoun	***WRB***	Wh-adverb

Tabla 3.3.1.1: English Tagset (Penn Treebank)

Tagsed Español: Etiquetas gramaticales Eagles		
Código	**Atributo**	**Valor**
Adjetivos		
AQ0CP0	Categoría	Adjetivo (A)
AQ0CP0	Tipo	Calificativo (Q), Ordinal (O)

AQ0CP0	Grado	Aumentativo (A), Diminutivo (D), Comparativo (C), Superlativo (S)
AQ0CP0	Género	Masculino (M), Femenino (F), Común (C)
AQ0CP0	Número	Singular (S), Plural (P), Invariable (N)
AQ0CP0	Función	- (0), Participio (P)
Adverbios		
RG	Categoría	Adverbio (R)
RG	Tipo	General (G), Negativo (N)
Determinantes		
DD0MS0	Categoría	Determinante (D)
DD0MS0	Tipo	Demostrativo (D), Posesivo (P), Interrogativo (T), Exclamativo (E), Indefinido (I), Artículo (A)
DD0MS0	Persona	Primera (1), Segunda (2), Tercera (3)
DD0MS0	Género	Masculino (M), Femenino (F), Común (C), Neutro (N)
DD0MS0	Número	Singular (S), Plural (P), Invariable (N)
DD0MS0	Poseedor	Singular (S), Plural (P)
Nombres		
NCMS000	Categoría	Nombre (N)
NCMS000	Tipo	Común (C), Propio (P)
NCMS000	Género	Masculino (M), Femenino (F), Común (C)
NCMS000	Número	Singular (S), Plural (P), Invariable (N)
NCMS000	Clasif. Semántica	Persona (SP), Lugar (G0), Organización (O0), Otros (V0)
NCMS000	Grado	Aumentativo (A), Diminutivo (D)
Verbos		
VMP00SF	Categoría	Verbo (V)

VMP00SF	Tipo	Principal (M), Auxiliar (A), Semiauxiliar (S)
VMP00SF	Modo	Indicativo (I), Subjuntivo (S), Imperativo (M), Infinitivo (N), Gerundio (G), Participio (P)
VMP00SF	Tiempo	Presente (P), Imperfecto (I), Futuro (F), Pasado (S), Condicional (C), - (0)
VMP00SF	Persona	Primera (1), Segunda (2), Tercera (3)
VMP00SF	Número	Singular (S), Plural (P)
VMP00SF	Género	Masculino (M), Femenino (F)
Pronombres		
PP1CSN00	Categoría	Pronombre (P)
PP1CSN00	Tipo	Personal (P), Demostrativo (D), Posesivo (X), Indefinido (I), Interrogativo (T), Relativo (R), Exclamativo (E)
PP1CSN00	Persona	Primera (1), Segunda (2), Tercera (3)
PP1CSN00	Género	Masculino (M), Femenino (F), Común (C), Neutro (N)
PP1CSN00	Número	Singular (S), Plural (P), Impersonal-Invariable (N)
PP1CSN00	Caso	Nominativo (N), Acusativo (A), Dativo (D), Oblicuo (O)
PP1CSN00	Poseedor	Singular (S), Plural (P)
PP1CSN00	Cortesía	Cortés (P)
Conjunciones		
CC	Categoría	Conjunción (C)
CC	Tipo	Coordinada (C), Subordinada (S)
Interjecciones		
I	Categoría	Interjección (I)
Preposiciones		
SPCMS	Categoría	Adposición (S)

SPCMS	Tipo	Preposición (P)
SPCMS	Forma	Simple (S), Contraída (C)
SPCMS	Género	Masculino (M)
SPCMS	Número	Singular (S)

Tabla 3.3.1.2: Etiquetas gramaticales para el Español (EAGLES)

- Reducción de palabras

Sea cual sea el lenguaje analizado, es muy común el uso de palabras derivadas o flexionadas, algo que debemos tener en cuenta para evitar duplicar palabras iguales, por ejemplo: *terraza* y *terrazas*. Esto suele presentar problemas sobre todo en textos de gran longitud y aunque los que trataremos no lo son, dado que nuestro objetivo es sacar palabras clave también tendremos que prestarle atención. Existen dos técnicas fundamentales:

En primer lugar la **lematización**, que consiste en hallar el lema correspondiente de una palabra, o lo que es lo mismo la entrada estándar de dicha palabra en un diccionario. De esta manera podremos sintetizar las palabras similares, ya sea por medio de una lematización puramente morfológica o, en caso de ubicarla según su contexto, a través de un análisis sintáctico.

En segundo lugar el **stemming**, que se basa en reducir una palabra a su raíz. Se utiliza sobre todo en implementaciones de búsqueda en bases de datos y en buscadores de navegador, para aumentar los resultados potenciales.

- Análisis de frecuencia

Consiste en determinar la importancia de una palabra a partir de su frecuencia de aparición en los textos, pudiendo realizarse bien en base a un término dentro de una colección de documentos, o bien según el número de documentos que contienen dicho término de entre toda la colección.

En nuestro caso, hemos desarrollado dicho análisis de manera manual, dado que las descripciones no constituyen ninguna colección indexada previamente, ni tampoco contamos a priori con un vocabulario inherente a las mismas.

- Presencia de anáforas

Es común en cualquier idioma encontrar el uso de pronombres o determinantes que sustituyan una entidad citada previamente, algo a tener en cuenta a la hora de analizar el objeto del cual se está hablando y la manera en que se hace. Por ejemplo:

"El chico trató de impresionarla, pero ella no le puso atención."

Así mismo, es frecuente hallar palabras omitidas dentro de una misma frase, que aunque en el habla y escritura natural no son necesarias, sí hacen una diferencia en cuanto a procesamiento computacional se refiere, ya que nos pueden indicar información adicional sobre el contenido.

- Análisis sintáctico (parsing)

Se lleva a cabo un análisis de cada oración presente en el texto, obteniendo como resultado un árbol describiendo la estructura sintáctica presente en cada una de ellas.

Por ejemplo, para la oración:

"One of our top picks in Adeje."

Obtendríamos el siguiente análisis sintáctico, donde cada palabra está además etiquetada con su función en la oración:

+(One Z)
+(of IN)
+(our PRP$)
+(top JJ)
+(picks NNS)
+(in IN)
+(Adeje NP)
+(. Fp)

3.3.2 Módulos de procesamiento

Una vez establecidas las bases teóricas y los conceptos del procesamiento del lenguaje, pasamos en el siguiente nivel al análisis de las diferentes herramientas que podríamos potencialmente utilizar para alcanzar los objetivos.

Existe una gran variedad de módulos de procesamiento de texto, en inglés mucho más amplia que en español, pero todos ellos diferentes respecto de ciertas características. A la hora de elegir el que será el definitivo, debemos tener en cuenta todos los detalles, la potencia y la profundidad que cada uno nos puede aportar, por lo que ha sido necesario un análisis completo de un conjunto amplio de ellos.

Además de ésto, aunque desde un principio se tenían ideas claras sobre diferentes formas de enfocar el problema, no se tenía el conocimiento para saber qué características iban a ser realmente necesarias dentro de las ofrecidas por cada módulo, con lo que además de su aprendizaje se llevó a cabo una clasificación de cada uno, según lo que pudiéramos obtener de ellos.

En general, los requisitos que en conjunto son capaces de facilitarnos todas estas herramientas son:

- Análisis morfológico
- Etiquetado gramatical
- Segmentación de frases
- Tokenización y lematización
- Analizador sintáctico
- Analizador de dependencias
- Correferencias
- Grafo semántico

Además, se ha tenido en cuenta dos variables adicionales ajenas al analizador en sí mismo, como son la potencia de la base de datos interna y el soporte para los idiomas. Éste último es un paso lógico, dado que buscamos actuar sobre el español y el inglés, y habrá herramientas más apropiadas para cada uno de forma independiente.

