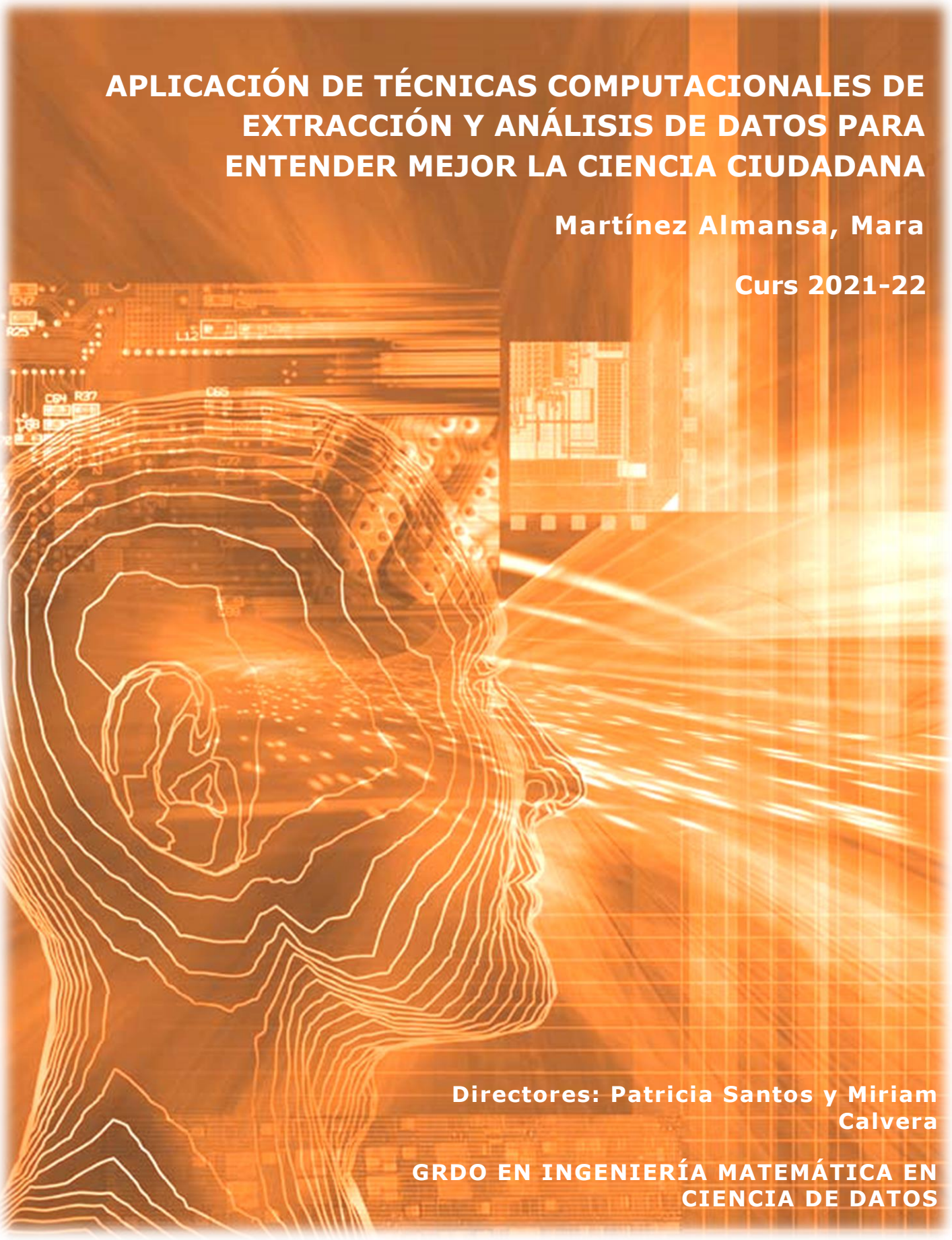
APLICACIÓN DE TÉCNICAS COMPUTACIONALES DE EXTRACCIÓN Y ANÁLISIS DE DATOS PARA ENTENDER MEJOR LA CIENCIA CIUDADANA

Martínez Almansa, Mara

Curs 2021-22



Directores: Patricia Santos y Miriam Calvera

GRDO EN INGENIERÍA MATEMÁTICA EN CIENCIA DE DATOS

upf. Universitat Pompeu Fabra Barcelona
Escola d'Enginyeria

Trabajo de fin de Grado

Agradecimientos

En primer lugar, quiero agradecer a Miriam y a Patricia su disposición y ayuda desde el primer momento de este proyecto. También agradecer a mi madre y amigos por estar siempre presentes y darme soporte en todo.

Resumen

La ciencia ciudadana es un nuevo proceso participativo que se utiliza en ciencia que permite a cualquier ciudadano participar en proyectos científicos. CS Track es un proyecto Europeo H2020 que tiene como objetivo ampliar el conocimiento sobre este modelo de participación ciudadana en ciencia.

Este trabajo tiene como objetivo principal extraer, automáticamente, información de las descripciones de proyectos de ciencia ciudadana almacenados en la base de datos de CS Track. Para ello se elaborarán diversos algoritmos donde se apliquen técnicas de extracción de la información. Posteriormente se procederá al análisis y visualización de los resultados. Gracias a este trabajo, CS Track podrá extraer y clasificar automáticamente la información almacenada, además de facilitar los futuros estudios y análisis con mayor profundidad.

Resum

La ciència ciutadana és un nou procés participatiu que s'utilitza en ciència que permet a qualsevol ciutadà participar en projectes científics. CS Track és un projecte Europeu H2020 que té com objectiu ampliar el coneixement sobre aquest model de participació ciutadana en ciència.

Aquest treball té com objectiu principal extreure, automàticament, informació a partir de les descripcions de projectes de ciència ciutadana emmagatzemats a la base de dades de CS Track. Per això es construiran diversos algoritmes on s'apliquin tècniques d'extracció de la informació. Posteriorment es procedirà a l'anàlisis i visualització dels resultats. Gràcies a aquest treball, CS Track podrà extraure i classificar automàticament la informació emmagatzemada, a més de facilitar els futurs estudis i anàlisis amb major profunditat.

Abstract

Citizen science is a new participatory process used in science that allows any citizen to participate in scientific projects. CS Track is a European H2020 project that aims to expand knowledge about this model of citizen participation in science.

The main objective of this work is to automatically extract information from the descriptions of citizen science projects stored in the CS Track database. For this, various algorithms will be developed where information extraction techniques are applied.

Subsequently, the analysis and visualization of the results will be carried out. Thanks to this work, CS Track will be able to automatically extract and classify the stored information, in addition to facilitating future studies and analysis in greater depth.

Índice

1. INTRODUCCIÓN

1.1. Contexto

"La ciencia ciudadana es el uso de métodos científicos por parte del público en general para hacer y responder preguntas y resolver problemas" [1]. Esta práctica está en auge desde hace varios años y crece con rapidez. Existen proyectos de ciencia ciudadana de todo tipo. Desde proyectos que correlacionan los hábitos de salud de una población con las enfermedades cardiovasculares e intentan preverlas (Heart Healthy Hoods[1]) hasta proyectos que analizan los movimientos de las medusas (MedusApp[2]). Lo importante en ciencia ciudadana es que la población participa activamente realizando diversos tipos de actividades (ej. recogiendo datos, tomando fotografías, colaborando con entidades científicas, …) en los proyectos. Cualquier persona es capaz de adquirir conocimientos para realizar las diferentes tareas, aprender a utilizar nuevas herramientas o consultar recursos necesarios y contribuir así al avance de la ciencia.

Actualmente existe un proyecto llamado CS Track[3] financiado por el programa de investigación e innovación Horizonte 2020 de la Unión Europea. Su objetivo principal es ampliar el conocimiento sobre ciencia ciudadana [2]. Para cumplir este gran y ambicioso objetivo, CS Track primero ha recolectado información disponible en internet sobre diversos proyectos de ciencia ciudadana que se han llevado a cabo o se están llevando a cabo en todo el mundo. Todos estos datos, se han anonimizado y almacenado en una base de datos propia.

Para llevar a cabo un estudio y análisis de la información de proyectos de ciencia ciudadana de su base de datos, es necesario automatizar el proceso de extracción y categorización de la información debido a la gran cantidad de datos recolectados. Para ello CS Track necesita una herramienta capaz de, automáticamente, extraer la información necesaria desde la descripción del proyecto y almacenarla en diferentes categorías descriptivas.

La extracción automática de información a partir de grandes volúmenes de datos es posible gracias a la ciencia de datos. La ciencia de datos es el campo que se ocupa de grandes volúmenes de datos utilizando herramientas y técnicas para encontrar patrones

[1] https://ciencia-ciudadana.es/proyecto-cc/heart-healthy-hoods/
[2] https://ciencia-ciudadana.es/proyecto-cc/medusapp/
[3] https://cstrack.eu/

invisibles, obtener información significativa y tomar decisiones [15]. Un científico de datos es capaz de recopilar y almacenar información de forma óptima, pudiendo ser esta de cualquier tipo (numérica, textual o multimedia) y capaz de ser afectada por las 5 Vs del Big Data (velocidad, variedad, volumen, valor y veracidad) [16].

La ciencia de datos incluye varias disciplinas como el aprendizaje automático, la preparación de los datos, la minería de datos, la visualización de datos, etc. Además, esta ciencia juega un papel importante en prácticamente todos los aspectos de las acciones y estrategias comerciales. Haciendo uso de los conceptos teóricos que incluyen todas estas disciplinas, se puede por ejemplo realizar un estudio sobre que clientes tienen más probabilidad de marcharse de una compañía telefónica.

Muchas de estas disciplinas deberán aplicarse a lo largo de este trabajo para efectuar con éxito el objetivo.

Este trabajo está supervisado por dos miembros del grupo de investigación TIDE-UPF (https://www.upf.edu/ca/web/tide) que pertenecen a CS Track y desde un principio se propuso automatizar el proceso de extracción de la información de diversos proyectos de ciencia ciudadana. Todos los avances conseguidos durante este trabajo irán destinados a mejorar los estudios y conclusiones que desarrolla CS Track.

1.2. Objetivos

El objetivo principal del trabajo es extraer, automáticamente, información, analizarla y categorizarla a partir de las descripciones de proyectos de ciencia ciudadana almacenados en la base de datos de CS Track. Así pues, los objetivos de este trabajo son:

1. Familiarizarse con la Ciencia Ciudadana y el tipo de proyectos en los que se organiza.
2. Identificar un listado de categorías a analizar a partir de la información contenida en la descripción de los proyectos de ciencia ciudadana en plataformas de ciencia ciudadana online.
3. Aplicar técnicas de procesado de texto, aprendizaje automático y redes neuronales para extraer información de las descripciones de proyectos de ciencia ciudadana descritos en inglés de acuerdo con los aspectos identificados previamente.

4. Analizar qué técnicas ofrecen resultados más eficientes y describir las ventajas y limitaciones de las técnicas empleadas.
5. Analizar la información extraída de los proyectos de ciencia ciudadana. Hacer un análisis global de los resultados obtenidos en relación con las categorías identificadas.
6. Alimentar la base de datos con la información extraída y categorizada.
7. Visualizar el análisis generando *dashboards*[4] interactivos para el usuario. La idea es crear una herramienta útil con la que el usuario pueda filtrar y buscar proyectos según las características que convengan para la búsqueda. Además de tener una visión general de la distinta información de cada proyecto de ciencia ciudadana.

1.3. Planificación

Se ha construido un diagrama de Gantt para planificar las tareas a realizar durante todo el trabajo.

[4] Un dashboard es una representación gráfica. Es una herramienta que nos permite visualizar los datos, es decir, transforma los datos en una información útil.

2. ESTADO DEL ARTE

2.1. Ciencia Ciudadana

2.1.1. Definición de ciencia ciudadana y su historia

La ciencia ciudadana es una metodología científica donde los participantes proporcionan datos experimentales, formulan nuevas preguntas y, junto con los investigadores, crean una nueva cultura científica. Mientras agregan valor a los proyectos de investigación, los voluntarios ganan nuevos conocimientos y habilidades [3].

