Trabajo de Fin de Grado

Extracción de información de textos legales y notas de prensa

Extracting information from legal texts and news releases

Carla Hernández Díaz

La Laguna, 4 de marzo de 2017

D. **Isabel Sánchez Berriel**, con N.I.F. 42.885.838-S profesora Contratada Doctora adscrita al Departamento de Departamento de Ingeniería Informática y de Sistemas de la Universidad de La Laguna, como tutora.

D. **Marcos Colebrook Santamaría**, con N.I.F. 43.787.808-V profesor Titular de Universidad adscrito al Departamento de Ingeniería Informática y de Sistemas de la Univeridad de La Laguna, como cotutor.

C E R T I F I C A (N)

Que la presente memoria titulada:

"Extracción de información de textos legales y notas de prensa"

ha sido realizada bajo su dirección por D. **Carla Hernández Díaz**,

con N.I.F. 54.048.936-V.

Y para que así conste, en cumplimiento de la legislación vigente y a los efectos oportunos firman la presente en La Laguna a 4 de marzo de 2017

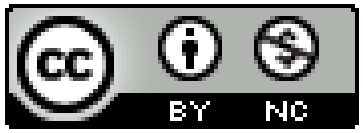
Agradecimientos

Agradecer a Isabel Sánchez Berriel y Marcos Colebrook Santamaría por la gran ayuda y constancia que me han ofrecido en todo momento.

A mis amigos más cercanos por el apoyo y los buenos ratos que hemos pasado durante esta etapa.

Y por último, a mi familia que gracias a su paciencia en los momentos más duros y su alegría en los buenos han conseguido que llegue a donde estoy ahora. Sin ellos esto no habría sido posible.

Licencia



Resumen

Este trabajo consiste en la creación de una herramienta que permita analizar documentos legales que cumplen unas determinadas características obtenidos del CENDOJ (Centro de Documentación Judicial). En concreto, se trata de sentencias judiciales sobre delitos relacionados con la corrupción que se procesarán con el fin de obtener información que ayude a entender la evolución de estos casos. Por otra parte, los datos extraídos son almacenados de manera estructurada para poder obtener con mayor facilidad las noticias publicadas en la prensa digital que estén relacionadas con ellas. Toda esta información, tanto los datos de las sentencias como las noticias, es usada para su posterior visualización a través de una página web.

Palabras clave: extracción de información, corrupción, documentos legales, CENDOJ, noticias.

Abstract

This project consists of the creation of a tool to analyze legal documents obtained from CENDOJ (Judicial Documentation Center) that meet certain characteristics. Specifically, these are judicial sentences on crimes related to corruption that will be processed in order to obtain information to help understand the evolution of these cases.

On the other hand, the extracted data are stored in a structured way to obtain more easily the news published in the digital press related to them. All this information, both the data of the sentences and the news, is used for later visualization through a web page-

Keywords: extraction of information, legal documents, corruption, news, visualization

Índice general

Índice de tablas

Capítulo 1
Introducción

Cuando consultamos un periódico tendemos a pensar que la información descrita en él es una fiel imagen de la realidad sin llegar a cuestionar si dicha información es no no del todo real o está incompleta. Además, dada la gran cantidad de noticias que día a día nos llegan no somos capaces de profundizar en las cuestiones que se tratan en ellas, lo que finalmente se traduce en un conocimiento superficial de los diferentes temas tratados por la prensa. Por otro lado, es un hecho que el perfil ideológico de la prensa actual queda reflejado en la orientación o incluso importancia que se de a una determinada noticia por parte de los diferentes medios. Debido a la gran relevancia que han ido adquiriendo las noticias de contenido político, y dentro de este campo las relacionadas con la corrupción, surge la idea central de este trabajo. A partir de estas consideraciones se plantea abordar el problema de contrastar las noticias sobre casos de corrupción proporcionadas por los periódicos frente a fuentes públicas que constituyen el origen de las mismas. También se confrontan los diferentes tratamientos que se hace de la noticia según el medio en el que se recupere. Se utilizan para ello herramientas que utilizan técnicas

de procesamiento de lenguaje natural para obtener y ampliar dicha información. En este capítulo se concretan los objetivos y alcance del trabajo, presentándose también los antecedentes que han contribuido a diseñar la solución final.

1.2 Objetivos

El objetivo de este proyecto es la creación de una herramienta que permite obtener información a partir de sentencias judiciales obtenidas en el buscador de la base de datos de jurisprudencia del Consejo General del Poder Judicial: CENDOJ (Centro de Documentación Judicial), y almacenarla en un formato estructurado para su posterior uso y análisis. Además, los datos extraídos se utilizan para construir filtros de búsqueda que permiten recuperar noticias relacionadas con cada uno de los casos que aparecieron en el tiempo en los periódicos digitales, integrándose así diversas fuentes de información.

Aunque algunos datos en la sentencias del CENDOJ están anonimizados, siendo totalmente ficticios los sujetos implicados en la sentencia, se pretende, a partir de aquellos datos que si son fidedignos construir filtros de búsqueda para localizar noticias en los periódicos respecto al caso en cuestión. Asimismo, y utilizando las fechas que aparecen en los diferentes textos tratados se pueden compilar históricos de noticias antes y

después de la publicación de la sentencia.

Si bien el objetivo principal del proyecto es la obtención en formato JSON de la información que se recupera, para su posterior uso en la realización de estadísticas y/o representación visual que facilite la obtención de conclusiones a partir de la información recuperada, también se marcó como objetivo del proyecto el desarrollo de un prototipo que facilita la presentación de resúmenes de los datos generados.

1.3 Alcance

La herramienta que se ha desarrollado debe llevar a cabo las siguientes tareas:

- Permitirá la extracción de información de sentencias judiciales, como las entidades públicas/privadas, figuras judiciales, fallo, delito y cargos.
- Permitirá almacenar la información en formato estructurado para su posterior análisis.
- Permitirá que esta información pueda ser utilizada como filtro de búsqueda para obtener la noticias relacionadas con la sentencia.
- Permitirán recuperar históricos de noticias publicadas en la prensa digital relacionadas con el caso en cuestión.
- Permitirá mostrar las notas de prensa obtenidas y las estadísticas relacionadas con cada sentencia.

1.4 Antecedentes

La idea de este proyecto surge a raíz del Trabajo Final de Grado [1] *Extracción y visualización de información de textos legales* realizado en el curso 2014-15 por el alumno Francisco Javier Rodríguez Dioniz.

Ambos trabajos tienen en común el uso del procesamiento del lenguaje natural en la extracción de información a partir de sentencias judiciales. En el trabajo citado se extraían los actores implicados en las sentencias, como jueces, procuradores, etc. y un conjunto de características que permitieron mediante técnicas de aprendizaje, clasificar cualquier caso en favorable, desfavorable o parcial. Se desarrolló también una herramienta para la visualización de estadísticas sobre las sentencias en cuestión.

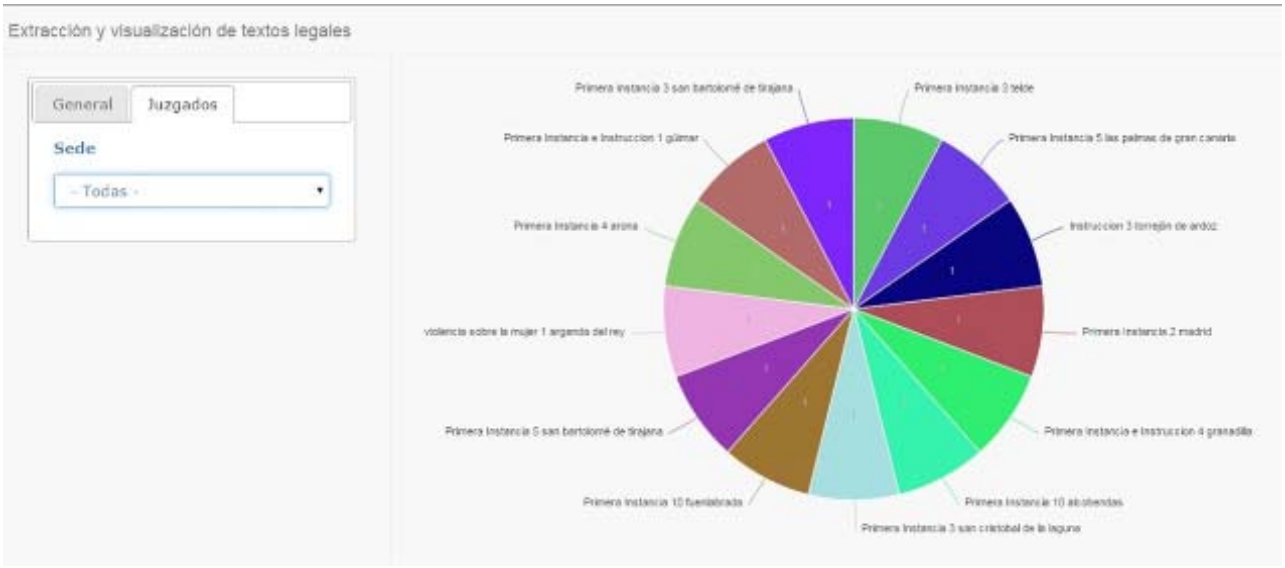


Figura 1.1: Interfaz de la aplicación (Fuente: Extracción y visualización de información de textos legales. Francisco Javier Rodríguez...)

Este trabajo también está fuertemente influenciado por los diferentes proyectos en Civio (www.civio.es), organización que impulsa la transparencia y acceso a los datos públicos, desarrollando herramientas que facilitan su difusión y análisis para una mejor comprensión de por parte de los ciudadanos. Algunos de los proyectos de Civio son: son *¿Dónde van mis impuestos?* (explora los Presupuestos Generales del Estado), *Digo Diego* (extrae los tuits más polémicos de los políticos que han sido borrados tras su publicación) y *El BOE*, (extrae y convierte en noticias información del BOE), *El Indultómetro* (clasifica la información sobre indultos en el BOE), o *Quién Manda* (crea un mapa público-privado de poder en España).

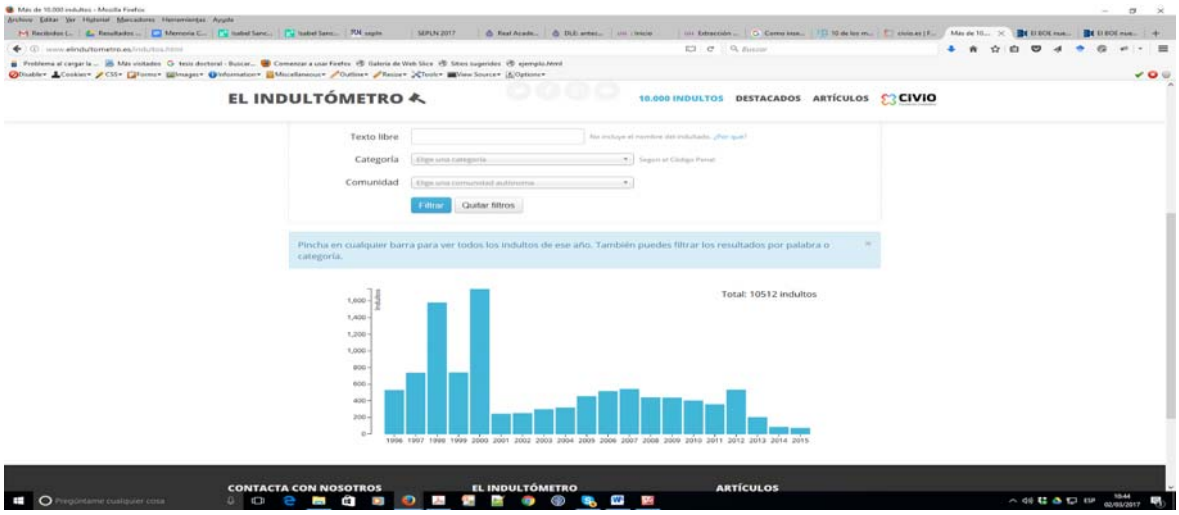


Figura 1.2 : Interfaz de El Indultómetro (Fuente: El Indultómetro. Fundación Civio (http://www.elindultometro....)

