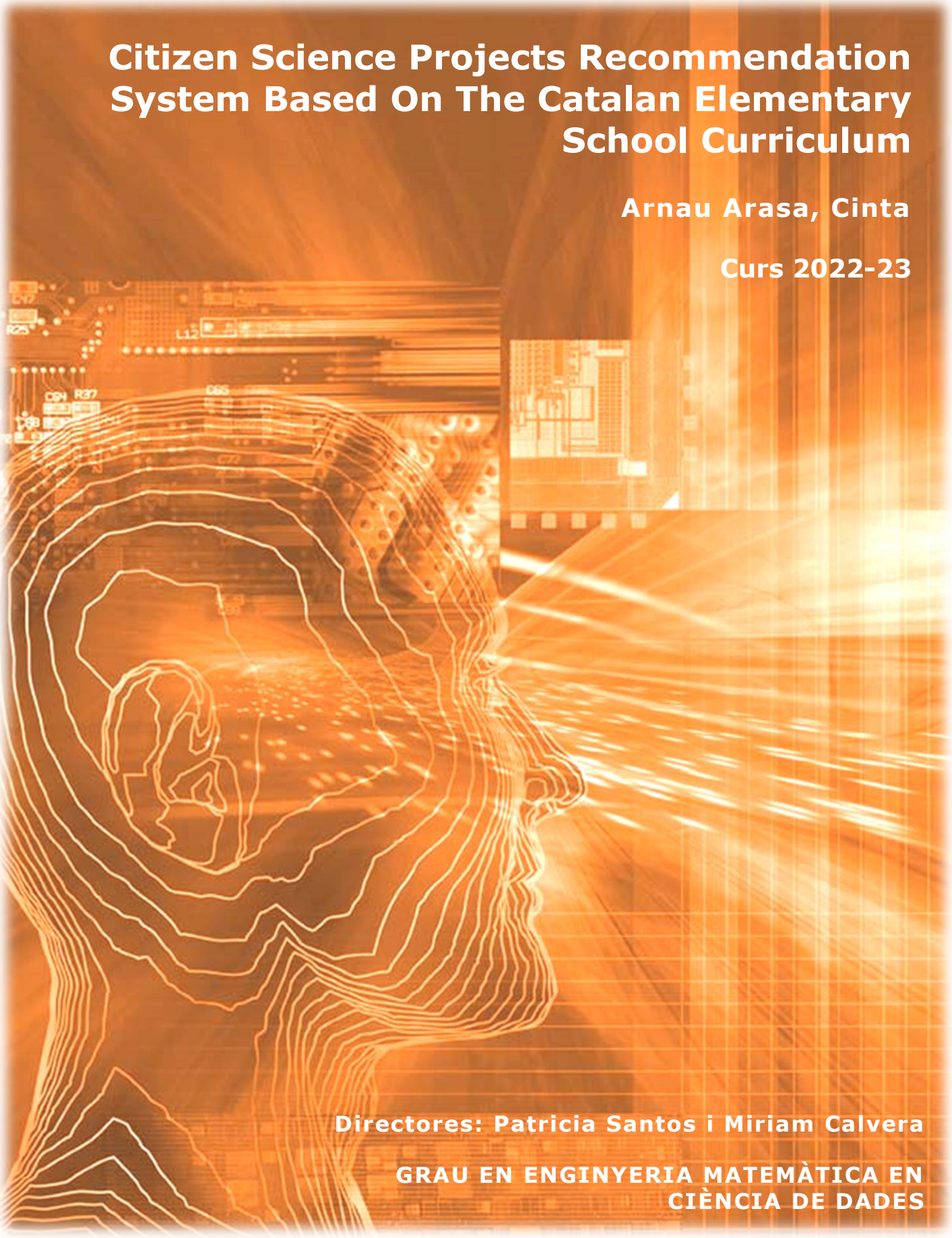
Citizen Science Projects Recommendation System Based On The Catalan Elementary School Curriculum

Arnau Arasa, Cinta

Curs 2022-23



Directores: Patricia Santos i Miriam Calvera

GRAU EN ENGINYERIA MATEMÀTICA EN CIÈNCIA DE DADES

Universitat Pompeu Fabra Barcelona
Escola d'Enginyeria

Treball de Fi de Grau

To my younger sister Núria
who I love the most

Acknowledgements

First of all, I would like to especially thank my two thesis mentors Miriam and Patricia for their availability, support, help and advice throughout this project.

I would also like to thank my family for always supporting me and believing in me during my whole life and also during these four years of the engineering degree. I truly appreciate the efforts my parents have done for me to get to where I am right now and all the love they have given me, how my sister looks up to me and the amount of love we have for each other, and how my godmother and Brigi, with care and love, always give me the best advice.

Finally, I would also like to thank the friends I have made in this degree, especially Oriol Gallego, for living this experience with me and making from it an unforgettable memory.

Abstract

Citizen Science (CS) is a field of research that allows citizens to engage in scientific projects. In the field of education, this type of science offers students practical and real-life opportunities to learn, as well as increase their interest in scientific research.
The goal of this project is to provide teachers with a tool to find citizen science projects that can be used in classrooms to help students achieve the key competences stated in Catalan elementary school curriculum. For this, various algorithms will be developed where information extraction techniques are applied to extract information about citizen science projects in order to create a recommender system that provides teachers with the best suited projects based on the key competences or any key words of their choice, and a web application to let users interact with the recommender system.
From an educational perspective, it is expected that by integrating information about CS projects in a formal education setting, teachers get inspired to create learning activities.

Resumen

La Ciencia Ciudadana (CC) es un campo de investigación que permite a los ciudadanos participar en proyectos científicos. En el campo de la educación, este tipo de ciencia ofrece a los estudiantes oportunidades prácticas y de la vida real para aprender, así como aumentar su interés por la investigación científica.
El objetivo de este proyecto es proporcionar a los profesores una herramienta para encontrar proyectos de ciencia ciudadana que se puedan utilizar en las aulas para ayudar a los estudiantes a alcanzar las competencias clave establecidas en el currículum de primaria de Catalunya. Para ello, se desarrollarán varios algoritmos donde se aplican técnicas de extracción de información para extraer información sobre proyectos de ciencia ciudadana con el fin de crear un sistema de recomendación que proporcione a los profesores los proyectos más adecuados en función de las competencias clave o palabras clave de su elección, y una aplicación web para que los usuarios puedan interactuar con el sistema de recomendación.
Desde una perspectiva educativa, se espera que al integrar información sobre proyectos de CC en un entorno educativo formal, se inspire a los maestros a crear actividades de aprendizaje.

Resum

La Ciència Ciutadana (CC) és un camp de recerca que permet als ciutadans participar en projectes científics. En l'àmbit de l'educació, aquest tipus de ciència ofereix als estudiants oportunitats pràctiques i reals per aprendre, així com augmentar el seu interès per la recerca científica.
L'objectiu d'aquest projecte és proporcionar al professorat una eina per trobar projectes de ciència ciutadana que es puguin utilitzar a les aules per ajudar els alumnes a assolir les competències bàsiques establertes en el currículum català de primària. Per a això, es desenvoluparan diversos algorismes on s'apliquen tècniques d'extracció d'informació per extreure informació sobre projectes de ciència ciutadana per tal de crear un sistema de recomanació que proporcioni al professorat els projectes més adequats a partir de les competències clau o de qualsevol paraula clau de la seva elecció, i una aplicació web que permeti als usuaris interactuar amb el sistema de recomanació.
Des d'una perspectiva educativa, s'espera que en integrar informació sobre projectes de CC en un entorn educatiu formal, s'inspiri els mestres a crear activitats d’aprenentatge.

TABLE OF CONTENTS

1. **INTRODUCTION** **10**
1.1 Context 10
1.2 Motivation 11
1.3 Objectives 11
1.4 Planning 12

2. STATE OF THE ART **14**
2.1 Citizen Science 14
2.1.1) Citizen Science definition and history 14
2.1.2) Citizen Science projects 15
2.1.3) Citizen Science as an educational tool 15
2.2 Catalan elementary school curriculum 16
2.2.1) Analysis of the Catalan elementary school curriculum 16
2.2.2) Key Competences 16
2.3 Automatic Extraction of Information 17
2.3.1) Natural Language Processing (NLP) 17
2.3.2) Term Frequency-Inverse Document Frequency (TF-IDF) 18
2.3.3) Semantic similarity 18
2.4 Recommendation System 19
2.4.1) Types of recommendation systems 19

3. DESIGN AND IMPLEMENTATION **22**
3.1 Environment set-up and selection of tools 22
3.1.1) Tools used 22
3.2 Spanish platform: “Observatorio de la Ciencia Ciudadana en España” 24
3.1.1) Platform Data Structure 24
3.1.2) Extracting the data from the Spanish platform 24
3.3 Barcelonian platform: “Oficina de la Ciència Ciutadana” 25
3.1.1) Platform Data Structure 25
3.1.2) Extracting the data from the Barcelonian platform 26

4. RECOMMENDATION SYSTEM **28**
4.1 Design of the Recommendation System 28
4.1.1) Preprocessing of the data 28
4.1.2) Text embeddings 29
4.1.3) Cosine similarity 29
4.2 Key Competences 29
4.3 Criteria of selection 31
4.4 Results 32
4.5 Analysis of the recommended projects 32

5. WEB APPLICATION **40**

6. CONCLUSIONS AND FUTURE WORK **44**

7. BIBLIOGRAPHY **46**

8. APPENDICES 48
8.1 Appendix A 48
8.2 Appendix B 53
8.2 Appendix C 53
8.3 Appendix D 54

1. INTRODUCTION

1.1 Context

"Citizen Science refers to the general public engagement in scientific research activities when citizens actively contribute to science either with their intellectual effort or surrounding knowledge or with their tools and resources" [1]. The participation in Citizen Science (CS) is emerging as a support to formal science. There are all types of CS projects, from projects that allow the study of human mobility by registering it through a mobile phone application (Beepath[1]) to projects that create a game to investigate the genomic alterations in cancer cells (Genigma[2]).

The communication and promotion of CS projects is done through social networks and web platforms that are accessible by all citizens and used by scientists and volunteers for scientific communications. Citizens can engage in different levels of the scientific process. They can participate actively by doing different types of activities including asking questions, reporting observations, conducting experiments, collecting data or taking photos. Everyone can contribute to the progress of Citizen Science.

Modern advances in technology have been one of the main factors for the increment in the number of CS projects, leading to the creation of many platforms such as the Barcelonian "Oficina de la Ciència Ciutadana" and the Spanish "Observatorio de la Ciencia Ciudadana en España". These platforms work as repositories of CS projects from Barcelona and Spain, respectively, with the goal of promoting citizen science between citizens by facilitating information about the projects and letting them join and participate.

For this project, it has been proposed to extract information from the Barcelonian platform since the objective of the project is to create a recommendation system of citizen science projects based on the key competences stated in the Catalan curriculum with the goal of finding the most suited projects that can be either used to participate in or to create similar learning activities based on the recommended project to accomplish the key competences. Therefore, it is adequate to extract information about projects that are carried out in the city of Barcelona. Additionally, it has been proposed to extract the data from the Spanish platform to have a larger database. However, this has been done as a proof of concept, and could be extended to other platforms.

The elementary school curriculum describes the objectives, contents and evaluation criteria of each area and subject. It contains information about the key competences, which are essential achievements for the students to ensure their successful progress in their education and their capability of facing local and global challenges and demands.

Data science is the study of vast volumes of data that uses modern techniques with the goal of finding unseen patterns, extracting meaningful insights, and making decisions [2]. It includes different disciplines such as machine learning, data preprocessing, data mining, data visualization, among many others. It allows the application of extraction

[1] Beepath https://www.barcelona.cat/barcelonaciencia/es/ciencia-en-la-ciudad/la-ciencia-y-la-ciudadania/ciencia-ciudadana/beepath.
[2] Genigma https://www.barcelona.cat/barcelonaciencia/es/ciencia-la-ciutat/la-ciencia-i-la-ciutadania/ciencia-ciutadana/genigma.

techniques and analysis to better understand how the practice of CS works, as well as getting to know the established relationships between the different involved agents.

These data science disciplines will have to be applied throughout this project in order to extract data about CS projects and work with it to accomplish the established objectives.

1.2 Motivation

This project intends to examine how citizen science can be used to incentivise teachers to use the information about Citizen Science projects as an educational tool by creating learning activities based on the existing CS projects. Therefore, information about these projects will be extracted from CS platforms to create a database that will be used in a recommender system with the objective to provide teachers with the best suited projects based on the key competences of the Catalan elementary curriculum. Therefore, the Catalan key competences will be analysed with the goal of better understanding the objectives to be accomplished by the students in this stage of their education and how teachers can use citizen science to achieve that goal.

By extracting the information from the CS platforms and analysing the key competences, a recommender system can be created with the goal of recommending teachers the best suited projects to carry out in the classroom to help students learn about a specific area, hence developing knowledge that helps them fulfill the needed skills.

Additionally, the project expects to enrich the knowledge people have about citizen science and support its divulgation.

From an educational perspective, it is expected that by integrating information about CS projects in a formal education setting, teachers get inspired to create learning activities.

