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

TFG - Mathematical Engineering in Data Science

03/07/2023
Cinta Arnau Arasa - cinta.arnau01@estudiant.upf.edu
Supervisors: Patricia Santos & Miriam Calvera

upf. Universitat
Pompeu Fabra
Barcelona
Escola
d'Enginyeria

OVERVIEW

- INTRODUCTION
- STATE OF THE ART
- IMPLEMENTATION
- RECOMMENDATION SYSTEM
- KEY COMPETENCES
- RESULTS & ANALYSIS
- WEB APPLICATION
- CONCLUSIONS
- FUTURE WORK

01

INTRODUCTION

Context & Objectives

CITIZEN SCIENCE

General Public

Participate

Research

Learn

Projects

Engage

Scientists

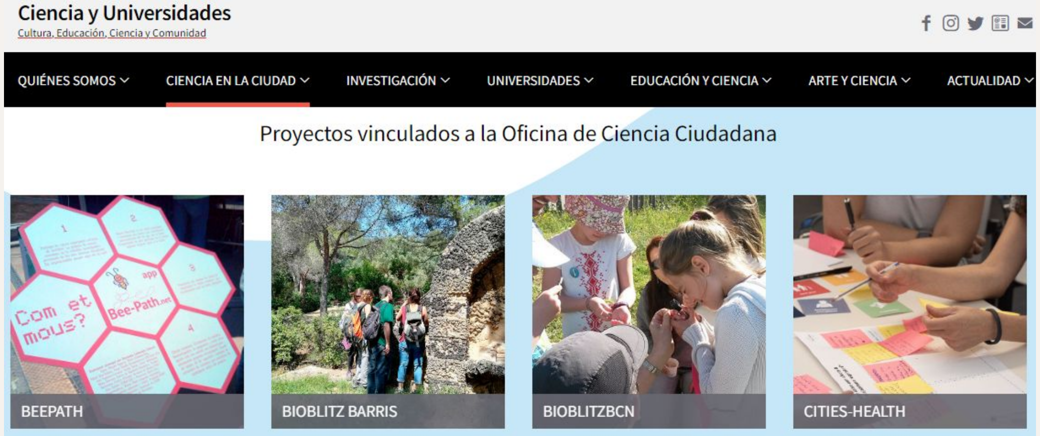
Inclusivity

Cooperate

Data Collection

Data Interpretation

Data Analysis



Ciencia y Universidades

Cultura, Educación, Ciencia y Comunidad

Proyectos vinculados a la Oficina de Ciencia Ciudadana

BEEPATH

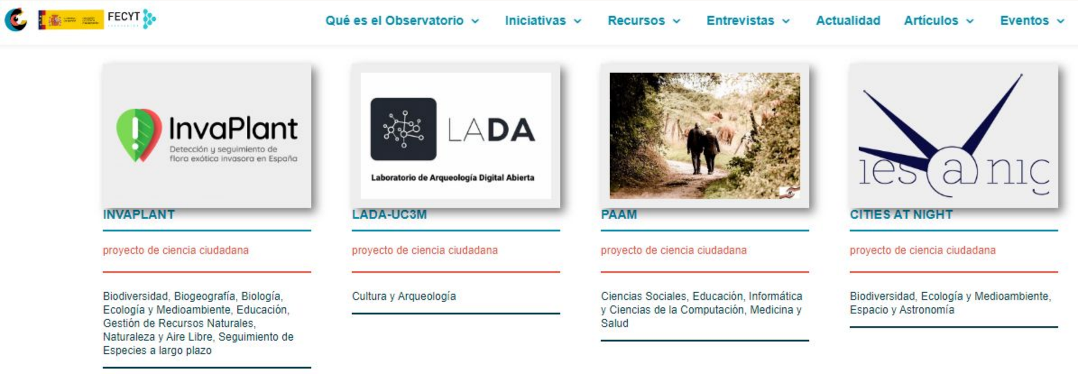
BIOBLITZ BARRIS

BIOBLITZBCN

CITIES-HEALTH

“Oficina de la Ciència Ciutadana”

“Observatorio de la Ciencia Ciudadana en España”



INVAPLANT

proyecto de ciencia ciudadana

Biodiversidad, Biogeografía, Biología, Ecología y Medioambiente, Educación, Gestión de Recursos Naturales, Naturaleza y Aire Libre, Seguimiento de Especies a largo plazo

LADA-UC3M

proyecto de ciencia ciudadana

Cultura y Arqueología

PAAM

proyecto de ciencia ciudadana

Ciencias Sociales, Educación, Informática y Ciencias de la Computación, Medicina y Salud

CITIES AT NIGHT

proyecto de ciencia ciudadana

Biodiversidad, Ecología y Medioambiente, Espacio y Astronomía

CATALAN ELEMENTARY SCHOOL CURRICULUM

Describes the **objectives**, **contents** and **evaluation criteria** of each subject..

The **goal** is to achieve the **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 and global challenges and demands.“

OBJECTIVES

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.

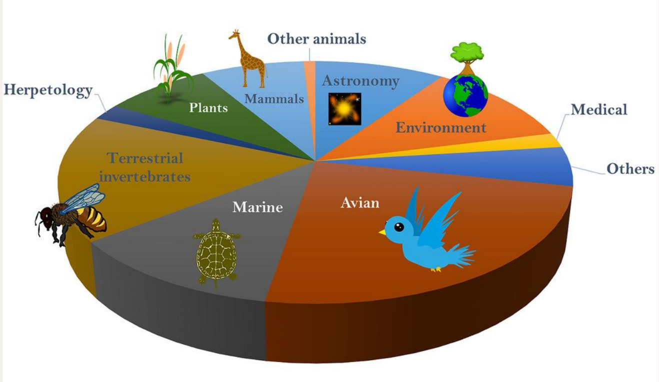
02

STATE OF THE ART

Citizen Science, Catalan elementary school curriculum, Automatic Extraction of Information & Recommendation Systems

CITIZEN SCIENCE

1. Observational citizen science.
2. Participatory citizen science.
3. Collaborative citizen science.
4. Citizen-led science.
5. Online citizen science.



> "Studies have shown that introducing interactive, research-based models of education can greatly improve classroom performance and retention"

– "Current Approaches in Implementing Citizen Science in the Classroom" by Shah and Martinez (2016)

KEY COMPETENCES

There are 11 key competences.

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. 1

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.

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. 3

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. 4

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

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. 6

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.

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

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.

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

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.

AUTOMATIC EXTRACTION OF INFORMATION

Allows identification and extraction of meaningful information from a document or text without the user having to read it. This can be achieved with the use of **Natural Language Processing (NLP)**.

Some NLP tasks include:

- Document summarization.
- Machine translation.
- Sentiment analysis.
- Speech-to-text and text-to-speech conversion.

AUTOMATIC EXTRACTION OF INFORMATION

TF-IDF (**term frequency-inverse document frequency**) evaluates how relevant a word is to a document in a collection of documents.