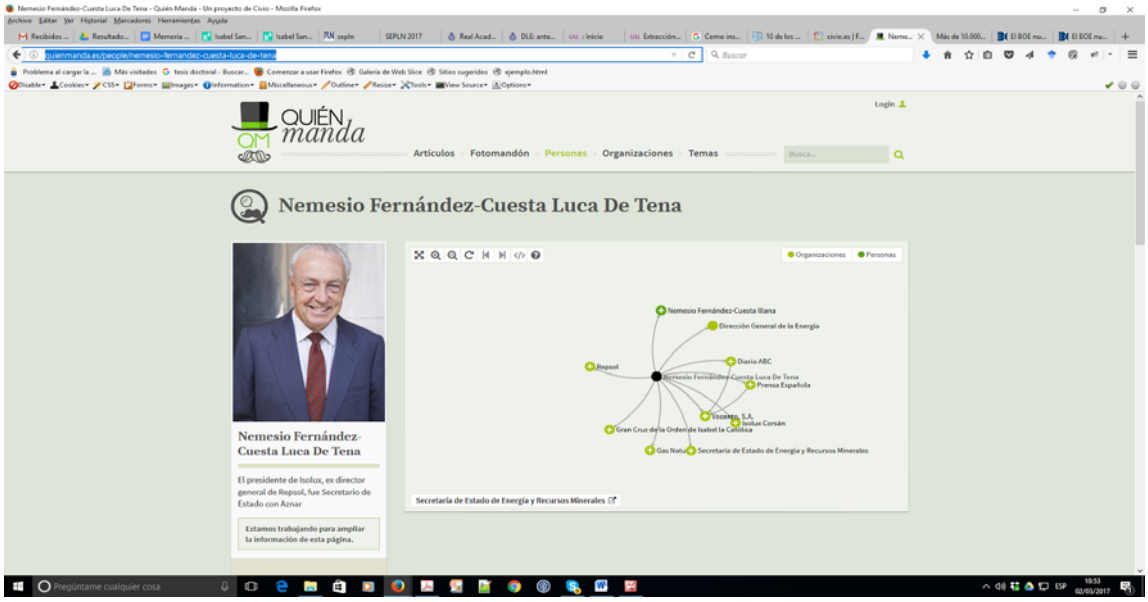


Figura 1.3: Interfaz de ¿Quién Manda?. (Fuente: ¿Quién Manda?. Fundación Civio (http://quienmanda.es/)

1.5 Destinatarios

Esta herramienta va dirigida a todas aquellos ciudadanos que quieran contrastar las noticias sobre casos de corrupción en la prensa frente a la fuente original de la información y en los diferentes periódicos.

Capítulo 2
Descripción del proyecto

2.2 Las fuentes de información

2.2.1 Sentencias judiciales

Una sentencia judicial, según el Diccionario del Español Jurídico (DEJ), es una "*Resolución que decide definitivamente el pleito o causa en cualquier instancia o recurso, o que, según las leyes procesales, debe revestir esta forma. Las sentencias, después de un encabezamiento, deben expresar en párrafos separados los antecedentes de hecho, los hechos que han sido probados, los fundamentos de Derecho y el fallo. Deben ir firmadas por el Juez, Magistrado o Magistrados. Asimismo, pueden ser dictadas de viva voz cuando lo prevea expresamente la legislación procesal aplicable*".

El CENDOJ o Centro de Documentación Judicial es un órgano técnico y centro tecnológico del Consejo General del Poder Judicial. Éste publica oficialmente, a través de su web *(poderjudicial.es)*, las sentencias que han generado jurisprudencia de todos los Tribunales colegiados en España de

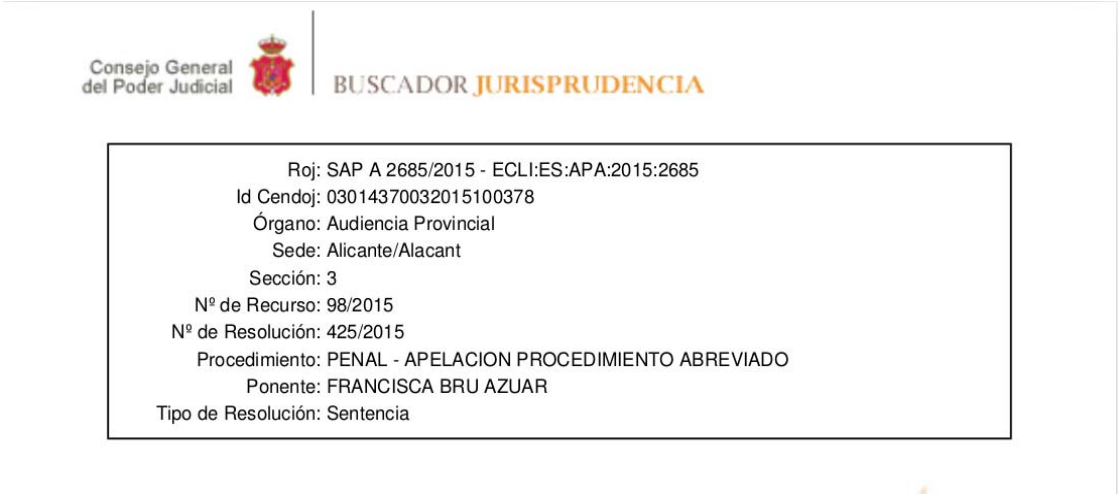
manera gratuita. Esta documentación tiene autorizado el uso particular, no se permite el uso comercial. También restringe la descarga masiva limitándola a un número determinado de documentos al día. Algunos de los criterios de búsqueda que permite especificar son los siguientes:

- Jurisdicción: Civil, Penal, Contencioso, Social, Militar y Especial.
- Tipo de resolución: Auto, Auto aclaratorio, Sentencia y Acuerdo.
- Tipo de órgano: Tribunal Supremo, Audiencia Nacional, Audiencia Provincial, Tribunal Superior de Justicia, Juzgado de Instrucción, etc.
- Localización: provincias españolas.

Las sentencias que nos ofrece el Fondo de Documentación del CENDOJ están en formato PDF, contando con un encabezamiento que incluye información general. Lo que resta del documento se divide en parte expositiva, la parte considerativa y la parte resolutiva. La información general recoge:

- Id Cendoj: identificador de cada sentencia.
- Órgano: tribunal para administrar la justicia.
- Sede: lugar (comunidad autónoma, provincia, etc.).
- Sección: número de sección.
- Número de Recurso.
- Número de Resolución.
- Procedimiento: tipo de procedimiento (penal, civil, etc.).

- Ponente: magistrado ponente.
- Tipo de Resolución: sentencia, auto, auto aclaratorio, etc.

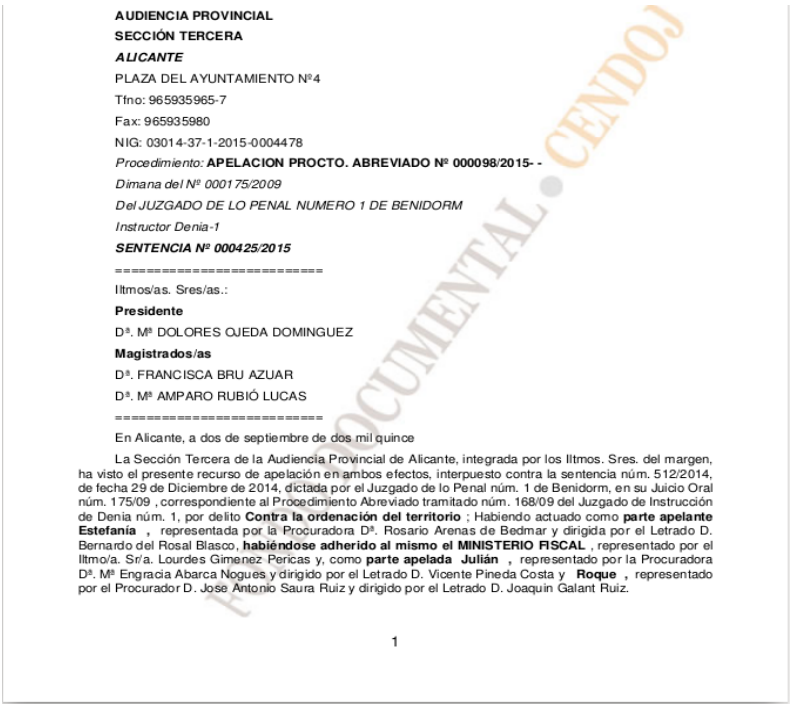

Consejo General del Poder Judicial

BUSCADOR JURISPRUDENCIA

Roj: SAP A 2685/2015 - ECLI:ES:APA:2015:2685
Id Cendoj: 03014370032015100378
Órgano: Audiencia Provincial
Sede: Alicante/Alacant
Sección: 3
Nº de Recurso: 98/2015
Nº de Resolución: 425/2015
Procedimiento: PENAL - APELACION PROCEDIMIENTO ABREVIADO
Ponente: FRANCISCA BRU AZUAR
Tipo de Resolución: Sentencia

Figura 2.1: Información general de la sentencia.

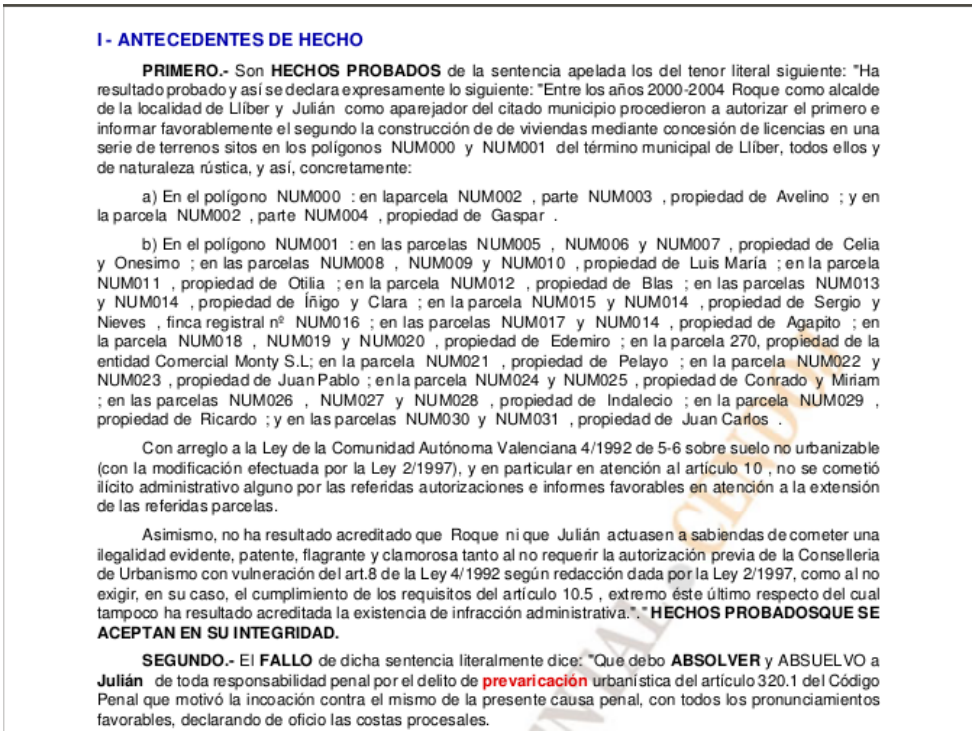
- La parte expositiva: figuras judiciales, juzgado, lugar y fecha en la que se dicta y los antecedentes.