Este enfoque participativo es posible gracias a los voluntarios. Según el *White Paper on Citizen Science for Europe*, "la ciencia ciudadana permite involucrar al público en actividades científicas gracias a la contribución activa de los ciudadanos en las investigaciones, con el aporte de su esfuerzo intelectual, su conocimiento general, o sus herramientas y recursos" [4].

El nombre de Ciencia Ciudadana como tal, se utilizó por primera vez en *Cornell's Laboratory of Ornithology* a principios del siglo XX [5]. Surgió como reflejo del gran número de naturalistas que observaban y recogían datos de poblaciones de diversas especies. Aún así, antes de que esta metodología tuviera nombre, muchos científicos ya hacían uso de este método para llevar a cabo sus investigaciones. Científicos famosos como Charles Darwin, Gregor Mendel y Benjamin Franklin hicieron posibles muchos de sus estudios gracias a personas que les prestaban su ayuda. De hecho, hace aproximadamente ciento cincuenta años, Charles Darwin redactó la teoría de la evolución basándose en las pruebas aportadas por cientos de personas de todo el mundo [6]. Actualmente la Biblioteca de la Universidad de Cambridge alberga cerca de mil quinientas cartas que el naturalista envió y recibió durante sus estudios [7]. Esas cartas no solo iban dirigidas a científicos y naturalistas, sino que también las enviaba a gente corriente. Hubiera sido difícil para Darwin reunir todas esas pruebas sin la ayuda de todos los que respondieron a sus cartas.

Si bien la colaboración en estudios científicos tiene su historia, el verdadero *boom* de la ciencia ciudadana ocurre a partir del siglo XXI. "*Es en la era de internet, cuando aumenta la participación ciudadana, pasando de los cientos o pocos miles de voluntarios en los proyectos naturalistas, a los cientos de miles de voluntarios de la era digital. De la misma manera creció la necesidad de divulgar y transferir los resultados de una investigación,*

para el entendimiento del ciudadano no experto o simplemente aquel que está fuera de su ámbito disciplinar." (Parelló, 2014) [19]

Gracias a las tecnologías, des de hace unos años somos capaces de ampliar nuestras capacidades y superar limitaciones. Así pues, el trabajo de recopilar información, transcribir y evaluar datos se ha transformado en una tarea mucho más amena. Ya no es necesario invertir tanto tiempo para terminar este trabajo.

Un claro ejemplo es el GPS. Este aparato es capaz de mostrarnos las coordenadas geográficas de cualquier lugar en cuestión de segundos cuando antes un experto tardaba varios minutos en encontrarlas. Ejemplos como este hay en gran cantidad. Actualmente la mayoría de la población tiene un teléfono inteligente que le permite tomar fotografías de gran calidad. Imaginemos el trabajo de los antiguos científicos cuando debían realizar un estudio a partir de documentación fotográfica, es decir, a partir de imágenes tomadas con cámaras analógicas. Y en cuanto a la búsqueda de información también poseían grandes desventajas. Ahora todos los ciudadanos tienen acceso a internet y con un solo clic pueden encontrar aquella información que necesiten. Antiguamente debían dirigirse a las universidades para obtener aquella información.

Por este motivo, podemos decir que el ciudadano de hoy en día no es el mismo que hace 50 años. La necesidad que tienen los ciudadanos de participar en todos los ámbitos de la vida pública es real. Las condiciones de hacer ciencia han cambiado y además la población se muestra más predispuesta a participar en la generación de conocimiento. Y hoy, más ilustrados y exigentes, las personas se esfuerzan por involucrarse más en los procesos de producción científica.

2.1.2. Proyectos de ciencia ciudadana

En ciencia ciudadana, todas las personas que participan en la investigación y no forman parte de su trabajo son consideradas voluntarias, por lo que científicos de otro campo, artistas, estudiantes, padres o niños pueden ofrecerse como voluntarios para participar en proyectos abiertos a la participación.

El papel de los voluntarios es de mayor importancia. De tal manera que los proyectos de ciencia ciudadana se pueden clasificar según el tipo de participación ciudadana [8]:

- Proyectos colaborativos. Son aquellos donde los participantes también analizan muestras y en ocasiones ayudan a diseñar el estudio, interpretar los datos, sacar conclusiones o difundir resultados.
- Proyectos contributivos. Son aquellos en que los participantes contribuyen en la recopilación de datos y puntualmente ayudan a analizarlos y difundir los resultados.
- Proyectos co-creados. A éstos también se les conoce como ciencia ciudadana extrema ya que los participantes colaboran en todas las etapas del proyecto, incluyendo la definición de preguntas, desarrollo de la hipótesis y discusión de resultados.

Esta clasificación fue definida por Bonney et al. en 2009. Tres años más tarde, Shirk et al desarrollaron la primera clasificación. Se definieron cinco categorías: proyectos conceptuales, proyectos de contribución, proyectos colaborativos, proyectos co-creados y proyectos independientes [9].

Siempre hay que tener en cuenta el tipo de proyecto y su objetivo ya que puede no ser adecuado o práctico para todos los fines científicos y todas las investigaciones pueden no siempre tener sentido. Es importante tener en cuenta si la ciencia ciudadana es el mejor método para esa investigación científica.

Existen proyectos de ciencia ciudadana de todo tipo. Según un análisis realizado en 2015 por un profesor de la Universidad de Sídney, existen campos populares donde se utiliza la ciencia ciudadana como método científico y de investigación [10]. La figura A.1 muestra la categorización que hizo Vladimir Strezov después de su estudio.

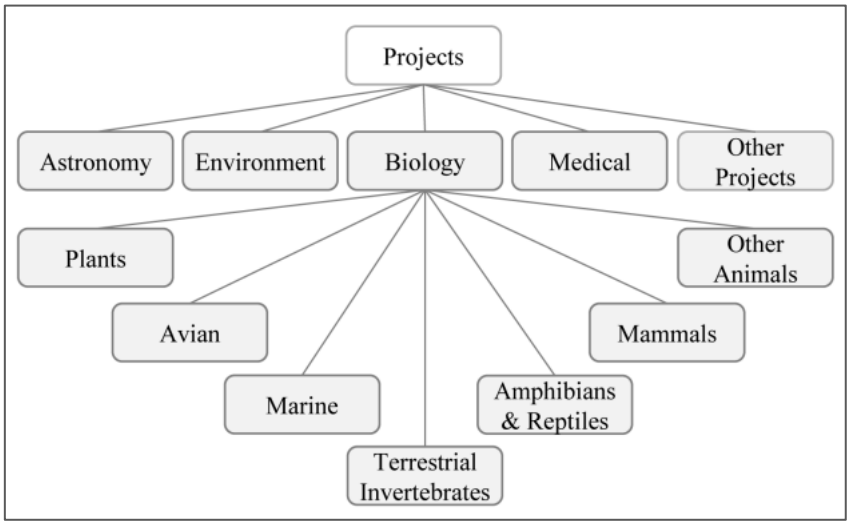


Figura A.1: Campos que abarcan los proyectos de ciencia ciudadana [10]

2.1.2.1. Proyectos recopilados por CS Track

Recordamos que el objetivo principal de CS Track es ampliar nuestro conocimiento sobre ciencia ciudadana. Por eso CS Track investiga todo tipo de proyectos de ciencia ciudadana que se han realizado o se están realizando principalmente en Europa o en línea.

Su base de datos ha sido creada extrayendo de forma automática información sobre miles de proyectos de ciencia ciudadana directamente desde páginas web que se utilizan como repositorios de estos. La tabla C.1 (ver anexo C) muestra la procedencia de todos los proyectos recopilados por CS Track. Existen muchas páginas web que albergan más de un proyecto de ciencia ciudadana como por ejemplo el caso de la *Oficina de Ciència Ciutadana de Barcelona*[5] que en su página web muestra diversos proyectos de ciencia ciudadana que se han realizado o se están realizando principalmente en Barcelona.

2.2. Extracción automática de la información

La explosión de la información vivida en la última década ha creado la demanda del procesamiento automático y el análisis de grandes volúmenes de datos en línea. En respuesta, la *Advanced Research Project Agency (ARPA)*[6] ha estado apoyando la investigación para desarrollar una nueva tecnología llamada la extracción de la información [11].

La extracción de la información (EI) es un tipo de procesamiento automático de documentos que captura y genera información contenida en un documento [18]. Similar a la recuperación de la información (IR), un sistema de extracción de información responde a la necesidad de información de un usuario. Mientras que un sistema IR identifica un subconjunto de documentos en una base de datos de texto o un subconjunto de recursos de una biblioteca, un sistema EI identifica un subconjunto de información dentro de un texto y/o documento.

Así pues, el objetivo de EI es extraer información de forma automática de un texto sin que el usuario tenga que leer el texto. Esto se puede conseguir con técnicas de aprendizaje profundo y algoritmos de procesamiento natural del lenguaje (NLP) y reconocimiento de entidades nombradas (NER). Así podemos automatizar la extracción de datos de toda la información requerida.

[5] La Oficina de Ciència Ciutadana de Barcelona es una plataforma multidisciplinar para el fomento de la ciencia ciudadana (https://www.barcelona.cat/barcelonaciencia/es/ciencia-ciudadana).
[6] https://www.darpa.mil/

Actualmente se utilizan técnicas de extracción de información en muchos ámbitos por ejemplo en el campo de la salud para transformar textos de notas médicas a bases de datos para ser consultadas de manera rápida y sencilla.

2.2.1. Natural Language Processing (NLP)

El procesamiento natural del lenguaje (NLP es el término y acrónimo en inglés) se refiere a la rama de la informática, más específicamente a la rama de la inteligencia artificial o IA, que se ocupa de proporcionar a los ordenadores la capacidad de comprender textos y palabras de la misma forma que un ser humano [12].

NLP combina la lingüística computacional (disciplina de la lingüística aplicada) con modelos estadísticos, aprendizaje automático y aprendizaje profundo. En conjunto, estas tecnologías permiten a los ordenadores procesar el lenguaje humano como texto o datos de voz y "comprender" su significado completo, junto con las intenciones y sentimientos del hablante o escritor.

El lenguaje humano está lleno de imprecisiones que hacen que sea extremadamente difícil escribir software que indique con precisión el significado previsto del texto. Existen homónimos, homófonos, sarcasmos, metáforas, etc. Estas son solo algunas de las anomalías del lenguaje humano que nosotros tardamos años en aprender, pero es exactamente lo que un programador necesita enseñar a las aplicaciones y comprender correctamente desde el principio para que éstas sean útiles.

Las diferentes tareas de NLP descomponen el texto humano de manera que ayudan a la computadora a dar sentido a lo que está leyendo o escuchando. Algunas de estas tareas incluyen lo siguiente [13]:

- El reconocimiento de voz es la tarea de convertir los datos de voz en datos de texto.
- El etiquetado del discurso, también llamado etiquetado gramatical, es el proceso de determinar la parte del discurso de una palabra o fragmento de texto en función de su uso y contexto.
- La desambiguación del sentido de las palabras es la selección del significado de una palabra con múltiples significados. A través de un proceso de análisis semántico se determina qué sentido tiene la palabra según el contexto dado.
- El reconocimiento de entidades nombradas (NER) identifica palabras o frases como entidades útiles.

- El análisis del sentimiento intenta extraer cualidades subjetivas (actitudes, emociones, confusión, etc.) del texto.
- La generación del lenguaje natural a veces se describe como lo opuesto al reconocimiento de voz; es la tarea de poner información estructurada en lenguaje humano.

La mejor herramienta para trabajar NLP en Python es la librería *Natural Language Toolkit* (NLTK)[7]. El lenguaje de programación Python proporciona una amplia gama de herramientas y librerías para atacar tareas específicas de NLP. Muchos de estos instrumentos se encuentran en NLTK, una librería *open-source*[8] que permite crear programas NLP.

NLTK incluye librerías para muchas de las tareas de NLP enumeradas anteriormente además de bibliotecas para subtareas, como análisis de oraciones, segmentación de palabras, derivación, lematización[9] y tokenización[10]. También incluye otras librerías para implementar capacidades como el razonamiento semántico, la capacidad de llegar a conclusiones lógicas basadas en hechos extraídos del texto.