En cuanto a la potencia de la base de datos, dentro del vocabulario específico relacionado con la hostelería que esperamos encontrar, obtendremos mejores o peores resultados dependiendo de la riqueza de los corpus de entrenamiento que se hayan utilizado en cada uno de los diferentes módulos, pudiendo resultar en palabras sin evaluación posible.

3.3.3 NLTK

La librería NLTK, o Natural Language Toolkit, es una plataforma para el procesamiento del lenguaje natural para el lenguaje de programación Python. Proporciona interfaces fáciles de usar, con más de 50 corpus y recursos léxicos como *WordNet*, una de las mayores bases de datos léxicas para el inglés, así como un conjunto de bibliotecas de procesamiento de texto multilingüe.

Se trata de un módulo muy potente y completo. Cuenta con una serie de corpus y colecciones de textos de una gran variedad de idiomas, precargados e indexados, que facilitan el uso y agilizan en gran medida las pruebas. Así mismo, es una librería muy extensa, con múltiples métodos de cálculo de frecuencias y probabilidades, búsquedas de términos y de textos, analizadores de frases, extracción y segmentación de información, etc. En cuanto a las características de procesamiento, encontramos:

- Clasificación de textos
- Tokenización del texto en palabras
- Segmentacion del texto en frases
- Análisis morfológico
- Etiquetado gramatical
- Tokenización y lematización
- Análisis semántico

Las pruebas llevadas a cabo con el conjunto de corpus que se proveen por defecto con la librería son excelentes. Sin embargo, las herramientas de parseo y análisis aplicadas a textos en español ajenos de los ya incorporados no son tan buenas, permitiendo sólo el tratamiento correcto en inglés. No obstante, suministra potentes herramientas y librerías que permiten el aprendizaje, toda vez que se le suministre un tagset apropiado y un corpus o un conjunto de ellos para su tratamiento.

3.3.4 Spaguetti Tagger

Se trata de una librería implementada a través de los recursos disponibles en la librería NLTK de Python para el tratamiento de textos en español.

El potencial que muestra podría escalar hasta el nivel que presenta el nativo propio de NLTK con el inglés, dado que el modelo de entrenamiento consta de las mismas herramientas. Sin embargo, el corpus de texto sobre el que se puede entrenar a priori y que también lo incorpora NLTK es cess_esp, el cual no es muy extenso y la gran mayoría de las palabras analizadas no consiguen identificarse.

Aunque sería viable una solicitud de corpus de texto en español para un entrenamiento más intensivo, y que con toda seguridad su potencial y sus resultados serían excelentes, se ha optado por seguir investigando otros modelos de procesamiento de texto para poder estudiar las diferentes opciones y elegir la mejor.

3.3.5 TreeTagger Wrapper

El módulo TreeTagger es una herramienta de procesamiento de textos que funciona sobre un total de 20 lenguajes, y que además es adaptable a nuevos si se provee un lexicón y un corpus de texto adecuados.

En cuanto a potencial, esta librería presenta unas características bastante más limitadas, siendo tan sólo disponibles la lematización y el etiquetado gramatical del texto, así como un stemming de palabras si se trabaja con inglés, alemán, francés o español.

Aunque con la información disponible a primera vista no parece ser una buena opción, se ha decidido tener conocimiento sobre él como parte de la investigación, además de contar con la gran variedad de idiomas para los que está implementada y que podría ser de utilidad en algún caso.

3.3.6 Stanford CoreNLP

Se trata de una herramienta de procesamiento del lenguaje natural desarrollada por la Universidad de Stanford. Está implementada en Java, aunque existen intervaces para otros lenguajes como Perl, Python o Ruby.

Este módulo presenta un amplio conjunto de características:

- Tokenización del texto en palabras
- Segmentación del texto en frases
- Etiquetado gramatical
- Lematización
- Reconocimiento de nombres de entidades (NER)
- Análisis sintáctico
- Grafo semántico

Tanto el potencial inherente a todas estas características como los resultados generados en las pruebas hacen que esta herramienta sea una opción bastante llamativa, casi tanto o más que la propia librería NLTK.

Por otra parte, cabe citar que el aspecto negativo es que no presenta soporte para otros idiomas, actuando únicamente sobre el Inglés. Aún así, dado que ambas implementaciones serán independientes no se busca que una misma herramienta cumpla con ambos lenguajes.

3.3.7 Ixa Pipes

Se trata de un conjunto de módulos para el procesamiento de lenguaje natural creada por el grupo IXA NLP de la Universidad del País Vasco. Proporciona una interfaz de fácil acceso y una notación eficiente, así como soporte para una cantidad limitada de lenguajes.

Las características principales que ofrece son:

- Tokenización del texto en palabras
- Segmentación del texto en frases
- Análisis morfológico
- Reconocimiento de nombres de entidades (NER)
- Chunker de palabras (sólo Vasco)
- Análisis sintáctico (sólo Español e Inglés)

Presenta una variedad de funcionalidades a tener en cuenta, y los resultados son bastante buenos, aunque no perfectos. Además tiene implementación para Español e Inglés salvo en el chunker, aunque a priori no parece ser necesario para nuestro propósito.

3.3.8 Freeling

Se trata de una librería implementada en C++ que provee un serie de herramientas para el análisis del lenguaje y el procesamiento de textos. Ha sido desarrollada por la Universitat Politècnica de Canalunya (UPC) bajo software libre, y permite trabajar además sobre una variedad considerable de idiomas.

Aunque no hay disponible una interfaz para Python, es posible compilar dicha librería por cuenta propia e implementar una interfaz que permita su utilización en este lenguaje, aunque con un esfuerzo considerable. Las funcionalidades disponibles en Freeling son:

- Tokenización del texto en palabras
- Segmentación
- Análisis morfológico
- Etiquetado gramatical
- Lematización
- Detección de nombres de entidades (NER)
- Análisis sintáctico
- Analizador de dependencias y correferencias
- Grafo semántico

Como podemos observar, presenta un abanico de características bastante amplio, y los resultados obtenidos son muy prometedores, aunque el análisis de dependencias, correferencias y grafo semántico no están disponibles para todos los idiomas.

Por otro lado, el idioma central sobre el que está implementada dicha librería es el español, siendo el único lenguaje que tiene implementadas y a pleno uso todas las funcionalidades mencionadas. Además, en la última versión (4.0) se ha añadido la implementación de los tres aspectos antes mencionados también para el inglés, lo cual completa la funcionalidad que buscamos.

3.3.9 Observaciones finales

Una vez estudiadas todas estas herramientas, las funcionalidades que serán necesarias para llevar a cabo nuestro análisis y atendiendo a los resultados obtenidos, se ha decidido utilizar el módulo Freeling para el procesamiento en ambos idiomas.

Para el Español, dada la menor cantidad de herramientas existentes y atendiendo a las funcionalidades que nos ofrecen las estudiadas, se ha determinado que no existe otra librería mejor tanto en potencial como en resultados y opciones de procesamiento, con lo que no ha habido dudas.

En cuanto al Inglés, los mejores resultados son los obtenidos con las librerías NLTK, Stanford Core NLP y Freeling. Las opciones disponibles son similares y la potencia de procesamiento de texto también ofrece resultados paralelos. Sin embargo, dado que la última versión de Freeling aporta análisis de dependencias y correferencias y que será utilizada ya para el Español, se ha decidido que sea de nuevo ésta la librería elegida debido a dichas funcionalidades y a la claridad en base al código que pueda aportar al desarrollar bajo una misma herramienta de trabajo.