AUDIENCIA PROVINCIAL
SECCIÓN TERCERA
ALICANTE
PLAZA DEL AYUNTAMIENTO Nº4
Tfno: 965935965-7
Fax: 965935980
NIG: 03014-37-1-2015-0004478
Procedimiento: **APELACION PROCTO. ABREVIADO Nº 000098/2015- -**
Dimana del Nº 000175/2009
Del JUZGADO DE LO PENAL NUMERO 1 DE BENIDORM
Instructor Denia-1
SENTENCIA Nº 000425/2015
==========
Iltmos/as. Sres/as.:
Presidente
Dª. Mª DOLORES OJEDA DOMINGUEZ
Magistrados/as
Dª. FRANCISCA BRU AZUAR
Dª. Mª AMPARO RUBIÓ LUCAS
==========
En Alicante, a dos de septiembre de dos mil quince

La Sección Tercera de la Audiencia Provincial de Alicante, integrada por los Iltmos. Sres. del margen, ha visto el presente recurso de apelación en ambos efectos, interpuesto contra la sentencia núm. 512/2014, de fecha 29 de Diciembre de 2014, dictada por el Juzgado de lo Penal núm. 1 de Benidorm, en su Juicio Oral núm. 175/09 , correspondiente al Procedimiento Abreviado tramitado núm. 168/09 del Juzgado de Instrucción de Denia núm. 1, por delito **Contra la ordenación del territorio** ; Habiendo actuado como **parte apelante Estefanía** , representada por la Procuradora Dª. Rosario Arenas de Bedmar y dirigida por el Letrado D. Bernardo del Rosal Blasco, **habiéndose adherido al mismo el MINISTERIO FISCAL** , representado por el Iltmo/a. Sr/a. Lourdes Gimenez Pericas y, como **parte apelada Julián** , representado por la Procuradora Dª. Mª Engracia Abarca Nogues y dirigido por el Letrado D. Vicente Pineda Costa y **Roque** , representado por el Procurador D. Jose Antonio Saura Ruiz y dirigido por el Letrado D. Joaquin Galant Ruiz.

1

Figura 2.2: Parte expositiva de la sentencia.



I - ANTECEDENTES DE HECHO

PRIMERO.- Son **HECHOS PROBADOS** de la sentencia apelada los del tenor literal siguiente: "Ha resultado probado y así se declara expresamente lo siguiente: "Entre los años 2000-2004 Roque como alcalde de la localidad de Llíber y Julián como aparejador del citado municipio procedieron a autorizar el primero e informar favorablemente el segundo la construcción de de viviendas mediante concesión de licencias en una serie de terrenos sitos en los polígonos NUM000 y NUM001 del término municipal de Llíber, todos ellos y de naturaleza rústica, y así, concretamente:

a) En el polígono NUM000 : en laparcela NUM002 , parte NUM003 , propiedad de Avelino ; y en la parcela NUM002 , parte NUM004 , propiedad de Gaspar .

b) En el polígono NUM001 : en las parcelas NUM005 , NUM006 y NUM007 , propiedad de Celia y Onesimo ; en las parcelas NUM008 , NUM009 y NUM010 , propiedad de Luis María ; en la parcela NUM011 , propiedad de Otilia ; en la parcela NUM012 , propiedad de Blas ; en las parcelas NUM013 y NUM014 , propiedad de Íñigo y Clara ; en la parcela NUM015 y NUM014 , propiedad de Sergio y Nieves , finca registral nº NUM016 ; en las parcelas NUM017 y NUM014 , propiedad de Agapito ; en la parcela NUM018 , NUM019 y NUM020 , propiedad de Edemiro ; en la parcela 270, propiedad de la entidad Comercial Monty S.L; en la parcela NUM021 , propiedad de Pelayo ; en la parcela NUM022 y NUM023 , propiedad de Juan Pablo ; en la parcela NUM024 y NUM025 , propiedad de Conrado y Miriam ; en las parcelas NUM026 , NUM027 y NUM028 , propiedad de Indalecio ; en la parcela NUM029 , propiedad de Ricardo ; y en las parcelas NUM030 y NUM031 , propiedad de Juan Carlos .

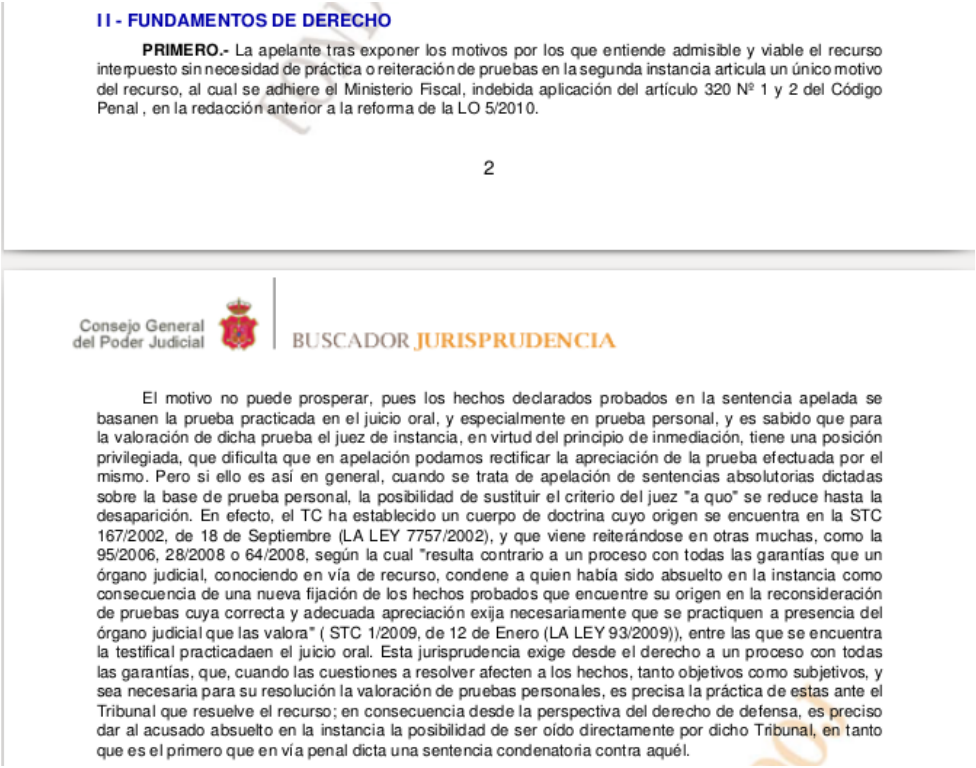
Con arreglo a la Ley de la Comunidad Autónoma Valenciana 4/1992 de 5-6 sobre suelo no urbanizable (con la modificación efectuada por la Ley 2/1997), y en particular en atención al artículo 10 , no se cometió ilícito administrativo alguno por las referidas autorizaciones e informes favorables en atención a la extensión de las referidas parcelas.

Asimismo, no ha resultado acreditado que Roque ni que Julián actuasen a sabiendas de cometer una ilegalidad evidente, patente, flagrante y clamorosa tanto al no requerir la autorización previa de la Conselleria de Urbanismo con vulneración del art.8 de la Ley 4/1992 según redacción dada por la Ley 2/1997, como al no exigir, en su caso, el cumplimiento de los requisitos del artículo 10.5 , extremo éste último respecto del cual tampoco ha resultado acreditada la existencia de infracción administrativa.". "**HECHOS PROBADOSQUE SE ACEPTAN EN SU INTEGRIDAD.**

SEGUNDO.- El **FALLO** de dicha sentencia literalmente dice: "Que debo **ABSOLVER** y ABSUELVO a **Julián** de toda responsabilidad penal por el delito de **prevaricación** urbanística del artículo 320.1 del Código Penal que motivó la incoación contra el mismo de la presente causa penal, con todos los pronunciamientos favorables, declarando de oficio las costas procesales.

Figura 2.3: Parte expositiva de la sentencia.

- Parte considerativa: fundamentos de derecho, jurídicos o de hecho que contienen los argumentos de las partes.



II - FUNDAMENTOS DE DERECHO

PRIMERO.- La apelante tras exponer los motivos por los que entiende admisible y viable el recurso interpuesto sin necesidad de práctica o reiteración de pruebas en la segunda instancia articula un único motivo del recurso, al cual se adhiere el Ministerio Fiscal, indebida aplicación del artículo 320 Nº 1 y 2 del Código Penal , en la redacción anterior a la reforma de la LO 5/2010.

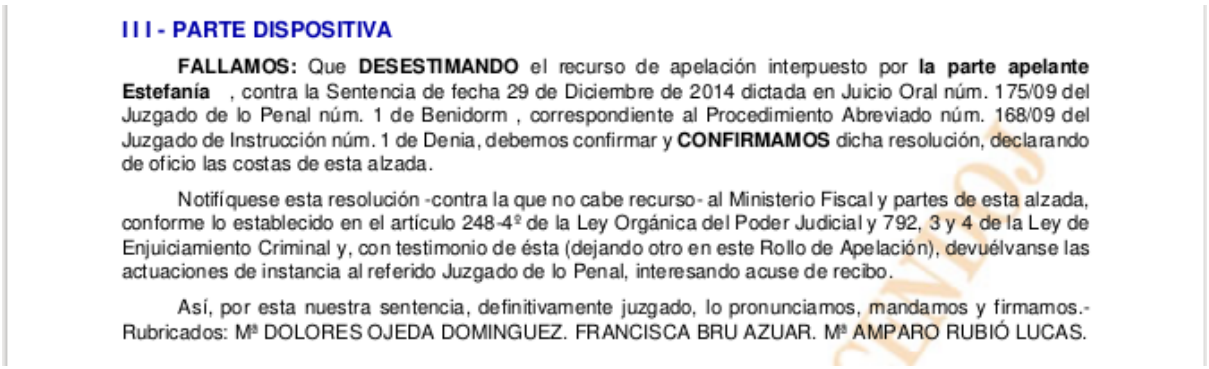
2

Consejo General del Poder Judicial BUSCADOR JURISPRUDENCIA

El motivo no puede prosperar, pues los hechos declarados probados en la sentencia apelada se basanen la prueba practicada en el juicio oral, y especialmente en prueba personal, y es sabido que para la valoración de dicha prueba el juez de instancia, en virtud del principio de inmediación, tiene una posición privilegiada, que dificulta que en apelación podamos rectificar la apreciación de la prueba efectuada por el mismo. Pero si ello es así en general, cuando se trata de apelación de sentencias absolutorias dictadas sobre la base de prueba personal, la posibilidad de sustituir el criterio del juez "a quo" se reduce hasta la desaparición. En efecto, el TC ha establecido un cuerpo de doctrina cuyo origen se encuentra en la STC 167/2002, de 18 de Septiembre (LA LEY 7757/2002), y que viene reiterándose en otras muchas, como la 95/2006, 28/2008 o 64/2008, según la cual "resulta contrario a un proceso con todas las garantías que un órgano judicial, conociendo en vía de recurso, condene a quien había sido absuelto en la instancia como consecuencia de una nueva fijación de los hechos probados que encuentre su origen en la reconsideración de pruebas cuya correcta y adecuada apreciación exija necesariamente que se practiquen a presencia del órgano judicial que las valora" (STC 1/2009, de 12 de Enero (LA LEY 93/2009)), entre las que se encuentra la testifical practicadaen el juicio oral. Esta jurisprudencia exige desde el derecho a un proceso con todas las garantías, que, cuando las cuestiones a resolver afecten a los hechos, tanto objetivos como subjetivos, y sea necesaria para su resolución la valoración de pruebas personales, es precisa la práctica de estas ante el Tribunal que resuelve el recurso; en consecuencia desde la perspectiva del derecho de defensa, es preciso dar al acusado absuelto en la instancia la posibilidad de ser oído directamente por dicho Tribunal, en tanto que es el primero que en vía penal dicta una sentencia condenatoria contra aquél.

