Modern Data Analytics

Horizon Europe Project Recommender

Group 25

David Ng (r0874153)
David O’Grady (r1032923)
Liren Xie (r0864366)

KU Leuven
Master of Statistics and Data Science
2024-2025

1 Project Background

The European Union's Horizon Europe programme, with a budget of €95.5 billion for the 2021-2027 period, stands as one of the world's largest research and innovation funding initiatives. It is designed to tackle global challenges, bolster the EU's industrial competitiveness, and strengthen the European Research Area (ERA). The programme is structured around three primary pillars: Excellent Science, Global Challenges & Industrial Competitiveness, and Innovative Europe, each targeting different facets of research and development.

To track and manage the multitude of projects funded under this ambitious framework, the CORDIS (Community Research and Development Information Service) datasets are maintained. Specifically, the "CORDIS EU Research Projects under Horizon Europe (2021–2027)" dataset offers extensive details on each project. This includes information such as project objectives, participating organizations, funding allocations, project durations, and thematic classifications, which may include terms from the European Science Vocabulary (EuroSciVoc).

The vastness and complexity of this data present both an opportunity and a challenge. While it is a rich resource for understanding the European research landscape, navigating it to identify relevant prior work, potential collaborators, or suitable funding schemes can be a significant undertaking for researchers preparing new proposals. This difficulty can impede the development of well-targeted, innovative, and competitive research proposals.

This project is motivated by the need to simplify access to and insights from this complex dataset. The aim is to develop an interactive web application that leverages Natural Language Processing (NLP) to semantically match a researcher's new topic proposal against the extensive CORDIS Horizon project database. By analyzing the textual information of project descriptions and objectives (which may implicitly include EuroSciVoc terms if they were part of the text processing for embeddings), the tool will identify comparable previously funded projects. It will then provide users with summary statistics and filtering capabilities related to these similar projects, such as funding schemes, award sizes, and participant counts. Ultimately, this system seeks to empower researchers to efficiently explore the EU research landscape, enabling them to gain crucial insights, strategically position their ideas, and formulate stronger, better-targeted proposals for Horizon Europe funding.

2 Data Foundation and Exploratory Analysis

This section outlines the data sources that form the basis of this project and presents the key findings from the Exploratory Data Analysis (EDA) conducted. Understanding the underlying data is crucial for developing an effective project recommendation system.

2.1 Data Sources

The foundational data for this project is derived from the CORDIS EU Research Projects under Horizon Europe (2021-2027) dataset, accessible via the EU Open Data Portal (`data.europa.eu`). This project utilized several core CSV files from the CORDIS repository to build a comprehensive understanding of the funded projects:

`euroSciVoc.csv`: This file links projects to thematic classifications using the European Science Vocabulary, typically including project identifiers, vocabulary codes, hierarchical paths, and descriptive titles for each scientific term.

`legalBasis.csv`: This dataset details the legal framework and specific programme parts under which projects are funded, listing for each project identifier the relevant legal basis codes and their corresponding descriptive titles.

`organization.csv`: This file provides detailed information about all organizations participating in the projects, including their identifiers, names, types, roles (e.g., coordinator, participant), country of origin, and financial contributions to specific projects.

`programme.csv`: This dataset offers descriptions, objectives, and keywords for the various Horizon Europe funding programmes, calls, and topics that projects are funded under.

`project.csv`: This is the central file containing core details for each funded project. Key information includes Project ID, Acronym, Title, a detailed Project Description (Objective), Start and End Dates, Total Project Cost, and the EU's Maximum Contribution.

`topics.csv`: This file links projects to specific Horizon Europe call topics, providing a code (e.g., ”ERC-2022-STG”) and title for each topic.

While EuroSciVoc terms (from `euroSciVoc.csv`) are part of the project data and are included in the textual information used for generating embeddings (as per the process explored during data analysis), they are not used as a separately weighted or structured input in the final similarity matching algorithm. The richness and granularity of these combined data sources are pivotal for the analytical and recommendation capabilities of the application.

2.2 Exploratory Data Analysis (EDA) Results and Visualizations

Exploratory Data Analysis (EDA) was performed to uncover key characteristics of the processed CORDIS project data. This step was vital for understanding the dataset's structure, informing data preprocessing, and contextualizing the information that the project recommendation system would utilize. The following presents a logical sequence of meaningful visualizations and their primary insights relevant to this project.

To begin, it was important to understand the current lifecycle stage of the projects within our dataset, as this influences the relevance of the data for researchers looking at current and future opportunities. We first examined the Project Status Distribution.