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

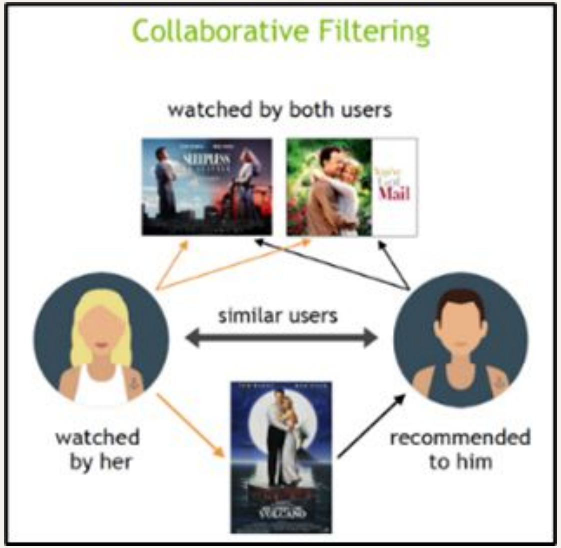
$$tf\,(t,\,d) = log\,(1 + freq\,(t,\,d))$$

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

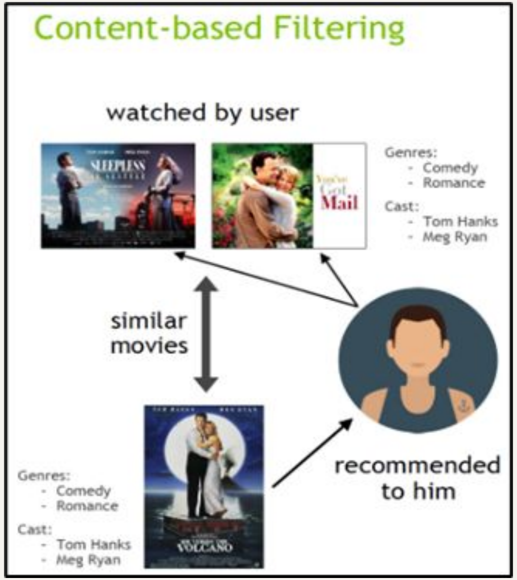
$$Similarity(A,B) = \frac{A.B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

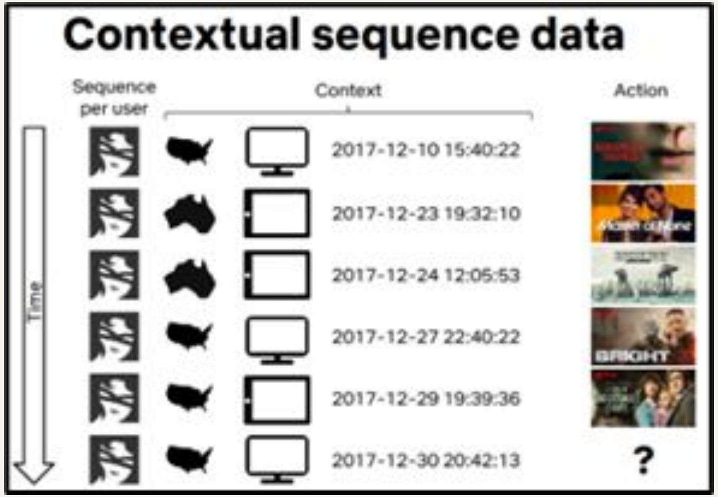
Text **semantic similarity** is used to identify if the meaning of two texts or words is similar.

RECOMMENDATION SYSTEMS



User-based
Item-based





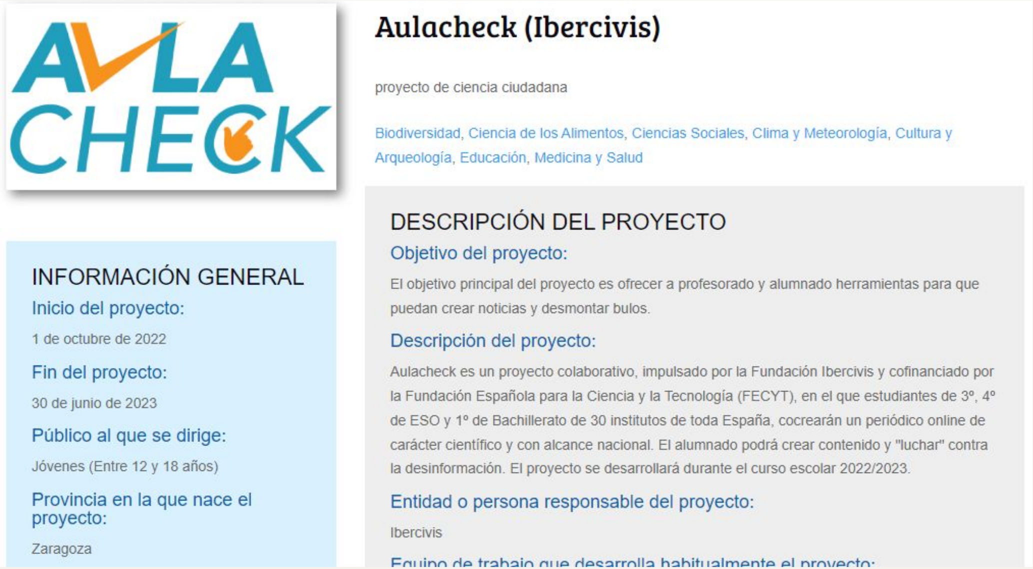
Context filtering

03

IMPLEMENTATION

Crawler for the Citizen Science online platforms

CITIZEN SCIENCE ONLINE PLATFORMS



Aulacheck (Ibercivis)

proyecto de ciencia ciudadana

Biodiversidad, Ciencia de los Alimentos, Ciencias Sociales, Clima y Meteorología, Cultura y Arqueología, Educación, Medicina y Salud

INFORMACIÓN GENERAL

Inicio del proyecto:

1 de octubre de 2022

Fin del proyecto:

30 de junio de 2023

Público al que se dirige:

Jóvenes (Entre 12 y 18 años)

Provincia en la que nace el proyecto:

Zaragoza

DESCRIPCIÓN DEL PROYECTO

Objetivo del proyecto:

El objetivo principal del proyecto es ofrecer a profesorado y alumnado herramientas para que puedan crear noticias y desmontar bulos.

Descripción del proyecto:

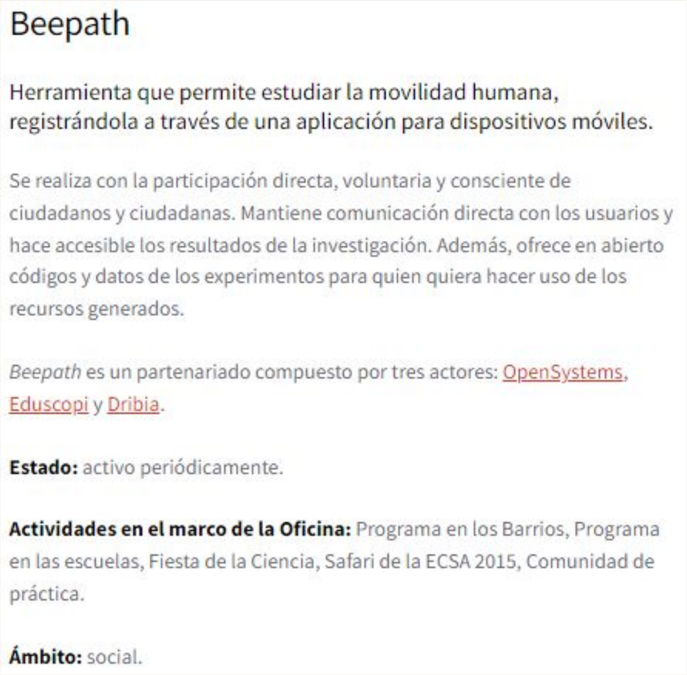
Aulacheck es un proyecto colaborativo, impulsado por la Fundación Ibercivis y cofinanciado por la Fundación Española para la Ciencia y la Tecnología (FECYT), en el que estudiantes de 3º, 4º de ESO y 1º de Bachillerato de 30 institutos de toda España, cocrearán un periódico online de carácter científico y con alcance nacional. El alumnado podrá crear contenido y "luchar" contra la desinformación. El proyecto se desarrollará durante el curso escolar 2022/2023.

Entidad o persona responsable del proyecto:

Ibercivis

Equipo de trabajo que desarrolla habitualmente el proyecto:

Project from Spanish platform: “Observatorio de la Ciencia Ciudadana en España”



Beepath

Herramienta que permite estudiar la movilidad humana, registrándola a través de una aplicación para dispositivos móviles.