Figura 2.4: Parte considerativa de la sentencia.

- Parte resolutiva: parte dispositiva o fallo que contiene la decisión de condena o absolución de cada una de las partes.



III - PARTE DISPOSITIVA

FALLAMOS: Que **DESESTIMANDO** el recurso de apelación interpuesto por **la parte apelante Estefanía** , contra la Sentencia de fecha 29 de Diciembre de 2014 dictada en Juicio Oral núm. 175/09 del Juzgado de lo Penal núm. 1 de Benidorm , correspondiente al Procedimiento Abreviado núm. 168/09 del Juzgado de Instrucción núm. 1 de Denia, debemos confirmar y **CONFIRMAMOS** dicha resolución, declarando de oficio las costas de esta alzada.

Notifíquese esta resolución -contra la que no cabe recurso- al Ministerio Fiscal y partes de esta alzada, conforme lo establecido en el artículo 248-4º de la Ley Orgánica del Poder Judicial y 792, 3 y 4 de la Ley de Enjuiciamiento Criminal y, con testimonio de ésta (dejando otro en este Rollo de Apelación), devuélvanse las actuaciones de instancia al referido Juzgado de lo Penal, interesando acuse de recibo.

Así, por esta nuestra sentencia, definitivamente juzgado, lo pronunciamos, mandamos y firmamos.- Rubricados: Mª DOLORES OJEDA DOMINGUEZ. FRANCISCA BRU AZUAR. Mª AMPARO RUBIÓ LUCAS.

Figura 2.5: Parte resolutiva de la sentencia.

Este trabajo está dirigido a sentencias relacionadas con casos de corrupción política, por lo que se ha añadido como criterio para la

recuperación de la base de datos del CENDOJ los delitos de malversación[1] o prevaricación[2]. Ambos delitos están directamente relacionados con la corrupción, como se comprobó en los preliminares del trabajo.

No toda la información que se puede extraer de la sentencia es real. Esto es debido a que los datos personales de los involucrados (encausados, imputados, investigados, acusados, etc.) no pueden ser desvelados. La información real a extraer de la sentencia es la información básica, también las figuras judiciales (procuradores, letrados, magistrados, presidente, etc.), el juzgado (número, lugar y tipo), la fecha y el lugar. Los cargos, organizaciones públicas o privadas y las localizaciones, además del delito son extraídos de los Antecedentes de Hecho. La parte formada por los Fundamentos de Derecho no es utilizada para la extracción de información. Por último se obtiene la decisión o fallo de la sentencia en la Parte Dispositiva.

2.2.2 Noticias en la prensa digital

Para la realización de este trabajo se ha elegido sentencias en tres Comunidades Autónomas: Valencia, Andalucía y Canarias, debido a que las dos primeras corresponden a comunidades en las que gobiernan los dos

1 Conducta delictiva que lleva a cabo la autoridad o el funcionario público que, en el ejercicio de sus funciones, realiza una administración desleal del patrimonio público o se apropia indebidamente de objetos que forman parte de dicho patrimonio.

2 Delito que realiza la autoridad o funcionario público que dicta una resolución arbitraria en un asunto administrativo, a sabiendas de su injusticia.

principales partidos políticos en España en los últimos años y el caso de Canarias por ser nuestra comunidad. Con esto se ha hecho una selección de periódicos de tirada nacional y periódicos locales a los ámbitos señalados en formato digital. Se han recopilado las direcciones para las consultas en la página http://www.prensaescrita.com/prensadigital.php en la que se mantiene un inventario de la prensa en la mayor parte de países del mundo. El compendio de periódicos compilados según su ámbito se muestra a continuación:

Nacionales:

- www.elpais.es
- www.abc.es
- www.laprovincia.es
- www.lavanguardia.com
- www.elmundo.es
- www.larazon.es
- www.20minutos.es
- www.elperiodico.com
- www.eleconomista.es

Canarias:

- www.eldia.es
- www.canarias7.es
- www.diariodeavisos.com

- www.laopinion.es

Andalucía:

- www.laopiniondesevilla.es
- www.abcdesevilla.es
- www.diariosur.es
- www.diariodecadiz.es
- www.diariodejerez.es
- www.europasur.es
- www.granadahoy.com
- www.diariocordoba.com
- www.lavozdealmeria.es
- www.diariojaen.es
- www.vivajaen.es
- www.vivahuelva.es

Valencia:

- www.levante-emv.com
- www.lasprovincias.es
- www.diarioinformacion.com
- www.elperiodicomediterraneo.com

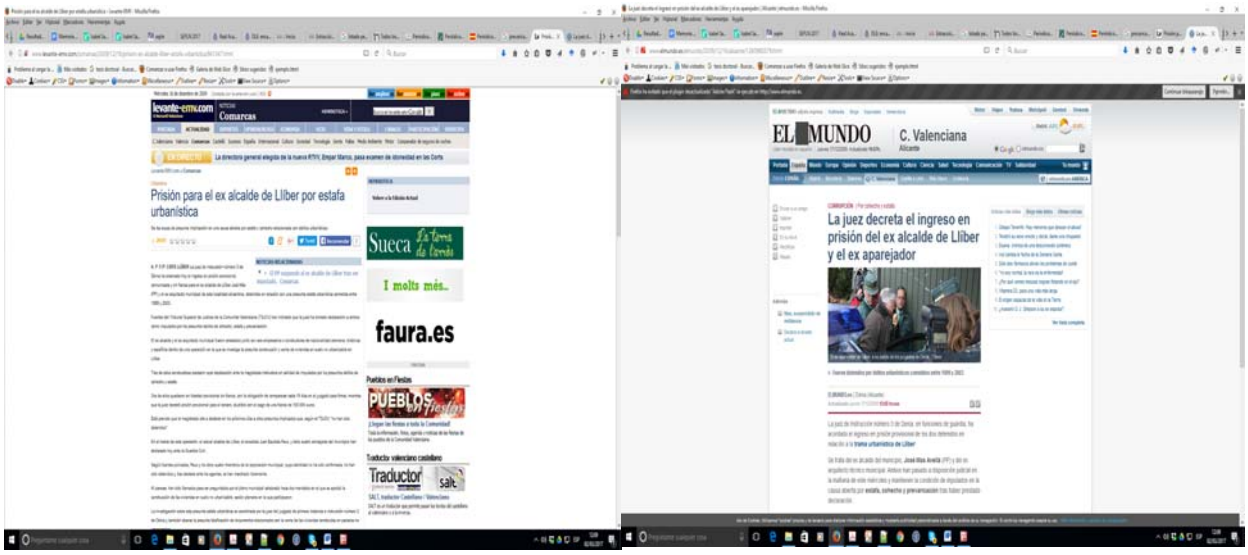


Figura 2.6: Tratamiento dado a una de las sentencias en la prensa. Fuente: Levante-emv, El Mundo.

2.3 Manejo de la información

2.3.1 Procesamiento de Lenguaje Natural

El procesamiento de lenguaje natural o PLN es un campo de las ciencias de la computación, inteligencia artificial y lingüística que estudia las interacciones con las computadoras en lenguaje humano. El objetivo de esta disciplina es hacer entender a los computadores el texto y se aplica en diferentes problemas como la traducción automática, los resúmenes automáticos, entre otros, y como en este caso en la extracción de información. Según el propósito para el que se emplee nos encontramos con diferentes técnicas y niveles de análisis posibles de los textos. Podemos hablar de identificación del lenguaje, análisis morfológico, sintáctico, semántico, etc.

Una gran parte de las tareas que conllevan aplicar PLN sigue un flujo en el que se concatenan diferentes subtareas, ya que para resolver la tarea actual se requieren los resultados de algún análisis previo. Un flujo de procesamiento típico en una aplicación de PLN incluye los siguientes procesos básicos:

- Tokenización: Se obtienen tokens o unidades en el texto, estos pueden ser palabras, signos de puntuación, etc.
- Splitting: Consiste en la división en frases de listas de tokens.
- Análisis morfológico: Se obtienen todas las formas canónicas posibles para una palabra, se identifican números, fechas, monedas, entidades. Las entidades son palabras o conjuntos de palabras con un significado especial en el texto, generalmente se utiliza para el reconocimiento de personas, localizaciones y organizaciones.

  • Análisis morfosintáctico: En este análisis se determina la categoría morfosintáctica de la palabras, asignándosele la etiqueta que corresponda a dicha categoría, este proceso también se conoce como POS Tagger (Part of Speech Tagger).

  • Análisis sintáctico: establece las relaciones estructurales y de dependencia entre las palabras dentro de la frase. Este nivel de análisis abarca tareas como la correferencia, árboles de dependencia, etc.

- Análisis semántico: Este nivel de análisis lingüístico abarca procesos como anotación de sentidos, desambiguación de sentidos, etc.

Se presenta como ejemplo un fragmento de texto extraído de la sentencia utilizada como ejemplo en esta memoria (ID CENDOJ 92875898275924758*)*: *Entre los años 2000-2004 Roque como alcalde de la localidad de Llíber y Julián como aparejador del citado municipio procedieron a autorizar el primero e informar favorablemente...* como ejemplo. El análisis básico dividirá el texto en tokens: palabras y signos de puntuación y reconocerá el ámbito de cada frase. Si realizamos el análisis morfológico obtendremos resultados del tipo: *aparejador* es un nombre o un adjetivo, sin embargo el resultado del etiquetado morfosintáctico nos indicará que en este texto *aparejador* es un nombre. *Julián* será identificado como un nombre propio, etc. Toda esta información se suele proveer mediante etiquetas a las que se les asigna un valor de la fiabilidad de ser la solución correcta.

Figura 2.7: Análisis morfosintáctico del texto de ejemplo usando la demo online de FreeLing

En la actualidad existen numerosas librerías de PLN para el idioma inglés (OpenNLP, Standford Core NLP, NLTK, MALLET, …). Aunque también pueden ser adaptadas a otros idiomas, incluido el español, no dejan de ser librerías orientadas a otra lengua. A diferencia de estas, FreeLing[3] e Ixa Pipes[21] son librerías para el procesamiento del lenguaje natural que han evolucionado con el español entre sus objetivos.

2.3.1.1 Reconocimiento y clasificación de Entidades

Como se ha mencionado previamente, de especial relevancia para este trabajo es el Reconocimiento y Clasificación de Entidades Nombradas en el texto. Estas tareas son denominadas NER y NEC respectivamente, siglas que vienen de su nombre en inglés, consiste en la identificación en el texto de palabras o expresiones con algún significado de interés, por ejemplo: nombres de personas, de organizaciones, de productos, de localizaciones, etc. Se distingue el proceso de identificarlas (NER) y el proceso de clasificarlas en alguno de los grupos de interés (NEC). Esta herramienta se aplicará en el trabajo para obtener los actores implicados en el delito: empresas y organismos públicos, los cargos políticos y también las localizaciones relacionadas.

Los algoritmos para el reconocimiento de entidades podemos englobarlos en dos categorías: reglas basadas en patrones de mayúsculas o aplicar un algoritmo de aprendizaje automático para clasificación.

- Reglas basadas en mayúsculas: por lo general se especifican en términos de palabras que empiezan por mayúscula y no son inicio de frase, secuencias de palabras en mayúsculas (se pueden considerar en la secuencia palabras funcionales que están en minúsculas pero forman parte de la entidad). Se debe considerar casos especiales, como palabras que se excluyen aunque estén en mayúsculas (Spanish en el idioma inglés), o palabras que pueden ser etiquetadas como una forma reconocida de la lengua pero también como una entidad (Campo, es un apellido que puede ser identificado como entidad y también un sustantivo masculino).
- Algoritmos de clasificación: Se aplican algoritmos de aprendizaje automático para decidir sobre cada palabra en el texto si es de alguna de las categorías: B, I, O. La categoría B se reserva para palabras con las que comienza una entidad, la categoría I para cualquier otra que no sea la inicial y forme parte de la entidad y la categoría O para palabras que no forman parte de una entidad.