3.4 Bases de datos

Una vez obtengamos los datos, debemos contar con algún tipo de persistencia para poder almacenar la información y acceder a la misma de forma rápida y efectiva, con lo que el uso de una base de datos se postula como la mejor opción.

3.4.1 PostgreSQL

Se trata de un sistema de gestión de bases de datos objeto-relacional, distribuido bajo licencia BSD y que constituye el más potente del mercado en código abierto, contando con un potencial a la altura de otras bases de datos comerciales.

PostgreSQL utiliza un modelo cliente/servidor y multiprocesos en lugar de multihilos para garantizar la estabilidad del sistema. Es completamente ACID, funciona muy bien con grandes cantidades de datos y una alta concurrencia de usuarios simultáneos, y cuenta con numerosas APIs para la programación en un gran número de lenguajes, entre los que se incluye Python.

Tras investigar acerca de las diferentes opciones, se ha decidido utilizar este sistema de gestión para nuestros datos, en parte por su gran potencia, y en otra parte por la posibilidad de instalar una extensión que permita tratar datos georreferenciados.

3.4.2 PostGIS

Se trata de una extensión que añade soporte de objetos geográficos a la base de datos PostgreSQL, proporcionando la funcionalidad de base de datos espacial y permitiendo su utilización en un sistema de información geográfica.

Ha sido desarrollado por la empresa canadiense Refraction Research, especializada en productos de código abierto y responsable entre otros del desarrollo de Udig, un framework de sistemas de información geográfica para eclipse.

Capítulo 4
Diseño e implementación

Una vez concluido el estudio previo de lenguaje y herramientas, se procede al desarrollo de un sistema que cumpla con los objetivos propuestos. Ha sido un proceso incremental, puesto que o bien han fallado algunas ideas o bien otras que sí han estado correctas han necesitado de un aumento constante de la funcionalidad. Se describe en este punto todo el trabajo realizado para alcanzar los objetivos.

4.1 Obtención de datos

El primer paso para poder tratar la información se basa en obtenerla. Dada una fuente que de forma general será una página de búsquedas de hoteles, se pretende obtener toda aquella información que pueda ser relevante para nuestros propósitos; nombre del hotel, descripción, valoraciones y comentarios de los clientes, así como características que puedan ser utilizadas para implementar posibles funcionalidades adicionales, como dirección y/o coordenadas, y los idiomas que se hablan en cada hotel.

Destacar que en este proyecto, se ha trabajado sobre los resultados de la página http://www.booking.com. Como la parte central del proyecto era el tratamiento de la información no se dio prioridad a la búsqueda en más webs, aunque el número de buscadores sobre los que se actuar puede escalar de forma significativa e indefinida.

Para recuperar las descripciones de las páginas se ha utilizado UrlLib y BeautifulSoup, descritas anteriormente. Por medio de estas herramientas y a través de un estudio previo manual del código de tres modelos de páginas: resultados de búsqueda de hoteles,

página del hotel, y página de sus comentarios; hemos podido desarrollar una serie de filtros que nos permiten obtener toda esta información. Por ejemplo, dados los datos del web scraping de cada hotel:

```
<!--Nombre-->
<span class="fn" id="hp_hotel_name">
    Europe Villa Cortes GL
</span>

<!--Descripción-->
<div class="hp_hotel_description_hightlights_wrapper ">
    <div class="hotel_description_wrapper_exp hp-description" >
        <div id="summary" class="">
            <div class="chain-content ">
            </div>
            <p><span class="hp-description--property-name-top-ranked"> <i class="bicon-acstar"></i>Está en
            nuestra selección para Adeje. </span> Este hotel ...lujoso y exclusivo para adultos está situado
            junto a la playa del Duque, en Costa Adeje. Ofrece suites junior elegantes con terraza o balcón
            amueblados, algunos con vistas al Atlántico. Hay conexión WiFi de alta velocidad gratuita en
            todas las instalaciones.</p>

            <p>Las habitaciones del Iberostar Grand Hotel El Mirador Adults Only son amplias y presentan una
            decoración de estilo colonial. Disponen de una cama con dosel, un vestidor independiente y un
            sofá cama.</p>

            <p>El edificio principal está rodeado de una piscina de agua dulce climatizada. También hay una
            bañera de hidromasaje. El restaurante de buffet principal del Hotel El Mirador propone cocina
            típica canaria y platos internacionales. El hotel también alberga un restaurante junto a la
            piscina, el restaurante gourmet El Cenador, además del bar La Tosca, que ofrece noches con música
            de piano en directo.
            <br /></p>
        </div>
    </div>
</div>

<!--Dirección-->
<span class="hp_address_subtitle jq_tooltip" rel="14" itemprop="address" data-source="top_link" data-coords=",
" data-node_tt_id="location_score_tooltip" data-bbox="-16.7503389716148,28.0053905069792,-16.7189946770668,28.
1071384713984" data-width="350" title="">
    Avenida Rafael Puig, s/n, 38660 Playa de las Américas, España
</span>
```

Vemos la facilidad de su extracción con las librerías estudiadas:

```
currentUrl = urlList.pop()
htmlSource = urllib.request.urlopen(currentUrl).read()
soup = BeautifulSoup(htmlSource, 'html.parser')
hotelName = soup.find('span', {'id' : 'hp_hotel_name'}).getText().strip()
hotelDescription = soup.find('div', {'id' : 'summary'}).getText().strip()
hotelAddress = soup.find('span', {'itemprop' : 'address'}).getText().strip()
```

No obstante cabe destacar que, por encima de la búsqueda propia de clases, identificadores o valores, se ha filtrado la información utilizando patrones de expresiones regulares sobre las url siempre que ha sido posible, pensando en que éstos últimos tienen menos probabilidad de ser modificados.

4.2 Almacenamiento de datos

Para su tratamiento, se hace necesario tener un registro de toda la información que hemos obtenido, para lo que se ha creado una clase Hotel que permita y facilite el análisis de la misma. Así mismo, se ha decidido alojar los datos en una base de datos PostgreSQL, de manera que no sobrecarguemos el sistema y tengamos facilidad de acceso y tratamiento de los mismos.

4.2.1 Clase Hotel

La clase Hotel contiene, además de los datos principales extraídos de la web, una serie de métodos internos para el procesamiento de la información, de manera que sea capaz de almacenar y sintetizar las palabras clave encontradas en la descripción y tener constancia de la singularidad de las mismas, y realizar el procesado de los comentarios según dicho vocabulario propio.

4.2.2 PostgreSQL - PostGIS

El almacenamiento de los datos obtenidos se llevará a cabo en una base de datos PostgreSQL con la extensión de PostGIS, de manera que podamos georreferenciar los hoteles insertados.

Para ello, se ha hecho uso del API Geocoder de Google, obteniendo las coordenadas correspondientes a través de las direcciones físicas extraídas. Se es consciente de que existe un límite diario de 2.500 solicitudes al Geocoder, pero es adecuado dado que la cantidad de datos que trataremos en este proyecto dista mucho de alcanzar esa cifra.

4.3 Análisis de la descripción

A medida que se ha ido avanzando en el problema, se han descartado medidas y se han pensado nuevas alternativas en un proceso que siempre ha sido incremental y en ocasiones iterativo, intentando completar de la mejor forma posible las tareas objetivo.

4.3.1 Extracción por repetición

Para poder extraer un vocabulario concreto de las descripciones debemos realizar un análisis completo de las mismas. Dado que no tenemos ningún tipo de referencia en cuanto a palabras clave, tenemos que analizar la información y tratar de obtenerlas en base a características puramente objetivas y estadísticas. Para ello, realizamos un etiquetado gramatical de las descripciones a tratar por medio de la librería Freeling, tanto en Español como en Inglés, y pasar posteriormente a su síntesis.