Se realiza con la participación directa, voluntaria y consciente de ciudadanos y ciudadanas. Mantiene comunicación directa con los usuarios y hace accesible los resultados de la investigación. Además, ofrece en abierto códigos y datos de los experimentos para quien quiera hacer uso de los recursos generados.

Beepath es un partenariado compuesto por tres actores: OpenSystems, Eduscopi y Dribia.

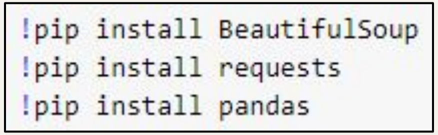
Estado: activo periódicamente.

Actividades en el marco de la Oficina: Programa en los Barrios, Programa en las escuelas, Fiesta de la Ciencia, Safari de la ECSA 2015, Comunidad de práctica.

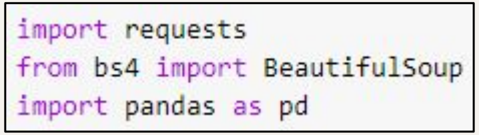
Ámbito: social.

Project from Barcelonian platform: “Oficina de la Ciència Ciudadana”

STEP 1. INSTALL & IMPORT LIBRARIES

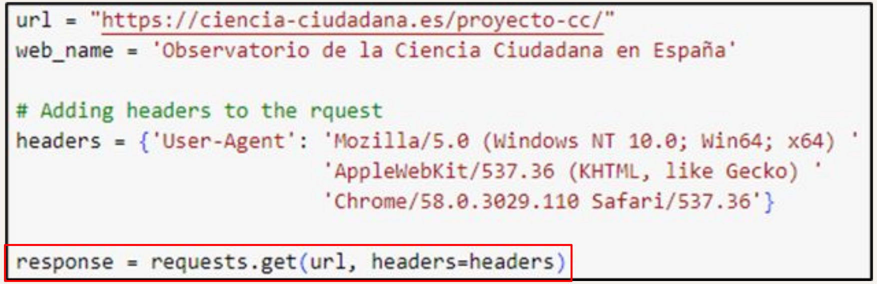


```
!pip install BeautifulSoup
!pip install requests
!pip install pandas
```



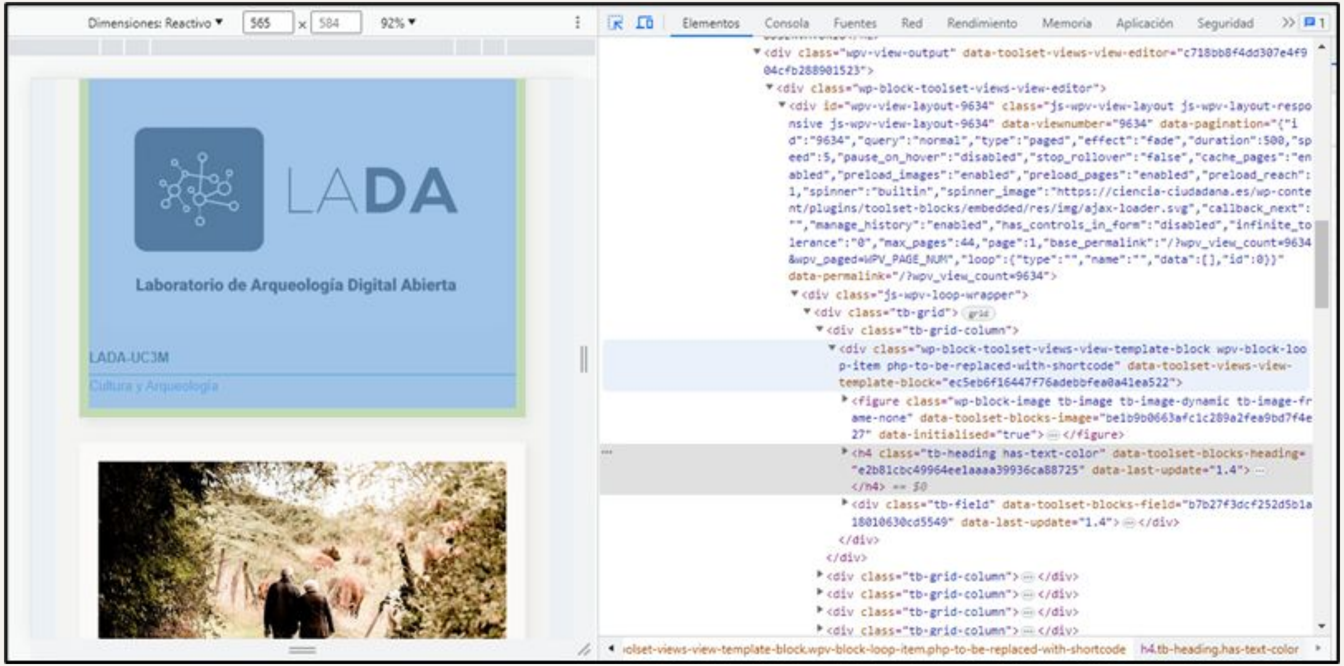
```
import requests
from bs4 import BeautifulSoup
import pandas as pd
```

STEP 2. SEND HTTP REQUEST



```
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)
```

STEP 3. PARSE HTML CONTENT



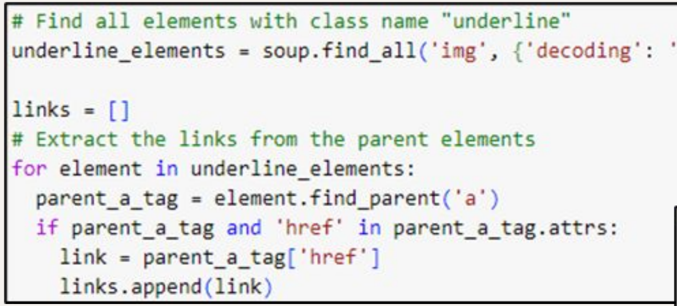
Dimensiones: Reactivo ▾ 565 × 584 92% ▾

```
▼<div class="wpv-view-output" data-toolset-views-view-editor="c718bb8f4dd307e4f904cfb288901523">
 ▼<div class="wp-block-toolset-views-view-editor">
  ▼<div id="wpv-view-layout-9634" class="js-wpv-view-layout js-wpv-layout-responsive js-wpv-view-layout-9634" data-viewnumber="9634" data-pagination="{"id":"9634","query":"normal","type":"paged","effect":"fade","duration":500,"speed":5,"pause_on_hover":"disabled","stop_rollover":"false","cache_pages":"enabled","preload_images":"enabled","preload_pages":"enabled","preload_reach":1,"spinner":"builtin","spinner_image":"https://ciencia-ciudadana.es/wp-content/plugins/toolset-blocks/embedded/res/img/ajax-loader.svg","callback_next":"","manage_history":"enabled","has_controls_in_form":"disabled","infinite_tolerance":"0","max_pages":44,"page":1,"base_permalink":"/?wpv_view_count=9634&wpv_paged=WPV_PAGE_NUM","loop":{"type":"","name":"","data":[],"id":0}}" data-permalink="/?wpv_view_count=9634">
   ▼<div class="js-wpv-loop-wrapper">
    ▼<div class="tb-grid"> grid
     ▼<div class="tb-grid-column">
      ▼<div class="wp-block-toolset-views-view-template-block wpv-block-loop-item php-to-be-replaced-with-shortcode" data-toolset-views-view-template-block="ec5eb6f16447f76adebbfea0a41ea522">
       ▶<figure class="wp-block-image tb-image tb-image-dynamic tb-image-frame-none" data-toolset-blocks-image="be1b9b0663afc1c289a2fea9bd7f4e27" data-initialised="true">…</figure>
       ▶<h4 class="tb-heading has-text-color" data-toolset-blocks-heading="e2b81cbc49964ee1aaaa39936ca88725" data-last-update="1.4">…</h4> == $0
       ▶<div class="tb-field" data-toolset-blocks-field="b7b27f3dcf252d5b1a18010630cd5549" data-last-update="1.4">…</div>
      </div>
     </div>
    ▶<div class="tb-grid-column">…</div>
    ▶<div class="tb-grid-column">…</div>
    ▶<div class="tb-grid-column">…</div>
    ▶<div class="tb-grid-column">…</div>
```

olset-views-view-template-block.wpv-block-loop-item.php-to-be-replaced-with-shortcode h4.tb-heading.has-text-color