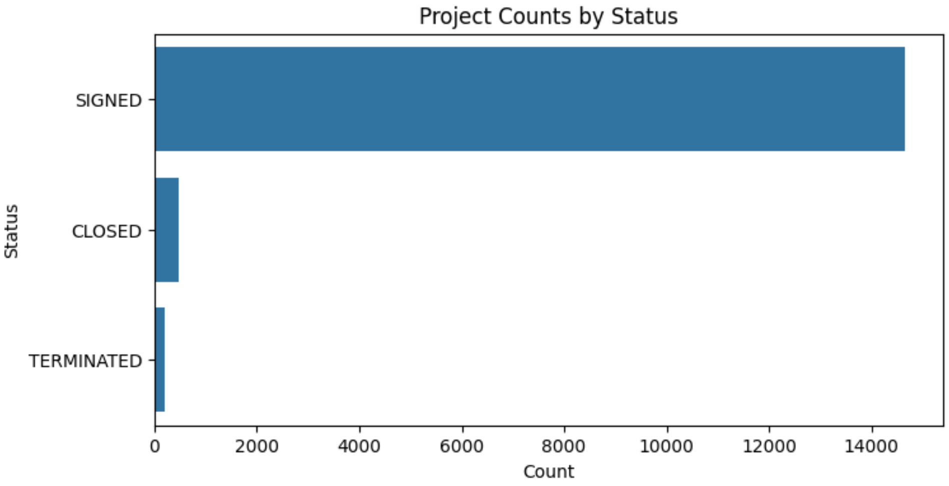


Figure 1: Distribution of Project Statuses.

This chart revealed that ”SIGNED” projects are by far the most prevalent category. This finding is significant as it confirms that the dataset predominantly comprises active or recently approved initiatives, making it highly relevant for providing up-to-date context for new research proposals.

Next, to understand the temporal trends and the recency of the projects, we analyzed the Annual Project Initiation Trends.

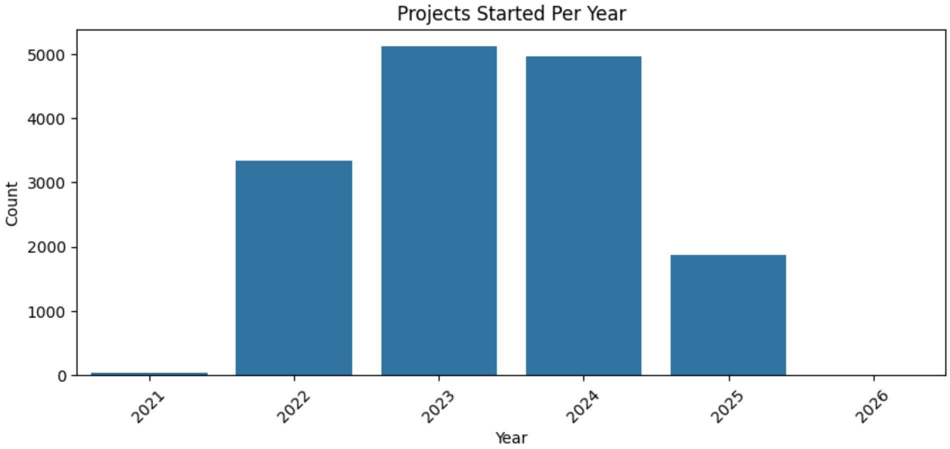


Figure 2: Annual Project Initiation Trends.

The data showed a clear trend of rapid growth in project initiations from 2021, peaking in 2023, and then indicating a decline in subsequent years (with 2025 data appearing incomplete and 2026 data likely being prospective or incomplete). This pattern offers users of our tool insights into recent funding cycles and periods of heightened activity within the Horizon Europe programme.

Understanding the financial scale of these projects is vital for researchers planning their own proposals. Therefore, we investigated the Project Total Cost Distribution.