1.3 Objectives

The main objective of this project is to create a recommendation system of Citizen Science projects with the goal of finding the most suited projects that can be either used to participate in or to create similar learning activities based on the recommended project to accomplish the key competences stated in the elementary school curriculum. To do so, it is necessary to analyze the citizen science projects developed in Barcelona and Spain. In a more systematic way, the objectives of this project are:

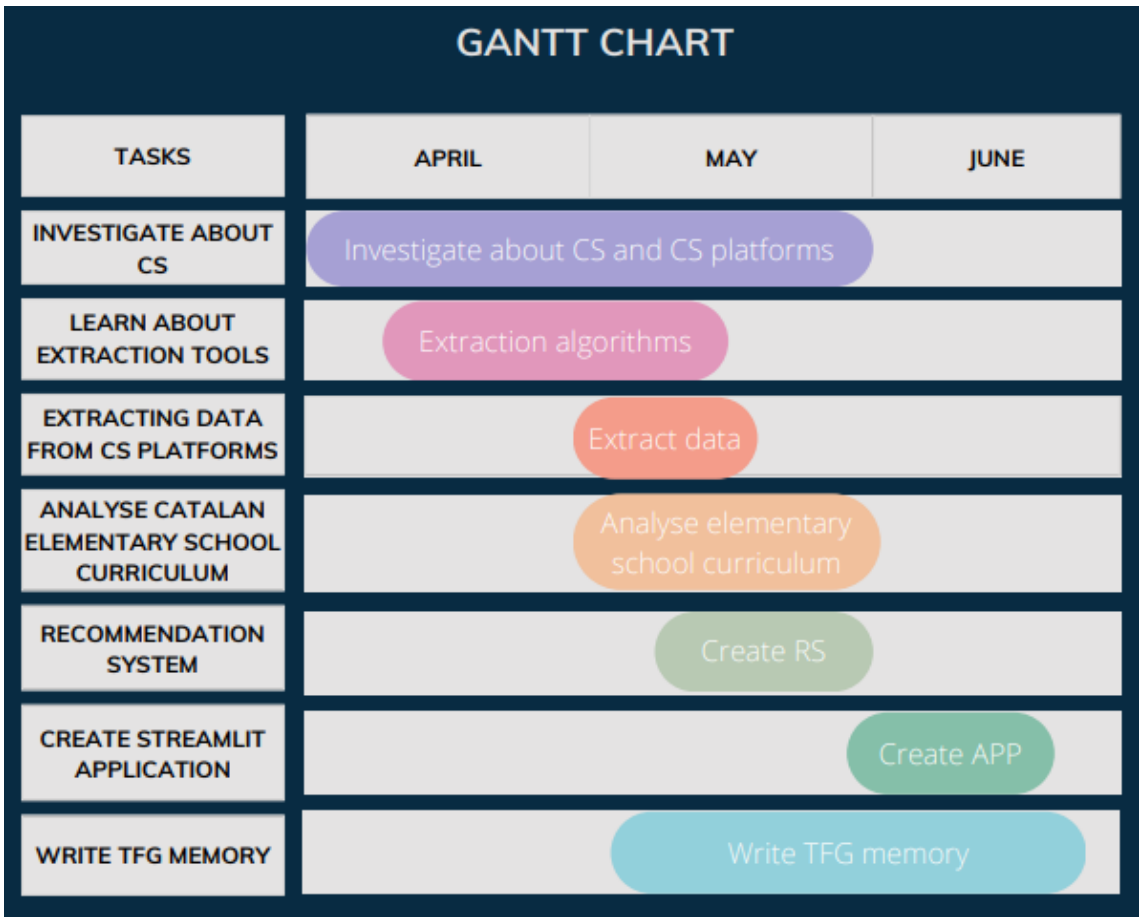
- Studying the Barcelonian citizen science online website “Oficina de la Ciència Ciutadana”[3] and the Spanish “Observatorio de la Ciencia Ciudadana en España”[4].
- Designing and developing a tool to extract information from the Barcelonian and Spanish platforms by programming a crawler and creating the corresponding databases.
- Analysing the Catalan elementary school curriculum and the key competences stated in it.
- Designing and developing a recommendation system that recommends CS projects based on the key competences of the curriculum.

[3] “Observatorio de la Ciencia Ciudadana en España” (2023, May 20) https://ciencia-ciudadana.es/.
[4] “Oficina de la Ciència Ciutadana” (2023, May 20) https://www.barcelona.cat/barcelonaciencia/es/.

- Analysing manually the recommended projects obtained from the recommendation system to check that the given recommendations are adequate.
- Extracting meaningful conclusions regarding the results obtained from the recommendation system.
- Give recommendations of the potential use of the recommendation system.

1.4 Planning

A Gantt diagram has been created to show the planned tasks that have been developed throughout this project.



2. STATE OF THE ART

2.1 Citizen Science

2.1.1) Citizen Science definition and history

"Citizen science is a field of research in which members of the general public participate in scientific projects, often in collaboration with professional scientists". It refers to the involvement of the general public in the scientific process. It is a collaborative approach that engages individuals in the different stages of scientific research, including data collection, analysis, and interpretation [3]. Citizen Science initiatives aim to take advantage of the collective power and skills of participants to contribute to scientific knowledge and address research questions on a large scale. It promotes inclusivity, scientific literacy, public participation, and scientific discovery.

The European Citizen Science Association[5] (ECSA) is a membership association that aims to encourage the growth of citizen science in Europe and support the active engagement of the general public in the research process. It states that the two main characteristics of this type of science are that "citizens are actively involved in research, in partnership or collaboration with scientists or professionals" and "there is a genuine outcome, such as new scientific knowledge, conservation action or policy change" [4]. The ECSA members have developed the '1O principles of citizen science'[6], which include the active engagement of citizens in the scientific research, the existence of a genuine science outcome from the CS projects, the possibility of both professional scientists and citizen scientists to benefit from taking part, and the right of citizen scientists to receive feedback from the projects, among many others.

For more than a century, citizen science has been quite popular, but in recent times, the internet has completely transformed scientists' capacity to connect with and involve citizen scientists in an extensive range of research projects. It was back in the 1990s when the field of citizen science was given its name by researchers at the 'Cornell Laboratory of Ornithology'. However, even before this science field had its name, many scientists already made use of this scientific method to carry out their investigations. "The roots of citizen science can be traced to 1900, perhaps, when the Audubon Society began its Christmas bird counts, or to the late 1800s, when the fledgling National Weather Service culled information from amateur meteorologists". In Europe it was in the mid-18th century when the Swedish botanist Carolus Linnaeus gave start to the record-keeping methodology with the objective of gathering data about flowering times in Sweden [5].

New and emerging technologies such as mobile applications, wireless sensor networks and online connectivity show great promise for the soon to come advances in citizen science. In the meanwhile, modern advances in technology have facilitated the scientific research process with the use of devices such as mobile phones, more specifically, with the help of GPS. The Global Positioning System provides users with positioning, navigation and timing services, making it easier for citizens and scientists to contribute to the gathering of the data in the research process, as it allows public to get the exact coordinates with just a click.

[5] ECSA website (2023, June 5) https://www.ecsa.ngo/.
[6] "Ten principles of citizen science", ECSA (2023, June 5) https://www.ecsa.ngo/ecsa-guidelines-and-policies/#documents.

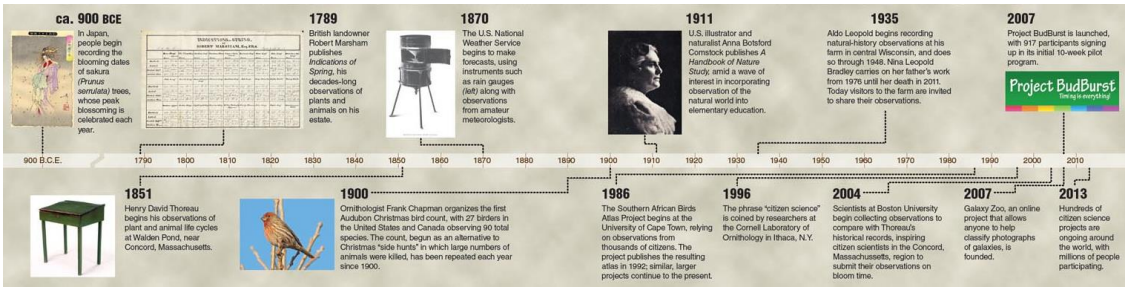


Figure 2.1. *Citizen Science chronology [26].*

2.1.2) Citizen Science projects

Citizen science can take many forms depending on the level of involvement required from the public as well as the type of data collected. As it is explained in the website "In depth: The rise of Citizen Science" (2023) [3], there are five main classifications:

1. Observational citizen science. It involves the participation of individuals from the general public engaging in activities like bird watching or monitoring air quality, where they make observations and gather data related to a particular subject.
2. Participatory citizen science. It involves active participation of individuals from the general public in both the design and execution of a scientific study, along with the responsibility of gathering and analysing data.
3. Collaborative citizen science. This involves direct collaboration between members of the public and professional scientists, working together to gather and analyse data through programs like community science initiatives.
4. Citizen-led science. It refers to a scenario in which citizens take the initiative to initiate, design, and lead research projects, with the involvement of scientists serving as advisors or collaborators whenever it is necessary.
5.
6. Online citizen science. It involves individuals from the public actively engaging in online projects, which may involve tasks like image classification, historical documents transcription, or the utilization of online tools to gather data, such as through surveys.

The main fields in which citizen science takes place are environmental research, biology and ecology, public health, and social science and humanities. However, all fields of science can benefit from the public's active participation, including psychology, computer science, astronomy, genetics, medicine, and many others [6].

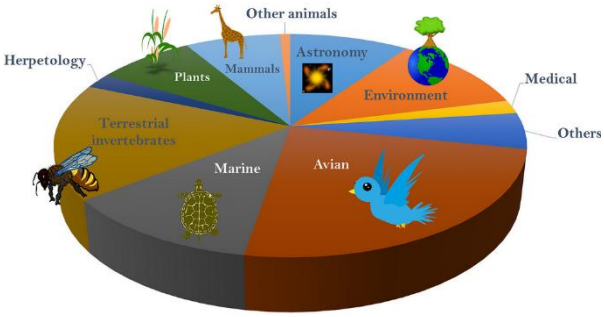


Figure 2.2. *Distribution of the CS projects' fields [6].*

2.1.3) Citizen Science as an educational tool

Teachers instruct students with learning activities along with their knowledge to change their course comprehension to a better understanding. Some of the factors that thrive teachers are their previous experience, their students' interest and the available educational material.

Citizen Science carries many benefits in the educational sector, one of them being the ability to offer students practical and real-life opportunities to learn, enabling them to address community challenges alongside their theoretical education. In the long run, the use of citizen science to create learning activities in the classrooms will increase the students' interest in scientific research, leading to greater interest in career options in these fields, such as STEM.

In the article "Current Approaches in Implementing Citizen Science in the Classroom" by Shah and Martinez (2016), it is mentioned how "studies have shown that introducing interactive, research-based models of education can greatly improve classroom performance and retention". There is a wide range of Citizen Science projects, meaning that the activities of these projects can be carried out by diverse educational levels and academic stages. All these projects share a common objective, which is to inspire young students to engage in scientific inquiry, enhance their scientific literacy, and promote their imagination as they explore the world around them [7].

2.2 Catalan elementary school curriculum

In today's rapidly changing and evolving society, elementary school education plays an important role in providing students with the necessary skills and mindset to thrive. Therefore, educational systems must anticipate and adapt to constant changes.

"Elementary school education must prepare students to provide innovative responses in a constantly changing and evolving society. The principles of equity, quality, and excellence determine and condition educational action since teaching and learning processes need to be personalized to the maximum extent and take into account the diversity of all students within an inclusive system" [8].

2.2.1) Analysis of the Catalan elementary school curriculum

The Catalan elementary school curriculum[7] describes the objectives, contents and evaluation criteria of each area and subject. It contains information about the key competences and the competency indicators at the end of stage, which constitute the profile upon completion.

The curriculum must serve as a guide to follow to choose certain methodologies or actions that teachers should propose so that students can pursue their personal and professional life aspirations, building upon their educational achievements, with the active engagement and support of families. Additionally, it should ensure further educational opportunities to promote lifelong learning [9]. The goal of the curriculum is to achieve the key competences that are developed throughout the knowledge areas and subjects.

2.2.2) Key Competences

"The key competences are the achievements that are considered essential for the students to progress successfully in their educational journey and to face the main global and local challenges and demands" [9]. The full list of the key competences stated in the Catalan elementary school curriculum can be checked in the Appendix C.

[7] DECRET 175/2022, de 27 de setembre, d'ordenació dels ensenyaments de l'educació bàsica.

2.3 Automatic Extraction of Information

The automatic extraction of information allows the identification and extraction of meaningful information from a document or text without the user having to read it. This extraction can be achieved with the use of techniques and algorithms such as the Natural Language Processing (NLP).

2.3.1) Natural Language Processing (NLP)

"Natural language processing (NLP) refers to the branch of computer science - and more specifically, the branch of artificial intelligence or AI - concerned with giving computers the ability to understand text and spoken words in much the same way human beings can" [10]. NLP combines the fields of computational linguistics with statistical, machine learning, and deep learning models in order for computers to process human language in the form of text or voice data and comprehend its complete meaning, intention and sentiment.

NLP tasks involve breaking down language into smaller, elemental components, aiming to comprehend the relationships between these components and exploring how they interact to create meaning. Some of the tasks that make use of NLP are [11]:

- Content categorization: linguistically-based summarization of documents, including search and indexing, content alerts, and duplication detection.
- Classification: capture contextual and semantic meaning of words in a text.
- Corpus Analysis: understanding the structure of a corpus and its documents through statistical analysis (sampling, data preparation and strategic modeling).
- Contextual extraction: automatically extracting structured information from text.
- Sentiment analysis: identifying the mood or subjective opinion in text, including sentiment analysis and opinion mining.
- Speech-to-text and text-to-speech conversion: converting spoken commands to text and vice versa.
- Document summarization: automatically generating a summary of extensive textual contents and detecting languages in multilingual documents.
- Machine translation: automated translation of text or speech from one language to another.