En primer lugar, por lo general en el vocabulario utilizado en la hostelería como en cualquier otro sector se utiliza una terminología recurrente, por lo que se ha llevado a cabo un análisis de todas las descripciones buscando los sustantivos más repetidos (por ejemplo, *piscina, habitación, limpieza, comida, playa, internet,* ...).

Sin embargo, los primeros resultados nos dejaron con luces y sombras. Bajo un análisis de 30 hoteles, las palabras relativas a un vocabulario hotelero sí coincidían en su mayoría con las más repetidas, pero entre ellas también entraban algunas que nada tenían que ver:

```
NOUN [piscina] => 22 times
NOUN [bar] => 16 times
NOUN [restaurante] => 12 times
NOUN [zona] => 12 times
NOUN [pista] => 12 times
NOUN [tv] => 10 times
NOUN [terraza] => 8 times
NOUN [balcón] => 7 times
NOUN [playa] => 7 times
NOUN [pie] => 7 times
NOUN [gimnasio] => 6 times
NOUN [bañera] => 6 times
NOUN [cocina] => 6 times
NOUN [spa] => 6 times
NOUN [baño] => 5 times
NOUN [servicio] => 5 times
NOUN [sala] => 5 times
NOUN [aire] => 5 times
NOUN [jardín] => 5 times
NOUN [salón] => 5 times
```

Como podemos observar en el ejemplo, palabras como *zona*, *pie* o *aire* tienen recurrencia moderada en las descripciones, y no tienen relación alguna con el vocabulario de hostelería, o al menos no directamente.

Así mismo, se ha llevado a cabo un intento de búsqueda también por repetición, pero teniendo en cuenta los verbos presentes más comunes para intentar extraer los más relevantes. Por ejemplo, la utilización de los verbos *ofrecer*, *contar* o *disponer* suele ser bastante alta en este tipo de vocabulario dado que dan información de lo que se proporcionará al cliente:

```
VERB [estar] => 62 times
VERB [ofrecer] => 33 times
VERB [encontrar] => 32 times
VERB [contar] => 30 times
VERB [disponer] => 29 times
VERB [haber] => 27 times
VERB [incluir] => 18 times
VERB [tener] => 15 times
VERB [poder] => 12 times
VERB [albergar] => 11 times
VERB [servir] => 10 times
VERB [presentar] => 9 times
VERB [disfrutar] => 6 times
VERB [proporcionar] => 4 times
VERB [descubrir] => 4 times
VERB [hallar] => 3 times
VERB [preparar] => 3 times
VERB [organizar] => 3 times
VERB [gozar] => 2 times
VERB [abrir] => 2 times
```

Dado que dentro de los errores en los sustantivos también hubo aciertos se llevó a cabo además un segundo intento de búsqueda por repetición, tratando de mejorar los resultados al tener constancia de los verbos que acompañan a los sustantivos, intentando obtener una especie de patrón que a simple vista pueda ser concurrente, obteniendo como resultado:

```
COMPOUND [minibar] => {'poder': 1, 'disponer': 1}
COMPOUND [aparcamiento] => {'contar': 1, 'ofrecer': 1, 'disponer': 1}
COMPOUND [cúpula] => {'encontrar': 1}
COMPOUND [paisaje] => {'incluir': 1}
```

Como podemos observar, esto nos proporcionó una idea un poco más clara que la anterior y algo más cercana al vocabulario hotelero que buscamos, pero aún lejana del nivel que nos gustaría.

Además, cabe destacar que una de las metas propuestas ha sido el identificar no sólo aquellas características afines, sino también las que puedan conformar la exclusividad del hotel. Nos percatamos de que realizando un filtro por repetición estos servicios podrían pasar desapercibidos. Por ejemplo, si atendemos a la oración:

"Presenta una decoración de estilo étnico."

Vemos que ni *presentar* es un verbo tan concurrente ni *decoración* se encuentra de forma habitual, y sin embargo una *"decoración de estilo étnico"* puede conformar perfectamente un rasgo único de un grupo de hoteles limitado.

Llegados a este punto y con todo lo estudiado, se ha decidido tomar otras vías para encontrar las palabras clave, quedándonos con lo positivo que hemos podido sacar para futuras implementaciones, pero descartando la búsqueda por medio de repeticiones, dado que será imposible casi con toda seguridad, al menos, tener resultados que puedan considerarse buenos.

4.3.2 Análisis sintáctico

En segundo lugar, se procedió a buscar un análisis más exhaustivo de cada frase dentro de la descripción. Por medio de un análisis sintáctico, además del etiquetado gramatical y lematización, es posible determinar qué función desempeña cada palabra dentro de una oración.

Para ejemplificar de forma correcta el desarrollo de ahora en adelante se trabajará sobre la siguiente descripción ejemplo, utilizando porciones de la misma según sea necesario para clarificar los problemas encontrados:

"Este complejo está situado en Guía de Isora, en Tenerife, y disfruta de vistas al océano Atlántico y a la vecina isla de La Gomera. Cuenta con 7 piscinas al aire libre, spa, campo de golf, pistas de tenis y playa.

El The Ritz-Carlton, Abama está compuesto por edificios de inspiración morisca con habitaciones y villas. El complejo se encuentra en un acantilado, rodeado de exuberantes jardines tropicales, con un funicular que llega hasta la playa exclusiva.

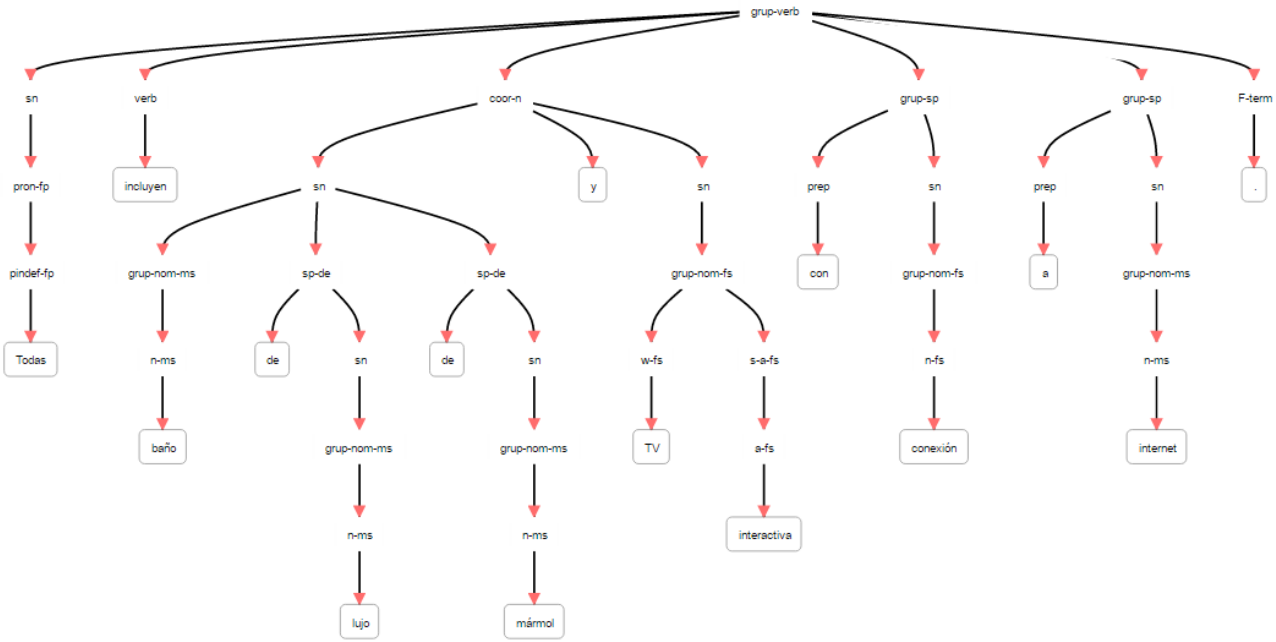
Las habitaciones y villas presentan una decoración de estilo étnico y la mayoría tienen vistas a los jardines y al océano Atlántico. Todas incluyen baño de lujo de mármol y TV interactiva con conexión a internet.