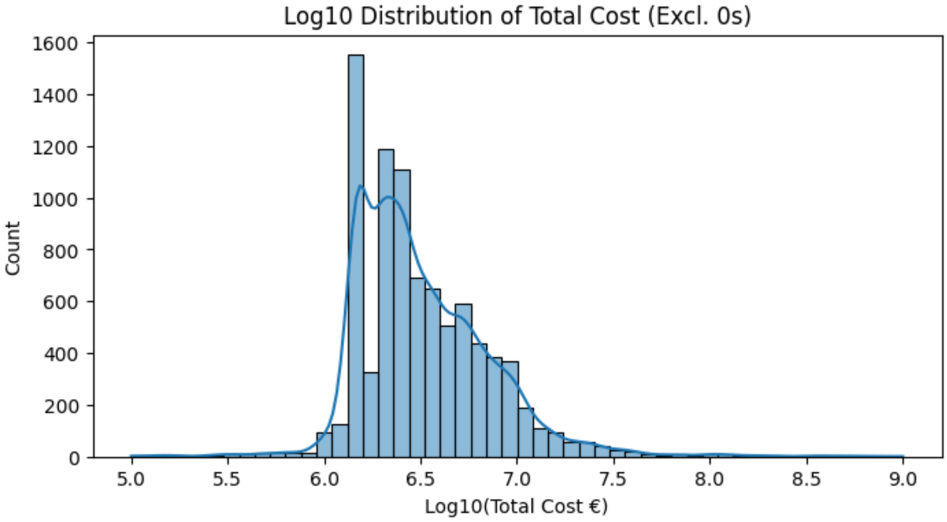


Figure 3: Log10 Distribution of Project Total Cost (Excluding Zeros).

After excluding zero-cost entries, the log10-transformed total project costs displayed a unimodal distribution. Most projects had costs concentrated in the millions of Euros, though the overall range was quite wide, signifying diverse project scales. This variability highlights the importance of providing users with context on typical funding levels for similar projects.

To further refine the financial understanding, we examined the Average Annual Cost Per Participant Distribution.

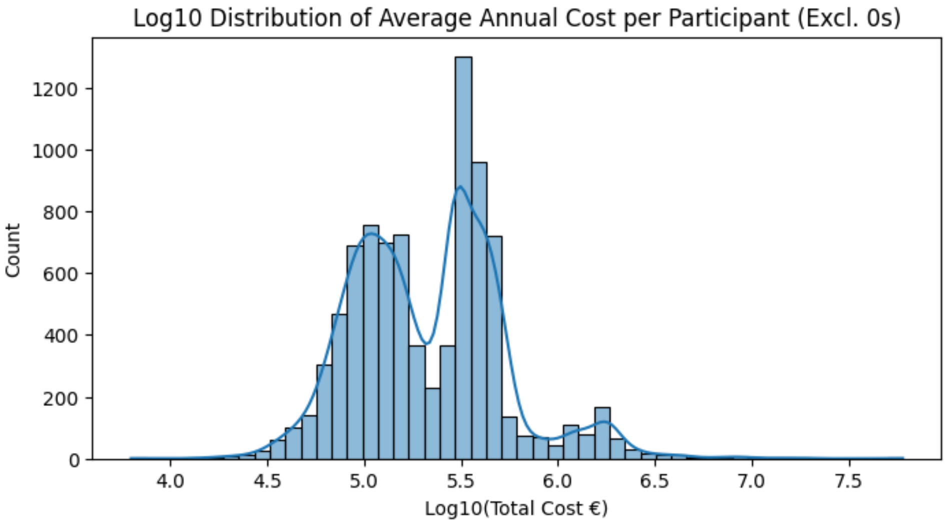


Figure 4: Log10 Distribution of Average Annual Cost per Participant (Excluding Zeros).

This distribution (log10 scale) was notably multimodal. This suggests the presence of distinct categories of projects or types of participant involvement, each associated with different characteristic annualized funding levels. This finding implies that the recommendation tool can potentially help users identify these differing project archetypes based on funding patterns.

Given that funding often relates to collaborative effort, we then analyzed the Consortium Size Distribution.

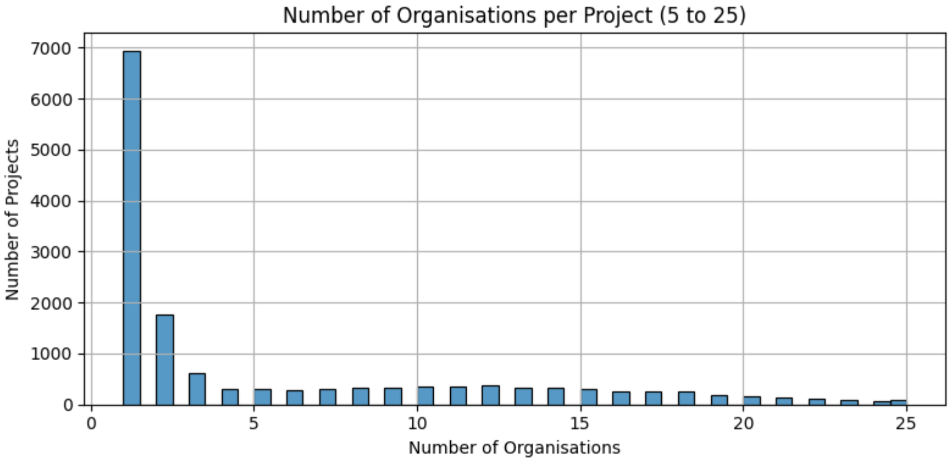


Figure 5: Distribution of Number of Organisations per Project.