The Natural Language Toolkit (NLTK[8]) is an amazing library to work with natural language in the Python programming language. This makes it a very suitable library for working with NLP, as it can be used in tasks such as tokenization[9], stemming[10], lemmatization[11], semantic reasoning, and more.

[8] NLTK Documentation (2023, May 31) https://www.nltk.org/.
[9] Tokenization consists of splitting up a larger body of text into smaller lines or words (tokens) that help the computer understand better the text.
[10] Stemming refers to the process to produce morphological variations of a word's original root.
[11] The goal of lemmatization is to reduce a word to its root form.

2.3.2) Term Frequency-Inverse Document Frequency (TF-IDF)

"TF-IDF (term frequency-inverse document frequency) is a statistical measure that evaluates how relevant a word is to a document in a collection of documents" [12]. This text analysis technique is a numerical measure used to get the importance of a word for a document in a collection.

The TF-IDF for a word is calculated by multiplying two different metrics, the first one being the amount of time a word has appeared in a document, and the second one being the inverse document frequency of the word across a set of documents.

- TF (term frequency). It counts the number of occurrences for a given word in a document (or text) [12].
- IDF (inverse document frequency). It measures if a term is common or not in a collection of documents. Its value can be obtained by first dividing the total number of documents in the set by the total number of documents from the set that contain the given word, and secondly taking the logarithm of the resulting division. Therefore, the closer the obtained IDF value is to 0, the more common the term is [12].

By multiplying the TF and IDF results, the TF-IDF score of a term in a document is obtained. The higher the score, the more relevant the term is in the document.

$$tf\ idf\ (t, d, D) = tf\ (t, d)\ .\ idf\ (t, D)$$

$$tf\ (t, d) = log\ (1 + freq\ (t, d))$$

$$idf\ (t, D) = log\ (\ \frac{N}{count\ (d \in D{:}t \in d)}\ )$$

Figure 2.3. *Formula to calculate the TF-IDF value of a term [12].*

2.3.3) Semantic similarity

Text semantic similarity is an active research area within NLP, and is used in many applications such as in sentiment analysis, natural language understanding , machine translation, question answering, chatbots, search engines, and information retrieval. Its goal is to identify if the meaning of two texts or words is similar [13].

To calculate text similarity, the process typically involves converting text into a vector of features. The algorithm then selects an appropriate representation of features, such as TF-IDF. Finally, the similarity is determined by comparing the vector representations of the texts. There are numerous techniques to calculate text similarity, being Jaccard similarity, cosine similarity, and K-Means the most used ones. The technique chosen for calculating the semantic similarity for the recommendation system process is the cosine similarity. The measure of semantic similarity is usually a score between 0 and 1, 0 meaning that the two texts or words are not similar at all, and 1 meaning they almost have identical meaning.

$$Similarity(A, B) = \frac{A.B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Figure 2.4. *Formula to calculate the cosine similarity, where n is the size of features vector [13].*

2.4 Recommendation System

Recommendation systems (or recommender systems) are a class of machine learning used to help the final users in the decisions taking. They can recommend or predict the expected results from a given set of data. Recommender systems are trained to understand the preferences, previous decisions and characteristics of users and products based on the data about their interactions such as impressions, clicks, likes, and purchases [14]. In general, the idea of a recommendation system is that, given data from user interests, including profiles, browsing behavior, item interaction behavior, ratings about various items, it uses such data to make recommendations to users about further interesting items.

Recommender systems have many benefits, such as improving retention, increasing sales, helping to form customer habits and trends, increasing user satisfaction, speeding up the pace of work and boosting cart value.

There are many companies that use recommender systems in order to achieve the previously mentioned benefits. Such companies are Amazon.com, Netflix, Spotify and LinkedIn, among many others.

- Amazon.com uses an item-to-item collaborative filtering recommender system in order to mainly boost the number of items purchased and hence the cart value and increase sales. "According to McKinsey, 35% of Amazon purchases are thanks to recommendation systems" [15].
- Netflix uses a collaborative filtering recommender system, where it shows some category and movie/show suggestions based on the user's preferences and other shows previously watched by the user. "The same McKinsey study highlights that 75% of Netflix viewing is driven by recommendations" [15].
- The Spotify music engine uses three different techniques: collaborative filtering, natural language processing (NLP) and audio file analysis [15].

2.4.1) Types of recommendation systems

The recommendation systems can be classified in these three broad categories: collaborative filtering, content-based filtering and context filtering.

2.4.1.1) Collaborative filtering

Collaborative filtering algorithms employ a filtering mechanism that recommends items by leveraging the preference data gathered from multiple users. This approach is based on the analysis of the similarities among user preference behavior given the past interactions between users and items [14]. This technique filters out items that may be liked by a user based on the items liked by similar users. Therefore, this algorithm constantly finds the relationships between the users and in result, it makes the recommendations.

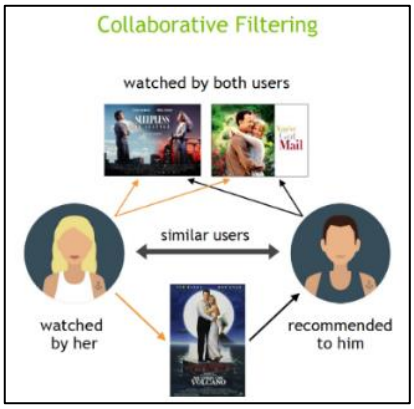


Figure 2.5. *Collaborative filtering [14].*

There are two types of collaborative filtering systems:

- User-based: measures the similarity between target users and other users.
- Item-based: measures the similarity between the items that target users rate or interact with and other items.

2.4.1.2) Content-based filtering

Content filtering algorithms recommend items similar to what the user preferences are based on the attributes or characteristics of a given item (referred to as the content), using the similarity between the item and the user features [14]. In content-based recommendations, users and items are associated with features, which are matched to infer interest. Some of the uses of this recommendation system are recommending other movies with the same director, age, genre, as viewed ones and recommending other products in the same category, brand, color, as purchased ones. This algorithm is able to recommend to users with very particular tastes, recommend new and obscure items, and provide explanations that are easily understandable.

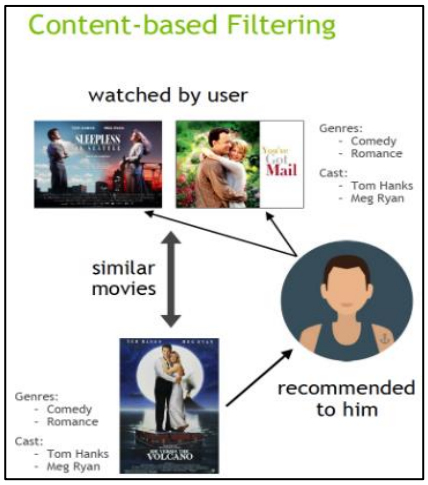


Figure 2.6. *Content-based filtering [14].*

2.4.1.3) Context filtering

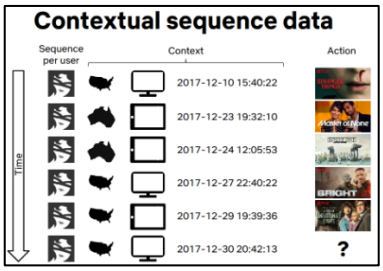


Figure 2.7. *Context filtering [14].*

"Context filtering includes users' contextual information in the recommendation process. This approach uses a sequence of contextual user actions, plus the current context, to predict the probability of the next action. In the Netflix example, given one sequence for each user—the country, device, date, and time when they watched a movie—they trained a model to predict what to watch next" [14].

3. DESIGN AND IMPLEMENTATION

In this section, the technologies, data sources, and designed extraction algorithms used for this project are detailed. More specifically, the extraction algorithm to obtain the data of the citizen science projects will be explained, along with the steps of the process.

3.1 Environment set-up and selection of tools

3.1.1) Tools used

In order to extract the data from the two platforms of Citizen Science projects and create the recommendation system various tools have been used throughout the entirety of the thesis process. The following descriptions of these tools aim to clarify why they are necessary and demonstrate their application within the project.

3.1.1.1) Python

Python[12] is a high-level programming language that is widely used for general-purpose programming. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming. Python's versatility and ease of use have made it a go-to language for a wide range of applications, including web development, data analysis, machine learning, artificial intelligence, scientific computing, and automation [16]. For this reason, and given the fact that Python has been the most used programming language throughout the degree, it has been the chosen language for the Citizen Science projects' extraction and the building of the recommendation system.

The libraries that have been used in the scripts are the following:

- Pandas[13]: Python library that provides data structures and data analysis tools for working with structured data, creating DataFrame objects and much more [17].
- Request[14]: Tool for making HTTP requests and interacting with web services [18].
- NLTK[15]: The Natural Language Toolkit offers functionalities for tasks such as tokenization, stemming, lemmatization, semantic reasoning, and more [19].
- Re[16]: The re (regular expression) library provides support for regular expressions, which are powerful tools for pattern matching and text manipulation [20].
- TfidfVectorizer[17]: Class from the scikit-learn library for text feature extraction, based on the term frequency-inverse document frequency weighting scheme [21].
- Cosine_similarity[18]: Method provided by the scikit-learn library to compute the cosine similarity between pairs of vectors or matrices [22].
- WordCloud[19]: Library to generate word clouds (visual representations of text data where the size of each word corresponds to its frequency within the text) [23].
- Matplotlib[20]: Python library for creating static, animated, and interactive visualizations. It provides a wide range of plots, charts and graphs [24].

[12] Python Website (2023, May 20) https://www.python.org/.
[13] Pandas Documentation (2023, May 20) https://pandas.pydata.org/docs/.
[14] Request Documentation (2023, May 20) https://docs.python-requests.org/en/latest/.
[15] NLTK Documentation (2023, May 31) https://www.nltk.org/.
[16] Python Documentation. re (2023, May 31) https://docs.python.org/3/library/re.html.
[17] Scikit-learn Documentation. TfidfVectorizer (2023, May 31) https://scikit-learn.org/TfidfVectorizer.
[18] Scikit-learn Documentation. cosine_similarity (2023, May 31) https://scikit-learn.org/cosine_similarity.
[19] Wordcloud Documentation (2023, May 31) https://amueller.github.io/word_cloud/.
[20] Matplotlib Documentation. (2023, May 31) https://matplotlib.org/.

3.1.1.2) BeautifulSoup

BeautifulSoup is a popular Python library used for web scraping and parsing HTML or XML documents [25]. It provides a convenient way to extract and navigate data from web pages by simplifying the process of locating and manipulating elements within the document structure. BeautifulSoup can be combined with other Python libraries, such as requests, to fetch web page content and then parse and extract the desired information.

3.1.1.3) Streamlit

Streamlit[21] is a Python library used for building interactive web applications and data dashboards with ease. [26]. It provides an API that allows users to write code and generate web-based visualizations and user interfaces. It also simplifies the deployment of applications by automatically managing the server and handling user interactions. It supports both local deployment for development and sharing applications as well as cloud-based deployment options. This Python library has been used previously for the creation of a web application for one of the university courses and, therefore, has been the chosen method for creating the app for this project.

3.1.1.3) Jupyter Notebook

"Jupyter Notebook[22] is an open-source web-based interactive computing environment widely used for data analysis, visualization, and prototyping" [27]. It supports many programming languages (including Python) and allows for interactive data analysis and visualization, enhancing the exploratory data analysis process.

In order to extract all the information about the Spanish and Catalan Citizen Science projects, the script "projects_extraction.ipynb" has been generated. All the information regarding the CS projects is extracted using the Python library BeautifulSoup. The script named "recommendation_system.ipynb" joins the information extracted from both the Spanish and the Barcelonian platform and the recommendation system to recommend CS projects based on the key competences stated in Catalan elementary school curriculum is created. The process to create the system is explained in section 4.1.

These Jupyter notebook scripts can be found in the Github repository so that anyone interested in Citizen Science or any teacher wanting to use the recommendation system can consult them (see Appendix B to consult the locations of the scripts).

3.1.1.4) Visual Studio Code

Visual Studio Code[23] (VSC) is a cross-platform source code editor that provides a lightweight yet powerful environment for coding, debugging and version control [28]. It supports many programming languages (including Python). Since VSC has been the used platform in university for using streamlit to create a web application, it has been the chosen one for the app creation (see Appendix B to consult the location of the code).

3.1.1.5) GitHub

"GitHub[24] is a web-based platform for version control and collaborative software development". It provides a centralized hub where developers can store, manage, and

[21] Streamlit Documentation. (2023, June 3) https://docs.streamlit.io/.
[22] Jupyter Notebook Website (2023, May 20) https://jupyter.org/.
[23] Visual Studio Code (2023, May 3) https://code.visualstudio.com/docs.
[24] GitHub Website (2023, May 20) https://docs.github.com/en.

collaborate on their code repositories [29]. A GitHub repository[25] has been created to store and manage all the scripts with the corresponding code of this thesis.

3.2 Spanish platform: "Observatorio de la Ciencia Ciudadana en España"

In this section, the data structure of the Spanish Citizen Science platform "Observatorio de la Ciencia Ciudadana en España"[26] and the extraction process of all the data related to the citizen science projects are explained in detail.

3.2.1) Platform data structure

The platform distinguishes between initiatives and resources. Initiatives can be citizen science projects, persons, or institutions. The initiatives can be found by field of knowledge, type (citizen science projects, persons, or institutions), or text/keyword. All three types share the same structure.

The initiatives all follow the same structure, no matter their type. The fields each initiative has are the following: project ID, project name, main organisation, subtitle, type of project, keywords, start date, end date, public, province, participants, URL, aim, project description, responsible entity, founding team, more entities, how to participate, results, results link, impact, impact examples, and motivation (why use citizen science).

3.2.1) Extracting the data from the Spanish platform

All the citizen science projects contained in the Spanish platform have been extracted directly from the website after having made sure that it allowed the automatic extraction through the use of robots.