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

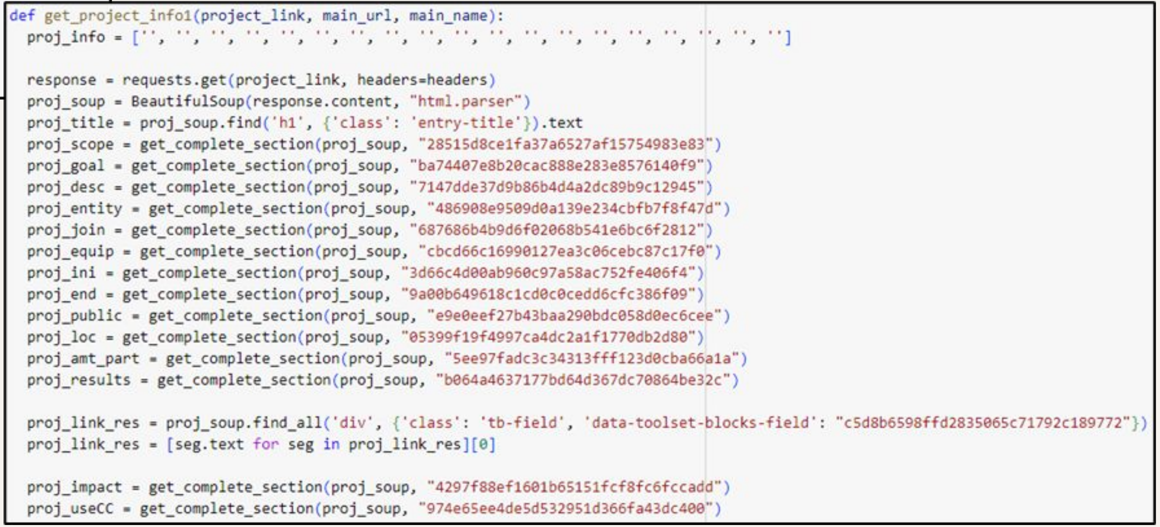
STEP 4. EXTRACT DATA



```
# 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)
```

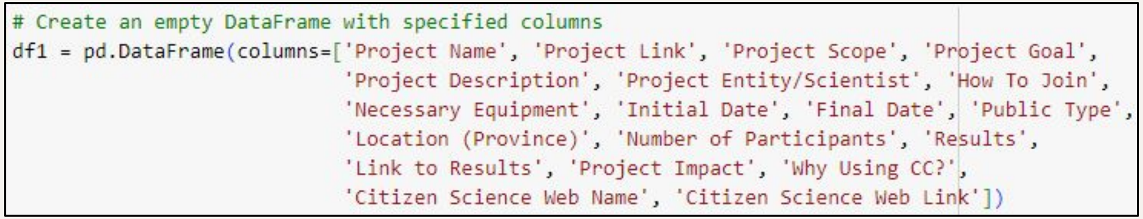
Extracting the project URLs

Extracting the necessary fields of a project

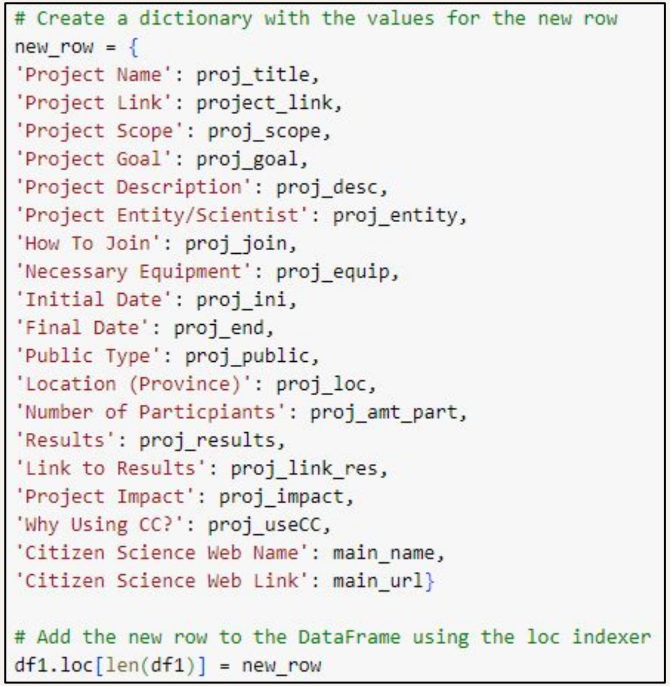


```
def get_project_info1(project_link, main_url, main_name):
  proj_info = ['', '', '', '', '', '', '', '', '', '', '', '', '', '', '', '', '', '', '']

  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")
```

STEP 5. STORE DATA IN DATAFRAME



```python
# 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'])
```



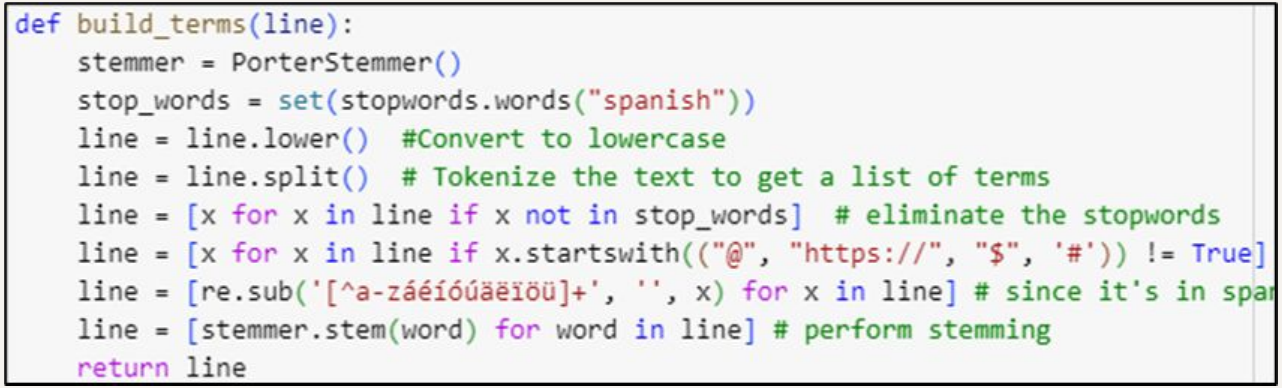
```python
# Create a dictionary with the values for the new row
new_row = {
'Project Name': proj_title,
'Project Link': project_link,
'Project Scope': proj_scope,
'Project Goal': proj_goal,
'Project Description': proj_desc,
'Project Entity/Scientist': proj_entity,
'How To Join': proj_join,
'Necessary Equipment': proj_equip,
'Initial Date': proj_ini,
'Final Date': proj_end,
'Public Type': proj_public,
'Location (Province)': proj_loc,
'Number of Particpiants': proj_amt_part,
'Results': proj_results,
'Link to Results': proj_link_res,
'Project Impact': proj_impact,
'Why Using CC?': proj_useCC,
'Citizen Science Web Name': main_name,
'Citizen Science Web Link': main_url}

# Add the new row to the DataFrame using the loc indexer
df1.loc[len(df1)] = new_row
```