2.2.1.1. Named Entity Recognition (NER)

El reconocimiento de entidades nombradas (NER) – a veces denominado fragmentación, extracción o identificación de entidades – se refiere a la tarea de identificar y categorizar información clave (entidades) en un texto [14]. NER no solo sirve como una herramienta independiente para la EI, sino que también realiza un papel importante en diferentes aplicaciones NLP, como la comprensión de textos, la recuperación de la información, el resumen automático de textos, la respuesta a preguntas, la traducción automática, etc.

Las entidades identificadas pueden ser cualquier palabra o serie de palabras que se refieran a la misma cosa. Cada entidad detectada se clasifica en una categoría predeterminada. Ejemplos de entidades nombradas son persona, organización, localización o fecha.

[7] https://www.nltk.org/

[8] Un software *open-source* es un código diseñado de manera que sea accesible al público: todos pueden ver, modificar y distribuir el código de la forma que consideren conveniente.

[9] La lematización es un conjunto de métodos para recortar palabras hasta llegar a su raíz.

[10] La tokenización se utiliza para dividir frases, oraciones y párrafos en tokens que ayudan a la computadora a comprender mejor el texto.

Formalmente, dada una secuencia de tokens[11] $w = [w_1, w_2, \ldots, w_N]$, NER es capaz de generar una lista de tuplas $[I_s, I_e, t]$, cada uno de los cuales es una entidad nombrada mencionada en w. Aquí, I_s e I_e son los índices inicial y final de una mención de entidad nombrada. La figura A.2 muestra un ejemplo en el que un sistema NER reconoce tres entidades de la oración dada.

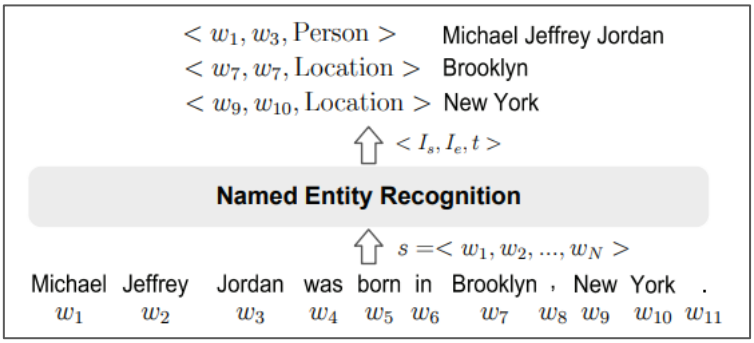


Figura A.2: Ilustración de la misión de NER [14]

Es bastante sencillo aplicar técnicas NER. Existen un gran número de excelentes librerías *open-source* que pueden ayudarnos con esta tarea. Una de las más conocidas es SpaCy[12]. Esta librería fue especialmente diseñada para construir sistemas EI.

Exactamente SpaCy es capaz de procesar textos de forma automática. Un texto procesado se define como un objeto Doc. Estos objetos tienen un atributo llamado *ents* que retornan las entidades nombradas del texto inicial. Para ver un listado de las entidades que reconoce SpaCy ver Anexo D. Además, SpaCy también proporciona una práctica librería de visualización llamada *displacy* para ver las entidades nombradas en un texto de forma más visual.

2.2.1.2 Similitud semántica

La similitud semántica es una de las muchas tareas de NLP. El objetivo de ésta es identificar cuando el sentido de dos textos o palabras es similar [17]. La similitud semántica comporta un gran reto para el procesamiento del lenguaje ya que se puede aplicar en diferentes faenas de NLP, tales como reconstrucción automática de textos, recuperación de la información, comparación y análisis de textos y muchas otras, que necesitan medir el grado de similitud entre dos textos dados.

[11] Un token es un solo elemento, en este caso cada palabra es un token
[12] https://spacy.io/

Existen varios métodos para calcular la similitud entre textos. Pero en general, podemos dividirlos en dos categorías: una se basa en cálculos estadísticos y la otra en cálculos de comprensión semántica. La principal diferencia entre estas dos categorías es que el método basado en cálculos estadísticos no tiene en cuenta la información de la estructura de la oración a la hora de hacer los cálculos.

De todos modos, uno de los métodos más populares es el modelo de espacio de vectores (VSM). En este método, cada palabra del texto constituye un vector. La similitud entre dos vectores puede calcularse mediante diferentes algoritmos como por ejemplo el coseno.

$$cos(\vec{q}, \vec{d}) = \vec{q} \cdot \vec{d} = \sum_{term\ i} q_i \cdot d_i$$

La librería SpaCy contiene un método para calcular la similitud entre dos palabras. Este método utiliza la fórmula anterior.

3. DISEÑO E IMPLEMENTACIÓN

En este apartado se explica con profundidad como se ha llevado a cabo la extracción de la información. Primero se introduce la base de conocimiento de la cual extraemos la información, en este caso la base de datos de CS Track. Segundo, se hace un análisis manual, necesario para pasar a la siguiente etapa, la aplicación de diversas técnicas de procesado y análisis de textos. En este segundo apartado se explican detalladamente los algoritmos utilizados.

Cabe decir que de primeras se aplicaron las técnicas de extracción definidas con anterioridad a textos escritos en inglés para luego, y ya, por último, aplicar técnicas de traducción automática para hacer posible esta extracción de información en varios idiomas.

3.1. Base de datos de CS Track

CS Track cuenta con una base de datos en MongoDB[13] compuesta por información sobre miles de proyectos (actualmente 4.737) de ciencia ciudadana extraídos de diferentes páginas web. Para este trabajo se ha consultado un archivo *.json* [14] extraído de la base de datos que almacena todos los datos a analizar.

Para entender la forma como están distribuidos los datos, se hizo un pequeño análisis de la base de datos. Este análisis se ha realizado con Python. El script (*BD_CSTrack.ipynb*) está en un repositorio en GitHub y el enlace directo se encuentra en el Anexo B.

Esta base de datos cuenta con 4737 proyectos de ciencia ciudadana recopilados desde marzo de 2020 y 26 atributos diferentes en los que se ha clasificado la información extraída. Estas veintiséis columnas describen diferentes características de los proyectos de ciencia ciudadana.

Todos los proyectos de ciencia ciudadana que contiene la base de datos han sido extraídos directamente de páginas web tras comprobar que estas permitían la extracción automática por medio de robots. En base a la estructura de los datos de estas webs, se han

[13] MongoDB es un sistema de base de datos NoSQL. https://www.mongodb.com/

[14] La extensión .json corresponde a un archivo que almacena estructuras de datos y objetos simples en formato JavaScript Object Notation (JSON), que es un formato estándar de intercambio de datos. Se utiliza principalmente para transmitir datos entre una aplicación web y un servidor.

categorizado los datos. Así pues, podemos clasificar los proyectos en estructurado, semiestructurado y no estructurado:

- Datos no estructurados. Son aquellos en los que la información no se encuentra clasificada por categorías.
- Datos semiestructurados. Son aquellos que cuentan con un nivel medio de estructuración y rigidez organizativa. Es decir, la información está parcialmente clasificada en categorías.
- Datos estructurados. Son aquellos en los que la información se encuentra clasificada en categorías.

Las figuras A.3.1, A.3.2 y A.3.3 son un ejemplo de cómo son las plataformas a partir de las cuales CS Track ha construido su base de datos (ver Anexo A). Las tres figuras muestran cómo está estructurada la información: datos estructurados, semiestructurados y no estructurados respectivamente.

Se han categorizado los registros de la base de datos de CS Track según los tipos de estructuras de datos. No se han podido categorizar todos los registros, aun así, la base de datos consta con 983 registros en los que sí que está definido este campo. La tabla B.1 muestra la frecuencia de los tipos de datos sí definidos. Se puede apreciar que la mayoría de los proyectos pertenecen a datos semiestructurados.

Tipo de dato	Frecuencia
No estructurado	13.53%
Semiestructurado	73.75%
Estructurado	12.72%

Tabla B.1: Porcentaje de los tipos de datos

3.2. Análisis manual

El análisis manual previo a la aplicación de algoritmos para la extracción de la información es totalmente necesario para seleccionar, con criterio, aquellos aspectos comunes e importantes de los proyectos de ciencia ciudadana.

El proceso descrito a continuación queda resumido en la siguiente figura:

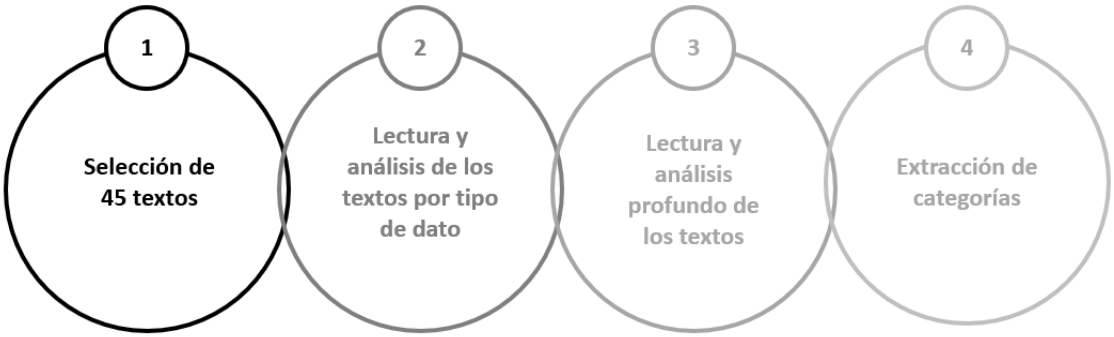


Figura 6: Flujo del proceso seguido para la selección de categorías comunes de los proyectos de ciencia ciudadana

Primero se debe definir qué información se quiere extraer para posteriormente poder aplicar diversas técnicas de extracción de la información. Para ello, se ha decidido hacer una selección aleatoria de 45 descripciones de textos en inglés de proyectos de ciencia ciudadana para su posterior análisis manual.

A partir de la base de datos de CS Track y Python[15] se creó un script que seleccionaba aleatoriamente un total de 45 textos en inglés tal y como muestran las figuras A.3.4 y A.3.5 (visitar el Anexo A para verlas figuras y el Anexo B para ver la localización del script). De estos 45 textos, 15 correspondían a datos no estructurados, 15 a datos estructurados y los 15 restantes correspondían a datos semiestructurados. Esta selección se hizo para posteriormente analizar las diferencias y similitudes entre los diferentes tipos de datos que contiene la base de datos de CS Track.

Una vez seleccionadas estas 45 muestras, se procede a guardar los textos en un archivo en formato *.txt* [16] para continuar con su lectura y análisis manual. El análisis se basa en la detección de características comunes entre textos. Se buscará información sobre la geolocalización, fechas, de que trata el proyecto de ciencia ciudadana, quién puede participar, etc.

Primero realicé un análisis por separado de los tres grupos de textos (según el tipo de datos) para posteriormente establecer semejanzas entre aspectos identificados en los diferentes grupos.

[15] Python es un lenguaje de programación interpretado de alto nivel que se utiliza para desarrollar aplicaciones de todo tipo.

[16] La extensión .txt corresponde a un archivo de texto simple o texto sin formato. Éste es un archivo informático que contiene únicamente texto formado solo por caracteres que son legibles por humanos, careciendo de cualquier tipo de formato topográfico.

Después de una primera lectura de varios textos se ha podido apreciar una clara diferencia entre los tres grupos de tipo de datos. Además, las descripciones que pertenecen a datos sin definir son mucho más pobres que las demás ya que los textos pueden llegar a ser de una sola frase y como consecuencia, no contienen demasiada información, es decir, son descripciones bastante vacías de contenido. Esto seguramente repercuta a la hora de extraer la información de este tipo de textos ya que será muy escasa y no se podrá recopilar toda la información deseada. Las descripciones de datos semiestructurados son bastante completas. Los textos son de una longitud considerable y además contienen bastante información. Las descripciones de datos estructurados también son completas, pero a diferencia de las anteriores éstas contienen ruido, al igual que las de datos no estructurados. Estos textos incluyen muchos carácteres que dificultan un poco la lectura y dificultarán más aún el trabajo de extracción de la información. En la figura A.3.6 (ver Anexo A) se pueden apreciar estos caracteres que dificultan la lectura de los textos mencionados como ruido.