In order to extract all the information from the Spanish platform, it is needed to create a crawler. The chosen pyhton library to do the web scraping is BeautifulSoup, which will make it possible to get the desired data throughout the examination of the website elements. The process to do so in a Jupyter Notebook environment, as it is done in the "projects_extraction.ipynb" script, is the following:

1. Install and import the BeautifulSoup library along with the other needed libraries pandas and request.
2. Sending an HTTP request to the web page by using the 'requests.get()' function to send a GET request to the URL of the Spanish website. The response from the server can be used to extract the HTML content of the page. (See *Figure 3.1* and *Figure 3.2*).
3. Parsing the HTML content. The HTML content obtained from the web page is passed to the BeautifulSoup constructor to create a BeautifulSoup object, which allows users to navigate and search through the HTML structure of the page. (See *Figure 3.3*).

[25] GitHub repository https://github.com/Cintaa1223/TFG.
[26] "Observatorio de la Ciencia Ciudadana en España" (2023, May 20) https://ciencia-ciudadana.es/.

4. Extracting data. This involves finding specific HTML elements such as tags, classes, or IDs, and accessing their attributes or text content. In order to do so, methods like 'find()' or 'find_all()' are used to locate and extract the desired data.

   To see the HTML structure of the webpage: "Ajustes" → "Más herramientas" → "Herramientas para desarrolladores"[27].

   - First of all, the home page of the Spanish platform contains all the projects which have to be accessed in order to extract the needed information. To do so, we first find where the URL to each project is found in the HTML structure and create a list that contains all the links to the citizen science projects. (See *Figure 3.4*).
   - Now we access each of the projects' URLs stored in the links array the same way as described in steps 2 and 3. The array can be iterated to get all the necessary information of each project. This needed data includes the following fields: *'Project Name', 'Project Link', 'Project Scope', 'Project Goal', 'Project Description', 'Project Entity/Scientist', 'How To Join', 'Necessary Equipment', 'Initial Date', 'Final Date', 'Public Type', 'Location (Province)', 'Number of Participants', 'Results', 'Link to Results', 'Project Impact', 'Why Using CC?', 'Citizen Science Web Name', 'Citizen Science Web Link'*.
     Given that each information to be extracted follows the same HTML structure, the function 'get_complete_section(project_soup, dtbf)' has been created to make it simple. (See *Figure 3.5*, *Figure 3.6* and *Figure 3.7*).

5. Storing the extracted data in a pandas DataFrame. In the same function 'get_project_info1()' as in *Figure 3.6*, all the fields extracted are stored in a dictionary so that it can be added as a new row to the created DataFrame 'df1'. (See *Figure 3.8*).

3.3 Barcelonian platform: "Oficina de la Ciència Ciutadana"

In this section, the data structure of the Barcelonian Citizen Science platform "Oficina de la Ciència Ciutadana"[28] and the extraction process of all the data related to the citizen science projects are explained in detail.

3.2.1) Platform data structure

The platform is used to advise, accompany and promote the city's citizen science projects. It includes projects, activities, news and a calendar. The information that is displayed about each individual citizen science project, including the project name and webpage, a brief description along with the objective, the state (whether it is active or not), the scope, a link of where to visualize the collected data, and different activities within the framework of the Office[29].

Additionally, this Barcelonian platform works together with the Barcelona Education Consortium to organize a program with the goal of introducing the idea of citizen science in schools, allowing the participation of the students in numerous projects in order to learn

[27] It can also be done by either right-clicking on any part of the webpage and selecting "Inspect" or by clicking 'fn'+'f12'.
[28] "Oficina de la Ciència Ciutadana" (2023, May 20) https://www.barcelona.cat/barcelonaciencia/es/.
[29] 'Office' ('Oficina' in Catalan) makes reference to the platform "Oficina de la Ciència Ciutadana".

how to collect data by using mobile applications and websites, as well as getting to know about the scientific method of sampling.

3.2.1) Extracting the data from the Barcelonian platform

Same way as done with the Spanish platform; all the citizen science projects contained in the Barcelonian platform have been extracted directly from the website after having made sure that it allowed the automatic extraction through the use of robots.
In order to extract all the information from the Spanish platform, it is needed to create a crawler, which has been done using BeautifulSoup. The process to do so do the web scraping and obtain the needed information about the citizen science projects is the following:

1. Install and import the BeautifulSoup library along with the other needed libraries pandas and request.
2. Sending an HTTP request to the web page by using the 'requests.get()' function to send a GET request to the URL of the Barcelonian website. The response from the server can be used to extract the HTML content of the page. (See *Figure 3.9* and *Figure 3.10*).
3. Parsing the HTML content. (See *Figure 3.3*).
4. Extracting data.
    - First of all, the home page of the platform contains all the projects that have to be accessed to extract the needed information. To do so, we first find where the URL to each project is found in the HTML structure and create a list that contains all the links to the citizen science projects. (See *Figure 3.11*).
    - Now we access each of the projects' URLs stored in the links array the same way as described in steps 2 and 3. The array can be iterated to get all the necessary information of each project. The needed data includes the fields: *'Project Name', 'Project Link', 'Project Description', 'Project Info', 'Project State', 'Link to Project Data', 'Activities within the framework of the Office', 'Project Scope', 'Citizen Science Web Name', 'Citizen Science Web Link'*.

      Given that the information regarding the state, description, activities, scope and results are all grouped in the same element of the HTML structure, the function 'get_segments(text)' has been created to obtain the corresponding information of each field from the extracted element. (See *Figure 3.12*, *Figure 3.13* and *Figure 3.14*).
5. Storing the extracted data in a pandas DataFrame. In the same function 'get_project_info2()' as in *Figure 3.12*, all the fields extracted are stored in a dictionary so that it can be added as a new row to the created DataFrame 'df2'. (See *Figure 3.15*).

4. RECOMMENDATION SYSTEM

The main reason for which a recommender system is the chosen tool for this projects is to provide the most adequate citizen science projects based on the key competences stated in the Catalan elementary school curriculum (see Appendix C), given that its use will speed up the pace of work since the process of having to look through a dense amount of CS projects to find the best one will be erased and substituted by the process of only having to introduce the objective to the system.

4.1 Design of the Recommendation System

In order to recommend the CS projects obtained from the Barcelonian and Spanish platforms based on the key competences of the Catalan elementary school curriculum, a content-based recommender system has been built using text-similarity.

The reason for choosing a content-based recommender system is that it leverages the description of the projects to make recommendations. By analysing the input key competence and finding projects with similar content, a content-based recommender can provide personalized recommendations based on the user's specific needs or interests. Additionally, this type of system does not rely on the user's historical behavior or data from other users, meaning that it can provide recommendations for new users by solely considering the project descriptions and the user's text input. Finally, the user can see the specific content features that influenced the recommendations, which can enhance trust and user satisfaction.

Previous to the creation of the recommender system, the database with the extracted CS projects had to be created. To do so, the dataframes that stored the projects were downloaded as comma-separated values (CSV) files, which were then imported in the recommender system script. A new column 'Project Full Description' was created by joining the 'Project Description' and the 'Project Goal' columns, which contained the two main parts of the description of the projects. Then, it was decided to eliminate most of the columns of both dataframes, leaving only the following: 'Project Name', 'Project Link', 'Project Scope', 'Project Description', 'Project Goal', 'Project Full Description', 'Citizen Science Web Name', 'Citizen Science Web Link'. This was done to merge both dataframes in one, which would be the final dataframe containing all the CS projects.

4.1.1) Preprocessing of the data

The first step in building a recommender system is preprocessing the data. The process of data preprocessing consists of cleaning and transforming the text data. This includes converting all text to lowercase, removing stop words, removing punctuation and other non-wanted symbols and applying stemming.

To preprocess all the data regarding project descriptions and the input key competence, the function 'build_terms()' has been created. This function receives a line of text as input and returns a list of the words contained in it after having removed the stopwords[30] and all non-wanted symbols, transforming the text to lowercase, tokenizing, and stemming. (See *Figure 4.1*).

[30] Stop words are a set of commonly used words in any language. For example, in English, "the", "is" and "and", would easily qualify as stop words.

All the project descriptions as well as the input key competence are preprocessed by using the mentioned function. (See *Figure 4.2* and *Figure 4.3*).

4.1.2) Text embeddings

"Word embedding is the collective name for a set of language modelling and feature learning techniques in natural language processing (NLP) where words or phrases from the vocabulary are mapped to vectors of real numbers. Words that are similar in a semantic sense have a smaller distance between them than words that have no semantic relationship" [30].

The second step is creating text embeddings. This process consists of converting the preprocessed text data into numerical representations that can be used for similarity calculations. The resulting word embeddings can then be combined to represent a document (in this case the project description) as a dense vector.

With the use of the Python library sklearn to import 'TfidfVectorizer', collection of project descriptions can be converted to a matrix of TF-IDF features. This created matrix is a representation where each row corresponds to a project description and each column corresponds to a TF-IDF score of a specific word. This is applied to both the project descriptions and the input key competence. (See *Figure 4.4*).

4.1.3) Cosine similarity

The third step is calculating the similarity. For this recommender system, it has been decided to use the cosine similarity as the similarity measure. The cosine similarity calculates the similarity between two document vectors, which in this case would be the key competence and the project description. The similarity measure quantifies how similar the two documents are based on their text embeddings.

Cosine similarity is a metric that measures the similarity between two non-zero vectors. It calculates the cosine of the angle between the vectors, indicating the degree of similarity or relatedness between them [22]. The cosine_similarity function measures the similarity between documents based on their vector representations, such as TF-IDF vectors or word embeddings. It takes as input two arrays representing the vectors or matrices for which the cosine similarity needs to be calculated, and it returns a matrix of similarity scores. In this case, the cosine_similarity function is applied as is shown in *Figure 4.5*.

Once these three steps are completed, the projects are ranked based on their similarity to the input sentence, which in this case corresponds to the key competence introduced to the recommendation system. When the ranking has been completed, the top 5 projects that are most similar to the input key competence are displayed as recommendations to the user.

4.2 Key Competences

The goal of the recommendation system is to recommend the user the most adequate citizen science projects based on the input key competence. The full list of key

competences stated in the Catalan elementary school curriculum can be found in Appendix C.

It has been observed how when a long text input is introduced, such as the whole sentence describing the key competence, the output recommended projects are not related at all with what was introduced to the recommendation system. This is due to two main factors:

- TF-IDF representation. The TF-IDF representation is based on the frequency of words within a document. The fact that the input text is long implies that it contains a higher number of words, from which some of them might occur more frequently, hence dominating the TF-IDF scores. Therefore, the similarity scores between the input text and the project descriptions may be biased toward those few dominant words.
- Cosine similarity. The cosine similarity computes the similarity between two vectors by considering the angle between them. If some words in the input text have higher TF-IDF scores, they can influence the similarity score resulting in redundant recommendations.

An additional reason for this to happen is that most of the project descriptions are relatively short, up to the point that in some cases the input text is longer than the descriptions, which also influences the results of the recommended projects.

There are many approaches to solve this problem, such as limiting the length of the input text, normalizing the TF-IDF scores, applying additional text processing techniques or using other similarity measures. From these solutions, given that the project descriptions are not too extended, the selected approach has been to limit the length of the input text.

To do so, the main idea and goal of each key competence has been analysed and a two-word equivalent has been proposed. The shortened key competences are the following, where each number corresponds to the respective number assigned to the key competence in the Appendix C:

1. **Environmental awareness** or just **environment**.
2. **Responsible consumption of local products** or just **local products**.
3. **Healthy lifestyle**.
4. **Inequality and exclusion** or **empathy** or **inclusion**.
5. **Gender equality** or **discrimination**.
6. **Conflicts in society**.
7. **Society opportunities** or **digital culture**.
8. **Cultural diversity** or **languages and culture**.
9. **Creativity**.
10. **Collective** or **collective engagement**.[31]
11. **Lifelong learning**.

After using the shortened version of the key competences as text input, the output recommended systems are more related to the competences introduced and, therefore, more accurate.

[31] Initially, for the key competence number 10, the short version proposed was **collective project**. However, since most of the projects contain the word 'project' in their description, some non-related projects were recommended. Therefore, the final proposed short version is just **collective** or **collective engagement**.

4.3 Criteria of selection

In this section, the method used for selecting and analysing the recommended projects after selecting the keywords for the key competences will be explained.

In the created dataset there are a total of 366 extracted projects, from which there are 363 left after eliminating the duplicates (those projects that were in both the Spanish and the Barcelonian citizen science platforms). It has been decided to analyse 20% of the projects, which would be approximately a total of 73 projects. Taking into account that there are 11 different key competences and we want to analyse different projects recommended by each of them, the process to select the projects to analyse consists of selecting the top 7 recommended projects for each competence, checking manually whether they would be useful for accomplishing the respective competence or not, that being whether the project recommendation falls into the category of 'True Positive' (TP) or 'False Positive' (FP).

However, checking whether the recommendations are True Positives or False Positives is not enough to get an accurate analysis, since it is not being taken into consideration whether there are projects that match the key competence and are not being recommended. In order to improve it, the recommended projects will also be analysed to check whether they could match any other key competence in which they were not one of the proposed projects, that being whether the project recommendation falls into the category of 'True Negative' (TN) or 'False Negative' (FN).

Knowing all this information obtained from the manual analysis, a confusion matrix can be built, hence allowing the computation of the precision and recall values, as well as the accuracy (F-measure). A confusion matrix is a performance metric used to evaluate machine learning classification problems where the output can have two or more classes. It consists of a table with 4 different combinations of predicted and actual values:

- True Positive (TP): predicted positive and it is true, that being a recommended project that has relation with the key competence.
- False Positive (FP): predicted positive and is not true, that being a recommended project that has no relation with the key competence.
- True Negative (TN): predicted negative and it is true, that being a project that has not been recommended and it has no relation with the key competence.
- False Negative (FN): predicted negative and it is not true, that being a project that has not been recommended and it has relation with the key competence.