04

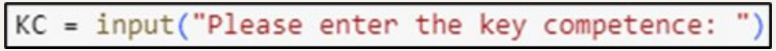
RECOMMENDATION SYSTEM

Design of the Recommendation System

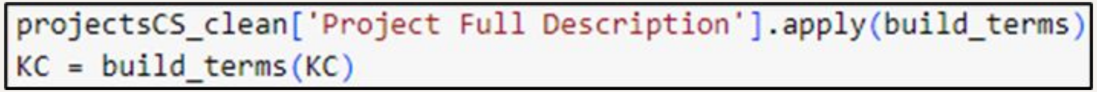
STEP 1. PREPROCESSING THE DATA



```
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-záéíóúäëïöü]+', '', x) for x in line] # since it's in spar
    line = [stemmer.stem(word) for word in line] # perform stemming
    return line
```

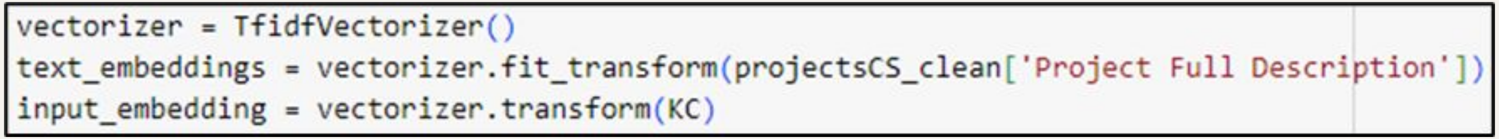


```
KC = input("Please enter the key competence: ")
```



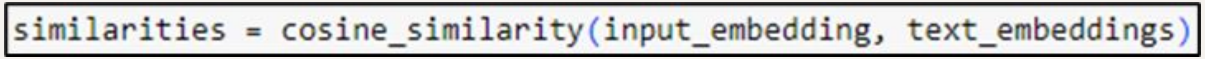
```
projectsCS_clean['Project Full Description'].apply(build_terms)
KC = build_terms(KC)
```

STEP 2. TEXT EMBEDDINGS



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

STEP 3. COSINE SIMILARITY



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

05

KEY COMPETENCES

Analysis of the Key Competences

KEY COMPETENCES

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. 1

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.

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. 3

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**. 4

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.

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. 6

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.

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

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**.

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

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.

06

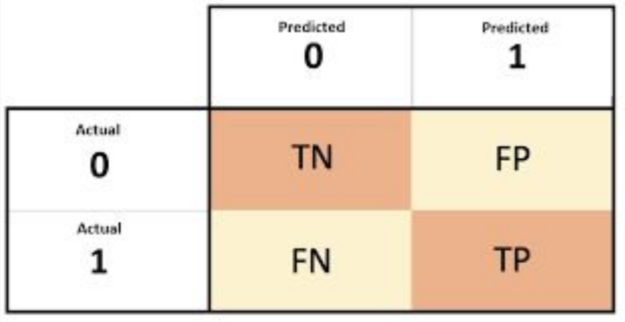
RESULTS & ANALYSIS

Analysis of the recommended projects

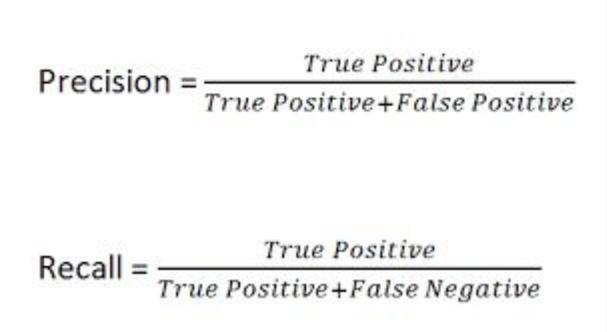
CRITERIA OF SELECTION

363 projects
11 key competences

Analysis of **20%** of the projects
7 top recommended projects per key competence



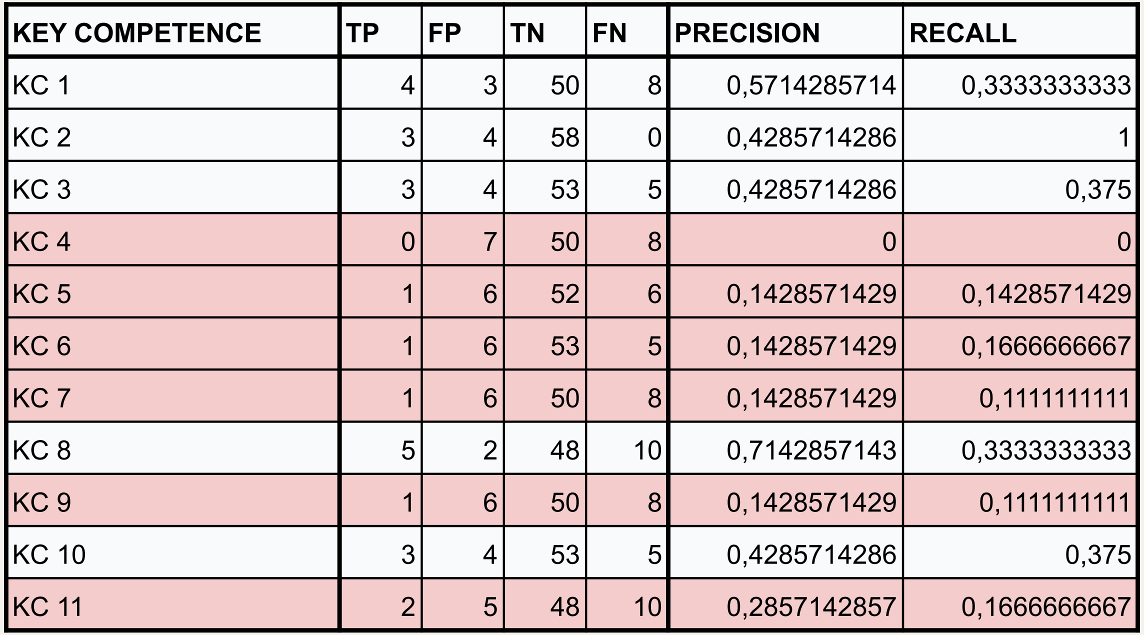
	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP



$$\text{Precision} = \frac{\mathit{True\ Positive}}{\mathit{True\ Positive} + \mathit{False\ Positive}}$$

$$\text{Recall} = \frac{\mathit{True\ Positive}}{\mathit{True\ Positive} + \mathit{False\ Negative}}$$

MANUAL ANALYSIS OF RECOMMENDED PROJECTS



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

KEY COMPETENCE #1

Initial keyword: “environment”

4 TP
3 FP

57% precision

8 FN

New keyword proposals:

“environmental awareness”

“environmental sensitization”

Final keyword: “**environmental awareness**”

6 TP
1 FP

86% precision

KEY COMPETENCE #2

Initial keyword: “**consumption of local products**”

3 TP
4 FP

43% precision

0 FN

NO new keyword proposals.