La clasificación de entidades se resuelve mediante algoritmos de aprendizaje automático aplicados a las entidades reconocidas previamente. Se definen características de las palabras a clasificar como su forma, los n-gramas o secuencias de palabras en una ventana alrededor de la actual, categoría gramatical, afijos, sufijos, etc. que serán usadas por algún algoritmo de clasificación multiclase. Por lo general las categoría que se tienen en cuenta

son: Persona, Organización, Localización y Miscelánea, esta última para aquellas entidades que no se clasifican en ninguno de los anteriores grupos. El rendimiento de estos sistemas decae cuando el dominio difiere del de entrenamiento, una alternativa en este caso puede ser los sistemas basados en reglas, o bien los sistemas híbridos. El conjunto de reglas consiste en patrones gramaticales, sintácticos y ortográficos combinados con diccionarios de entidades. Si se dispone de vocabularios (gazetters) de entidades clasificadas se utilizan para determinar la clase de la entidad sin necesidad de aplicar el algoritmo de aprendizaje. En este caso, se asigna la clase por estar recogida la entidad en el vocabulario o porque es reconocida mediante alguna expresión regular correspondiente a las reglas que cumplen las entidades. Este tipo de soluciones mejora el rendimiento pero pierde generalidad, son dependientes del dominio.

Si se aplica este tipo de análisis al ejemplo anterior, se reconocen como entidades de tipo persona: *Roque*, *Julián* y se les asigna la etiqueta *NP00SP0.* Por otra parte, también se reconoce la entidad *Llíber* y es clasificada como localización (eitqueta *NP00G00*).

Figura 2.8: Análisis morfosintáctico del texto de ejemplo usando la demo online de FreeLing, incluyendo la clasificación de entidades (NEC)

2.3.1.2 Diccionarios

Dado que la información que se va a extraer corresponde a administraciones locales, autonómicas y estatales, se puede asegurar que estamos ante un dominio específico para el que existen numerosos registros públicos y fiables[3]. Esto se traduce en la garantía de disponer de recursos útiles para establecer vocabularios sobre esta materia que permitirán mejorar la precisión y eficiencia del proceso de clasificación de entidades. Por esta razón, se han usado diccionarios para almacenar el vocabulario que puede aparecer en las sentencias judiciales que vamos a analizar. Por un lado, un vocabulario para el tipo de fallo que interesa obtener, y otro con un catálogo exhaustivo de las entidades nombradas. Éste último se crea con objeto de clasificar cada una de las entidades en organizaciones, localizaciones o cargos. Hay que notar que se ha incluido una categoría adicional a las personas, localizaciones y organizaciones que generalmente reconocen los módulos implementados para el análisis NEC en las librerías de PLN.

Estos diccionarios permite a los usuario añadir, modificar o eliminar elementos según su criterio para el análisis de sentencias legales. A continuación se detallan los vocabularios que se han considerado para cada

3 Cabe destacar la proliferación de portales de Transparencia de las diferentes administraciones que han surgido como consecuencia de la La Ley 19/2013, de 9 de diciembre, de Transparencia, Acceso a la Información Pública y Buen Gobierno. La mayor parte de vocabularios se han elaborado a partir de publicaciones en estos portales.

uno de los tipos de entidades que interesan:

- Organizaciones: En este caso corresponderá a los organismo públicos, en cualquiera de los ámbitos estatal, autonómico o local. Estos se dividen en organismos autónomos, entidades públicas empresariales, agencias y entidades a los que se les reconozca por una Ley la independencia funcional o una especial autonomía respecto de la Administración en su ámbito. Este vocabulario se ha obtenido del Inventario de Entes del Sector Público (INVENTE), publicado en la página del Ministerio de Hacienda y Función Pública [22]. A partir de este catálogo se han incluido en los vocabularios: Ayuntamientos, Agencias, Organismos Estatales, Universidades, Sociedades mercantiles, Entidades públicas empresariales, Diputaciones Provinciales, Consejos, Cabildos, Comunidades Autónomas,, etc.
- Localizaciones: Las localizaciones son todas aquellas provincias, pueblos, ciudades, etc., que pertenezcan a las tres comunidades autónomas elegidas. Este vocabulario está formado por los municipios y provincias de España, la información se ha recuperado en la página del INE [23]
- Cargos: Los cargos públicos son aquellos empleos públicos y puestos de la administración: funcionarios, asesores, contratados, cargos políticos (concejales, diputados, consejeros, ministros, etc.). En este caso como fuente se ha tomado la Relación de Puestos de Trabajo (RPT) de los

diferentes departamentos del Gobierno de Canarias [24] y del Gobierno Vasco [25]. Se comprobó que la similitud en la denominación de la mayor parte de puestos de trabajo en ambos casos, por lo que se elaboró, a partir de ambos, un vocabulario general que fuese válido para cualquier Comunidad Autónoma.

2.3.2 Expresiones regulares

Una expresión regulares es una secuencia de caracteres que forma un patrón utilizada para buscar cadenas de texto. Las expresiones regulares están disponibles en casi cualquier lenguaje de programación, pero aunque su sintaxis es relativamente uniforme, cada lenguaje usa su propio dialecto. Por ejemplo, la expresión: \b\w{4}\b se refiere a todas las palabras de 4 letras.

En este proyecto las expresiones regulares juegan un papel importante ya que con ellas se extrae la información básica de la sentencia, las figuras judiciales, delito, y juzgado. Mediante expresiones regulares se especifican los diferentes lemas que corresponderían a tales casos, por ejemplo: *letrado, letrada, letrados, letradas*, ... El motivo por el que se ha decidido aplicar este método sobre estos datos es la reducción de tiempo de cómputo respecto a otras técnicas más generales propias del PLN como la lematización o el aprendizaje automático.

Capítulo 3
Herramientas del proyecto

En este capítulo se describen los aspectos más relevantes respecto a las herramientas utilizadas en el desarrollo del proyecto. Para el desarrollo se ha utilizado Eclipse como entorno y Java como lenguaje de programación por el conocimiento que ya se tenía de los mismos.

3.1. FreeLing

FreeLing es una librería de código abierto para el procesamiento multilíngüe automático desarrollada y mantenida en el Centro de Investigación TALP de la Universidad Politécnica de Cataluña. La primera versión de FreeLing fue lanzada en 2003 y se distribuye bajo una licencia GNU General Public License3 y bajo licencia dual a empresas que deseen incluirlo en sus productos comerciales.

Se estructura como una librería que puede ser llamada desde cualquier aplicación de usuario que requiera servicios de análisis del lenguaje. Proporciona diversas funcionalidades para el análisis lingüístico (análisis

morfológico, detección y clasificación de entidades, análisis sintáctico, etc.) para una gran variedad de idiomas (español, inglés, italiano, alemán, ruso, catalán, gallego, croata, etc.).

FreeLing es personalizable y ampliable, y está fuertemente orientado a aplicaciones del mundo real en términos de velocidad y robustez. Una de la grandes ventajas de esta herramienta es que permite a los desarrolladores ampliarla y adaptarla a dominios particulares, desarrollar otros nuevos para idiomas específicos o necesidades especiales de las aplicaciones. Ofrece módulos potentes para la aplicación de técnicas de Procesamiento de Lenguaje Natural (PLN) con los diferentes servicios lingüísticos para cada idioma.

Además de las técnicas necesarias en el flujo de las técnicas de PLN (Tokenización, Splitting, Análisis Morfológico) también ofrece otras que se han empleado en esta herramienta como Part-of-Speech tagging y reconocimiento y clasificación de entidades. En la siguiente figura se observan las diferentes tareas disponibles para cada idioma en Freeling.

	as	ca	cy	en	es	gl	it	pt	ru
Tokenization	X	X	X	X	X	X	X	X	X
Sentence splitting	X	X	X	X	X	X	X	X	X
Number detection		X		X	X	X	X	X	X
Date detection		X		X	X	X		X	X
Morphological dictionary	X	X	X	X	X	X	X	X	X
Affix rules	X	X	X	X	X	X	X	X	
Multiword detection	X	X	X	X	X	X	X	X	
Basic named entity detection	X	X	X	X	X	X	X	X	X
B-I-O named entity detection				X	X	X			
Named Entity Classification				X	X				
Quantity detection		X		X	X	X		X	X
PoS tagging	X	X	X	X	X	X	X	X	X
WN sense annotation		X		X	X				
UKB sense disambiguation		X		X	X				
Shallow parsing	X	X		X	X	X		X	
Full/dependency parsing	X	X		X	X	X			
Coreference resolution					X				

*F*igura 3.1: Análisis disponibles en Freeling. Fuente: Página de Freeling [3]

Las etiquetas que asigna Freeling en el análisis morfosintáctico corresponden a las de EAGLES, sistema de etiquetado definido para abarcar todas las lenguas europeas. Estas son de longitud variable y dependiente de la categoría, que siempre se especifica para la primera posición, en las restantes posiciones se codifican atributos propios de la categoría, el 0 se reserva para valores del atributo desconocidos. Por ejemplo, si se identifica un nombre común singular masculino se le asigna la etiqueta: *NCMS*, y si se desconociera el género: *NC0S*

Los módulos NER y NEC de Freeling permiten el reconocimiento y clasificación de las entidades, respectivamente. Las entidades no pueden ser clasificadas sin antes ser reconocidas. Pero para llegar a esto es necesario primero etiquetar gramaticalmente cada palabra (PoS tagger), es decir, necesitamos saber cuáles son las palabras de alrededor

de las entidades para poder clasificarlas correctamente. Aunque sólo se usa la clasificación de organizaciones (etiqueta *NP00O00*) y localizaciones (etiqueta *NP00L00*), también se puede clasificar por personas (etiqueta *NP00SP0*) y otros (etiqueta *NP00V00*).

El módulo NER de Freeling realiza el reconocimiento basado en reglas para las palabras en mayúsculas, aunque también dispone del módulo BioNER basado en aprendizaje automático. Se ha utilizado el primero por cuestiones de rendimiento en términos de tiempo de procesamiento.

3.2.JSON

JSON (JavaScript Object Notation) es un formato de texto ligero para el intercambio de datos. JSON es un lenguaje de texto independiente que utiliza algunas características similares a los lenguajes C, C++, C#, Java, JavaScript, Python, etc. Soporta un gran número de tipos de datos como cadenas, números, booleanos, nulos, vectores y objetos. Está basado en un subconjunto del Lenguaje de Programación JavaScript, Standard ECMA-262 3rd Edition - Diciembre 1999.

Además de los diferentes tipos de datos se pueden utilizar dos tipos de estructuras complejas:

- Coolección de pares de clave/valor. Corresponde a arreglos

asociativos, se especifican con llaves.

- Arrays indexados, se especifican con corchetes.

En el ejemplo de la siguiente figura se ha codificado un objeto con un campo lib que es una array de objetos con que incluye los campos libro, año, un array de autores y el campo editorial.