As mentioned, the precision, recall and accuracy values can then be computed from the above combinations.

- Precision: represents the percentage of actual positive values from all the ones that have been predicted, representing therefore the percentage of correctly recommended projects. It is calculated by dividing the total TP values by the sum of TP and FP values (*Precision = TP/(TP+FP)*). The precision value is between 0 and 1, and should be as high as possible.
- Recall: represents the percentage of predicted positive values from all the actual positive values, representing therefore the percentage of correctly recommended projects from all the existing projects that are related with the key competence. It

is calculated by dividing the total TP values by the sum of TP and FN values (*Recall = TP/(TP+FN)*). The recall value is between 0 and 1, and should be as high as possible.

- Accuracy: represents the percentage of correctly predicted values from both positive and negative classes. The accuracy value should be as high as possible, and it is calculated as: *F-Measure = (2*Precision*Recall)/(Precision + Recall)*.

After carrying out the mentioned analysis, which will be explained in detail in section 4.5, a threshold regarding the cosine similarity value will de defined, so that only those projects that have a similarity value greater than the threshold are recommended, hence being accurate recommendations.

4.4 Results

In this section, one of the results of the recommendation system will be shown.
The key competence with most accurate recommendations is the eighth competence, where the top 5 projects have been correctly recommended, as they are accurate and can be used to create learning activities. This competence aims to promote creativity among the students. The correctly recommended projects are the following, including the project name and the relevant part of the project description and goal:

- **Kid's KitCar**. Through the design, construction and competition of electric cars we make the kids learn team management, project management, financial management, creativity, design, ...
- **Zaragoza Activa**. We are a public ecosystem of people, companies and projects to promote entrepreneurship, creativity and citizen innovation.
- **Convocatoria CeSAr-Etopia Labs**. During the past year, the Etopia laboratories were equipped with technical equipment from the Institute for Biocomputing and Physics of Complex Systems (BIFI) with the aim of promoting research in citizen science, bringing science, technological creativity and art closer to new media to citizens, promote collaborative knowledge and consolidate Etopia as a production center for multidisciplinary projects.
- **SMART OPEN LAB**. SOL is an open technology development space, focused on digital manufacturing technologies and rapid prototyping. Currently, the SOL community is made up of around a hundred people with very different degrees of involvement, developing their creativity and satisfying their curiosity. It is a space to learn and to do, each contributing their knowledge and skills. A space for students and any restless apprentice, a space for teachers and designers. A space that aims to mix art and technology.
- **«CIUDADES que CUIDAN»**. A caring city must be a benchmark where its citizens can age healthily and participate actively co-creating the conditions, services and structures, in improving the common good. It must allow the integration of values and processes to address the end of life in peace and dignity, framed in an environment of innovation and knowledge based on creativity and high technology, and committed to promoting the health of its citizens.

4.5 Analysis of the recommended projects

In this section, the analysis of the top 7 recommended projects for each of the key competences will be shown, and the results will be explained in detail.

The recommended projects have been analysed following the mentioned method in section 4.3, resulting in the following confusion matrix:

KEY COMPETENCE	TP	FP	TN	FN	PRECISION	RECALL	F-MEASURE
KC 1	4	3	50	8	0,5714285714	0,3333333333	0,4210526316
KC 2	3	4	58	0	0,4285714286	1	0,6
KC 3	3	4	53	5	0,4285714286	0,375	0,4
KC 4	0	7	50	8	0	0	0
KC 5	1	6	52	6	0,1428571429	0,1428571429	0,1428571429
KC 6	1	6	53	5	0,1428571429	0,1666666667	0,1538461538
KC 7	1	6	50	8	0,1428571429	0,1111111111	0,125
KC 8	5	2	48	10	0,7142857143	0,3333333333	0,4545454545
KC 9	1	6	50	8	0,1428571429	0,1111111111	0,125
KC 10	3	4	53	5	0,4285714286	0,375	0,4
KC 11	2	5	48	10	0,2857142857	0,1666666667	0,2105263158

Table 4.1. *Confusion matrix and precision, recall and accuracy values of the recommended projects.*

To check which have been the top 7 recommended projects from each key competence and the analysis performed, the excel sheet 'Analysis of the recommended projects' can be checked (see location of the excel sheet in Appendix B).

It can be observed in *Table 4.1* that the recommended projects from the key competences 4, 5, 6, 7, 9 and 11did not have good results given that there are almost no True Positive values, and an improvement in the recommendation system should be thought of in order to get better results and increase the number of correctly recommended projects.

In this scenario in which we are talking about a recommendation system, it is important to take special attention to the precision value, as this metric is the one informing about how many recommended projects have been correctly recommended from all the predicted ones. Given that only the top 7 recommended projects – those 7 projects with higher value of the cosine similarity – have been taken into account for the analysis, the recall value should not be taken into consideration. The reason for doing so is that those projects that have fallen into the False Negative category, which are affecting negatively to the recall value, could actually have been recommended by the system for the respective key competence, but have not been shown in the recommended projects since they are not in the top 7. Therefore, only those projects that have obtained a cosine similarity of 0 and have fallen into the category of FN should be considered as it. Consequently, this affects the accuracy value as well. Therefore, the only metric that can be reliable without taking into account the cosine similarity values is the precision.

Key Competence #1. Keyword used: environment.

With 4 of the recommended projects being TP and 3 FP, the precision value is 57%.

The recommended projects for this key competence were all related to the environment, such as the project 'Parlamento joven'[32] that consists of carrying out a bioblitz in the natural environment of the river, or the project 'Public Lab'[33] that finances initiatives related to the environment. However, these projects do mention the environment, but do not help accomplish the key competence, as they do not help understand the causes that contribute to the environment degradation, worsen it, or improve it. On the other hand, the other recommended projects do help understand the causes and help raise environmental awareness.

To analyse how this recommendation could be improved, the FN cases have been analysed and it has been observed that most project descriptions that match the key competence contain the keywords 'awareness' and 'sensitization' along the word 'environment'. Therefore, it has been checked how the recommended projects change if the used keyword is changed to 'environment awareness' and it results in 6 projects being TP and only 1 being FP, resulting in a 86% precision.

Final keyword: 'environmental awareness'.

Key Competence #2. Keyword used: consumption of local products.

With 3 of the recommended projects being TP and 4 FP, the precision value is 43%.
Because the keyword contained the word 'consumption' in it, some of the projects that appeared as recommended contained ideas such as 'energy consumption' or 'tobacco consumption', which were not related with the main idea of the key competence.

Given that there is only 1 project that has been identified as FN, there is no possibility of analysing more deeply with the objective to find what the incorrectly non-recommended projects have in common in order to improve the query. However, given that the mistakes in the recommended projects all originate from the keyword containing the word 'consumption', it has been checked whether there are more correctly recommended projects if the word is eliminated from the keyword. However, when doing so, all the recommended projects turn out to be False Positives. Therefore, given that in this case there is only one FN, the initial keyword is kept.

Key Competence #3. Keyword used: healthy lifestyle.

With 3 of the recommended projects being TP and 4 FP, the precision value is 43%.
The recommended projects for this key competence all contained the word 'life'. However, not all of them used it in the concept the keyword intends. For example, the project 'Banco de Imágenes del Mundo rural' aims to preserve, know and spread the rural heritage of Burgos, so it talks about concepts such as 'the life of our towns', 'traditional ways of life' and 'moments of everyday life'. Other non-related projects talk about the 'insects life cycle' and 'how to improve life in cities'.

To analyse how this recommendation could be improved, the FN cases have been analysed and it has been observed that most project descriptions that match the key competence contain the keywords 'diet', 'nutrition', 'health', 'physical activity' and

[32] Parlamento joven https://ciencia-ciudadana.es/proyecto-cc/parlamento-joven/.
[33] Public Lab https://ciencia-ciudadana.es/proyecto-cc/public-lab/.

‘wellbeing’. First, it has been tried to use the keyword ‘healthy diet’, resulting in only 2 TP cases. At the same time, the keyword ‘physical activity’ does not produce any good recommendation, as the word ‘activity’ affects the result since all the proposed projects are related to participating ‘actively’ in the project or proposing and participating in citizen science ‘activities’.

Since the option that gives the best precision value is the actual keyword, it is not changed.

Key Competence #4. Keyword used: inequality, exclusion and empathy.

With 0 of the recommended projects being TP, the precision value is 0%.
Given that no projects have matched the concept implied by the key competence, those projects that fall into the category of FN have been analysed. The ideas that are contained in the description of those projects that make them a good match are ‘social problems’, ‘reflection space’, ‘civic society’, ‘social awareness’, ‘limitations’ and ‘barrier’ or ‘disability’ or ‘functional diversity’. Because of the word ‘problem’ being in many project descriptions referring to all different aspects but the social ones, ‘social problems’ cannot be used as a keyword. Similarly, in ‘social awareness’, the results all refer to environmental awareness. On the other hand, when using ‘disability’ as the keyword, the first 3 recommended projects are TP, since they talk about social inclusion. Also, when using the keyword ‘social inclusion’, 4 TP projects are obtained, from which 3 are the ones also recommended when using the word ‘disability’. This results in a 57% precision.

Final keyword: ‘social inclusion’.

Key Competence #5. Keyword used: gender equality.

With 1 of the recommended projects being TP and 6 FP, the precision value is 14%.
Given the scarcity of True Positives, the FN projects have been analysed to see what they all have in common and figure out why the actual keyword is not effective. Some of the concepts that appear in the False Negative projects are ‘reflection space’, ‘female’, ‘accessible’, ‘people rights’, ‘gender’, ‘functional diversity’, ‘sign language’ and ‘citizen biodiversity’. By using the keyword ‘gender’, most of the recommended projects turn out to be True Positives, having only 2 FN in the top 7, which implies a precision value of 72%, which is a noticeable increment.

Final keyword: ‘gender’.

Key Competence #6. Keyword used: conflicts in society.

With 1 of the recommended projects being TP and 6 FP, the precision value is 14%.
Given the scarcity of True Positives, an analysis has been carried out to check the FN projects for any possible keywords that would improve the results of the recommendation system for this key competence. Those possible keywords found by analysing the FN projects are ‘social problems’, ‘social awareness’, ‘civic society’, ‘solution proposal’, ‘people rights’, ‘problem detection’, ‘social perspective’ and ‘support’. As previously mentioned in the analysis of the 4th key competence, ‘social problems’ (similar to ‘problem detection’) and ‘social awareness’ do not provide adequate recommendations. However, when using ‘social perspective’ as the keyword, 5 of the top 7 recommended

projects are TP, as it is used in the concept of how is the citizens' perspective of the society, which includes and makes reference to aspects such as gender and functional diversity, gender violence and social movements, among others.

Final keyword: 'social perspective'.

Key Competence #7. Keyword used: digital culture.

With 1 of the recommended projects being TP and 6 FP, the precision value is 14%.
The main reason for why project recommendations are not adequate when using this keyword is because the word 'culture' is associated with the concept of 'scientific culture' instead of the technological aspect. Some proposals based on the project descriptions of those projects falling into the category of FN are using as just the word 'digital' as the keyword, or the concept of 'artificial intelligence (AI)', 'fablabs', 'technology', 'decision taking', or 'scientific advances'. Since the concepts of AI and FabLabs are rarely used in the descriptions of the projects, these keywords produce only 3 correct recommendations. However, by using the keyword 'technology', 5 of the 7 recommended projects turn out to be TP, as they talk about promoting the use of technology and teaching people how to correctly use it, as well as showing how technology affects society.

Final keyword: 'technology'.

Key Competence #8. Keyword used: creativity.

With 5 of the recommended projects being TP and 2 FP, the precision value is 72%.
Given that most of the recommended projects have been correctly proposed, since they match the key competence, it has been decided to maintain the keyword. Most of the projects promote activities that will help the participating public put their creativity in use, whether it is in what refers to technology or society.

Key Competence #9. Keyword used: languages and cultures.

With 1 of the recommended projects being TP and 6 FP, the precision value is 14%.
Since the words 'language' and 'tongue' are written the same in Spanish, which is the language the project descriptions are written in, two of the recommended projects talk about a scientific research of finding out the presence of bacteria and fungi in people's tongues ('Saca La Lengua'[34]), which is not related to the key competence. The only correctly recommended project is 'Milmots'[35], whose objective is to collect all the words and phrases in the different variants of the Catalan language.

In order to improve the results, the False Negative projects of this competence have been analysed to gather other possible keywords. The resulting proposal contains keywords such as 'diversity', 'collaborative participation', 'cooperation' and 'world connection'. Since most of the projects are related to the environment, the diversity concept mostly refers to biodiversity, resulting in recommendations that are all wrong. Similarly, when using 'collaborative participation' all recommended projects talk about participating in

[34] Saca La Lengua https://ciencia-ciudadana.es/proyecto-cc/saca-la-lengua/.
[35] Milmots https://ciencia-ciudadana.es/proyecto-cc/milmots/.

the citizen science project no matter whether it is collaborative or not. However, when using 'cooperation' as the keyword, 3 of the 7 recommended projects talk about how people can cooperate with each other in the projects, resulting in an improvement in society as it refers to the idea of cooperating and coexisting in open and evolving societies, as the key competence states.

Final keyword: 'cooperation'.

Key Competence #10. Keyword used: collective.