Este pequeño análisis y resultados se han comunicado a dos miembros de CS Track para una posterior mejora del algoritmo de extracción de la información de las diferentes plataformas.

Para seguir trabajando en esta primera prueba de concepto se seleccionaron solo los textos en inglés. Estos suman un total de 2637 proyectos de ciencia ciudadana.

3.2.1. Extracción de categorías a partir del análisis manual

Después de un análisis más profundo de los 45 textos en inglés, se han identificado las categorías que son interesantes a extraer:

- **Objetivo del proyecto**. Finalidad del proyecto. Es un aspecto que lo encontramos en la mayoría de los textos, tanto en textos extraídos de datos no estructurados, como estructurados como de semiestructurados.
- **Promotor o quién lo lleva a cabo**. Quién financia el proyecto de ciencia ciudadana o qué asociación u organización lo lleva a cabo.
- **Localización geográfica**. Cualquier información geolocalizada que pueda utilizarse para entender dónde se lleva a cabo el proyecto.
- **Fecha de inicio**. Cuando empieza el proyecto
- **Fecha final**. Cuando acaba el proyecto

- **Duración**. Duración temporal del proyecto. Si la fecha de inicio y final del proyecto están definidas, implícitamente se define la duración de éste. Además, existen textos en los cuales no se definen estas fechas, pero sí que se define la duración del proyecto.
- **Estado**. El estado del proyecto, si el proyecto ha finalizado o sigue en proceso.
- **Participantes.** Este aspecto es un campo categórico[17]. No en todos los proyectos puede participar cualquier persona. Se definirá una lista para poder clasificar los proyectos de ciencia ciudadana.

No todas estas categorías se encuentran en todos los textos. Como se ha explicado anteriormente de la mayoría de los textos no sabemos el tipo de dato de la web a la que pertenecen. Las descripciones de estos proyectos son bastante escuetas, pero no se puede olvidar aquellos textos que sí que contienen toda esta información. Para cumplir el objetivo principal del trabajo no solo se tienen que considerar dos características comunes que sí que se encuentran en todos los textos, se debe extraer de forma automática la mayor cantidad de información posible.

3.3. Extracción automática de la información

Este punto del desarrollo está dividido en ocho bloques que corresponden con las categorías descritas en el apartado anterior. En cada uno de ellos se explican al detalle las técnicas aplicadas sobre las descripciones de los proyectos de ciencia ciudadana con el objetivo de extraer una categoría determinada.
Los algoritmos construidos están definidos en el script de Python *InformationExtraction.ipynb*. El enlace hacia el script se encuentra en el Anexo B.

3.3.1. Preprocesado

Antes de empezar a hacer pruebas sobre varios textos se preparó el entorno. Primero se definieron dos funciones:

- La primera función extrae un registro aleatorio de la base de datos. Para optimizar la visualización y el trabajo se seleccionaron solo las categorías más importantes,

[17] Una variable categórica es aquella que permite clasificar una serie de datos por medio de valores fijos asociados a una cualidad o categoría concreta.

es decir, la propia descripción del proyecto, su título, el tipo de dato de la página web de la que se ha extraído y el país al que pertenece esta página web. La figura A.3.7 muestra esta función (ver Anexo A).

- La segunda función tiene la misión de limpiar el texto. A medida que se ha ido avanzando en el trabajo se ha ido modificando esta función para obtener los mejores resultados. Básicamente la función *clean()* que se muestra en la figura A.3.8 (ver Anexo A) elimina posibles caracteres que dificulten el posterior análisis del texto.
- Una vez definidas estas dos funciones se seleccionó un proyecto de ciencia ciudadana escrito en inglés de forma aleatoria y se limpió el texto de la descripción de éste. Con el texto limpio en formato *string* pasamos a trabajar con algoritmos y librerías de Python para extraer la información necesaria para completar las ocho categorías mencionadas antes.

3.3.2. Objetivo del proyecto

El objetivo del proyecto es la finalidad del proyecto, así como que iniciativa se promueve. Tras el análisis manual, podemos concluir que esta información se encuentra en la mayoría de las descripciones de proyectos de ciencia ciudadana que se encuentran en la base de datos de CS Track. El fin de un proyecto de ciencia ciudadana está descrito mayoritariamente en una oración. Por esta razón, el resultado que se espera es la obtención de una sola frase que describa el objetivo de dicho proyecto siempre que sea posible.

Después del previo análisis manual se identificaron palabras clave que aparecen en la mayoría de las oraciones que contienen el objetivo. El primer listado de palabras identificado fue el siguiente: *'goal'*, *'purpose'*, *'objective'*, *'intention'*, *'ambition'*, *'promote'*, *'work'*. Las palabras de la lista anterior podían identificarse en varias de las oraciones que describen el objetivo de los proyectos de ciencia ciudadana.

A partir de este listado de palabras se procede a extraer aquellas oraciones que contienen dichas palabras. Este es el método más obvio y seguro para extraer automáticamente el objetivo de un proyecto de ciencia ciudadana. Pero aquí surge el primer problema. Aplicando este método se extraían oraciones que contenían la palabra *"work"*, pero éstas no correspondían con aquella parte del texto que se debía seleccionar. Por este motivo el listado de palabras inicial se modificó.

La siguiente propuesta de palabras clave fue la siguiente: *"goal", "purpose", "objective", "intention", "ambition", "promote", "dedicate"*. Este conjunto de palabras minimiza el error ya que todas ellas se ajustan a la hora de describir el objetivo del proyecto.

El siguiente paso fue ampliar esta lista de palabras de forma automática. Con solo siete palabras clave las frases extraídas eran muy escasas. Así que se decidió añadir los sinónimos de estas siete palabras para ampliar el listado.

La librería *nltk* incluye la función *synsets()*. Esta función es capaz de devolver una lista de sinónimos de la palabra que le indiques. Aplicando esta función a las siete palabras del listado de palabras clave se consiguió ampliar la lista hasta 66 palabras diferentes. Ahora la probabilidad de extraer la frase que contenía el objetivo del proyecto de ciencia ciudadana era mayor.

Pero también se tuvo en cuenta el aumento de la probabilidad de error. Cuantas más palabras, mayor es la posibilidad de extraer una frase equivocada. Por ejemplo, la librería *nltk* retorna la palabra "*use*" como sinónima de "*purpose*". El verbo usar es muy común en nuestro vocabulario y se puede utilizar en muchos contextos. Por este motivo se decide eliminar la palabra "*use*" de la lista. La figura A.3.9 (ver Anexo A) muestra una descripción de un proyecto de ciencia ciudadana extraído de la base de datos. El objetivo que extrae el algoritmo hasta ahora es el siguiente: *"You can use the same user account in the CrowdWater game and the CrowdWater app"*. La frase extraída no se corresponde con el objetivo del proyecto, pero el algoritmo lo identifica como tal porque simplemente contiene la palabra "*use*" que es sinónima de "*purpose*".

Eliminando esta palabra del listado se vuelve a minimizar el error. Pero el algoritmo todavía era muy exigente ya que solo se extraían aquellas frases que contenían dichas palabras.

El último paso fue añadir el concepto de similitud. La similitud entre las palabras del texto y el listado de palabras solo se calcula si el algoritmo no extrae ninguna frase que contenga las seis palabras clave o bien sus sinónimos.

Calcular la similitud entre dos palabras resulta extremadamente fácil con la librería SpaCy y su función *similarity()*. La similitud es un número de 0 a 1 por este motivo se definió un umbral. Si la similitud entre dos palabras supera ese umbral, se considera que son similares y por lo tanto se extrae la frase en la que se encuentra esa palabra similar. Es decir, si cualquier palabra del texto es similar a una de las palabras de la lista ampliada (palabras clave más sus sinónimos) se extrae la frase que contiene dicha palabra. De esta

forma lo que se hace es extraer aquellas oraciones que contengan una palabra de significado parecido al de las palabras clave.
El primer umbral definido fue 0.8 pero después de varias pruebas se pudo observar que esta es una similitud muy exigente con la que no se conseguían mejoras del algoritmo. Así que se volvió a definir este número en 0.7. Con el nuevo umbral se ha conseguido extraer aquellas oraciones que de primeras no se identificaban como objetivo del proyecto.
La figura A.3.10 (ver Anexo A) muestra un ejemplo de extracción automática del objetivo de un proyecto aleatorio de ciencia ciudadana de la base de datos de CS Track. El objetivo del proyecto se extrae de forma correcta.
Finalmente se deben tener en cuenta aquellas ocasiones donde el algoritmo no encuentra un objetivo a extraer. Delante de estas situaciones el algoritmo devolverá *None*. De esta manera, cuando se obtenga *None,* querrá decir que el objetivo del proyecto de ciencia ciudadana no se explica en ese texto.

3.3.3. Localización

La localización geográfica se refiere a donde se realiza o se ha realizado el proyecto de ciencia ciudadana. Esta categoría permitirá geolocalizar todos los proyectos de ciencia ciudadana almacenados en la base de datos de CS Track.
Dentro de la descripción de los proyectos de ciencia ciudadana no encontramos un criterio único que determine la geolocalización de los proyectos. Es decir, en algunos proyectos se nombra el país en el que se realiza dicho proyecto, otros se nombra la ciudad solo o la ciudad y el país, pero también existen proyectos que se pueden realizar de forma online en todo un continente o incluso en todo el mundo. Finalmente, también se contempla la opción de que no se mencione ningún tipo de localización. Por este motivo, existe la posibilidad de que el resultado a la hora de extraer la geolocalización sea nulo.
Para extraer esta categoría se ha utilizado la librería SpaCy en Python para así aplicar la técnica de identificación de entidades nombradas (NER). Después de procesar un texto a través de la función nlp() de esta librería, se obtienen las entidades nombradas de este texto. En este caso se debían extraer aquellas entidades etiquetadas como ‘GPE’ y ‘LOC’. La etiqueta ‘GPE’ (Geopolitical Entities) señala todas las ciudades, países y estados que puedan nombrarse en un texto. La etiqueta ‘LOC’ extrae todas aquellas localizaciones no geopolíticas como por ejemplo montañas, ríos, lagos, etc. Un posible ejemplo de entidad

nombrada etiquetada como GPE podría ser Barcelona y un ejemplo de entidad etiquetada con ‘LOC’ podría ser Norte América.

De esta manera resulta bastante fácil extraer la localización del proyecto de ciencia ciudadana, así que el algoritmo implementado es bastante sencillo. Simplemente se necesita localizar en el texto aquellas entidades nombradas etiquetadas como ‘GPE’ y ‘LOC’.

Aunque ya se ha comentado que en varias descripciones de proyectos de ciencia ciudadana no se nombra la ciudad, existen otros en los que se nombra tanto el país como la ciudad o la región. En estos casos, en los que aparece más de una geolocalización, el resultado que obtenemos después de aplicar el algoritmo a un texto es uno solo. Es decir, se hace un recuento de las veces que aparecen las diferentes localizaciones y se extrae la más nombrada en el texto. De este modo, si en la descripción de un proyecto de ciencia ciudadana aparece la entidad New York cinco veces y la entidad North America solo dos veces, el resultado de la extracción de la geolocalización es New York. La figura A.3.11 (ver Anexo A) muestra una descripción de un proyecto de ciencia ciudadana de la base de CS Track y el resultado obtenido después de aplicar el algoritmo de extracción de la geolocalización.

En el caso de que en las descripciones de proyectos de ciencia ciudadana aparezcan varias localizaciones, se ha extraído la primera identificada. En la figura A.3.12 (ver Anexo A) podemos ver un claro ejemplo de este caso.

Finalmente, también se debe tener en cuenta aquellos textos en los que no aparece ninguna geolocalización. Cuando el algoritmo no encuentre ninguna etiqueta nombrada con ‘GPE’ o ‘LOC’, éste devolverá *None*. Obtener *None* querrá decir que no se determinada la geolocalización en ese texto.