KEY COMPETENCE #3

Initial keyword: “healthy lifestyle”

3 TP
4 FP } **43%** precision

5 FN

New keyword proposals:

“diet” “nutrition” “health” “physical activity” “wellbeing”

Final keyword: “**healthy lifestyle**”

3 TP
4 FP } **43%** precision

KEY COMPETENCE #4

Initial keyword: “inequality, exclusion and empathy” **0** TP **7** FP } **0%** precision

8 FN

New keyword proposals:

“social problems” “civic society” “social awareness” “limitations”

“reflection space” “barrier” or “disability”

Final keyword: “**social inclusion**” **4** TP **3** FP } **57%** precision

KEY COMPETENCE #5

Initial keyword: “gender equality”

1 TP
6 FP

14% precision

6 FN

New keyword proposals:

“functional diversity” “people rights” “female” “accessible”

“gender” “reflection space” “citizen biodiversity” “sign language”

Final keyword: “**gender**”

5 TP
2 FP

72% precision

KEY COMPETENCE #6

Initial keyword: “conflicts in society”

1 TP
6 FP

14% precision

5 FN

New keyword proposals:

“social awareness” “social problems” “solution proposal” “support”

“people rights” “civic society” “social perspective” “problem detection”

Final keyword: “**social perspective**”

5 TP
2 FP

72% precision

KEY COMPETENCE #7

Initial keyword: “digital culture”

1 TP
6 FP

14% precision

8 FN

New keyword proposals:

“artificial intelligence” “fablabs” “scientific advances” “digital”

“technology”

Final keyword: “**technology**”

5 TP
2 FP

72% precision

KEY COMPETENCE #8

Initial keyword: “**creativity**” **5** TP **2** FP } **72%** precision

10 FN

NO new keyword proposals.

KEY COMPETENCE #9

Initial keyword: “languages and cultures”

1 TP
6 FP

14% precision

8 FN

New keyword proposals:

“collaborative participation”

“world connection”

“diversity”

“cooperation”

Final keyword: “**cooperation**”

3 TP
4 FP

43% precision

KEY COMPETENCE #10

Initial keyword: “collective”

3 TP
4 FP
} **43%** precision

5 FN

New keyword proposals:

“collaborate”

Final keyword: “**collaborate**”

5 TP
2 FP
} **72%** precision

KEY COMPETENCE #11

Initial keyword: “lifelong learning”

2 TP
5 FP
} **29%** precision

10 FN

New keyword proposals:

“knowledge” “education”

Final keyword: “**education**”

5 TP
2 FP
} **72%** precision

FINAL COSINE SIMILARITY SCORES

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

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

Cosine similarity **threshold**:
0,08160556

07

WEB APPLICATION

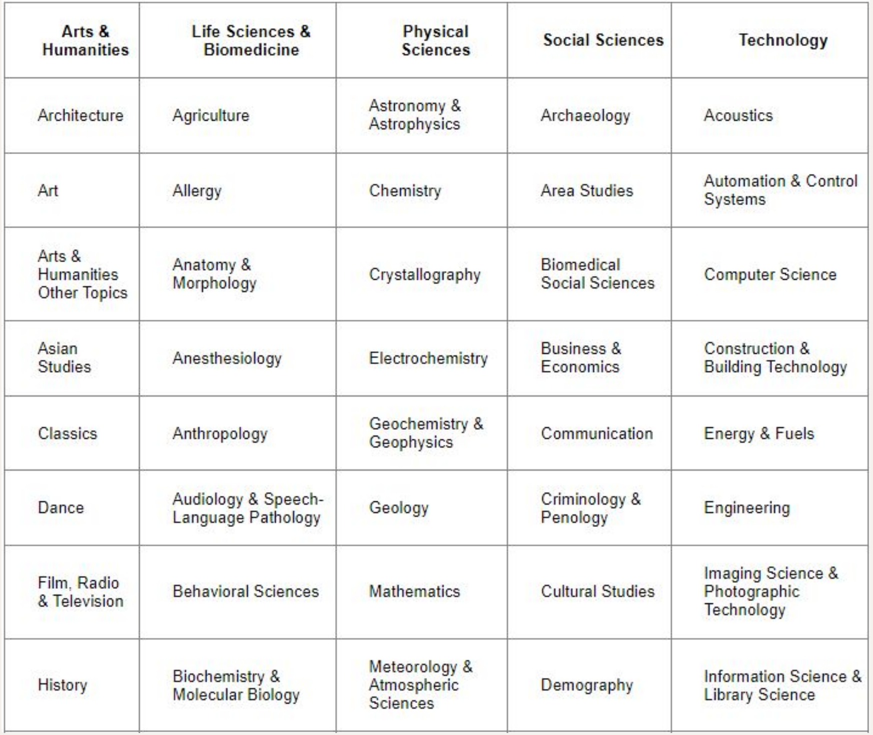
Design of the Web Application

WEB APPLICATION

The web application is in Spanish!

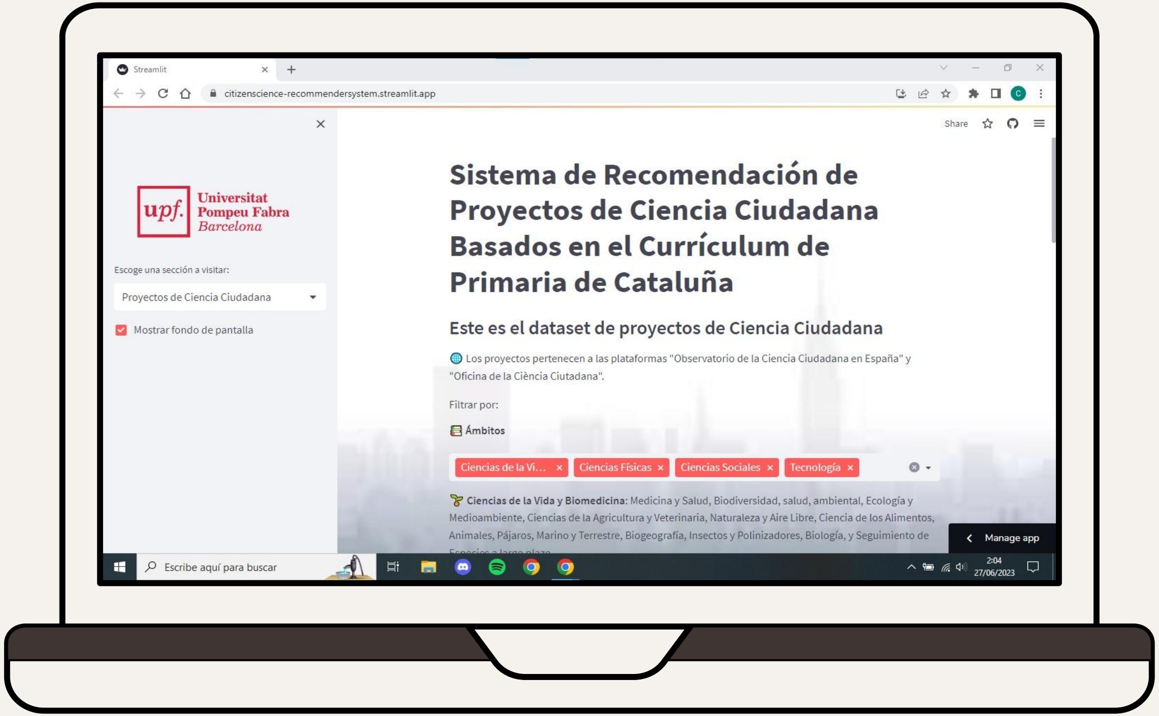
4 major categories:

- Life Sciences & Biomedicine
- Physical Sciences
- Social Sciences
- Technology



Arts & Humanities	Life Sciences & Biomedicine	Physical Sciences	Social Sciences	Technology
Architecture	Agriculture	Astronomy & Astrophysics	Archaeology	Acoustics
Art	Allergy	Chemistry	Area Studies	Automation & Control Systems
Arts & Humanities Other Topics	Anatomy & Morphology	Crystallography	Biomedical Social Sciences	Computer Science
Asian Studies	Anesthesiology	Electrochemistry	Business & Economics	Construction & Building Technology
Classics	Anthropology	Geochemistry & Geophysics	Communication	Energy & Fuels
Dance	Audiology & Speech-Language Pathology	Geology	Criminology & Penology	Engineering
Film, Radio & Television	Behavioral Sciences	Mathematics	Cultural Studies	Imaging Science & Photographic Technology
History	Biochemistry & Molecular Biology	Meteorology & Atmospheric Sciences	Demography	Information Science & Library Science