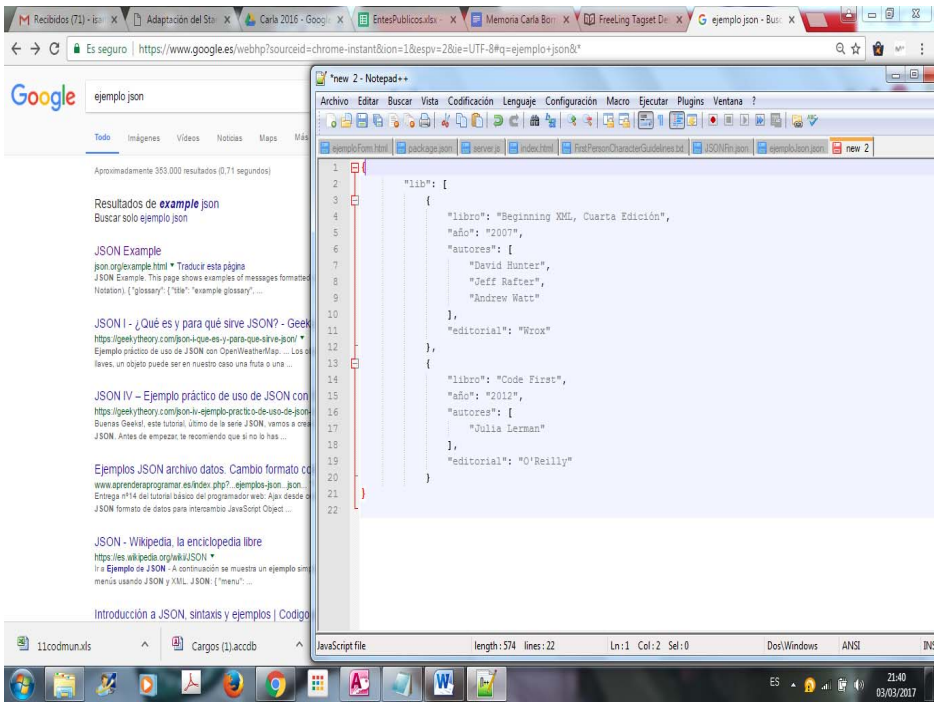


Figura 3.2: ejemplo de fichero de tipo JSON

3.3. Apache PDFBox

PDFBox es una librería Java de código abierto que permite trabajar con documentos en formato pdf. Nos permite crear, manipular y extraer el contenido de documentos en formato pdf. Más detalladamente permite:

- Extraer texto de un pdf
- Mezclar varios documentos pdfs en uno
- Dividir en páginas un documento pdf
- Gestionar los formularios de un documento pdf
- Validar un documento pdf contra el estándar PDF/A ISO
- Dado un pdf generar un fichero que puede ser gestionado por el API de impresión de Java.
- Convertir un pdf a una imagen
- Crear un documento pdf a partir de un texto
- Cifrar/descifrar documentos pdf

También incluye varias utilidades de línea de comandos y se publica bajo la Licencia Apache v2.0.

Esta librería ha sido integrada en el desarrollo con objeto de extraer los textos de las sentencias que provee el CENDOJ en formato pdf.

3.4.Google Custom Search API

Google Custom Search (antes conocida como Google Co-op) es una plataforma proporcionada por Google. Permite a los desarrolladores web ofrecer información especializada en búsquedas en la web, refinar y categorizar consultas y crear motores de búsqueda personalizados basados en la Búsqueda de Google. Anteriormente era conocida como Google Web

Search API y se lanzó el 23 de octubre de 2006.

Para utilizar esta API es necesario crear un motor de búsqueda a través de una cuenta de google en la aplicación: https://cse.google.es/cse/all. Se puede configurar dándole un nombre, indicar el idioma, los sitios web donde buscar: páginas sueltas, todo un sitio, partes de un sitio o todo un dominio.

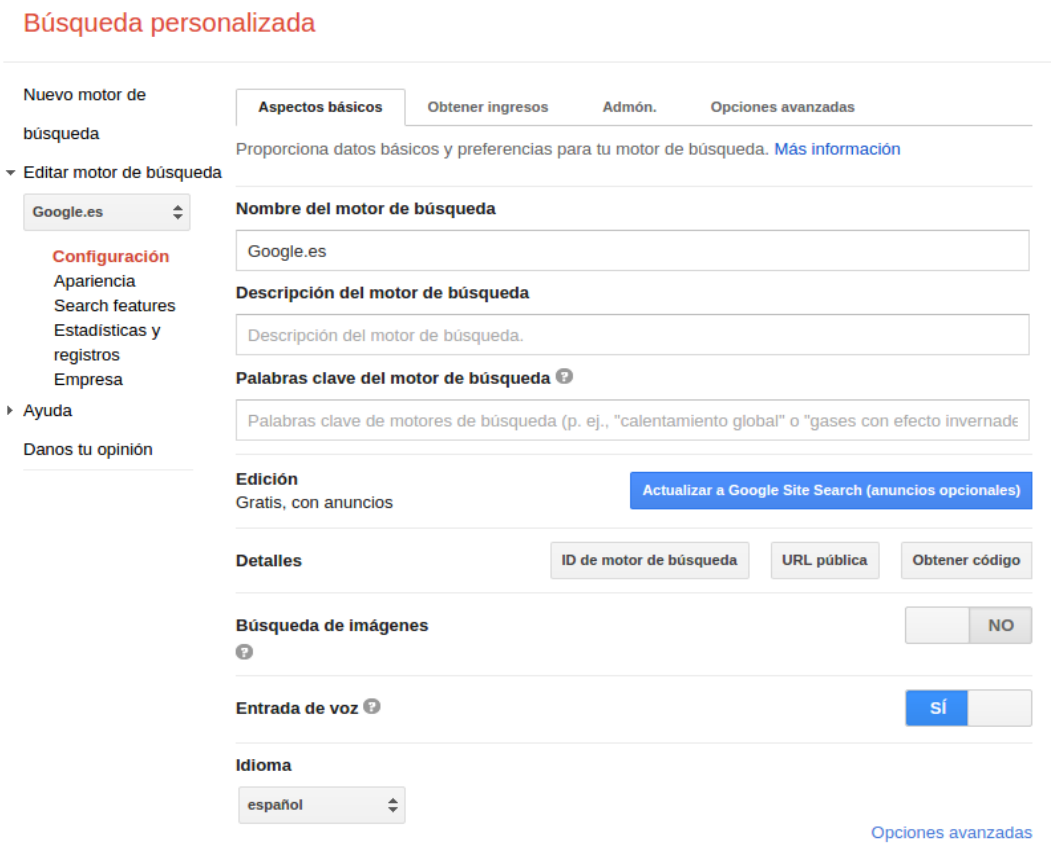


Figura 3.3: Nombre, idioma, detalles, etc. del Motor de búsqueda de GCS.

3.5.Regex Java

Regex es una librería de Java para definir expresiones regulares,

permitiendo definir patrones de búsqueda para la manipulación de cadenas. También se usa para para definir restricciones en cadenas como la validación de contraseña y correo electrónico. Esta librería proporciona clases e interfaces para el uso de estas expresiones regulares.

- Interfaz MatchResult
- Clase Matcher
- Clase Pattern
- Clase PatternSyntaxException

De estas clases se han utilizado Matcher y Pattern. La clase Pattern permite la definición del patrón y la clase Matcher compara el patrón definido con la cadena de cararacteres que se le pasa por parámetro.

Capítulo 4
Implementación y resultados

En este capítulo se describen los aspectos a destacar respecto a la implementación del la herramienta: el funcionamiento de ésta en general, cómo se ha realizado las extracción de datos, la obtención de las noticias y el resultado de todo este proceso.

4.1 Requisitos

Para utilizar este entorno es necesario tener instalado previamente Java (JRE o Java Runtime Environment) y el JDK (JAVA Development Kit), la librería FreeLing y las sentencias judiciales a analizar. Como último requerimiento se debe disponer de conexión a Internet para poder mostrar de forma amigable la web con el resultado final de este sistema.

4.2 Arquitectura del sistema

Se ha desarrollado una aplicación local Java para generar los ficheros JSON en los que se almacenan todos los datos relativos a cada sentencia tratada. Una aplicación web analiza y muestra los resultados registrados en

dicho fichero.

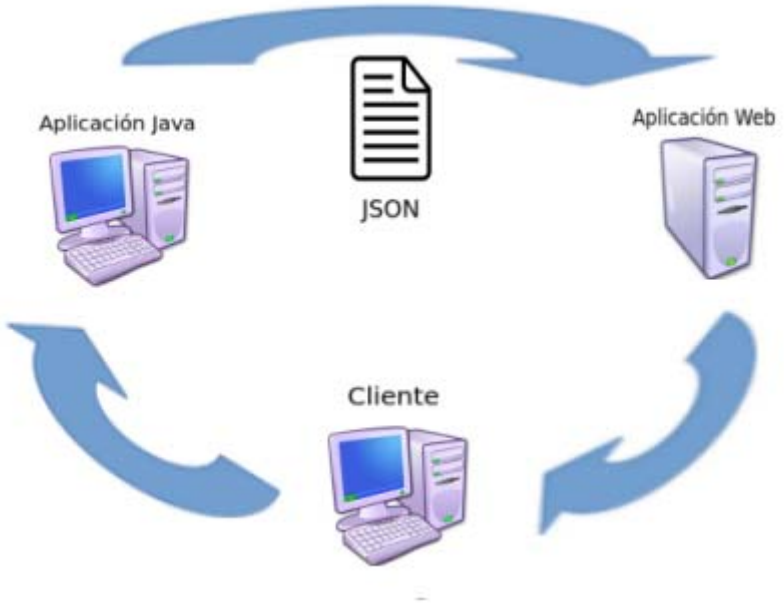


Figura 4.1 : Arquitectura de Software del sistema.

4.3 Extracción de la información

4.4 Información general

La información general como se describió anteriormente son aquellos datos principales que identifican y describen la sentencias. Estos datos, el ID Cendoj (identifica cada una de las sentencias), órgano, sede (lugar), sección, nº de recurso, nº de resolución, procedimiento, ponente y tipo de resolución son extraídos mediante expresiones regulares.

```
System.out.println("-------- Información básica ----------");

p = Pattern.compile("(Id\\sCendoj:)(\\s)(.*)");
m = p.matcher(basic);
if (m.find()) {
    setIdCendoj(m.group(3));
    System.out.println("ID Cendoj: " + getIdCendoj());
    jsonFile.put("ID Cendoj", getIdCendoj());
}

p = Pattern.compile("(Organo:)(\\s)(.*)");
m = p.matcher(basic);
if (m.find()) {
    setOrgano(m.group(3));
    System.out.println("Organo: " + getOrgano());
    jsonFile.put("Organo", getOrgano());

}

p = Pattern.compile("(Sede:)(\\s)(.*)");
m = p.matcher(basic);
if (m.find()) {
    setSede(m.group(3));
    System.out.println("Sede: " + getSede());
    jsonFile.put("Sede", getSede());
}

p = Pattern.compile("(Seccion:)(\\s)(.*)");
m = p.matcher(basic);
if (m.find()) {
    setSeccion(m.group(3));
    System.out.println("Seccion: " + getSeccion());
    jsonFile.put("Seccion", getSeccion());
}
```

Figura 4.2: Expresiones regulares de la Información General.

4.5 Figuras judiciales

Las figuras judiciales extraídas son: magistrados, procuradores, letrados y el presidente. Podemos encontrarlos en todo el texto excepto en la información principal o general (en ella sólo podemos extraer el Ponente) y en el encabezamiento o parte expositiva.

```
p = Pattern.compile("(L?l?etrado?a?s?\\s*)(dª|Dª.|D.|Sr.|sr.|don|doña|Don|Doña|Dña.)?(\\s*)?" +
        "(((Mª[.]*|[A-ZÁÉÍÓÚ][a-záéíóúñ]*)\\s*){1}(([A-ZÁÉÍÓÚ][a-zñáéíóú-]*|M[.]|(de)\\s*(la)?)\\s*){1,})" +
        "(\\s*y\\s*|,|.|;|:|\\n)");
m = p.matcher(content);
```

Figura 4.3: Expresión regular para los Letrados.