With 3 of the recommended projects being TP and 4 FP, the precision value is 43%.
The reason why only 3 of the top 7 recommended projects turn out to be correctly predicted is that the word 'collective', which meaning is 'a group of people acting together', is mostly understood as an exclusive group, so it does not fulfill the goal of the key competence, which is to fell part of a collective project. Therefore, it has been decided to change the keyword to one with a similar definition but more inclusive, such as 'collaborate'. With this one, the recommended projects are more accurate, being 5 out of 7 True Positives.

Final keyword: 'collaborate'.

Key Competence #11. Keyword used: lifelong learning.

With 2 of the recommended projects being TP and 5 FP, the precision value is 29%.
Since 'learning' is a word used in most project descriptions, it is quite ambiguous and does not produce adequate recommendations. Therefore, a good option would be to use a word similar to 'learning' that implies the same meaning but is not so common, such as 'knowledge', or 'education'. By using 'knowledge' as the keyword, 4 True Positives are obtained, as it is used in the project descriptions as in gaining knowledge through the use of citizen science, but also in others with the goal of only recruiting volunteers with a certain knowledge. On the other hand, the word 'education' is used to refer to how citizen science can be used in the education field to improve education itself and to educate people in a certain field or area. With this last keyword, there are a total of 5 TP projects.

Final keyword: 'education'.

As it has been observed, many of the initial keywords did not provide accurate recommendations since they were either:

- Words that appeared in many of the project descriptions and, therefore, were not specific enough.
- Keywords that contained a verb that can be used in other scenarios that are not related to the competence.
- Words that were homonyms[36].

Taking these limitations into account, the keywords have been improved to obtain better results by making them more specific to the content that is wished to be obtained, making sure that the word and its meaning cannot be used in other scenarios and making sure it makes reference to the respective key competence.

[36] Homonyms are words which sound alike or are spelled alike but have different meanings.

Additionally, after having stated the final keywords, the cosine similarity for each of the recommended projects has been calculated, resulting in the following table:

	KC 1	KC 2	KC 3	KC 4	KC 5	KC 6
Project #1	0,18536508	0,22103930	0,18536508	0,19049071	0,30264455	0,11955611
Project #2	0,14104352	0,08347518	0,14570430	0,08513638	0,18589796	0,10083722
Project #3	0,13960967	0,08160556	0,13859927	0,06979678	0,12667906	0,09787030
Project #4	0,09059667	0,07406310	0,10301485	0,06432633	0,09929353	0,08772262
Project #5	0,09054423	0,07327929	0,10157705	0,05392965	0,06567730	0,07974859
Project #6	0,06812860	0,04952730	0,10122714	0,03302789	0,00000000	0,07799059
Project #7	0,06247686	0,04569420	0,09680735	0,00000000	0,00000000	0,07140684

	KC 7	KC 8	KC 9	KC 10	KC 11
Project #1	0,34158426	0,19064696	0,14750869	0,25221323	0,25009614
Project #2	0,33268796	0,10692479	0,08828293	0,21607832	0,18166138
Project #3	0,22178164	0,09102290	0,08133678	0,18377470	0,17111941
Project #4	0,16413042	0,07893194	0,07135871	0,15225796	0,15895923
Project #5	0,12620316	0,07369620	0,00000000	0,11639503	0,15631386
Project #6	0,11079269	0,00000000	0,00000000	0,08996711	0,15545278
Project #7	0,10705904	0,00000000	0,00000000	0,08736982	0,14635925

Table 4.2. *Cosine similarity values for each of the top 7 recommended projects.*

Given that the second key competence has been the one with less correctly recommended projects from the top 7, having only 3 TP and only 1 FN, it has been decided to set the threshold at the cosine similarity value of the third recommended project of that competence. Therefore, the cosine similarity threshold is **0,08160556**, meaning that only those projects with equal or higher cosine similarity value will be recommended.

Although the keywords have been refined, the recommender system could be improved as well. Improving a recommender system involves enhancing its accuracy, relevance, and overall user experience. Some improvement proposals that can be taken into consideration for the recommendation system are:

- **Enhanced Similarity Measures**. Taking into consideration other similarity measure techniques such as Jaccard similarity, K-Means, Euclidean distance, or BM25 ranking, as they may produce more accurate and diverse recommendations.
- **Hybrid Approaches**. This is a content-based recommendation system that has been built using text-similarity. However, considering combining multiple recommendation techniques such as content-based filtering and collaborative filtering, in a hybrid approach, could lead to better results.
- **User Feedback and Personalization**. Incorporating user feedback mechanisms like ratings, reviews, or explicit user preferences to personalize the recommendations so that they can better suit what the user is looking for.

5. WEB APPLICATION

A web application[37] has been designed in order to have a web platform where people, especially teachers, can interact with the recommender system. In this section, the structure of the platform will be explained, as well as how the user can interact with the recommendation system.

The application has three different pages: 'Citizen Science Projects', 'Key Competences' and 'About the project'. The 'Key Competences' page (See *Figure 5.2*) contains the list with all the key competences stated in the Catalan elementary school curriculum (same competences stated in Appendix C) so that any user can take a look at them before using the recommender system. The 'About the project' page (See *Figure 5.3*) contains a short introduction about what this project is about and its objective. The main page, 'Citizen Science Projects', contains the recommendation system itself so that any user can make use of it to find any citizen science projects that fulfill his or her needs.

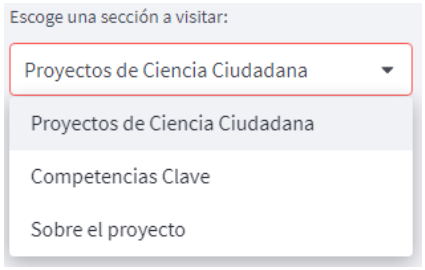


Figure 5.1. *Dropbox to access the pages.*

Before getting to the explanation of the different features of the web application, it is important to be aware that the web application has been designed in Spanish. There are two main reasons for that choice being taken:

1. The projects' descriptions are in Spanish.
2. The recommendation system is based on the key competences of the **Catalan** elementary school curriculum, meaning that the main public to which this application is aimed is Catalan/Spanish teachers/users.

The first feature that can be found in the web application is that the user can filter the dataset containing all the extracted web projects based on the project scopes. This feature is useful as it helps the user find projects that are more adequate or related to the field that the user wants to work with, hence the recommendation system can output projects that are more suitable to what the user is looking for.

Given that there are many project scopes, they have been classified into four major categories, following the research areas defined in the Web of Science Core Collection Help[38]. Classifying all the project scopes in these areas makes the web application visually more attractive to the user since it takes a smaller space to show them, as well as easier to filter the projects. The final four categories and the project scopes included in each of them are:

- Life Sciences and Biomedicine: medicine and health, biodiversity, environmental, ecology and environment, agricultural and veterinary sciences, nature and outdoors, food science, animals, birds, marine and terrestrial, biogeography, insects and pollinators, biology, and long-term monitoring of species.

[37] Citizen Science Recommender System. https://citizenscience-recommendersystem.streamlit.app/.
[38] Web of Science Core Collection Help (2023, June 8) https://images.webofknowledge.com/images/help/WOS/hp_research_areas_easca.html.

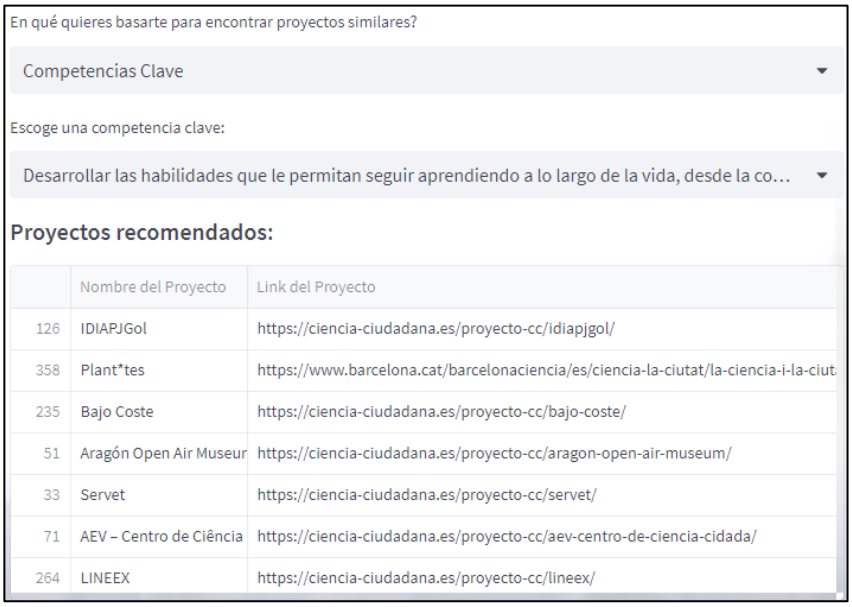
- Physical Sciences: oceans, water, physics, space and astronomy, climate and meteorology, natural resources management, geology and earth sciences, chemical sciences, and geography.
- Social Sciences: culture and archaeology, social sciences, education, social, political sciences, and indigenous cultures.
- Technology: technology and computer science, transportation, and sound.

When the web application is deployed, all the categories are selected in the filter by default (See *Figure 5.4*). Whenever the user decides to modify the filter, the displayed dataset containing the projects is modified accordingly. The dataset used for the recommendation system is modified as well, following the selected scopes requirement. Additionally, if the user decides to delete all the scopes from the filter, a message will be displayed saying 'You must choose at least one option' and the dataset a recommender system will not show until at least one option is selected.

Afterwards, the user can choose which method to use in order to find similar projects. The options to choose from are 'Key Competences' and 'Others'. By default, the selected option is 'Key Competences'.

With the 'Key Competences' option being chosen as the desired method, a dropdown list is displayed, showing a total of eleven options, each of them being a different key competence. By default, the key competence number one in the list is the selected one Whenever a user chooses one of these eleven competences, the top seven recommended projects with respect to the chosen competence are displayed in a dataframe format containing information about the name of the project, its link, the description and goal of the given project, the name of the platform from where it has been extracted, and the link to the platform.

On the other hand, if the user chooses 'Others', a space for the user to input the desired keywords to find CS projects related to it will be displayed, along with the text 'Introduce the keywords to find similar projects (e.g., wild animals)'. Same way as with the previous option, the top 7 recommended projects will be displayed in a dataframe.



	Nombre del Proyecto	Link del Proyecto
126	IDIAPJGol	https://ciencia-ciudadana.es/proyecto-cc/idiapjgol/
358	Plant*tes	https://www.barcelona.cat/barcelonaciencia/es/ciencia-la-ciutat/la-ciencia-i-la-ciut
235	Bajo Coste	https://ciencia-ciudadana.es/proyecto-cc/bajo-coste/
51	Aragón Open Air Museur	https://ciencia-ciudadana.es/proyecto-cc/aragon-open-air-museum/
33	Servet	https://ciencia-ciudadana.es/proyecto-cc/servet/
71	AEV – Centro de Ciência	https://ciencia-ciudadana.es/proyecto-cc/aev-centro-de-ciencia-cidada/
264	LINEEX	https://ciencia-ciudadana.es/proyecto-cc/lineex/

Figure 5.5. *Filtrating by key competences and result of the recommended projects.*

Then, a word cloud generated from the displayed top 7 projects' descriptions is shown (see Appendix D to look at the word clouds obtained for each one of the key competences).

"Word clouds or tag clouds are graphical representations of word frequency that give greater prominence to words that appear more frequently in a source text. The larger the word in the visual the more common the word was in the document(s)" [31]. Word clouds help analyze the text faster and easier. Therefore, when used to analyze the resulting recommended CS projects, it is less challenging to check whether the output projects are adequate and related to the input key competence or not. It has been decided to display the word cloud along the dataframe containing the recommended projects because it is a visual aid that enhances trust and user satisfaction.

Additionally, and mainly with the idea of making the application visually more attractive, a background has been added to the webpage, which can be taken off or activated again by checking the checkbox 'Show background'.

Therefore, the idea of this web application is to provide users with Citizen Science projects that fulfill their requests in order to either participate in them or create learning activities to bring to the classrooms so that the students can learn more about the searched topic and/or develop the needed key competences, at the same time that it creates a space to learn more about citizen science and encourages people to participate in the research areas.

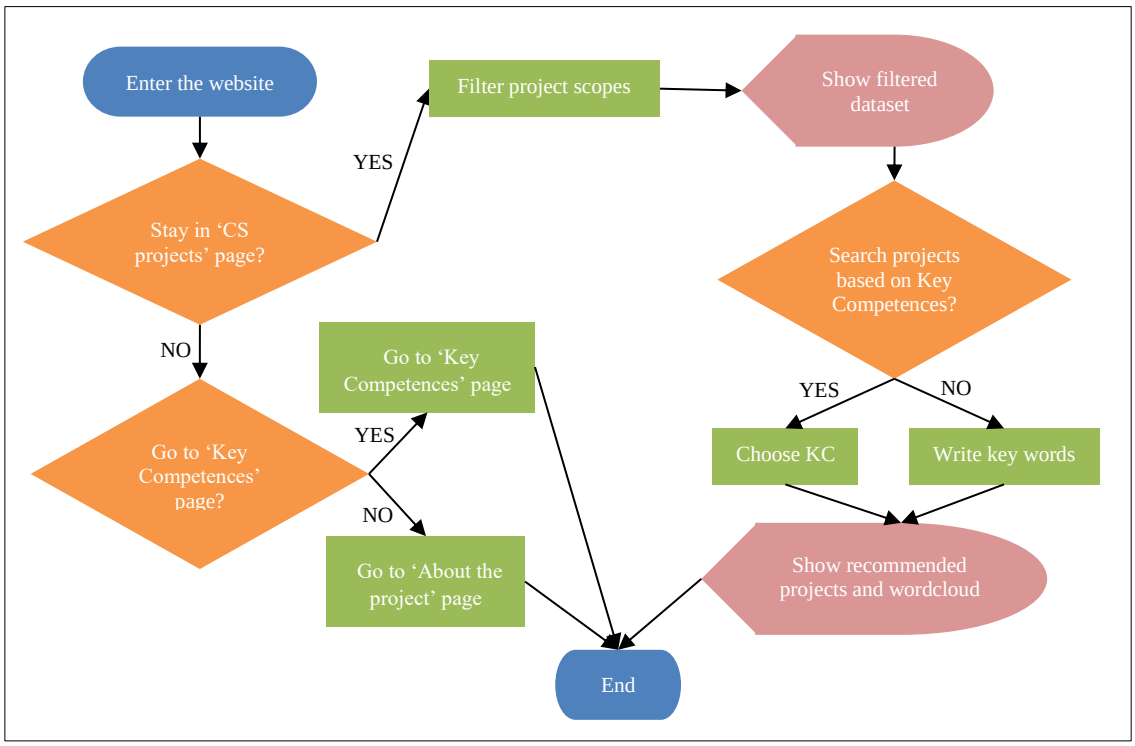


Figure 5.6. *Flux Diagram of the web application.*

6. CONCLUSIONS AND FUTURE WORK

The main objective of this project has been to create a recommendation system of citizen science projects with the goal of finding the most suited projects that can be either used to participate in or to create similar learning activities based on the recommended project to accomplish the key competences stated in the elementary school curriculum. To do so, it was necessary to analyze the citizen science projects developed in Barcelona and Spain, so that they could be extracted from the Spanish and Barcelonian as well as the key competences from the Catalan elementary school curriculum. Other objectives were to extract meaningful conclusions regarding the results obtained from the recommendation system and to give recommendations of its potential use.