WEB APPLICATION


Streamlit
citizenscience-recommendersystem.streamlit.app
Share
Universitat Pompeu Fabra Barcelona
Manage app
Escribe aquí para buscar
2:04
27/06/2023

Escoge una sección a visitar:

Proyectos de Ciencia Ciudadana

☑ Mostrar fondo de pantalla

Sistema de Recomendación de Proyectos de Ciencia Ciudadana Basados en el Currículum de Primaria de Cataluña

Este es el dataset de proyectos de Ciencia Ciudadana

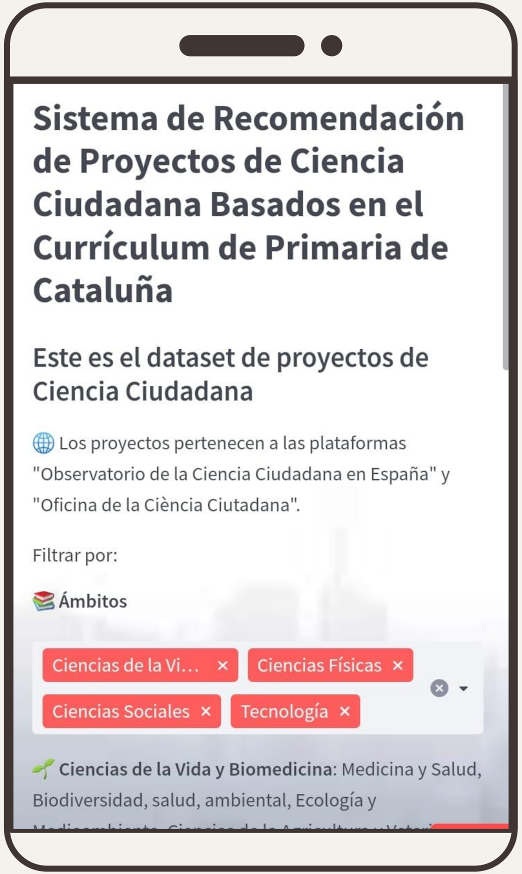
🌐 Los proyectos pertenecen a las plataformas "Observatorio de la Ciencia Ciudadana en España" y "Oficina de la Ciència Ciutadana".

Filtrar por:

📚 Ámbitos

Ciencias de la Vi... × Ciencias Físicas × Ciencias Sociales × Tecnología ×

🌱 **Ciencias de la Vida y Biomedicina:** Medicina y Salud, Biodiversidad, salud, ambiental, Ecología y Medioambiente, Ciencias de la Agricultura y Veterinaria, Naturaleza y Aire Libre, Ciencia de los Alimentos, Animales, Pájaros, Marino y Terrestre, Biogeografía, Insectos y Polinizadores, Biología, y Seguimiento de Especies a largo plazo



Sistema de Recomendación de Proyectos de Ciencia Ciudadana Basados en el Currículum de Primaria de Cataluña

Este es el dataset de proyectos de Ciencia Ciudadana

🌐 Los proyectos pertenecen a las plataformas "Observatorio de la Ciencia Ciudadana en España" y "Oficina de la Ciència Ciutadana".

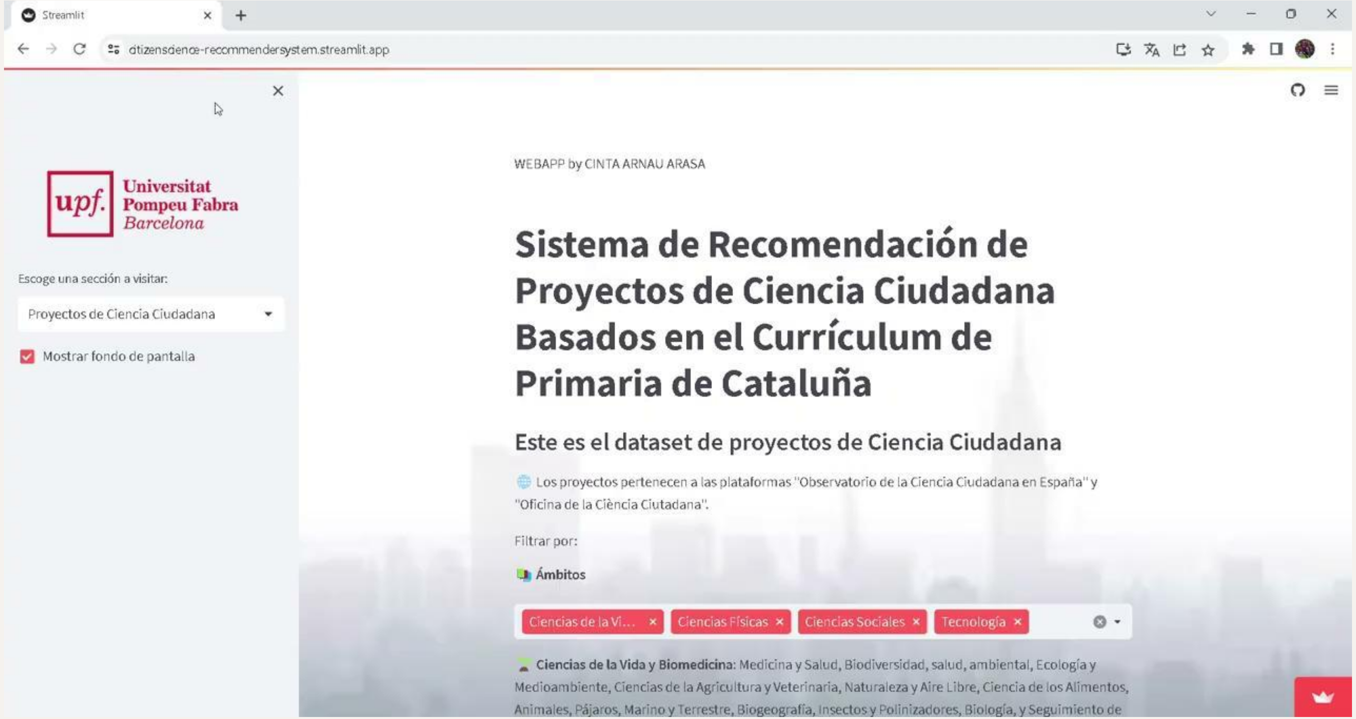
Filtrar por:

📚 Ámbitos

Ciencias de la Vi... × Ciencias Físicas × Ciencias Sociales × Tecnología ×

🌱 **Ciencias de la Vida y Biomedicina:** Medicina y Salud, Biodiversidad, salud, ambiental, Ecología y Medioambiente, Ciencias de la Agricultura y Veterinaria

WEB APPLICATION - DEMO



Streamlit
citizenscience-recommendersystem.streamlit.app
Universitat Pompeu Fabra Barcelona

Escoge una sección a visitar:

Proyectos de Ciencia Ciudadana

☑ Mostrar fondo de pantalla

WEBAPP by CINTA ARNAU ARASA

Sistema de Recomendación de Proyectos de Ciencia Ciudadana Basados en el Currículum de Primaria de Cataluña

Este es el dataset de proyectos de Ciencia Ciudadana

Los proyectos pertenecen a las plataformas "Observatorio de la Ciencia Ciudadana en España" y "Oficina de la Ciència Ciutadana".