El The Ritz-Carlton, Abama alberga 10 restaurantes. El MB, con 2 estrellas Michelin, sirve platos creados por el chef Martín Berasategui. El Kabuki, con 1 estrella Michelin, prepara cocina japonesa innovadora.

Hay un gran spa con 10 salas de tratamientos, circuito de agua y gimnasio. El hotel también cuenta con un Ritz Kids para los más pequeños."

Cuando hablamos de descripciones de hoteles, es una práctica habitual encontrar calificativos que encumbren cada uno de los servicios estrella ofertados, más aún si tenemos en cuenta que el objetivo final es atraer a un cliente. Por medio de un análisis sintáctico podemos ampliar en una medida considerable el abanico de posibilidades del objeto de estudio, dado que podemos determinar qué términos actúan como sujeto de una frase y cuáles lo complementan, desde la perspectiva a priori de que serán en su mayoría datos de interés para nosotros.

Así pues, el análisis sintáctico obtenido se conforma:



Trabajando sobre esta idea, podemos observar la división de los diferentes sintagmas nominales, preposicionales y verbales, aunque no se trata de forma específica las relaciones entre ellos.

Comprobamos como la información obtenida aunque es menor cantidad, es más rica en cuanto a tratamiento, aunque conteniendo un considerable número de errores y datos no determinantes para nosotros, encontrando diversos sustantivos que aunque bien calificados no son relativos a la hostelería.

Así mismo, la división por bloques establecida en el análisis sintáctico no nos proporciona los datos suficientes sobre las referencias propias de cada elemento, y por tanto se deriva el trabajo hacia un análisis más en profundidad que nos pueda aportar más información.

4.3.3 Análisis de depencencias y correferencias

En tercer lugar, se ha llevado a cabo un análisis de las dependencias de cada entidad presente en la oración, y las referencias entre las mismas.

Este proceso nos da un resultado mucho más detallado, de manera que podamos conocer a dónde deriva cada sección de la frase, qué términos dentro de un bloque se encuentran al mismo nivel (sea por conjunciones, signos de puntuación, o separadores de algún tipo), a qué referencia cada uno de los adjetivos y complementos preposicionales, además del etiquetado gramatical y la lematización llevadas a cabo.

Por tanto, podemos hablar de que se establece una jerarquía entre los términos presentes en la oración, a través de la cual nosotros podemos extraer la información de manera más concreta.

Para facilitar la explicación, se presentan a continuación un ejemplo de correferencia sobre una porción de la descripción modal:

"Cuenta con 7 piscinas al aire libre, spa, campo de golf, pista de tenis y playa. Hay un gran spa con salas de tratamientos, circuito de agua y gimnasio."

```
grup-verb/top/(Cuenta contar VMIP3S0 -) [
  grup-sp/sp-obj/(con con SPS00 -) [
    sn/obj-prep/(piscinas piscina NCFP000 -) [
      numero-nopart/sn-mod/(7 7 Z -)
    ]
  ]
  grup-sp/sp-obj/(a a SPS00 -) [
    sn/obj-prep/(aire aire NCMS000 -) [
      espec-ms/espec/(el el DA0MS0 -)
      s-a-ms/adj-mod/(libre libre AQ0CS0 -)
    ]
  ]
  coor-n/dobj/(y y CC -) [
    Fc/term/(, , Fc -)
    sn/co-n/(spa spa NCFS000 -)
    Fc/term/(, , Fc -)
    sn/co-n/(campo campo NCMS000 -) [
      sp-de/sp-mod/(de de SPS00 -) [
        sn/obj-prep/(golf golf NCMN000 -)
      ]
    ]
    Fc/term/(, , Fc -)
    sn/co-n/(pistas pista NCFP000 -) [
      sp-de/sp-mod/(de de SPS00 -) [
        sn/obj-prep/(tenis tenis NCMS000 -)
      ]
    ]
    sn/co-n/(playa playa NCFS000 -)
  ]
  F-term/term/(. . Fp -)
]
```

```
grup-verb/top/(Hay haber VMIP3S0 -) [
  sn/dobj/(gran gran AQ0CS0 -) [
    indef-ms/espec/(un uno DI0MS0 -)
  ]
  sn/modnomatch/(spa spa NCFS000 -)
  grup-sp/cc/(con con SPS00 -) [
    sn/obj-prep/(salas sala NCFP000 -) [
      numero-nopart/sn-mod/(10 10 Z -)
      sp-de/sp-mod/(de de SPS00 -) [
        sn/obj-prep/(tratamientos tratamiento NCMP000 -)
      ]
    ]
  ]
  coor-n/modnomatch/(y y CC -) [
    Fc/term/(, , Fc -)
    sn/co-n/(circuito circuito NCMS000 -) [
      sp-de/sp-mod/(de de SPS00 -) [
        sn/obj-prep/(agua agua NCCS000 -)
      ]
    ]
    sn/co-n/(gimnasio gimnasio NCMS000 -)
  ]
  F-term/term/(. . Fp -)
]
```

La primera aproximación desarrollada se basó en un algoritmo que fuera capaz de identificar de forma precisa todos aquellos sustantivos y su correspondiente adjetivo, partiendo de la base de que el análisis de correferencias nos conectará de forma efectiva ambos valores.

Como podemos observar en la imagen anterior, el resultado de la correferencia se expone en forma de árbol, donde cada nuevo nivel de profundidad es separado por [corchetes]. La ejecución se orientará por tanto según puntos de control que serán configurados en base a la profundidad actual del árbol y que nos indicará los límites de cada relación y manteniendo los valores encontrados hasta que acabe su referencia, con el consiguiente resultado:

```
['isla', 'vecina']
['aire', 'libre']
['inspiración', 'morisca']
['jardines', 'exuberantes']
['jardines', 'tropicales']
['playa', 'exclusiva']
['estilo', 'étnico']
['cocina', 'japonesa']
['cocina', 'innovadora']
```

Posteriormente se amplió la funcionalidad de esta primera aproximación para que actuara también sobre complementos preposicionales. Se tuvo especial cuidado en la implementación, dado que un complemento preposicional puede abarcar cualquier término en su interior, sea sustantivo adjetivo o incluso otro complemento en sí mismo, de manera que hay que distinguir bien qué hace referencia al conjunto preposicional y qué al punto de control externo. Se decidió enfocar como una especie de “modo preposicional”, almacenando todos los valores útiles dentro de un mismo enclave, y asociarlo a lo correspondiente una vez terminado dicho bloque de contenidos. Los resultados obtenidos fueron como sigue:

```
['isla', 'vecina']
['aire', 'libre']
['jardines', 'exuberantes']
['jardines', 'tropicales']
['playa', 'exclusiva']
['cocina', 'japonesa']
['cocina', 'innovadora']
['isla', 'de La_Gomera']
['campo', 'de golf']
['pistas', 'de tenis']
['edificios', 'de inspiración morisca']
['decoración', 'de estilo étnico']
['baño', 'de lujo']
['baño', 'de mármol']
['salas', 'de tratamientos']
['circuito', 'de agua']
```

Podemos observar que los resultados obtenidos analizando ambos complementos, aunque están lejos de la perfección, toman mejor forma que lo tratado hasta ahora.