It was posteriorly established as another objective the creation of a web application so that teachers or any other person interested in making use of citizen science projects to create learning activities could interact with the recommendation system.

Although all these objectives have been accomplished, there have been some limitations. It was initially thought that there would be a greater variety of project scopes, hence being able to recommend projects that could be applied in any of the subjects of the elementary school. However, there has seemed to be a scarcity of project fields, since although there are a lot of scopes, they mostly fall into the categories of science and environment. Also, the database had a total of 363 projects, which limits the diversity of recommendations.

Another limitation has been that most of the project descriptions were too short and not too explicit. This could have been fixed by extracting the information from the projects' websites instead of the citizen science online platforms so that more information could be obtained. However, this would imply having to analyse the HTML structure of each website individually as each one has its own structure and, hence, the websites would have to be parsed in a different way one from another. This would have taken more time and less projects could have been added to the database.

Despite the fact that the recommendation system and the web application have been successfully developed, it has not been possible to test the application with teachers. However, this application could have a great potential as it can easily be accessed by anyone and it provides accurate citizen science projects that can help teachers create learning activities to help students accomplish the key competences, as well as precise projects based on any keyword of their choice.

Some proposals of future work are:

- Taking into consideration other similarity measure techniques such as Jaccard similarity, K-Means, Euclidean distance, or BM25 ranking, as they may produce more accurate and diverse recommendations.
- Incorporate user feedback mechanisms like ratings, reviews, or explicit user preferences to personalize the recommendations to get more accurate results.
- Allow the creation of user accounts in the web application. With this implementation, an interface between users could be developed so that a collaborative filtering recommendation system can be implemented along with the incorporation of the user feedback, so that users that have provided similar ratings to the same projects can be recommended more accurate and interesting projects. A feature that allows users to save a project could also be implemented.

7. BIBLIOGRAPHY

1. UAB - Universitat Autònoma de Barcelona (n.d.). What is citizen science? What is Citizen Science? - Universitat Autònoma de Barcelona - UAB Barcelona. Retrieved May 28, 2023, from https://www.uab.cat/web/research/responsible-research-and-innovation/citizen-science/what-is-citizen-science-1345869944451.html.
2. Biswal, A. (2023) *Data Science unveiled: Prerequisites, applications, tools, and more in 2023*, *Simplilearn.com*. Retrieved June 4, 2023, from https://www.simplilearn.com/tutorials/data-science-tutorial/what-is-data-science.
3. In depth: The rise of citizen science (2023) Three o'clock. Retrieved June 5, 2023, from https://threeoclock.co/in-depth-the-rise-of-citizen-science.
4. European Citizen Science Association (n.d.) European Citizen Science Association (ECSA). Retrieved June 5, 2023, from https://www.ecsa.ngo/.
5. Havens, K. and Henderson, S. (2013) Citizen science takes root, Northwestern Scholars. Retrieved June 6, 2023, from https://www.scholars.northwestern.edu/en/publications/citizen-science-takes-root.
6. Meyrueix, L. (2018) Science is not just for scientists, The Pipettepen. Retrieved June 6, 2023, from http://www.thepipettepen.com/science-is-not-just-for-scientists/.
7. Shah, H.R. and Martinez, L.R. (2016) 'Current approaches in implementing citizen science in the classroom', Journal of Microbiology & Biology Education, 17(1), pp. 17–22. doi:10.1128/jmbe.v17i1.1032.
8. Educació Primària (n.d.) XTEC. Retrieved May 25, 2023, from https://xtec.gencat.cat/ca/curriculum/primaria.
9. El Decret d'educació Bàsica (September 27, 2022) – El nou currículum. Retrieved May 25, 2023, from https://projectes.xtec.cat/nou-curriculum/educacio-basica/decret-educacio-basica/.
10. What is natural language processing? (n.d.) IBM. Retrieved June 8, 2023, from https://www.ibm.com/topics/natural-language-processing.
11. Natural language processing (NLP): What it is and why it matters (n.d.) SAS. Retrieved June 8, 2023, from https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html.
12. Understanding TF-ID: A simple introduction (2019) MonkeyLearn Blog. Retrieved June 9, 2023, from https://monkeylearn.com/blog/what-is-tf-idf/.
13. Baeldung, W. by: (2023) Semantic similarity of two phrases, Baeldung on Computer Science. Retrieved June 9, 2023, from https://www.baeldung.com/cs/semantic-similarity-of-two-phrases.
14. What is a recommendation system? (n.d.) NVIDIA Data Science Glossary. Retrieved May 30, 2023, from https://www.nvidia.com/en-us/glossary/data-science/recommendation-system/.
15. Recommendation systems: Applications and examples in 2023 (January 17, 2023) AIMultiple. Retrieved May 30, 2023, from https://research.aimultiple.com/recommendation-system/.
16. Python Software Foundation. (n.d.). Python. Retrieved May 20, 2023, from https://www.python.org/.
17. Pandas Documentation. (n.d.). Retrieved May 20, 2023, from https://pandas.pydata.org/docs/.
18. Requests Documentation. (n.d.). Retrieved May 20, 2023, from https://docs.python-requests.org/en/latest/.
19. NLTK Documentation. (n.d.). Retrieved May 31, 2023, from https://www.nltk.org/.

20. Python Documentation. (n.d.). re — Regular expression operations. Retrieved May 31, 2023, from https://docs.python.org/3/library/re.html.
21. Scikit-learn Documentation. (n.d.). TfidfVectorizer. Retrieved May 31, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.
22. Scikit-learn Documentation. (n.d.). cosine_similarity. Retrieved May 31, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html.
23. Wordcloud Documentation. (n.d.). Retrieved May 31, 2023, from https://amueller.github.io/word_cloud/.
24. Matplotlib Documentation. (n.d.). Retrieved May 31, 2023, from https://matplotlib.org/.
25. BeautifulSoup Documentation. (n.d.). Retrieved May 20, 2023, from https://www.crummy.com/software/BeautifulSoup/bs4/doc/.
26. Streamlit Documentation. (n.d.). Retrieved June 3, 2023, from https://docs.streamlit.io/.
27. Project Jupyter. (n.d.). Jupyter Notebook. Retrieved May 20, 2023, from https://jupyter.org/.
28. Visual Studio Code Documentation. (n.d.). Retrieved June 3, 2023, from https://code.visualstudio.com/docs.
29. GitHub. (n.d.). About GitHub. Retrieved May 20, 2023, from https://github.com/about.
30. Bujokas, E. (2022) Text classification using word embeddings and deep learning in python-classifying tweets from Twitter. Retrieved May 31, 2023, from https://towardsdatascience.com/text-classification-using-word-embeddings-and-deep-learning-in-python-classifying-tweets-from-6fe644fcfc81.
31. Word cloud (n.d.) Better Evaluation. Retrieved May 31, 2023, from https://www.betterevaluation.org/methods-approaches/methods/word-cloud.

8. APPENDICES

8.1 APPENDIX A

In this appendix there are all the code screenshots that show the CS projects' extraction process and the creation of the recommendation system. There will also be found in this section the screenshots of the web application.

```
url = "https://ciencia-ciudadana.es/proyecto-cc/"
web_name = 'Observatorio de la Ciencia Ciudadana en España'

# Adding headers to the rquest
headers = {'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) '
                         'AppleWebKit/537.36 (KHTML, like Gecko) '
                         'Chrome/58.0.3029.110 Safari/537.36'}

response = requests.get(url, headers=headers)
```

Figure 3.1. *Code to send HTTP request to the Spanish website.*

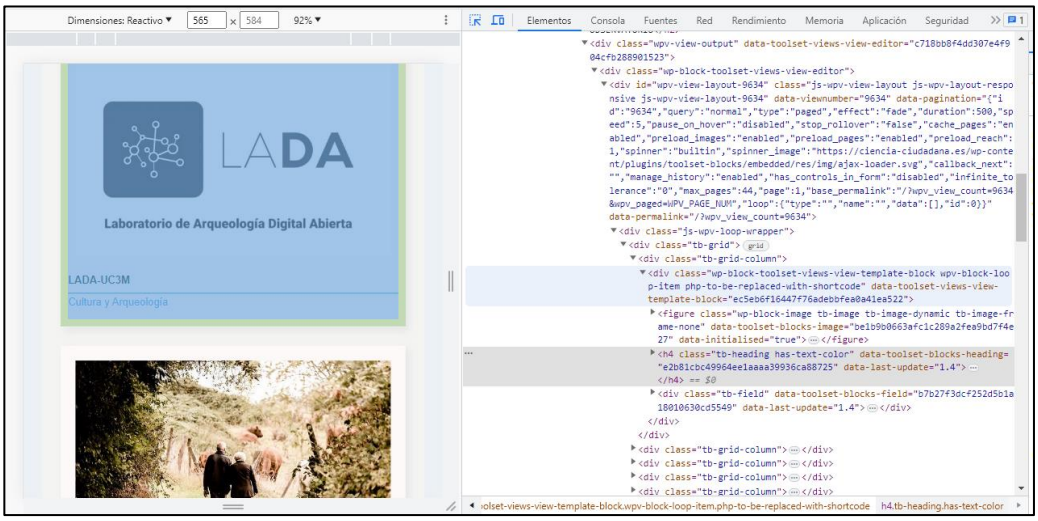


Figure 3.2. *Elements of the Spanish platform website.*

```
soup = BeautifulSoup(response.content, "html.parser")
```

Figure 3.3. *Code to parse HTML content.*

```
# Find all elements with class name "underline"
underline_elements = soup.find_all('img', {'decoding': 'async'})

links = []
# Extract the links from the parent elements
for element in underline_elements:
  parent_a_tag = element.find_parent('a')
  if parent_a_tag and 'href' in parent_a_tag.attrs:
    link = parent_a_tag['href']
    links.append(link)
```

Figure 3.4. *Code to extract projects' URLs (Spanish platform).*

```
def get_complete_section(proj_soup, dtbf):
  proj_seg = proj_soup.find_all('div', {'class': 'tb-field', 'data-toolset-blocks-field': dtbf})
  return ''.join([seg.text for seg in proj_seg])
```

Figure 3.5. *Function get_complete_section(proj_soup, dtbf).*

```
def get_project_info1(project_link, main_url, main_name):
  proj_info = ['', '', '', '', '', '', '', '', '', '', '', '', '', '', '', '', '', '', '']

  response = requests.get(project_link, headers=headers)
  proj_soup = BeautifulSoup(response.content, "html.parser")
  proj_title = proj_soup.find('h1', {'class': 'entry-title'}).text
  proj_scope = get_complete_section(proj_soup, "28515d8ce1fa37a6527af15754983e83")
  proj_goal = get_complete_section(proj_soup, "ba74407e8b20cac888e283e8576140f9")
  proj_desc = get_complete_section(proj_soup, "7147dde37d9b86b4d4a2dc89b9c12945")
  proj_entity = get_complete_section(proj_soup, "486908e9509d0a139e234cbfb7f8f47d")
  proj_join = get_complete_section(proj_soup, "687686b4b9d6f02068b541e6bc6f2812")
  proj_equip = get_complete_section(proj_soup, "cbcd66c16990127ea3c06cebc87c17f0")
  proj_ini = get_complete_section(proj_soup, "3d66c4d00ab960c97a58ac752fe406f4")
  proj_end = get_complete_section(proj_soup, "9a00b649618c1cd0c0cedd6cfc386f09")
  proj_public = get_complete_section(proj_soup, "e9e0eef27b43baa290bdc058d0ec6cee")
  proj_loc = get_complete_section(proj_soup, "05399f19f4997ca4dc2a1f1770db2d80")
  proj_amt_part = get_complete_section(proj_soup, "5ee97fadc3c34313fff123d0cba66a1a")
  proj_results = get_complete_section(proj_soup, "b064a4637177bd64d367dc70864be32c")

  proj_link_res = proj_soup.find_all('div', {'class': 'tb-field', 'data-toolset-blocks-field': "c5d8b6598ffd2835065c71792c189772"})
  proj_link_res = [seg.text for seg in proj_link_res][0]

  proj_impact = get_complete_section(proj_soup, "4297f88ef1601b65151fcf8fc6fccadd")
  proj_useCC = get_complete_section(proj_soup, "974e65ee4de5d532951d366fa43dc400")
```