3.3.4. Promotor

El promotor se refiere a quién organiza o quién lleva a cabo el proyecto de ciencia ciudadana. Normalmente se corresponde con asociaciones u organizaciones. Esta categoría permitirá filtrar proyectos de ciencia ciudadana de la base de datos de CS Track según la organización que lo promueve o realiza.

Para identificar estas organizaciones se han utilizado los mismos métodos que en el apartado anterior con la única diferencia que a la hora de extraer las entidades nombradas, solo se seleccionan las etiquetadas con el tag ‘ORG’.

Una vez identificadas todas las organizaciones se descartan aquellas que coinciden con el nombre del propio proyecto de ciencia ciudadana. Este nombre, al ser un nombre propio, el algoritmo lo detecta como una organización.
Finalmente, el resultado que obtenemos es la organización con más frecuencia, es decir la que aparece más veces en el texto. Aunque si no se encuentra ninguna organización, el algoritmo devolverá *None*.

3.3.5. Fecha de inicio y final y duración

La fecha de inicio y final del proyecto de ciencia ciudadana indica cuándo empezó y terminó (si procede) dicho proyecto. De esta manera, la duración de cada proyecto se calcula a partir de las fechas de inicio y final.
Para identificar esta categoría también se han utilizado los mismos métodos que en el apartado 3.3.3., haciendo uso de la librería SpaCy. Así pues, las entidades nombradas que se debían seleccionar eran aquellas etiquetadas con el tag ‘TIME’ y ‘DATE’.
Una vez identificadas todas aquellas entidades que hacen referencia a fechas, se debía seleccionar aquellas que realmente corresponden con las fechas de inicio y final de los proyectos de ciencia ciudadana.
Esta tarea resultó más costosa de lo planificado ya que actualmente hay muy pocas descripciones de proyectos de ciencia ciudadana en la base de datos de CS Track que contengan explícitamente las fechas de inicio y final de cada proyecto.
Pero el principal problema se encuentra en identificar de forma automática la fecha exacta de inicio y final de proyecto. Ya que las descripciones de proyectos sí que contienen bastantes fechas, es decir, que sí que se identifican entidades etiquetadas con el tag ‘DATE’, pero no se ha podido extraer automáticamente a partir de un algoritmo aquellas fechas que se corresponden con el inicio y final (si esta existe) del proyecto.
Por consecuencia, la duración de los proyectos de ciencia ciudadana tampoco se ha podido calcular al no contar con las fechas de inicio y final.

3.3.6. Estado

El estado del proyecto se refiere a si dicho proyecto de ciencia ciudadana ha finalizado o aún está en curso. Por lo tanto, esta categoría sólo puede tomar dos valores: terminado o vigente.
Existen dos formas de identificar el estado de cada proyecto de ciencia ciudadana:

- A partir de las fechas de inicio y final del proyecto. Si estas fechas están definidas y la fecha que se corresponde con el final del proyecto ya ha pasado, quiere decir que el proyecto ya ha terminado. De forma contraria, si la fecha final es una fecha futura, quiere decir que el proyecto está en vigor.
- Si lo especifica explícitamente en la descripción del proyecto. Al analizar el texto se puede encontrar una oración o palabra dónde quede muy claro el estado del proyecto. Normalmente, se hallan frases que explícitamente indican que el proyecto ya ha finalizado.

Extraer el estado de cada proyecto de ciencia ciudadana hubiera resultado fácil si se hubieran extraído las fechas de inicio y final. Así pues, la única forma de identificar el estado es a partir de un análisis explícito de las descripciones de los proyectos. Pero esta segunda forma de extraer el estado tampoco ha tenido éxito.

Muy pocas de las descripciones de ciencia ciudadana almacenadas en la base de datos de CS Track contienen explícitamente una oración o palabra que haga referencia a su estado, es decir si ha terminado o no. Por este motivo tampoco se ha podido construir un algoritmo que extraiga de forma automática el estado de cada proyecto.

3.3.7. Participantes

La categoría Participantes hace referencia al tipo de personas que pueden participar en un proyecto de ciencia ciudadana. Después del primer análisis manual, se pudo detectar que los proyectos de ciencia ciudadana pedían la participación de un perfil concreto de personas. Existen proyectos que están destinados a determinados grupos de población. Así pues, los Participantes es una variable categórica que solo puede tomar ciertos valores.

Para decidir estos valores se realizó otro análisis de cuarenta y cinco descripciones de proyectos de ciencia ciudadana de la base de datos de CS Track. El análisis se centró en la detección de diferentes grupos de población para así poder categorizar estos participantes.

Después de este estudio se detectaron siete categorías posibles: *anyone, adults, students, university students, kids, area community* y *not specified*.

El algoritmo diseñado identifica como participantes de proyectos de ciencia ciudadana aquellas palabras que aparecen en la lista anterior. Es decir, si una de las siete palabras aparece en un texto, el algoritmo la extrae de forma automática. Pero además, después de

realizar varios testeos, se añadieron *18 years* (siendo lo mismo que *adults)* y *group of* a la lista anterior. Con el conjunto *group of* se extraen aquellos grupos de personas más específicos como por ejemplo un grupo de excursionistas o grupo de propietarios.

El principal problema del algoritmo está en identificar cuando el texto se refiere a que puede participar cualquier persona del mundo en ese proyecto, cualquier persona que se encuentre físicamente en la zona donde se realiza el proyecto o simplemente no se especifica esta información. Para ajustar el algoritmo a los datos, se realizó otro análisis. En este caso se tomaron 100 muestras de descripciones de proyectos de ciencia ciudadana en los que el algoritmo no diferenciaba entre *Anyone* y *Not specified*, pero haciendo una lectura sí que se distinguían las dos categorías.

De estas 100 muestras, un 43% de las descripciones de proyectos se referían a *Anyone*, es decir, que cualquier persona podía participar. Un 37% de las descripciones corresponden a *Area Community*, donde solo pueden participar aquellas personas que pasen por la zona donde se realiza el proyecto de ciencia ciudadana. Y finalmente, un 20% de las descripciones se corresponden con la categoría *Not specified*. Por mayoría, se decidió añadir al algoritmo que cuando no se detectara ninguna de las palabras de la lista, se extrajera *Anyone* como participante de ese proyecto.

4. RESULTADOS

Una vez elaborados todos los algoritmos de extracción automática de la información, se procede a evaluar y analizar los resultados obtenidos. Esta evaluación y análisis se realiza sobre la muestra de 2637 descripciones de proyectos de ciencia ciudadana en inglés, dejando a un lado los 2100 registros que completan la base de datos. Aunque para calcular el error de los algoritmos se ha extraído una muestra menor, de 150 textos.

El error de los algoritmos se refiere a la diferencia entre el resultado verdadero (el que obtendríamos leyendo cada una de las descripciones) y el resultado obtenido a partir del algoritmo. Éste se ha calculado de forma manual. Es decir, leyendo cada uno de los ciento cincuenta textos y valorando si la información extraída por los algoritmos era la correcta o no. La tabla 4.1 muestra el error obtenido de cada una de las categorías de las que se ha elaborado el algoritmo de extracción automática.

Categoría	Error (%)
Objetivo	20%
Geolocalización	16,67%
Organización	18,67%
Participantes	26,67%

Tabla 4.1: Porcentaje de error de los algoritmos

4.1 Análisis

Una vez cuantificado el error, se ha realizado un análisis de éste para determinar y evaluar sus causas. Para ello es conveniente analizar las cuatro categorías por separado. Aunque, de todas maneras, se ha realizado también un análisis general englobando los cuatro algoritmos:

- Como ya se ha explicado en el punto 3.2., existen descripciones de proyectos de ciencia ciudadana bastante escuetas. De forma general, se ha observado que cuando esta descripción es de una sola frase, no suele contener mucha información. En este tipo de situación, los algoritmos devuelven *None* (los algoritmos de extracción del objetivo, geolocalización y organización) o *Anyone* (en el caso de la extracción de participantes). La mayoría de los resultados obtenidos en este caso son correctos ya que la única frase que contiene el texto no suele definir ninguna de las categorías a extraer.

  Por ejemplo, a partir de la descripción: *Planting of native saplings on 3 ha of degraded land in NW Tasmania*, obtenemos los siguientes resultados:

- Objetivo: *None*
- Geolocalización: NW Tasmania
- Organización: *None*
- Participantes: *Anyone*

En este caso solo está definida la geolocalización del proyecto y el algoritmo la extrae correctamente, pero, en el caso de la extracción de participantes, se obtiene un resultado erróneo. Este tipo de resultados están explicados en el apartado 4.1.4.

- Por lo contrario, de los pocos textos extensos que han aparecido en la muestra de 150 descripciones de proyectos de ciencia ciudadana, se puede decir que los algoritmos sí que extraen la información correcta de todas las categorías. Los textos extensos sí que contienen la información de las categorías a extraer.

La figura 4.1 muestra el porcentaje de error de los algoritmos según el número de palabras del texto. Se ha dividido en cuatro grupos: descripciones de proyectos de 0 a 20 palabras, de 26 a 150, de 151 a 500 y, por último, las descripciones de más de 500 palabras.

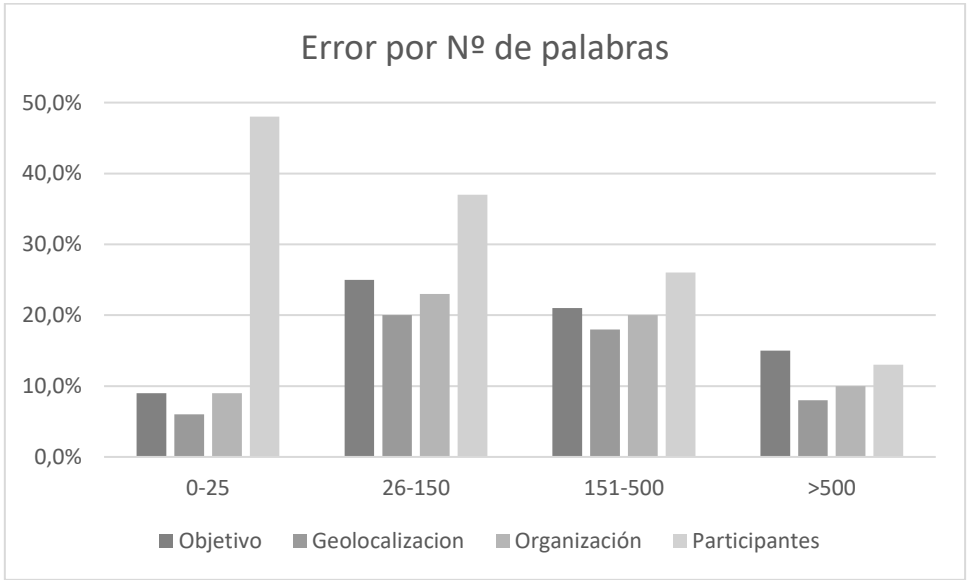


Figura 4.1: Porcentaje de error de los algoritmos según el número de palabras de las descripciones de proyectos

Con este gráfico se puede apreciar lo descrito anteriormente. El mínimo error se consigue (en todos los algoritmos menos el de Participantes) cuando las descripciones son de una sola frase (0-25 palabras) ya que éstas no contienen información relevante y los algoritmos devuelven *None*. En cambio, el algoritmo de participantes devuelve *Anyone* y este resultado muchas veces es incorrecto ya que tampoco se define que tipo de ciudadanos pueden participar en el proyecto.

Finalmente, se puede apreciar que cuantas más palabras contenga el texto, el error es menor. Pero el error no va únicamente relacionado con el número de palabras, sino

también con la calidad del texto. Existen textos muy extensos y de calidad pésima con los que los algoritmos no extraen resultados correctos y en cambio, hay descripciones que contienen menos palabras, pero ganan en calidad y esto se ve reflejado en los resultados de los algoritmos.

4.1.1. Objetivo

El algoritmo de extracción del objetivo se podría decir que es el que utiliza más tipos de técnicas de extracción. Es un algoritmo que contempla las estructuras más utilizadas para describir el objetivo de un proyecto de ciencia ciudadana. Por este motivo, nos encontramos este error:

- Existen casos de estructuras de palabras que explican el objetivo pero que no se han definido en el algoritmo porque son mínimas o incluso únicas. Estas estructuras se han ido identificando al realizar el cálculo del error, pero no se han añadido al algoritmo porque pueden dar pie a confusión en la extracción. Incluir dichas estructuras significaría aumentar el error ya que solo son útiles para casos concretos.