Como punto siguiente, se tratará de ampliar la funcionalidad para que abarque las conjunciones de elementos, sea por signos de puntuación o por conjunciones copulativas “y”, y poder así agrupar diversos elementos que aún nos quedan pendientes:

“Cuenta con 7 piscinas al aire libre, spa, campo de golf, pistas de tenis y playa.”

Comprobamos que de los elementos citados, obtenemos correctamente aquellos que están complementados, como *campo de golf* y *pistas de tenis*, no tomando el resto (el caso concreto de *piscinas al aire libre*, el cual tampoco se recoge, se estudiará más adelante).

Se desarrolló por tanto una segunda implementación algorítmica, tratando de abarcar todo lo anterior y buscando la mejor forma posible de extraer las palabras clave. Esto conllevó algunos problemas, dado que debíamos mantener localizado el sustantivo base y aislarlo de los posibles sustantivos contenidos dentro de la propia conjunción, de manera que no mezcláramos complementos equívocos ni información falsa a nuestro filtro de datos. Como observamos en la siguiente oración:

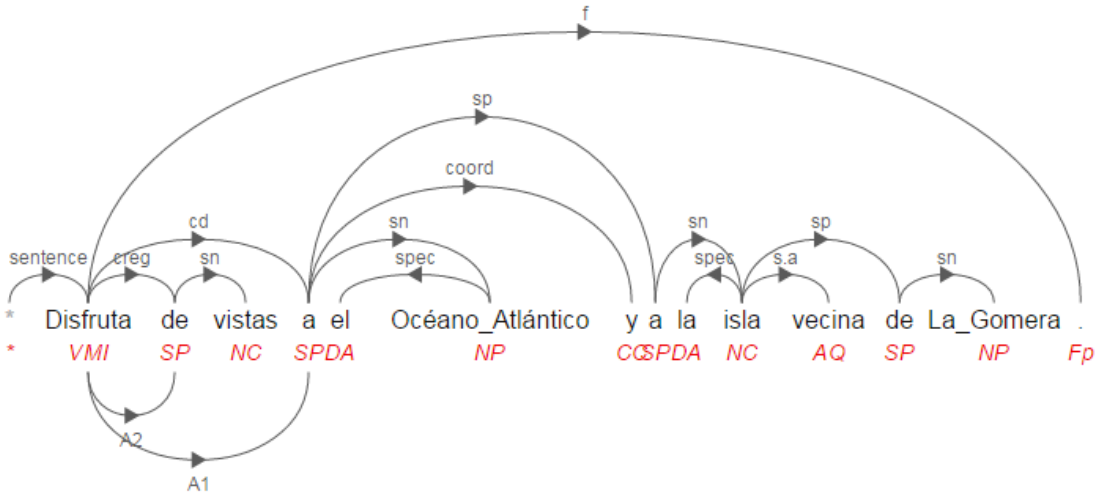
“Este complejo está situado en Guía de Isora, en Tenerife, y disfruta de vistas al océano Atlántico y a la vecina isla de La Gomera.”

Teniendo en cuenta los resultados del algoritmo hasta ahora, comprobamos que sólo recoge el atributo *"isla de La Gomera"*, donde trata *isla* como el nombre y *de La Gomera* como su complemento, mientras que la solución correcta pasa por un sólo complemento *a la vecina isla de La Gomera* que referencia a *vistas*.

La forma de lidiar con ello se basó en un nuevo punto de control, que en su base actúa igual que en el resto del análisis pero con la salvedad de estar en un "ámbito independiente", tratando sus sustantivos y complementos de manera aislada al resto del documento, cuyos resultados fueron los siguientes:

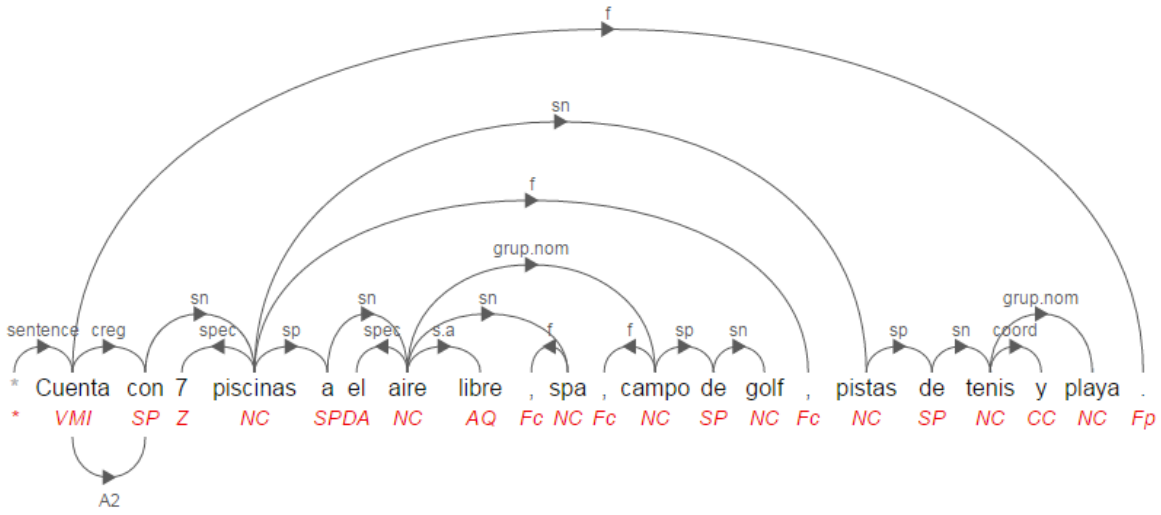
```
['aire', 'libre']
['spa', '']
['jardines', 'exuberantes']
['jardines', 'tropicales']
['playa', 'exclusiva']
['baño', 'interactiva']
['TV', 'interactiva']
['cocina', 'japonesa']
['cocina', 'innovadora']
['gimnasio', '']
['complejo', 'en Guía_de_Isora']
['complejo', 'en Tenerife']
['vistas', 'a océano Atlántico']
['vistas', 'a isla vecina de La_Gomera']
['campo', 'de golf']
['pistas', 'de tenis']
['edificios', 'de inspiración morisca']
['decoración', 'de estilo étnico']
['jardines', 'a océano Atlántico']
['baño', 'de lujo']
['baño', 'de mármol']
['salas', 'de tratamientos']
['circuito', 'de agua']
['gimnasio', 'con Ritz_Kids']
['gimnasio', 'para pequeños']
```

Ningún sistema es completamente exacto en el procesamiento del texto, y en el caso concreto de Freeling el porcentaje de acierto aproximado es de un 97-98% para el etiquetado gramatical, lo que implica aproximadamente un error cada 50 palabras, o un error en una de cada cuatro frases. Transportado al análisis sintáctico o de dependencias, donde ya entra en juego además el sentido que el emisor quiera darle a la frase y el orden de las palabras, es bastante más probable encontrar relaciones mal ejecutadas:



Como podemos observar en el ejemplo anterior, las secciones: *"de vistas"* y *"a el océano Atlántico..."* se establecen en un mismo nivel (ambas preposiciones son enlazadas al verbo *disfruta*, como si se tratase de *"disfruta de vistas y disfruta a el oceáno Atlántico..."*), mientras que la referenciación correcta pasa porque el complemento *"a el océano Atlántico..."* haga referencia a *vistas*.

Veamos otro ejemplo que anteriormente citamos:



Comprobamos que las relaciones establecidas en las conjunciones *","* e *"y"* se encuentran en dos niveles diferentes, una compuesta por *"piscinas, pistas de tenis y playa"* y otra *"aire libre, spa, campo de golf"*. En el análisis del árbol resultante, por tanto, se separa el servicio *piscinas* de su complemento correspondiente *al aire libre* como si de dos características diferentes *piscina* y *aire* se tratara, dando como resultado la obtención del sustantivo *aire* complementado por *libre*.