4.6 Juzgado

Después de la información general, en el encabezamiento podemos encontrar el juzgado junto con otra información como las figuras judiciales, la fecha, etc. Del juzgado de origen extraemos tres datos, el tipo, el número y el lugar.

```
p3 = Pattern.compile("(Juzgado\\s*)(de\\s*)(lo\\s*)?" +
                     "(.*)" + // Tipo Juzgado
                     "(número|nº|Nº|num.)(\\s*)" +
                     "([A-ZÁÉÍÓÚ][a-zñáéíóú]*|[0-9]*)" + // Número Juzgado
                     "(\\s*)(de\\s*)" +
                     "([A-ZÑÁÉÍÚÓ][a-záéíóú]*)(\\s*)" + // Lugar Juzgado
                     "(,|.|;|:|\\n)");
m3 = p3.matcher(content);
```

Figura 4.4:Expresión regular para el Juzgado

Se ha generado varias expresiones regulares con diferentes patrones debido a la variedad de formas que aparecen. Aspectos a destacar en este sentido y que fue necesario contemplar: los números pueden aparecer con letras (“seis” o “6”), puede aparecen puntos, guiones, y el orden no es fijo.

4.7 Fecha

La fecha en la que se dicta la sentencia es un dato importante ya que se compara con la de cada noticia relacionada para comprobar si ésta fue

publicada antes o después. Esta información la mayoría de las veces se encuentra con el mismo formato en el encabezamiento del documento. Sin embargo aunque esté en el mismo formato, ocurre los mismo que con el juzgado, el día, mes y año algunas veces vienen con números y otras con letras, contiene símbolos como puntos, guiones, etc. Es por ello que tras extraerlas se les da un formato uniforme para poder realizar la comparación.

```
p = Pattern.compile("(En\\s)?(.*)" +
        "(\\sa\\s)([a-záéíóú]*\\s*[a-záéíóú]*\\s*[a-záéíóú]*\\s*[0-9]*\\s*)" +
        "(de\\s)([A-ZÁÉÍÚÓ]*[a-záéíóúñ]*\\s)" +
        "(del?\\s)([a-záéíóú]*\\s*[a-záéíóú]*\\s*[a-záéíóú]*\\s*[0-9]*\\s*)");
m = p.matcher(content);
```

Figura 4.5: Expresión regular para la Fecha.

4.8 Delito

Los delitos se pueden encontrar en el encabezamiento, también en la parte considerativa y en la resolutiva. Para la obtención de los delitos se ha recurrido a expresiones regulares por la necesidad de reducir el tiempo de procesamiento, ya que su extracción requiere procesar todo el documento.

```
p = Pattern.compile("(malversación)");
m = p.matcher(content);

p1 = Pattern.compile("(prevaricación)");
m1 = p1.matcher(content);
```

Figura 4.6: Expresión regular para el Delito.

4.9 Fallo

La Parte Dispositiva o Fallo contiene la decisión de la sentencia

(condena o absolución). Ésta es la última parte o parte resolutiva del documento de la que sacamos las palabras más frecuentes a usar, las cuales son desestimar, absolver y condenar. Para la detección de estas palabras se utiliza la lematización.

Con la detección de entidades de la librería FreeLing podemos obtener el lema de cada una de las palabras que aparecen en la sentencia. Estos lemas son comparados con las que están contenidas en el diccionario para el fallo. Como resultado de esta comparación se obtienen las coincidencias entre las palabras del diccionario y las que aparecen en la sentencia judicial.

```
Forma: ABSOLVER- Lema: absolver
```

Figura 4.7: Resultado de la obtención del Fallo.

4.10 Extracción de las entidades

Los cargos, localizaciones y organizaciones son las entidades que se extraen de los textos. Con la clasificación de entidades de FreeLing (módulo NEC) las obtenemos clasificadas con las etiquetas *NP00O0* (Organizaciones) y *NP00L0* (Localizaciones). Para la identificación se usa el módulo NER, basado en las mayúsculas, lo que hace posible catalogar los cargos como entidades, ya que estos datos aparecen en las sentencias siempre en mayúsculas, sin embargo, son clasificados como organizaciones (con la etiqueta *NP00O0*).

Debido a esto creamos una serie reglas que, junto con los diccionarios de cargos y organizaciones, se usan para separar las organizaciones identificadas por Freeling en cada una de estas dos categorías. Es decir aquellas organizaciones y cargos públicos de las comunidades de Valencia, Andalucía y Canarias. Al igual que con la extracción del fallo, se compara las entidades obtenidas que pasan las reglas creadas con las palabras del diccionario, y así obtenemos las entidades del documento.

```
Clave: las_palmas_de_gran_canaria % Etiqueta: NP00G00
Clave: gobierno_municipal_de_telde % Etiqueta: NP00O00
Clave: ciudad_jardín % Etiqueta: NP00G00
Clave: comisión_de_ordenación_de_el_territorio_de_canarias % Etiqueta: NP00O00
Clave: canary_quality_business % Etiqueta: NP00O00
Clave: grupo_de_cooperativas_europa % Etiqueta: NP00O00
Clave: municipio_de_telde % Etiqueta: NP00G00
Clave: diputado_de_el_parlamento_de_canarias % Etiqueta: NP00O00
Clave: arquitecto_municipal_estanislao_celso % Etiqueta: NP00O00
Clave: agencia_estatal_de_la_administración_tributaria % Etiqueta: NP00O00
Clave: servicio_de_ordenación_territorial_de_el_ayuntamiento_de_telde % Etiqueta: NP00O00
Clave: grupo_europa_s.a. % Etiqueta: NP00O00
Clave: sala_de_el_tribunal_superior_de_justicia_de_canarias % Etiqueta: NP00O00
Clave: isla_de_gran_canaria % Etiqueta: NP00G00
Clave: plan_general_de_ordenación_urbana_de_telde % Etiqueta: NP00O00
Clave: consejera_delegada_de_la_empresa_municipal_de_la_vivienda % Etiqueta: NP00O00
Clave: tribunal_superior_de_justicia_de_canarias % Etiqueta: NP00O00
Clave: caja_madrid % Etiqueta: NP00O00
Clave: ayuntamiento_de_telde % Etiqueta: NP00O00
Clave: grupo_de_cooperativas_europa_sau % Etiqueta: NP00O00
Clave: marisa_informática_s.l. % Etiqueta: NP00O00
Clave: junta_de_gobierno_de_la_corporación_municipal_de_telde % Etiqueta: NP00O00
```

Figura 4.8: Resultado de la obtención de las Entidades y su clasificación.

4.11 Extracción de las noticias

Para obtener las noticias utilizamos los datos almacenados en el fichero JSON a partir de las sentencias además de la API de Google Custom Search. En concreto se usan para configurar las consultas que conducirán a las noticias

relacionadas con la sentencia.

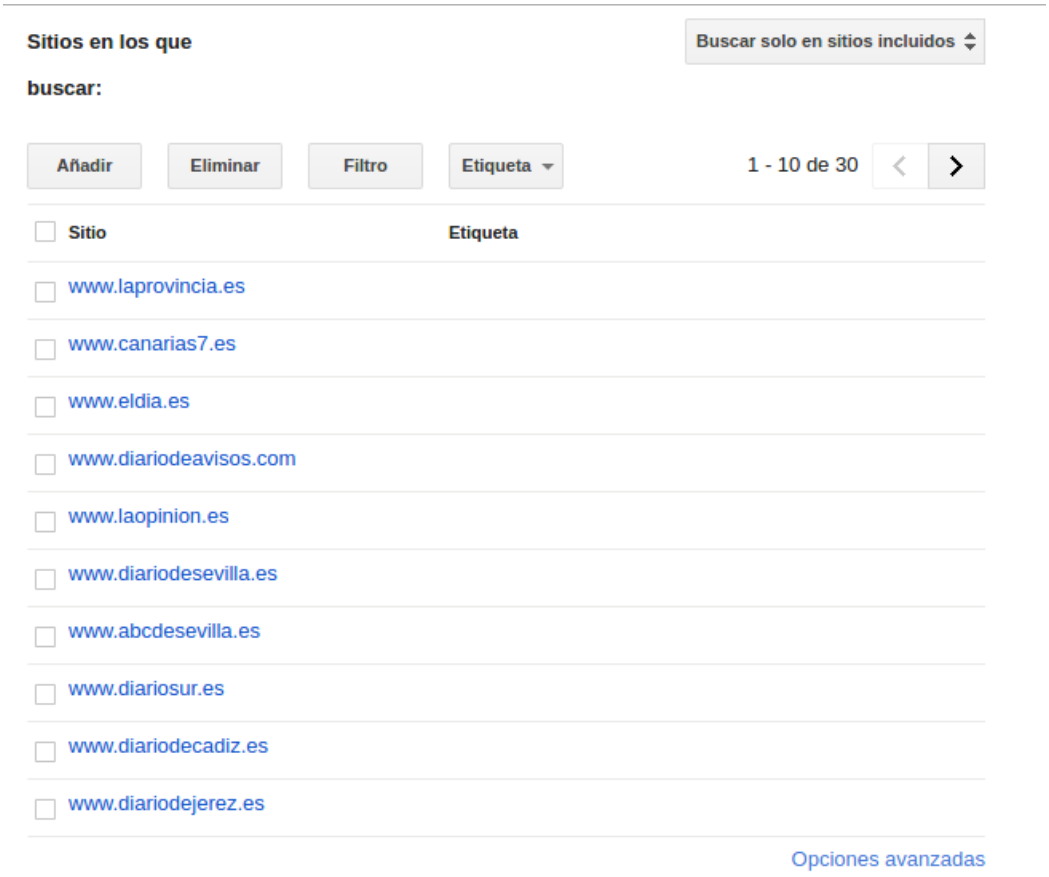


Figura 4.9: Sitios para el Motor de búsqueda de Google Custom Search.

Y por último restringir páginas mediante tipos de schema (org, net, etc.).

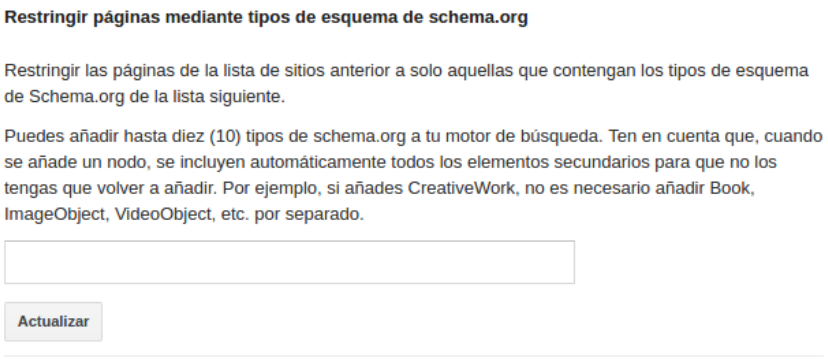
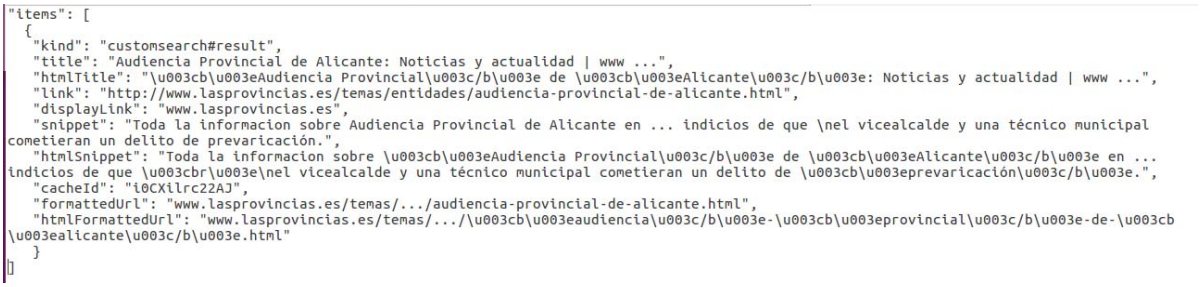


Figura 4.10: Restricción de páginas mediante schemas de GCS.