4.1.2. Geolocalización

El algoritmo de extracción de la geolocalización es con el que se ha obtenido el menor porcentaje de error. Aun así, hay ocasiones donde no se extrae de forma correcta la localización:

- Se ha observado que existen textos donde no aparece un lugar exacto como por ejemplo podría ser Francia o el norte de EE. UU., pero sí que aparecen gentilicios. Estos gentilicios indican donde se realiza el proyecto de ciencia ciudadana.
- Uno de los errores más comunes es extraer la geolocalización más repetida en el texto sin que esa sea la ubicación real del proyecto de ciencia ciudadana. Por ejemplo, si en un texto se repite cinco veces Barcelona y dos veces París, el algoritmo extraerá como geolocalización la ciudad Barcelona. Pero puede darse el caso en que el proyecto se desarrolle en la ciudad de París y no en Barcelona y que simplemente durante todo el texto se haga referencia a Barcelona por algún motivo.
- Por último, se ha definido como error también la extracción de una sola geolocalización cuando el proyecto de ciencia ciudadana se realiza en más de una

ubicación. La gran mayoría de los proyectos solo se llevan a cabo en una zona del mundo, pero existen proyectos que se desempeñan en varios países, por ejemplo.

4.1.3. Organización

El algoritmo de extracción de la organización utiliza la técnica de las entidades nombradas para extraer la organización, asociación o entidad que lleva a cabo el proyecto. Se ha identificado el siguiente error:

- Al extraerse una única organización (la más repetida en el texto), puede darse el caso de que justo esa no sea la correcta. Es el mismo error analizado en el apartado anterior con la extracción de la geolocalización.

4.1.4. Participantes

El algoritmo de extracción de los participantes es el que cuenta con un mayor porcentaje de error. Se han identificado dos principales problemáticas:

- El primer problema está en que cuando no se especifica el tipo de participantes, el algoritmo devuelve *Anyone* por defecto. Esto se definió así después de realizar un análisis manual de una pequeña muestra de textos. Y aunque con esta implementación se reduce el error general, no hay que olvidar ese porcentaje donde el algoritmo no se ajusta.
- El segundo problema más común está en extraer *Anyone* cuando lo correcto sería Area Community. En este tipo de casos, los textos no tienen ninguna palabra clave para que el algoritmo detecte que debería devolver *Area Community*, pero leyendo estas descripciones de proyectos de ciencia ciudadana se comprende que solo pueden participar aquellas personas que se encuentran físicamente en la zona donde se realiza el proyecto. Es decir, el resultado que más se ajustaría aquí sería una combinación de *Area Community* y *Anyone* ya que puede participar cualquier tipo de persona de esa zona.

4.2. Visualización

Se han construido tres gráficos de barras que muestran las diez geolocalizaciones, organizaciones y tipo de participantes principales.

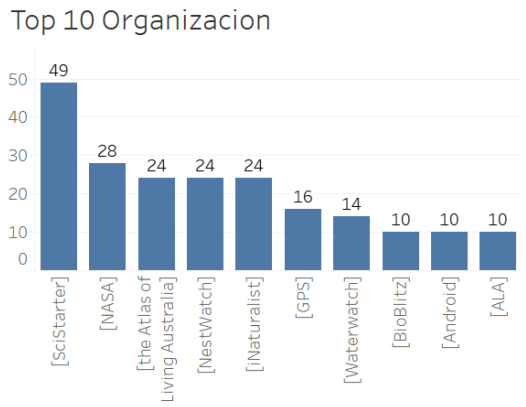


Figura 4.2: Top 10 organizaciones extraídas automáticamente

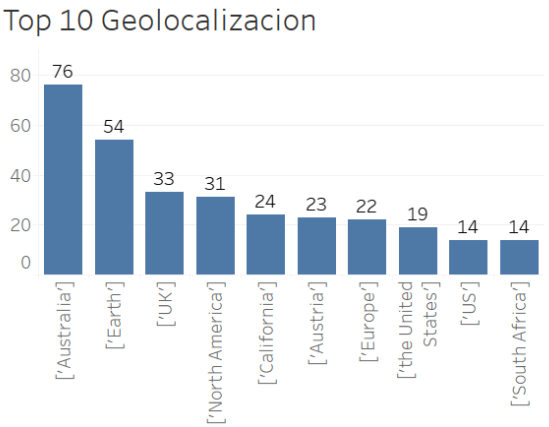


Figura 4.3: Top 10 geolocalizaciones extraídas automáticamente

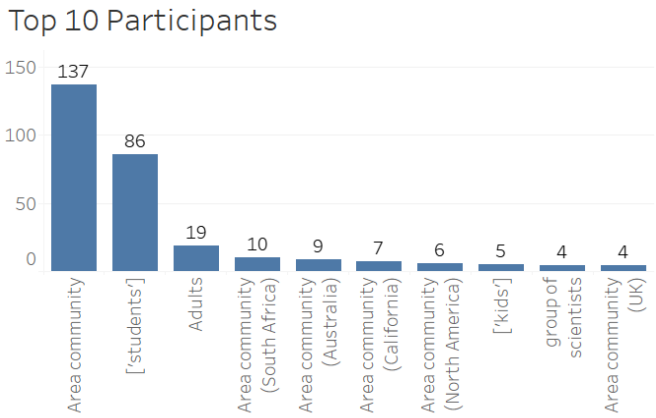


Figura 4.4: Top 10 participantes extraídos automáticamente

Cabe decir, que de la figura 4.4 se ha ocultado la primera posición: *Anyone*, ya que sumaban un total de 1755. Sobre esta visualización es interesante ver los tipos de participantes que sí que están definidos y no lo que se asigna por defecto cuando el algoritmo no encuentra el tipo de participante. Pero hay que seguir teniendo en cuenta que dentro de estos 1755 *Anyone* no todos son por defecto, es decir, una parte sí que realmente corresponde con que cualquier tipo de persona puede participar en ese proyecto de ciencia ciudadana.

Por último, se han querido visualizar las palabras más utilizadas en los objetivos de los proyectos de ciencia ciudadana. La figura X.X muestra un *WordCloud* (nube de palabras), una técnica para mostrar qué palabras son más frecuentes en un texto. Cuanto más presente esté una palabra en el texto considerado, más grande aparecerá en la nube de palabras.

Figura X.X: Palabras más utilizadas en la definición de los objetivos de proyectos de ciencia ciudadana

Antes de crear esta figura se han tenido que limpiar los textos (los objetivos extraídos de forma automática a partir del algoritmo) así como eliminar las palabras irrelevantes ("*stopwords*[18]").

Ahora sí, analizando estas figuras, se pueden extraer varias conclusiones sobre estos resultados:

- Las organizaciones que se muestran en el Top 10 corresponden con grandes proyectos o grandes organizaciones de ciencia ciudadana como iNaturalist, NestWatch, etc. Aunque también hay errores como GPS o Android. Sí que son organizaciones, pero no indican el promotor en sí del proyecto de CC sino la tecnología que se usa para llevar a cabo ese proyecto.
- En cuanto a la geolocalización, aunque los proyectos son europeos, mucha información que contiene la base de datos es de Norte América. Esto indica que los proyectos de CC que se pueden realizar de forma remota son de allí (además de los categorizados como '*Earth'*) o simplemente en la base de datos de CS Track se han filtrado plataformas como SciStarter, que se centra en EE. UU.
- Sobre los participantes, principalmente se observa que se mezcla el perfil con el área. Es decir, se mezcla el tipo de perfil (adultos, estudiantes, niños, etc) con la zona geográfica donde se realiza el proyecto de CC.

[18] Las *stopwords* o palabras vacías son básicamente conjunciones, artículos, proposiciones y adverbios.

5. CONCLUSIONES

El objetivo principal era extraer, automáticamente, información, analizarla y categorizarla a partir de las descripciones de proyectos de ciencia ciudadana almacenados en la base de datos de CS Track. Para cumplir este objetivo, se propusieron diversos subobjetivos que se han cumplido en gran parte.

Tanto los subobjetivos que comportan la familiarización con la ciencia ciudadana y con la base de datos proporcionada (subobjetivos 1 y 2), como los que engloban la aplicación y análisis de técnicas de extracción automática de la información (subobjetivos 3, 4 y 5) se han alcanzado. En resumen, tanto el estudio previo, la creación de los diferentes algoritmos de extracción automática y el posterior análisis de estos se han terminado con éxito en el tiempo establecido.

También se estableció como subobjetivo la creación de *dahsboards* interactivos. Éstos no han sido creados por la limitación de los datos y tiempo de cómputo del algoritmo de extracción del objetivo. En un principio se creía que los datos iban a ser diferentes y sí que se podrían realizar visualizaciones en diversos mapas, por ejemplo. Después de desarrollar y aplicar los algoritmos de extracción de la información no se han podido realizar este tipo de visualizaciones. Así pues, las diferentes plataformas de donde se extrajo la información en un principio podrían añadir descripciones de mayor calidad de los proyectos que ofrecen. De esta manera el error de los algoritmos disminuiría. Otra posible recomendación para las plataformas es el uso de datos estructurados en su página web. Así los datos son más fáciles de categorizar y extraer.

Finalmente, en cuanto a la alimentación de la base de datos de CS Track con la información extraída no se ha cumplido, pero podría ser una línea de trabajo futura explicada en el siguiente punto.

Pero, aunque la mayoría de los objetivos se hayan cumplido, se pueden añadir mejoras como la optimización del algoritmo de extracción del objetivo o la aplicación de diferentes modelos de aprendizaje automático o redes neuronales que entrenen una pequeña parte de los datos para después aplicarlos a toda la base de datos.

Finalmente, solo queda hacer una pequeña valoración propia sobre los conocimientos aprendidos y aplicados. Gracias a los conocimientos previos de Python me ha resultado bastante fácil trabajar con este lenguaje. La dificultad la he encontrado al aplicar las diferentes técnicas. Aún así, agradezco haber cursado diferentes asignaturas durante el grado donde aprendí de forma teórica los conceptos de NLP y varias técnicas de

extracción de la información. Gracias a estos conceptos y la búsqueda de más información durante el desarrollo de este trabajo, pude pensar y probar por mí misma qué técnicas debía aplicar en cada caso. Podría decir que ha sido un trabajo en su mayoría de prueba y error donde también he tenido que aprender a gestionar el tiempo y los recursos. Por último, dejando de lado los aspectos técnicos, he conocido que es la ciencia ciudadana (término que no conocía antes) y su importancia y que hace CS Track.

6. TRABAJO FUTURO

Se han considerado tres posibles líneas de trabajo futuro:

- El vuelco de resultados en la base de datos de CS Track. La información extraída a partir de los algoritmos elaborados durante este trabajo se podría almacenar y relacionar con los datos que ya posee CS Track. Esta tarea comportaría el trabajo de preparación y procesado de estos datos, así como la elaboración de un formato adecuado. Además de adquirir conocimientos mínimos sobre el funcionamiento de MongoDB, ya que es el sistema de base de datos que utiliza CS Track.
- Optimización del algoritmo de extracción del objetivo. El cálculo de la similitud lleva bastante tiempo. Aunque no es necesario hacer este cálculo de todos los textos que pasan por el algoritmo, los que sí que lo necesitan, se aprecia un aumento del tiempo de cómputo considerable. Esta parte del algoritmo se podría intentar optimizar para así conseguir trabajar con mayor número de datos. Otra opción sería paralelizar el algoritmo para poder así ejecutar en paralelo.
- Trabajar en la traducción automática. Conseguir extrapolar la extracción de la información de textos en inglés a otros idiomas como por ejemplo el español o francés aplicando técnicas de traducción automática. Todo el trabajo está centrado en descripciones de proyectos de ciencia ciudadana en inglés, pero realizando un análisis de palabras clave en otros idiomas y con la ayuda de diferentes librerías públicas en Python (como nltk o la misma SpaCy) se puede crear un algoritmo capaz de extraer la información a partir de textos de diferentes idiomas.