Figure 3.6. *Code to extract all the necessary fields of a project (Spanish platform).*

```
for link in links:
  get_project_info1(link, url, web_name)
```

Figure 3.7. *Code to iterate through the links array (Spanish platform).*

```
# Create an empty DataFrame with specified columns
df1 = pd.DataFrame(columns=['Project Name', 'Project Link', 'Project Scope', 'Project Goal', 'Project Description',
                            'Project Entity/Scientist', 'How To Join', 'Necessary Equipment', 'Initial Date',
                            'Final Date', 'Public Type', 'Location (Province)', 'Number of Participants', 'Results',
                            'Link to Results', 'Project Impact', 'Why Using CC?', 'Citizen Science Web Name',
                            'Citizen Science Web Link'])
```

Figure 3.8. *Creation of the DataFrame to store the Spanish platform projects.*

```
url = "https://www.barcelona.cat/barcelonaciencia/es/ciencia-en-la-ciudad/la-ciencia-y-la-ciudadania/ciencia-ciudadana"
web_name = 'Ciencia Ciudadana Ayuntamiento de Barcelona'

# Adding headers to the request
headers = {
    'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) '
    'AppleWebKit/537.36 (KHTML, like Gecko) Chrome/58.0.3029.110 Safari/537.36'}

response = requests.get(url, headers=headers)
soup = BeautifulSoup(response.content, "html.parser")
```

Figure 3.9. *Code to send HTTP request to the Barcelonian website.*

Figure 3.10. *Elements of the Barcelonian platform website.*

```
links = []
# Find all elements with class name "underline"
underline_elements = soup.find_all('a', {'class': 'item-link'})

# Extract the links from the parent elements
for element in underline_elements:
    link = element['href']
    links.append('https://www.barcelona.cat'+link)
```

Figure 3.11. *Code to extract projects' URLs (Barcelonian platform).*

```
def get_project_info2(project_link, main_url, main_name):

    response = requests.get(project_link, headers=headers)
    soup = BeautifulSoup(response.content, "html.parser")

    proj_name = soup.find('h2', {'class': 'page-title'})
    if proj_name is not None:
      proj_name = proj_name.text.replace('\n  ', '').replace('\n', '')
    else:
      print("Project Name not found or is None")

    proj_info = soup.find('div', {'class': "field__item entradeta"})
    if proj_info is not None:
      proj_info = proj_info.text.replace('\n  ', '').replace('\n', '')
    else:
      print("Project Info not found or is None")

    proj_desc = soup.find('div', {'class': "field field--name-body field--type-text-with-summary field--label-hidden"})
    elements = proj_desc.find_all('p')
    text = [p.get_text(strip=True) for p in elements]

    desc, state, activities, scope, results = get_segments(text)
```

Figure 3.12. *Code to extract all the necessary fields of a project (Barcelonian platform).*

```
def get_segments(text):
  state = ''
  desc = ''
  activities = ''
  scope = ''
  results = ''

  for i in text:
    if i.startswith('Estado'):
      state = i.replace('Estado:', '')
    elif i.startswith('Actividades en el marco de la Oficina:'):
      activities = i.replace('Actividades en el marco de la Oficina:', '')
    elif i.startswith('Ámbito:'):
      scope = i.replace('Ámbito:', '')
    elif i.startswith('Dónde visualizar los datos recogidos:'):
      results = i.replace('Dónde visualizar los datos recogidos:', '')
    elif i == '':
      continue
    else:
      desc = desc + i

  return desc, state, activities, scope, results
```

Figure 3.13. *Code to extract the state, description, scope and results (Barcelonian platform).*

```
for link in links:
  get_project_info2(link, url, web_name)
```

Figure 3.14. *Code to iterate through the links array.*

```
df2 = pd.DataFrame(columns=['Project Name', 'Project Link', 'Project Description', 'Project Info',
                            'Project State', 'Link to Project Data', 'Activities within the framework of the Office',
                            'Project Scope', 'Citizen Science Web Name', 'Citizen Science Web Link'])
```

Figure 3.15. *Creation of the DataFrame to store the Barcelonian platform projects.*

```python
def build_terms(line):
    stemmer = PorterStemmer()
    stop_words = set(stopwords.words("spanish"))
    line = line.lower()  #Convert to lowercase
    line = line.split()  # Tokenize the text to get a list of terms
    line = [x for x in line if x not in stop_words]  # eliminate the stopwords
    line = [x for x in line if x.startswith(("@", "https://", "$", '#')) != True]
    line = [re.sub('[^a-zàéíóúäëïöü]+', '', x) for x in line] # since it's in spa
    line = [stemmer.stem(word) for word in line] # perform stemming
    return line
```

Figure 4.1. *Code to preprocess text data.*

```python
KC = input("Please enter the key competence: ")
```

Figure 4.2. *Code to input the key competence.*

```python
projectsCS_clean['Project Full Description'].apply(build_terms)
KC = build_terms(KC)
```

Figure 4.3. *Preprocessing project descriptions and key competence.*

```python
vectorizer = TfidfVectorizer()
text_embeddings = vectorizer.fit_transform(projectsCS_clean['Project Full Description'])
input_embedding = vectorizer.transform(KC)
```

Figure 4.4. *Code to create the text embeddings.*

```python
similarities = cosine_similarity(input_embedding, text_embeddings)
```

Figure 4.5. *Code to calculate the cosine similarity.*

Competencias Clave del currículum de primaria de Cataluña

Las competencias clave que se definen en el currículum son las siguientes:

1. Desarrollar una actitud responsable a partir de la toma de conciencia de la degradación del medio ambiente basada en el conocimiento de las causas que la provocan, agravan o mejoran, desde una visión sistémica, tanto local como global.

2. Identificar los distintos aspectos relacionados con el consumo responsable y de productos de proximidad, valorando sus repercusiones sobre el bien individual y el común, juzgando críticamente las necesidades y los excesos y ejerciendo un control social ante la vulneración de sus derechos como consumidor.

3. Desarrollar hábitos de vida saludable a partir de la comprensión del funcionamiento de el organismo y la reflexión crítica sobre los factores internos y externos que inciden, asumiendo la responsabilidad personal en la promoción de la salud pública, incluido el conocimiento de una sexualidad positiva, respetuosa e igualitaria.

Figure 5.2. *'Key Competences' page of the web application.*

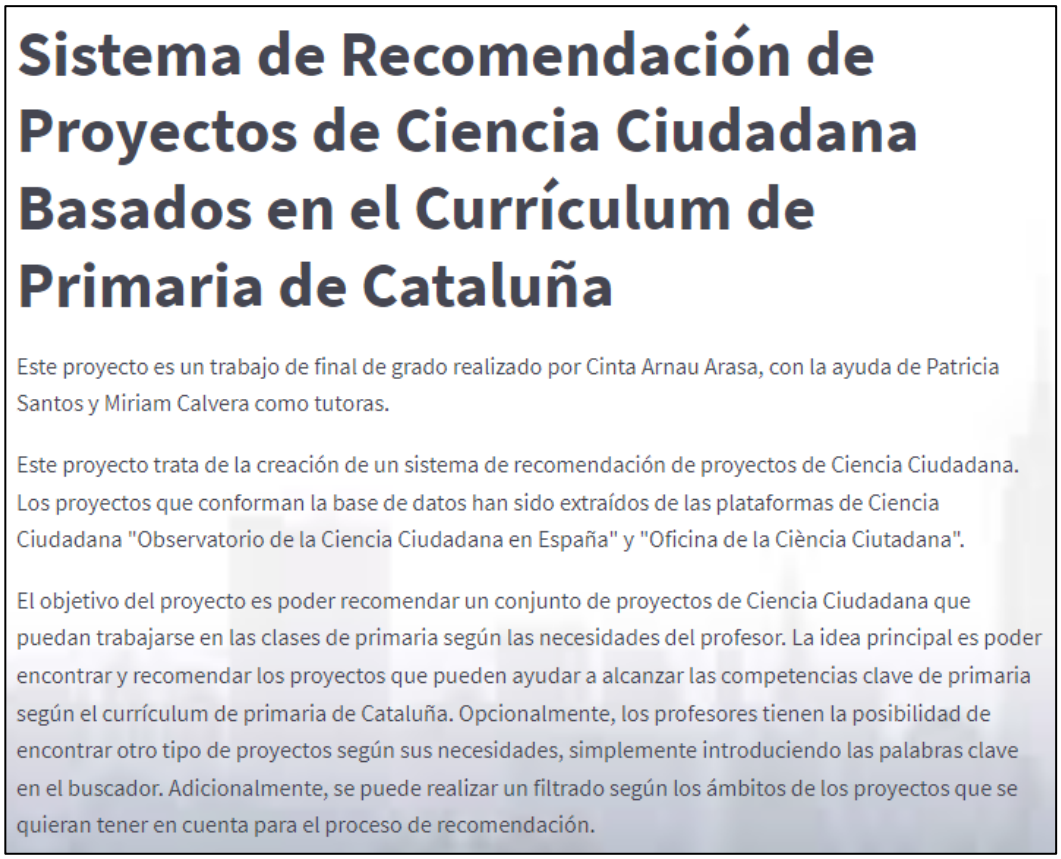


Figure 5.3. *'About the project' page of the web application.*

Figure 5.4. *'Citizen Science Projects' page. Filtrating by project scopes.*

8.2 APPENDIX B

Script	Location of the Script (GitHub)
Project Extraction	https://github.com/Cintaa1223/TFG/blob/main/projects_extraction.ipynb
Recommender System	https://github.com/Cintaa1223/TFG/blob/main/CS_Projects_Rec_Sys.ipynb
Web application	https://github.com/Cintaa1223/TFG/tree/main/Webapp
URL of the Web Application	https://citizenscience-recommendersystem.streamlit.app/
Analysis of the recommended projects	https://github.com/Cintaa1223/TFG/blob/main/Analysis%20Recommended%20Projects.xlsx

8.3 APPENDIX C

List of the key competences stated in the Catalan elementary school curriculum:

1. To develop a responsible attitude based on the awareness of environmental degradation, based on the understanding of the causes that contribute to it, worsen it, or improve it, from a systemic perspective, both locally and globally.

2. To identify the different aspects related to responsible consumption and local products, assessing their repercussions on individual and common good, critically judging the needs and excesses.

3. To develop healthy lifestyle habits based on the understanding of how the body functions and the critical consideration of the internal and external factors that influence it, taking personal responsibility for promoting public health, including the knowledge of a positive, respectful, and egalitarian sexuality.

4. To exercise the sensibility to detect situations of inequality and exclusion from the comprehension of the complex causes behind them to develop feelings of empathy.

5. To develop an active commitment to gender equality, equal treatment, and non-discrimination, knowing the historical journey towards achieving human rights for all individuals and groups.

6. To understand conflicts as inherent elements of life in society that need to be resolved peacefully and rejecting any expression of misogynistic, LGBTQ+-phobic, racist violence, motivated by any type of personal or socioeconomic circumstances.

7. To analyze critically and take advantage of all types of opportunities offered by today's society, particularly those related to digital culture, assessing their benefits and risks, and making an ethical and responsible use of them that contributes to the improvement of both personal and collective life quality.

8. To accept uncertainty as an opportunity to generate more creative responses, learning to manage the anxiety it may bring.

9. To cooperate and coexist in open and evolving societies, valuing personal and cultural diversity as a source of enrichment and promoting the interest in other languages and cultures.

10. To feel part of a collective project, both locally and globally, developing empathy and generosity.

11. To develop the skills that allow lifelong learning, based on the confidence in knowledge as a driving force for development and the critical evaluation of the risks and benefits of this knowledge.

8.3 APPENDIX D

In this section there are all the word clouds obtained from the top 7 recommended projects of each one of the key competences.

Figure 8.3.1. *Word cloud for key competence 1. 'Environmental awareness'.*

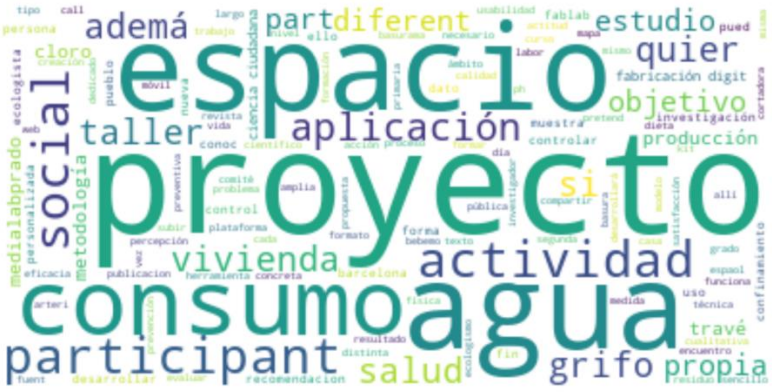


Figure 8.3.2. *Word cloud for key competence 2. 'Consumption of local products'.*

Figure 8.3.3. *Word cloud for key competence 3. 'Healthy lifestyle'.*

Figure 8.3.4. *Word cloud for key competence 4. 'Social inclusion'.*

Figure 8.3.5. *Word cloud for key competence 5. 'Gender'.*

Figure 8.3.6. *Word cloud for key competence 6. 'Social perspective'.*

Figure 8.3.7. *Word cloud for key competence 7. 'Technology'.*

Figure 8.3.8. *Word cloud for key competence 8. 'Creativity'.*

Figure 8.3.9. *Word cloud for key competence 9. 'Cooperation'.*

Figure 8.3.10. *Word cloud for key competence 10. 'Collaborate'.*

Figure 8.3.11. *Word cloud for key competence 11. 'Education'.*