Al igual que los datos de los que dispongo no son 100% exactos, tampoco la salida de un algoritmo basado en ellos lo podrá ser. Sin embargo, se ha intentado afinar al máximo el tratamiento de los complementos preposicionales, tratando de tener constancia de las diferentes entidades analizadas a lo largo de cada frase en cada uno de estos casos, e intentando reasignarlos de manera correcta al sustantivo al que se refiere. Este desarrollo ha sido el más extenso en cuanto a diseño y optimización y ha conllevado un esfuerzo notable, obteniéndose unos resultados muy satisfactorios:

```
NOUN: cocina
Values: japonesa
Values: innovadora
NOUN: edificio
Values: de inspiración morisca
Values: con habitaciones
NOUN: spa
Values: gran
Values: con 10 salas de tratamientos
NOUN: baño
Values: de lujo
Values: de mármol
NOUN: jardín
Values: a el océano Atlántico
NOUN: decoración
Values: de estilo étnico
NOUN: campo
Values: de golf
NOUN: piscina
Values: a el aire libre
NOUN: playa
NOUN: gimnasio
NOUN: vista
Values: a el océano Atlántico
Values: a la vecina isla
Values: de La_Gomera
NOUN: pista
Values: de tenis
NOUN: circuito
Values: de agua
NOUN: plato
Values: por el chef Martín_Berasategui
NOUN: tv
Values: con conexión
Values: a internet
NOUN: restaurante
NOUN: acantilado
Values: de exuberantes jardines tropicales
Values: con un funicular que llega hasta la playa exclusiva
```

Citar que para cada uno de los servicios se ha tomado como clave el lema de dicha palabra, para así evitar nombres duplicados y filtrar de manera óptima todos los complementos que los referencian.

Finalmente, se ha implementado un algoritmo para la interacción

de todas las descripciones y las palabras clave obtenidas, de manera que para cada hotel podamos averiguar qué servicios de entre los ofrecidos son comunes en la zona, a la vez que identificamos qué aspectos le dan exclusividad dentro del rango hotelero analizado. A modo de resultado, mostramos:

```
<<< HOTEL The Ritz-Carlton, Abama >>>
UNIQUE:
[cocina] => ['innovadora', 'japonesa']
[vista] => ['a la vecina isla', 'de La_Gomera']
[decoración] => ['de estilo étnico']
[spa] => ['con 10 salas de tratamientos']
[plato] => ['por el chef Martín_Berasategui']
[baño] => ['de mármol']
COMMON:
[piscina] => ['a el aire libre']
[tv] => ['con conexión', 'a internet']
[jardín] => ['a el océano Atlántico']
[baño] => ['de lujo']
[circuito] => ['de agua']
[campo] => ['de golf']
[spa] => ['gran']
[pista] => ['de tenis']
[vista] => ['a el océano Atlántico']
```

Así mismo, se ha generado un registro de todas las características disponibles de entre todos los hoteles y su porcentaje de ocurrencia:

```
[piscina] => 5.82%
[restaurante] => 4.18%
[bar] => 3.82%
[selección] => 3.64%
[zona] => 3.64%
[servicio] => 2.91%
[tv] => 2.73%
[playa] => 2.36%
[balcón] => 2.36%
[pista] => 2.36%
[baño] => 2.18%
[conexión] => 2.18%
[cocina] => 2.00%
[coche] => 2.00%
[terraza] => 2.00%
[jardín] => 1.82%
[gimnasio] => 1.82%
[bañera] => 1.45%
[spa] => 1.45%
[decoración] => 1.27%
```

4.3.4 Análisis de las valoraciones

El segundo gran apartado se basaba en el análisis de los comentarios dejados por los clientes de cada hotel

respectivamente, pudiendo ser éstos positivos o negativos. El primer objeto de estudio, por tanto, consta del sentimiento que puedan mostrar en la reseña expuesta, determinando dicha polaridad a través de un análisis a nivel de frase de cada comentario.

En el caso del Inglés, esto se ha llevado a cabo a partir de un lexicón de sentimientos creado por Bing Liu, profesor de la Universidad de Illinois en Chicago, y el cual consta de una cantidad aproximada de 6800 palabras en Inglés correctamente etiquetadas.

Por otro lado, para el Español se ha hecho uso del lexicón iSOL, desarrollado por el Departamento de Informática de la Escuela Politécnica Superior de Jaén a partir del lexicon anteriormente mencionado de Bing Liu, completamente revisado y testeado, y el cual cuenta con 8135 palabras.

A excepción de algún caso aislado, la aplicación de ambos lexicones a los comentarios ha resultado muy satisfactoria, evaluando correctamente casi su totalidad. Citar que desde la web tratada (http://booking.com) los comentarios de los clientes aparecen ya separados en reseña positiva y negativa, pero aún así se aplica el análisis a cada uno de ellos, acertando en un 94% de los casos.

Para llevarlo a cabo, se ha hecho un análisis sintáctico de cada uno de los comentarios, comparando todos aquellos términos que complementen un sustantivo con el índice de términos en cada lexicón respectivamente, y obteniendo como resultado la orientación positiva o negativa de dicha valoración.

Dado que ambos lexicones se encuentran ordenados alfabéticamente, también se ha creado un índice sencillo dentro del propio programa para facilitar la tarea de búsqueda, indicando en cada caso en qué línea del fichero se inicia cada letra y buscando sólo en el sector correspondiente según la primera letra del término buscado. Aunque no tiene un gran impacto para el volúmen de datos tratado, sí es algo que puede mejorar su rendimiento si éste se multiplica.

Vemos algunos ejemplos de los resultados obtenidos:

Positivos:

"La piscina estaba ideal."

"Las habitaciones muy bonitas, luminosas y elegantes."

"La ubicación mejor de lo que creíamos, en dos pasos estás en las Verónicas."

Negativos:

"El poco aparcamiento y los ruidos de la calle hasta las 6 de la mañana."

"El bar era lento (más de dos horas para comer), ineficiente, inconcebible."

"Desayuno saturado y sala ruidosa."

A continuación, una vez obtenidos los comentarios y su clasificación, se ha desarrollado un algoritmo que permita la búsqueda de cada característica dentro de dichas opiniones, de manera que podemos conocer cuáles son valoradas de forma positiva y cuáles de forma negativa.

Con la orientación obtenida anteriormente, evaluamos cada uno de los servicios dentro de cada comentario, teniendo constancia del número de veces que aparecen de forma positiva y de forma negativa, y haciendo un balance final según dicha valoración acerca del mismo. Los resultados finales obtenidos son los siguientes:

```
 === Positive Services ===
[campo] found 1 times
[jardín] found 2 times
[restaurante] found 12 times
[piscina] found 12 times
[vista] found 7 times
[playa] found 13 times
[servicio] found 14 times
[baño] found 2 times
 === Negative Services ===
[villa] found 4 times
[plato] found 1 times
[jardín] found 2 times
[gimnasio] found 1 times
[edificio] found 1 times
[spa] found 3 times
```

4.3.5 Optimización del sistema y resultados

Se ha buscado en todo momento maximizar la capacidad posible del programa, a la vez que se tome el menor tiempo posible para

ejecutar las solicitudes de procesamiento tanto de la petición html, como el análisis del código y los algoritmos de procesamiento del lenguaje sobre descripciones y comentarios.

Por una parte, se ha hecho uso del modo servidor de Freeling, lo cual ha mejorado de forma significativa el tiempo transcurrido para cada una de las peticiones de correferencia, aumentando en gran medida la velocidad del programa. Al establecerlo de esta manera, se consiguen levantar las opciones de configuración una sola vez, tener las herramientas de correferencia precargadas y la salida definida, y de esta manera aumentar la velocidad de procesamiento de las peticiones cliente.