7. BIBLIOGRAFÍA

[1] Chari, R., Blumenthal, M. S., & Matthews, L. J. (2019). *Community Citizen Science: From Promise to Action*. RAND Corporation. https://www.rand.org/pubs/research_reports/RR2763.html

[2] Project, C. S. | C. T. (s. f.). *About* (World) [Text]. Https://Cstrack.Eu/; Citizen Science | CS Track Project. Recuperado 8 de junio de 2022, de https://cstrack.eu/about/

[3] Gordienko, Y. (2013). *Green Paper on Citizen Science*.

[4] *Socientize_white_paper_on_citizen_science.pdf*. (s. f.). Recuperado 8 de junio de 2022, de https://ec.europa.eu/futurium/en/system/files/ged/socientize_white_paper_on_citizen_science.pdf

[5] University (BA), W., & University (MDiv), H. (s. f.). *What Is Citizen Science? History, Practices, and Impact*. Treehugger. Recuperado 8 de junio de 2022, de https://www.treehugger.com/what-is-citizen-science-history-practices-and-impact-5189634

[6] *Guia-para-conocer-la-ciencia-ciudadana.pdf*. (s. f.). Recuperado 8 de junio de 2022, de https://ciencia-ciudadana.es/wp-content/uploads/2019/01/guia-para-conocer-la-ciencia-ciudadana.pdf

[7] (S. f.). Recuperado 8 de junio de 2022, de https://www.dw.com/es/ciencia-ciudadana-la-salvaci%C3%B3n-de-la-biodiversidad/a-17488487

[8] *Revista Comunicar*. (s. f.). Recuperado 8 de junio de 2022, de https://www.revistacomunicar.com/index.php?contenido=detalles&numero=54&articulo=54-2018-03

[9] Shirk, J., Ballard, H., Wilderman, C., Phillips, T., Wiggins, A., Jordan, R., McCallie, E., Minarchek, M., Lewenstein, B., Krasny, M., & Bonney, R. (2012). Public Participation in Scientific Research: A Framework for Deliberate Design. *Ecology and Society*, *17*(2). https://doi.org/10.5751/ES-04705-170229

[10] Follett, R., & Strezov, V. (2015). An Analysis of Citizen Science Based Research: Usage and Publication Patterns. *PLOS ONE*, *10*, e0143687. https://doi.org/10.1371/journal.pone.0143687

[11] Turmo, J., Ageno, A., & Català, N. (2006). Adaptive information extraction. *ACM Computing Surveys*, *38*(2), 4. https://doi.org/10.1145/1132956.1132957

[12] *Deep_learning_for_nlp.pdf*. (s. f.). Recuperado 8 de junio de 2022, de http://ling.snu.ac.kr/class/AI_Agent/deep_learning_for_nlp.pdf

[13] *What is Natural Language Processing? An Introduction to NLP*. (s. f.). SearchEnterpriseAI. Recuperado 8 de junio de 2022, de https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP

[14] Li, J., Sun, A., Han, J., & Li, C. (2020). *A Survey on Deep Learning for Named Entity Recognition* (arXiv:1812.09449). arXiv. http://arxiv.org/abs/1812.09449

[15] *What Is Data Science? The Ultimate Guide*. (s. f.). SearchEnterpriseAI. Recuperado 8 de junio de 2022, de https://www.techtarget.com/searchenterpriseai/definition/data-science

[16] *Las 5 V del Big Data: Volumen, velocidad, veracidad, variedad y valor*. (s. f.). Recuperado 8 de junio de 2022, de https://empresas.blogthinkbig.com/5-v-big-data/

[17] *Semantic Similarity Using Transformers | by Raymond Cheng | Towards Data Science*. (s. f.). Recuperado 8 de junio de 2022, de https://towardsdatascience.com/semantic-similarity-using-transformers-8f3cb5bf66d6

[18] Sarawagi, S. (2008). *Information Extraction*. Now Publishers Inc.

[19] Ciencia ciudadana: Conocimiento al poder. (2014, abril 24). *CCCB LAB*. https://lab.cccb.org/es/ciencia-ciudadana-conocimiento-al-poder/

8. ANEXOS

8.1. Anexo A

Figura A.3.1: https://scistarter.org/save-the-waves-app

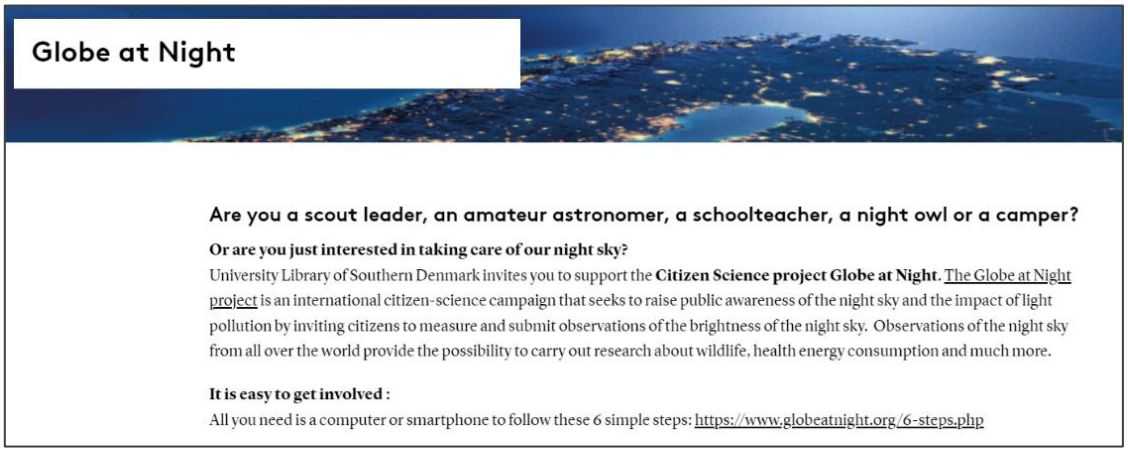


Figura A.3.2: https://www.sdu.dk/en/forskning/forskningsformidling/citizenscience/globeatnight

Figura A.3.3: https://www.barcelona.cat/barcelonaciencia/es/bioblitzbcn

```
english_data = data[data['Language'] == 'English']
english_data = english_data.reset_index(drop=True)
manual_analysis = english_data[['_id', 'DESCRIPTION', 'Plat Id']]
```

Figura A.3.4: Selección de textos en inglés

```
structured_sample = manual_analysis[manual_analysis['Data_type'] == 'structured']
structured_sample = structured_sample.sample(15)
non_structured_sample = manual_analysis[manual_analysis['Data_type'] == 'not-structured']
non_structured_sample = non_structured_sample.sample(15)
semi_structured_sample = manual_analysis[manual_analysis['Data_type'] == 'semi-structured']
semi_structured_sample = semi_structured_sample.sample(15)
```

```
sample = pd.concat([structured_sample, non_structured_sample, semi_structured_sample], axis=0)
sample = sample.reset_index(drop=True)
print(len(sample))
```

```
45
```

Figura A.3.5: Selección de los 45 textos

['Generation Solar acts as a database of photovoltaic installations. It has various purposes including data exchange between PV installation owners. Generation Solar is a multi-faceted app: It boosts the acceptance of solar energy among the users and their community. And it helps researchers increase their knowledge in photovoltaic research. Generation Solar is also fun: It includes a gaming section where you can be a solar energy detective: For instance, you report as many photovoltaic installations as you can.', 'The first platform to create a unique solar energy community and promote data exchange between photovoltaic installation owners and scientists.Sign up or log in to share as many photovoltaic installations as you can and...• Take part in an environmentally committed international community• Provide valuable open data for research and management• Face new individual and community challenges• Chat with other European citizens to share interests• Invite friends to join the network• Help put the community squarely on the map• Get public data to analyse the behaviour of decentralized energy system', 'The first platform to create a unique solar energy community and promote data exchange between photovoltaic installation owners and scientists.\r\n\r\n Sign up or log in to share as many photovoltaic installations as you can and...\r\n\r\n• Take part in an environmentally committed international community\r\n\r\n• Provide valuable open data for research and management\r\n\r\n• Face new individual and community challenges\r\n\r\n• Chat with other European citizens to share interests\r\n\r\n• Invite friends to join the network\r\n\r\n• Help put the community squarely on the map\r\n\r\n• Get public data to analyse the behaviour of decentralized energy system\r\n']

Figura A.3.6: Descripción de un proyecto de ciencia ciudadana extraído de la base de datos de CS Track

```
def select_random_sample(data):
    random_sample = data.sample(1)
    df = pd.DataFrame()
    #title
    df['Title'] = random_sample['TITLE']
    #data type
    df['Data_type'] = random_sample['Data_type']
    #Plat country
    df['Country'] = random_sample['Plat country']
    #description
    text_list = random_sample['DESCRIPTION']
    text_list = text_list.tolist()
    text_list = text_list[0]
    text = ''
    for sentence in text_list:
        text = text + sentence
    print(text)
    df['Description'] = text

    # = df.reset_index()
    return df
```

Figura A.3.7: Función cuyo objetivo es seleccionar un proyecto aleatorio

```python
def clean(text):

    # removing new line characters
    text = re.sub('\n ','',str(text))
    text = re.sub('\n',' ',str(text))
    text = text.replace('\\n',' ')
    # removing hyphens
    text = re.sub("-",' ',str(text))
    text = re.sub("— ",'',str(text))
    # removing quotation marks
    text = re.sub('\"','',str(text))
    # removing salutations
    text = re.sub("Mr\.",'Mr',str(text))
    text = re.sub("Mrs\.",'Mrs',str(text))
    # removing any reference to outside text
    #text = re.sub("<\[\(\[].*?[\)\]]>", "", str(text))
    text = re.sub("</a>", "", str(text))
    text = re.sub("<a", "", str(text))
    return text
```

Figura A.3.8: Función cuyo objetivo es limpiar un texto

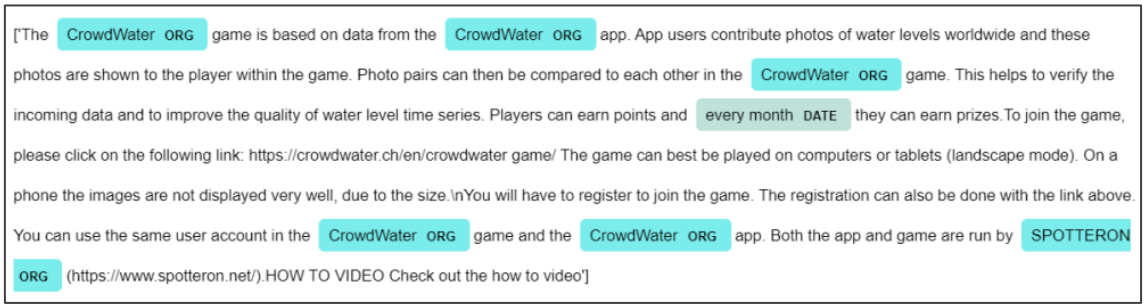


Figura A.3.9: Descripción de un proyecto de ciencia ciudadana aleatorio

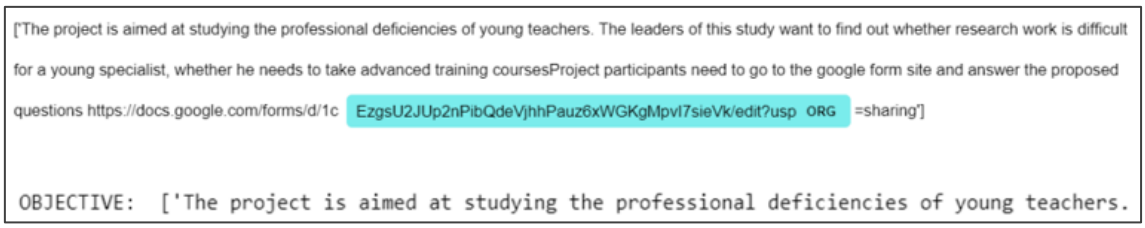


Figura A.3.10: Extracción automática del objetivo de un proyecto aleatorio de ciencia ciudadana