Una vez creado necesitamos el ID del motor de búsqueda, y en el Panel de Control (Console) de las APIs de Google obtenemos en credenciales la clave de la API. Con estas dos claves, la herramienta y las palabras del JSON que se utilizan como filtro en la consulta se construye la URL que se pasa al motor de búsqueda. Los parámetros en la URL son la sede, una entidad (cargo, organización, o localización) y el delito. El motivo de esta elección se basa en las pruebas que se han realizado para cuáles son con las que se obtienen las noticias más exactas.

El resultado de esto es un fichero JSON con información de la búsqueda y cada uno de los links encontrados junto a otros datos como el snippet, displaylink, title, datepublished, etc. El title es el título de la noticia, el displaylink es la url de la web principal, es decir, del periódico (www.elpais.es, www.abc.es, etc.), el snippet es la descripción de la noticia, también nos muestra la fecha de publicación o datepublishied y por último el enlace o link

a la noticia.



```
"items": [
  {
   "kind": "customsearch#result",
   "title": "Audiencia Provincial de Alicante: Noticias y actualidad | www ...",
   "htmlTitle": "\u003cb\u003eAudiencia Provincial\u003c/b\u003e de \u003cb\u003eAlicante\u003c/b\u003e: Noticias y actualidad | www ...",
   "link": "http://www.lasprovincias.es/temas/entidades/audiencia-provincial-de-alicante.html",
   "displayLink": "www.lasprovincias.es",
   "snippet": "Toda la informacion sobre Audiencia Provincial de Alicante en ... indicios de que \nel vicealcalde y una técnico municipal
cometieran un delito de prevaricación.",
   "htmlSnippet": "Toda la informacion sobre \u003cb\u003eAudiencia Provincial\u003c/b\u003e de \u003cb\u003eAlicante\u003c/b\u003e en ...
indicios de que \u003cbr\u003e\nel vicealcalde y una técnico municipal cometieran un delito de \u003cb\u003eprevaricación\u003c/b\u003e.",
   "cacheId": "i0CXilrc22AJ",
   "formattedUrl": "www.lasprovincias.es/temas/.../audiencia-provincial-de-alicante.html",
   "htmlFormattedUrl": "www.lasprovincias.es/temas/.../\u003cb\u003eaudiencia\u003c/b\u003e-\u003cb\u003eprovincial\u003c/b\u003e-de-\u003cb
\u003ealicante\u003c/b\u003e.html"
  }
]
```


Figura 4.11: JSON con resultado de GCS.

Finalmente los links y atributos de cada una de las sentencias se extraen y se guardan en el JSON con los datos extraídos de éstas .

4.12 Resultados

Para la visualización del resultado obtenido de esta herramienta se ha creando una Aplicación Web en la que se muestra de cada sentencia judicial las noticias que se han encontrado relacionadas con estas. También se puede ver las estadísticas que se ha realizado de los datos obtenidos.

Para el diseño se ha utilizado HTML5, CSS3 junto con Bootstrap (framework o conjunto de herramientas de código abierto para diseño de sitios y aplicaciones web) se han utilizado para el diseño, jQuery (biblioteca multiplataforma de JavaScript) para el desarrollo. jQuery nos provee de un método llamado getJSON() el cual se utiliza para obtener datos JSON mediante una solicitud AJAX tipo HTTP GET.

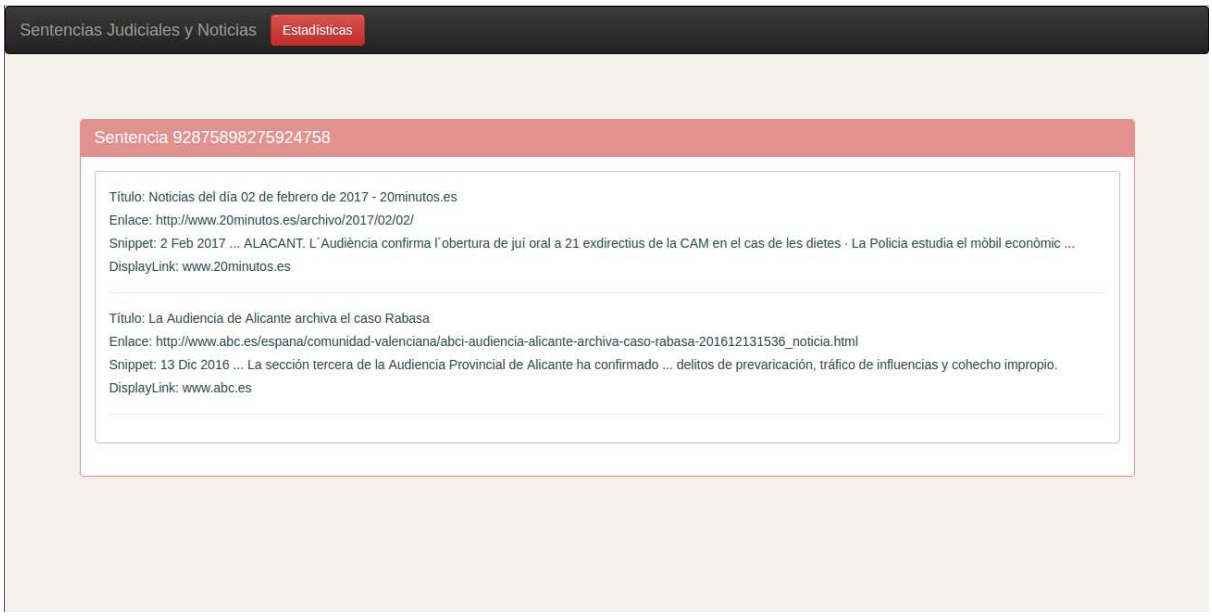


Figura 4.12: Página web para la visualización de noticias.

Capítulo 5
Conclusiones y líneas futuras

Cada vez hay mas información pública a la cual es difícil de acceder y de entender por parte de los ciudadanos. Debido a esto es muy importante la creación de herramientas que faciliten el acceso a dicha información y expliquen de manera sencilla de qué trata.

En este proyecto se ha conseguido tratar la información de sentencias judiciales para proporcionar a los usuarios mayor información mostrando las noticias de la prensa digital que hacen referencia a dichas sentencias. En su desarrollo se ha adquirido conocimientos de PLN especialmente el reconocimiento y clasificación de entidades. En este sentido, si bien se ha utilizado una librería que incorpora un módulo específico para ello, la adaptación para la mejora de los resultados al dominio en que se ha aplicado ha requerido gran parte del esfuerzo de este proyecto.

Este trabajo se ha centrado principalmente en resolver los aspectos relacionados con la transformación de la información de textual en un conjunto organizado de información que se pueda utilizar para diversos fines

posteriores. Por esta razón como trabajos futuros es claro que se abre una línea respecto al análisis de los datos registrados en el fichero JSON que se genera:

- Realizar análisis exploratorio de los datos.
- Análisis predictivos.
- Visualización de los datos obtenidos.
- Incluir sentencias judiciales de otras comunidades.

Capítulo 6
Summary and Conclusions

There is more and more public information that is difficult for citizens to access and understand. Because of this, it is very important to create tools that facilitate access to this information and explain in a simple way what it is about.

This project has managed to treat information from court judgments to provide users with more information showing the news of the digital press that refer to these sentences. In its development has acquired knowledge of PLN especially the recognition and classification of entities. In this sense, although a library has been used that incorporates a specific module for this, the adaptation for the improvement of the results to the domain in which it has been applied has required a great part of the effort of this project.

This work has mainly focused on solving aspects related to the transformation of textual information into an organized set of information that can be used for various purposes. For this reason as future work it is

clear that a line is opened regarding the analysis of the data recorded in the JSON file that is generated:

- Perform exploratory analysis of the data.
- Predictive analysis.
- Visualization of the data obtained.
- Include judicial decisions from other communities.

Capítulo 7
Presupuesto

En este capítulo se describe los tiempos de trabajo, las herramientas usadas (libres o de pago). El presupuesto se calcula en base a las horas de trabajo y del software a utilizar.

7.2 Trabajo y herramientas

Tipos	**Descripción**
Horas de trabajo	200
Licencias de herramientas	0
*API Google Custom Search	Opcional

Tabla 7.1: Resumen de tipos

7.3 Costes

Tipos	Coste
Horas de trabajo	5,00 €
Licencias de herramientas	0, 00 €
*API Google Custom Search	94,13 €/año (100 USD)

Tabla 7.2: Resumen de costes

Bibliografía

[1] Extracción y visualización de información de textos legales. Francisco Javier Rodríguez Dióniz. Trabajo de Fin de Grado de Ingeniería Informática, Escuela Superior de Ingeniería y Tecnología de la ULL, 2015 http://riull.ull.es/xmlui/handle/915/845

[2] Analizadores Multilingües en Freeling - Linguamatica – 115-454-1-PB.pdf

[3] Página de FreeLing http://nlp.lsi.upc.edu/freeling/node/1

[4] CENDOJ http://www.poderjudicial.es/search/indexAN.jsp

[5]CENDOJ http://www.poderjudicial.es/cgpj/es/Temas/Documentacion-Judicial/El-Centro-de-Documentacion-Judicial--Cendoj--/

[6] Eclipse https://es.wikipedia.org/wiki/Eclipse

[7] Java https://es.wikipedia.org/wiki/Java_(lenguaje_de_programaci%C3%B3n)

[8] Apache PDFBox https://www.adictosaltrabajo.com/tutoriales/pdfbox/

[9] Apache PDFBox https://pdfbox.apache.org/

[10] Google Custom Search https://developers.google.com/custom-search/

[11] Lematización https://es.wikipedia.org/wiki/Lematizaci%C3%B3n

[12] Expresiones regulares

https://es.wikipedia.org/wiki/Expresi%C3%B3n_regular

[13] Regex http://www.javatpoint.com/java-regex

[14] Juzgado http://definicion.de/juzgado/

[15] Diccionario del Español Jurídico (DEJ)

http://www.rae.es/obras-academicas/diccionarios/diccionario-del-espanol-juridico

[16] Fundación Civio - http://www.civio.es/

[17] Organismos Públicos

https://es.wikipedia.org/wiki/Organismo_p%C3%BAblico_de_Espa%C3%B1a

[18] Funcionarios públicos

https://es.wikipedia.org/wiki/Funcionario

[19] El Indultómetro. Fundación Civio

http://www.elindultometro.es/indultos.html

[20] ¿Quién Manda?. Fundación Civio.

http://quienmanda.es/

[21] Ixa Pipes IXA NLP Group, Universidad del País Vasco. http://ixa2.si.ehu.es/ixa-pipes/

[22] INVENTE, Inventario de Entes Públicos. Ministerio de Hacienda y Función Pública.

http://www.igae.pap.minhafp.gob.es/sitios/igae/es-ES/invente/Paginas/inicio.aspx

[23] Instituto Nacional de Estadística, INE http://www.ine.es

[24] Relación de Puestos de Trabajo del Gobierno de Canarias http://www.gobiernodecanarias.org/cpj/dgfp/index.jsp?idc=4131

[25] Relación de Puestos de Trabajo del Gobierno Vasco . http://opendata.euskadi.eus/catalogo/-/relaciones-de-puestos-de-trabajo-de-los-departamentos-y-organismos-autonomos-de-la-administracion-de-la-comunidad-autonoma/

[26] Xavier Carreras, Lluís Màrquez, and Lluís Padró. A simple named entity extractor using adaboost. *Proceedings of CoNLL-2003 Shared Task:* Edmonton, Canada, June 2003.

[27] Pablo Gamallo, Juan Carlos Pichel, Marcos García, José Manuel Abuín, Tomás Fernández Pena. Análisis morfosintáctico y clasificación de entidades nombradas en un entorno Big Data. Revista de la Sociedad Española de Procesamiento de Lenguaje Natural (SEPLN), Vol. 53, 2014