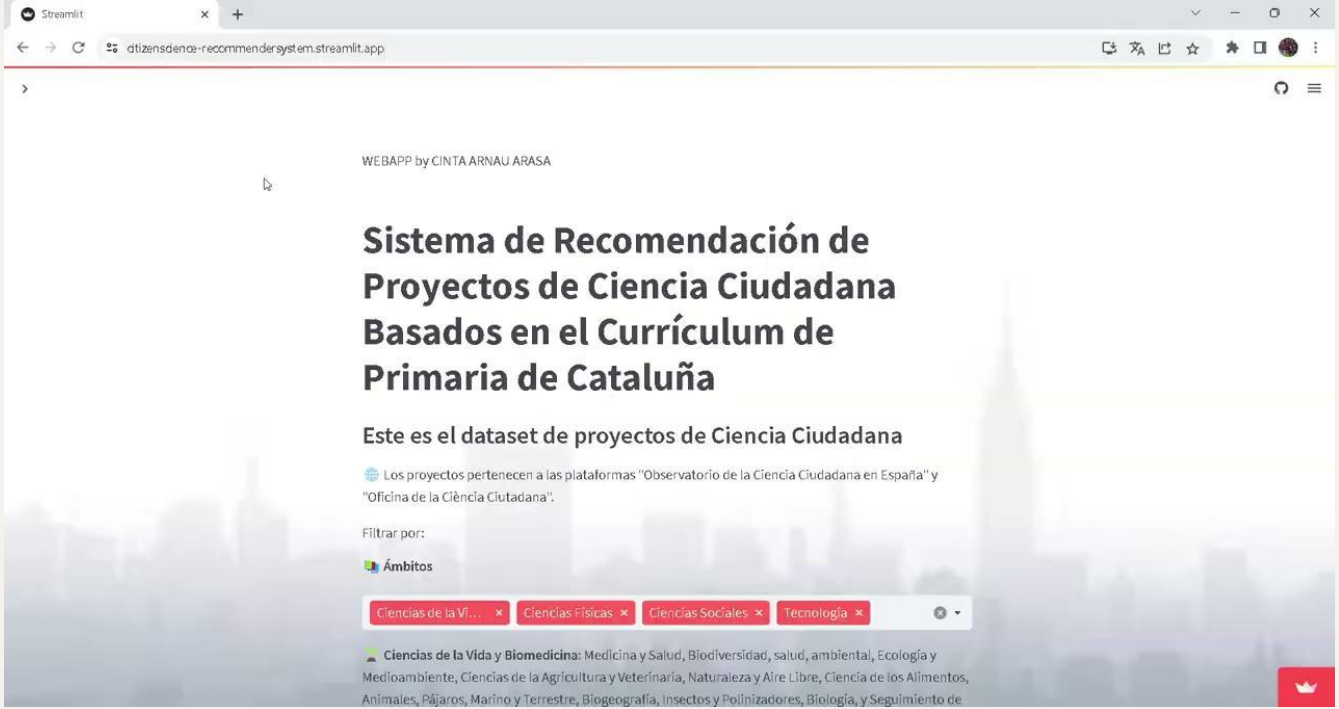
Filtrar por:

Ámbitos

Ciencias de la Vi… × Ciencias Físicas × Ciencias Sociales × Tecnología ×

Ciencias de la Vida y Biomedicina: Medicina y Salud, Biodiversidad, salud, ambiental, Ecología y Medioambiente, Ciencias de la Agricultura y Veterinaria, Naturaleza y Aire Libre, Ciencia de los Alimentos, Animales, Pájaros, Marino y Terrestre, Biogeografía, Insectos y Polinizadores, Biología, y Seguimiento de

WEB APPLICATION - DEMO



WEBAPP by CINTA ARNAU ARASA

Sistema de Recomendación de Proyectos de Ciencia Ciudadana Basados en el Currículum de Primaria de Cataluña

Este es el dataset de proyectos de Ciencia Ciudadana

🌐 Los proyectos pertenecen a las plataformas "Observatorio de la Ciencia Ciudadana en España" y "Oficina de la Ciència Ciutadana".

Filtrar por:

Ámbitos

Ciencias de la Vi... × | Ciencias Físicas × | Ciencias Sociales × | Tecnología ×

Ciencias de la Vida y Biomedicina: Medicina y Salud, Biodiversidad, salud, ambiental, Ecología y Medioambiente, Ciencias de la Agricultura y Veterinaria, Naturaleza y Aire Libre, Ciencia de los Alimentos, Animales, Pájaros, Marino y Terrestre, Biogeografía, Insectos y Polinizadores, Biología, y Seguimiento de

08

CONCLUSIONS

CONCLUSIONS

Objectives have been accomplished

Limitations

Scarcity of project fields
Project descriptions too short
Application not tested

09

FUTURE WORK

FUTURE WORK

Other Similarity Measures

Jaccard Similarity
K-Means
Euclidean Distance
BM25 Ranking

User Feedback Mechanisms

User Accounts in Web Application

Collaborative filtering
(Hybrid approach)
Projects saving

THANK YOU FOR YOUR ATTENTION!

TFG - Mathematical Engineering in Data Science

03/07/2023

Cinta Arnau Arasa - cinta.arnau01@estudiant.upf.edu

Supervisors: Patricia Santos & Miriam Calvera

Universitat Pompeu Fabra Barcelona

Escola d'Enginyeria

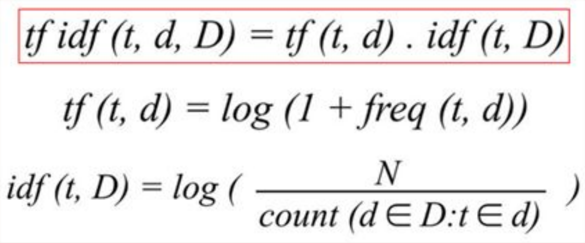
Types of Recommendation Systems

- In a collaborative filtering algorithm, items are recommended 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.
- Content filtering algorithms recommend items similar to what the user preferences are based on the attributes or characteristics of a given item, using the similarity between the item and the user features.
- 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.

How is TF-IDF calculated?

- TF (term frequency). It counts the number of occurrences for a given word in a document (or text).
- 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.

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.



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

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

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

Process to calculate text similarity

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 TFIDF. 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.

Why using a content-based recommendation system?

For this projects, 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.

Example of recommended projects (KC 8)

- **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.

Manual Analysis

https://docs.google.com/spreadsheets/d/1ZhxyN8hx6Hg7udQyKvMxNmFyudYxLu_CbzJDSs8DpfE/edit#gid=583161345

Why did previous keywords not work well?

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.

* Homonyms are words that sound alike or are spelled alike but have different meanings.

Why is cosine similarity the best option?

If your projects have rich textual content and you want to capture the semantic similarity, cosine similarity with text embeddings like TF-IDF or word embeddings may be more appropriate.

In recommendation systems, cosine similarity is generally more commonly used than Euclidean distance. Unlike Euclidean distance, cosine similarity is not affected by the magnitude of the vectors and focuses on the direction or orientation of the vectors. This makes cosine similarity more robust for high-dimensional data or when the magnitudes of the feature vectors vary significantly.

How would other similarity measures improve the recommendation system?

- **Jaccard similarity**. Jaccard similarity can be useful when representing projects as sets of categorical or binary features.
- **Euclidean distance**. Euclidean distance can be used to measure the similarity between project feature vectors. However, it is necessary to preprocess the data and scale the features appropriately before applying Euclidean distance because if he magnitudes of the features are not normalized or standardized, attributes with larger value ranges may dominate the distance calculation.
- **BM25 ranking**. BM25 is advantageous for recommendation systems that involve textual data. It captures the relevance between the query (input sentence) and the projects by considering the statistical properties of the terms in the documents.

How would other similarity measures improve the recommendation system?

- **K-Means**. K-means can be used as a preprocessing step to extract features from the project data. By running k-means clustering on project attributes, you can derive cluster labels or cluster assignments as additional features for each project. These cluster labels can then be used as input features in a recommendation model, providing an extra level of information and potentially improving the recommendation accuracy.

  In recommendation systems, k-means is often used in conjunction with other recommendation techniques, such as collaborative filtering or content-based filtering, to create hybrid or ensemble approaches. By combining multiple recommendation strategies, you can leverage the strengths of different methods and potentially improve the accuracy and relevance of recommendations.