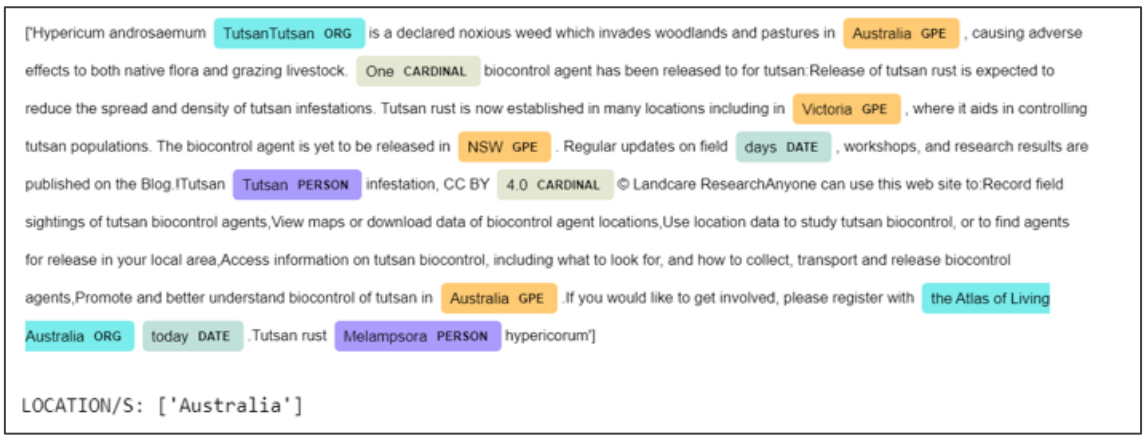


Figura A.3.11: Resultado después de la extracción de la geolocalización

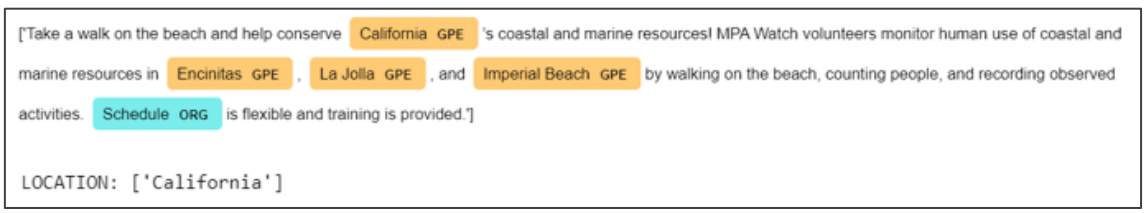


Figura A.3.12: Resultado extracción geolocalización

8.2. Anexo B

Localización de los scripts:

Nombre del script	Link
BD_CSTrack.ipynb	https://github.com/maramartiez00/TFG_MaraMartinez/blob/main/BD_CSTrack.ipynb
InformationExtraccion.ipynb	https://github.com/maramartiez00/TFG_MaraMartinez/blob/main/InformationExtracction.ipynb

8.3. Anexo C

Country	Name	URL
AUT	Österreich forscht	https://www.citizen-science.at/projekte
CHE	Schweiz forscht	https://www.schweiz-forscht.ch/de/citizen-science-projekte
AUT	Zentrum für Citizen Science	https://www.zentrumfuercitizenscience.at/en/projects
USA	USA Citizen science platform	https://www.citizenscience.gov/catalog/#
BEL	Citizen Science Vlaanderen	https://www.scivil.be/en/projects
DNK	Citizen Science DK	https://citizenscience.dk/portfolio/
USA	Scistarter	https://scistarter.org/finder
EU	Digital Earth Lab	https://digitalearthlab.jrc.ec.europa.eu/csp/topics
USA	Scientific American	https://www.scientificamerican.com/citizen-science/
EU	Erasmus+ project results	https://ec.europa.eu/programmes/erasmus-plus/projects/?pk_kwd=180150#search/project/keyword=citizen%20science&matchAllCountries=false
EU	CORDIS	https://cordis.europa.eu/search/en?q=%27citizen%20science%27%20AND%20(contenttype%3D%27project%27%20OR%20%2Farticle%2Frelations%2Fcategories%2Fcollection%2Fcode%3D%27brief%27%20OR%20%2Fresult%2Frelations%2Fcategories%2Fcollection%2Fcode%3D%27pubsum%27,%27deliverable%27,%27publication%27)&p=1&num=10&srt=Relevance:decreasing
UK	Natural History Museum UK	https://www.nhm.ac.uk/take-part/citizen-science.html
USA	North Carolina Museum of Natural sciences	https://naturalsciences.org/research-collections/citizen-science/current-projects
BEL	Idereen Wetenschapper	https://www.iedereenwetenschapper.be/

World wide	inaturalista network	https://www.inaturalist.org/projects/browse
UK	Earthwatch	https://earthwatch.org.uk/get-involved
UK	Earthwatch	https://earthwatch.org.uk/our-science/research-topics-and-projects/past-projects
UK	Earthwatch	https://earthwatch.org.uk/get-involved/2-uncategorised/25-projects-and-activities
UK	University of Dundee: Leverhulme Research Centre for Forensic Science	https://www.dundee.ac.uk/leverhulme/citizenscience/
ES	Ciencia Ciudadana Espana	https://ciencia-ciudadana.es/proyecto-cc/
World wide	Critical Ecosystem Partnership Fund	https://www.cepf.net/grants/grantee-projects
World wide	Conservation Leadership Programme	http://www.conservationleadershipprogramme.org/our-projects/supported-projects/
AUS	Australian Citizen Science Project Finder	https://biocollect.ala.org.au/acsa#isCitizenScience%3Dtrue%26isWorldWide%3Dfalse%26max%3D20%26sort%3DdateCreatedSort
CL	Ciencia Ciudadana Chile	http://cienciaciudadana.cl/proyectos/
USA	Citizen Science Competitive Funding Program	https://www.fs.usda.gov/working-with-us/citizen-science/projects
ARG	Cientópolis	https://www.cientopolis.org/participar/#proyectos
BRA	Ciência Cidadã	https://www.sibbr.gov.br/cienciacidada/projetos.html
NZL	Participatory Science Platform	https://www.curiousminds.nz/projects/?fund=participatory-science-platform&start=0
DEU	Science Cite	https://www.science-et-cite.ch/de/home/projekte/projekte-national-und-deutschschweiz?limit=20&tag_list_language_filter=de-DE&types[0]=1&start=0
World wide	Zooniverse	https://www.zooniverse.org/projects
BEL	Observations.be	https://observations.be/projects/
DEU	Bürger schaffen Wissen	https://www.buergerschaffenwissen.de/projekte
DEU	Citizen Science Germany	http://www.citizen-science-germany.de/citizen_science_germany_projekte.html
DEU	citizen science working group	https://www.uni-muenster.de/AFO/CS/cs_mitforschen.html

FR	Jagis pour la nature	http://www.vigienature.fr/fr
FR	Observatoires Participatives des Espèces et de la nature	https://www.open-sciences-participatives.org/ecosysteme-sciences-participatives/
LVA	Latvian Fund for Nature	https://ldf.lv/en/list_of_projects
FR	J'agis pour la nature	https://www.jagispourlanature.org/
ES	Oficina ciència ciutadana Barcelona	https://www.barcelona.cat/barcelonaciencia/es/proyectos-ciencia-ciudadana
ES	Open Systems	http://www.ub.edu/opensystems/projectes/
ES	CREAF	http://www.creaf.cat/research/citizen-science
ES	Desqbre	https://fundaciondescubre.es/ciencia-ciudadana/
ES	Victoria-Gasteiz Ayuntamiento	https://www.vitoria-gasteiz.org/wb021/was/contenidoAction.do?idioma=es&uid=u25e08f9d_14a56aaea69__7f88
UK	nQuire	https://nquire.org.uk/discover
World wide	Instant wild	https://instantwild.zsl.org/projects
ES	Parcs de catalunya	https://parcs.diba.cat/es/web/conservacio-de-la-biodiversitat/ciencia-ciutadana
ES	ICM Divulga	http://icmdivulga.icm.csic.es/proyectos-divulgativos/
ES	Cities Health	https://citieshealth.eu/get-involved/
AUT	ZAMG	https://www.zamg.ac.at/cms/de/forschung/citizen-science/citizen-science?searchterm=citize
LVA	Dabas dati	https://dabasdati.lv/en/cat/7/?links=en/cat/7/
World wide	National Geographic	https://www.nationalgeographic.org/idea/citizen-science-projects/
IRL	Ireland Environmental protection agency	["http://www.epa.ie/irelandsenvironment/getinvolved/citizenscience/epacitizenscienceinitiatives/" "http://www.epa.ie/irelandsenvironment/getinvolved/citizenscience/nationalcitizenscienceinitiatives/" "http://www.epa.ie/irelandsenvironment/getinvolved/citizenscience/completedprojects/"]
IT	Iteritalia	http://www.lteritalia.it/?q=content/citizenscience
UK	Scotland's environment	https://www.environment.gov.scot/get-involved/submit-your-data/citizen-science-portal/
World wide	Environment Live's Citizen Science portal	https://environmentlive.unep.org/citizen
DEU	Citizen Science within the research institute Leibniz	https://www.leibniz-gemeinschaft.de/en/research/citizen-science.html

DEU	Citizen Science within the research institute Helmholtz	https://www.helmholtz.de/transfer/wissenstransfer/citizen_science/
SWE	ARCS	https://medborgarforskning.se/lankar/
World wide	Environment Live's Citizen Science portal	https://environmentlive.unep.org/citizen
World wide	Zooniverse	https://www.zooniverse.org/projects?page=1&status=paused
World wide	Zooniverse	https://www.zooniverse.org/projects?page=1&status=finished
UK	CEH	https://www.ceh.ac.uk/citizen-science#poms
CZE	Institute of Botany of the Czech Academy of Sciences	https://www.ibot.cas.cz/en/public-relations/citizen-science/
FIN	Luonto-Liiton Kevätseuranta	http://kevatseuranta.fi/osallistu/
AUT	BOKU Universität für Bodenkultur Wien	https://boku.ac.at/citizen-science/projekte
LVA	Citizen Science Initiatives Latvia 2020	https://data.gov.lv/dati/eng/dataset/citizen-science-initiatives-in-latvia/resource/9925676d-938d-4d33-b9c1-759dfb87ca50
PRT	Rede Portuguesa de Ciência Cidadã	https://www.cienciacidada.pt/index2.php
Europe (EU)	Eu citizen science platform	https://eu-citizen.science/projects
UK	Flusurvey	https://flusurvey.net/en/
USA	Track Together	https://tracktogether.org/
SGP	TraceTogether	https://www.gov.sg/article/help-speed-up-contact-tracing-with-tracetogether
CAN	Operation COVID-19 (OpCovID)	https://opcovid.com/about-us
UK	NHS 111 online	https://111.nhs.uk/covid-19
DNK	CoronAPP	https://www.coronapp.dk/
DNK	Virus Simulator (COVID-19)	https://corona-land.org/
UK	COVID Symptom Study	https://covid.joinzoe.com/
ISR	Daily report on the fight against the corona virus (יומי דיווח בקורונה למאבק)	https://coronaisrael.org/
EU	ODYSSEA: OPERATING A NETWORK OF INTEGRATED OBSERVATORY SYSTEMS IN THE	http://odysseaplatform.eu/

	MEDITERRANEAN SEA	
NLD	Burgerwetenschap	https://www.rivm.nl/burgerwetenschap
AUT	Young science AT (OEAD)	https://youngscience.at/de/angebote/projekte-zum-mitforschen/mitforschen-von-daheim/
UK	UCL university	https://www.ucl.ac.uk/library/research-support/open-science/citizen-science

8.4. Anexo D

TYPE	DESCRIPTION
PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.
LAW	Named documents made into laws.
LANGUAGE	Any named language.
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day.
PERCENT	Percentage, including "%".
MONEY	Monetary values, including unit.
QUANTITY	Measurements, as of weight or distance.
ORDINAL	"first", "second", etc.
CARDINAL	Numerals that do not fall under another type.