De esta manera se ha conseguido rebajar el tiempo total en algo más de la mitad, pasando de una ejecución de casi 3 minutos a 1 minuto 26-31 segundos.

Por otro lado, también se ha hecho un estudio acerca del uso de hilos en Python. Una vez se obtienen las URL de los hoteles y su información relativa, se crea un hilo para cada uno de ellos que será el encargado de gestionar la petición de correferencias al servidor de Freeling, el análisis de la descripción parseada resultante y la posterior evaluación de los comentarios almacenados. Se tiene especial cuidado en el control de la finalización de los mismos, de manera que no se acceda en ningún momento a datos que aún no han sido completamente extraídos. Este ha sido el otro gran logro en cuanto a velocidad de procesamiento se refiere, aportando un potencial extra al sistema y minimizando aún más los tiempos de respuesta.

El hecho de mantener la funcionalidad del sistema en paralelo ha supuesto un aumento del rendimiento y la consecuente bajada de tiempos hasta 1 minuto 2-6 segundos.

Ambas referencias anteriores se han llevado a cabo en base a 10 hoteles y el rango de tiempos variará levemente en función de la cantidad de datos extraídos, dado que habrá hoteles con más o menos información en su descripción y sobre todo unas cantidad muy variable de valoraciones de los clientes.

A continuación se muestra un cálculo completo con los tiempos finales:

```
ELAPSED TIMES
URL Search: 3.3392 seconds
Hotel Analysis: 64.9152 seconds
Global Statistics Analysis: 0.0078 seconds
Hotel Features Analysis: 0.0028 seconds
```

Vemos desglosado los tiempos de acuerdo a la búsqueda y almacenamiento de las url respectivas a los hoteles, el análisis de cada hotel en profundidad, el cálculo de estadísticas globales, y la segmentación de los servicios de cada uno en únicos y comunes. Cabe destacar que para optimizar el funcionamiento, las variables estadísticas se conforman según se suceden los hoteles (en el mismo procesamiento del análisis), lo que da como resultado unos tiempos de ejecución muy cortos en ambas funcionalidades.

Capítulo 5
Conclusiones y líneas futuras

El procesamiento del lenguaje natural constituye un objeto de estudio en constante progreso, pero también en constante aumento. Las redes sociales y la investigación a nivel general hacen que incluso los idiomas más establecidos sigan creciendo, que se añadan nuevos términos y nuevas formas de citar a diferentes elementos, o incluso el origen de nuevas palabras.

Sin embargo, es un área que ha crecido mucho en los últimos años, y que actualmente aunque le queda bastante camino por recorrer, muestra un gran potencial en cuanto a procesamiento del lenguaje se refiere. Es sin duda un trabajo titánico el desarrollo de tecnologías y todos los avances realizados hasta la fecha en cuestión de tratamiento del lenguaje natural, el cual podemos ver reflejado fácilmente en los últimos desarrollos en torno al análisis sintáctico y de dependencias funcionales, clara muestra del importante avance que se está sucediendo.

En líneas futuras, dada la importancia de las redes sociales y del Big Data en el mundo actual, es fácil presumir que encontraremos cada vez más herramientas destinadas al análisis del lenguaje, así como mucha más polivalencia de idiomas que puedan ser tratados por las mismas, puesto que se hará una investigación “casi” necesaria.

A nivel personal, la investigación es un aspecto de la informática que me apasiona enormemente, así como todo lo relacionado con Inteligencia Artificial y desarrollo y optimización de algoritmos, con lo que he disfrutado realmente llevando a cabo este proyecto. Tengo que hacer especial mención al aspecto principal relacionado con la extracción del vocabulario, porque el hecho de tener presente un problema sin tomar ningún tipo de referencia para su solución y con la libertad de poder investigar, hace que el trabajo a llevar a cabo de pensamiento, ejecución y optimización sea bastante importante, y no hay nada con lo que yo disfrute más que con un reto de estas características.

Capítulo 6
Summary and Conclusions

The natural language processing is an object of study in constant progress, but also steadily rising. Social networking and generally any research make even the more established languages continuing to grow, adding new terms and new ways of quoting different elements, or even the source of new words.

However, it is an area that has grown considerably in recent years, and although there's still a long way ahead, a great potential in terms of language processing is achieved. It is certainly a herculean task developing technologies and all the progress made so far within natural language processing, which can easily be reflected in recent developments around the parsing and functional dependencies and coreferences, clearly showing the important current progress.

In future lines, given the importance of social networks and Big Data nowadays, it is easy to assume that we will find a lot of new for language analysis, and so even more languages that can be treated by them, since an investigation will be almost “necessary”.

About myself, research is an aspect of computing that I love immensely, and everything related to Artificial Intelligence, and algorithms development and optimization, so I really enjoyed carrying out this project. Special mention to the main tasks related to the extraction of vocabulary from the hotel descriptions, because the fact of solving this problem without any reference to its solution and the free self-investigating research, makes the task and relative optimization a very important and hard work, and there is nothing I enjoy more than a challenge of this nature.

Capítulo 7
Presupuesto

7.1 Tiempos de ejecución del proyecto

Tipos	Descripción
Estudio previo	30 horas
Tiempo de análisis	55 horas
Tiempo de desarrollo	42 horas

Tabla 7.1: Resumen de tiempos

7.2 Presupuesto

Características	Coste Aproximado
Hora de trabajo	4,00 €
Pc (Intel Core2-Duo E6600, 4Gb RAM DDR2, 500Gb SATA HDD)	100,00 €
Licencia de herramientas	0,00 €

Tabla 7.2: Resumen de presupuesto

Bibligrafía

- http://www.nltk.org/book/
- http://textminingonline.com/dive-into-nltk-part-i-getting-started-with-nltk
- http://www.ling.helsinki.fi/kit/2009s/clt231/NLTK/book/
- http://cesaraguilar.weebly.com/uploads/2/7/7/5/2775690/pln_uc_09.pdf
- http://www.clips.ua.ac.be/pages/using-wikicorpus-nltk-to-build-a-spanish-part-of-speech-tagger
- https://github.com/dav009/awesome-spanish-nlp
- https://github.com/alvations/spaghetti-tagger
- http://treetaggerwrapper.readthedocs.io/en/latest/
- http://www.nltk.org/api/nltk.tag.html#module-nltk.tag.stanford
- http://stanfordnlp.github.io/CoreNLP/other-languages.html
- http://ixa2.si.ehu.es/ixa-pipes/
- http://nlp.lsi.upc.edu/freeling/node/1
- http://nlp.lsi.upc.edu/freeling-old/index.php?option=com_content&task=view&id=15&Itemid=44
- http://www.smo.uhi.ac.uk/~oduibhin/oideasra/interfaces/winfreeling.htm
- https://talp-upc.gitbooks.io/freeling-user-manual/content/analyzer.html
- https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
- http://nlp.lsi.upc.edu/freeling-old/doc/tagsets/tagset-es.html
- https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon
- http://gplsi.dlsi.ua.es/attos/?q=system/files/publicaciones/5090-4381-1-PB.pdf
- http://docs.qgis.org/2.2/es/docs/user_manual/working_with_pro

jections/working_with_projections.html

- http://mappinggis.com/2016/04/los-codigos-epsg-srid-vinculacion-postgis/
- http://postgis.net/docs/using_postgis_dbmanagement.html#Geography_Basics
- https://www.python.org/
- http://www.tutorialspoint.com/python/
- https://docs.python.org/3/howto/urllib2.html
- https://www.crummy.com/software/BeautifulSoup/bs4/doc/
- https://pymotw.com/2/threading/
- http://www.swig.org/Doc1.1/HTML/Python.html