The analysis of consortium sizes showed that projects undertaken by a single organization were the most common. The frequency of projects decreased sharply as the number of participating organizations increased. This provides a baseline understanding of typical research team structures and collaboration scales within the dataset.

To explore the interplay between collaboration and funding, we investigated the Relationship Between Consortium Size and Total Cost.

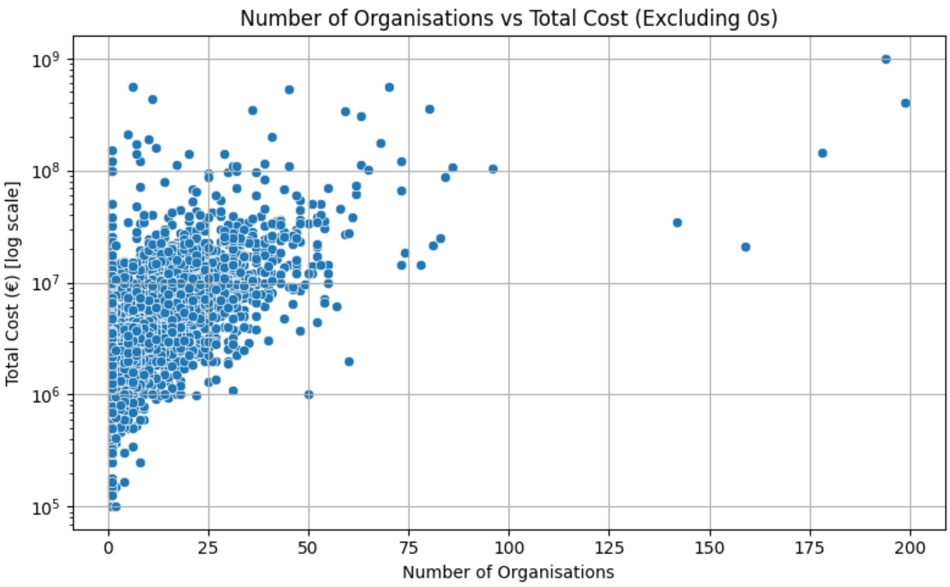


Figure 6: Relationship Between Consortium Size and Total Cost (Log Scale).

A general positive trend was observed, suggesting that projects with more participating organizations tend to manage higher total budgets. However, the considerable cost variance at all consortium sizes clearly indicates that while collaboration size is a factor, other variables also heavily influence overall project funding, a useful insight for proposal planning.

In summary, this EDA process provided a multifaceted understanding of the Horizon Europe project landscape as represented in the dataset. Key trends in project status, timing, financial structures, and collaboration patterns were identified. These insights not only guided subsequent data processing steps but also affirmed the dataset's suitability for building a meaningful project recommendation tool for researchers.

3 Project Overview

3.1 General Description

This project centers on the creation of an interactive web application aimed at assisting researchers in effectively navigating the extensive database of projects funded under the EU's Horizon Europe programme. Users can input their research topic or proposal text into the application. The system then employs Natural Language Processing (NLP) techniques, specifically sentence embeddings, to identify semantically similar existing projects within the CORDIS dataset. The core matching process relies on comparing the textual content of project objectives and descriptions. Upon identifying these similar projects, the application presents users with relevant summary statistics, including typical funding schemes, average funding amounts, and the number of participating organizations. The primary goal is to furnish researchers with an intelligent tool that recommends comparable Horizon projects, thereby offering insights into funding avenues, historical award patterns, and collaborative footprints to support the development of more strategically informed research proposals.

3.2 Tech Stack and Frameworks Used

The project was developed using a Python-based technology stack, selected for its capabilities in data manipulation, machine learning, web development, and visualization:

Data Processing and Exploratory Data Analysis: Core data operations, including loading, cleaning, and transforming diverse datasets, were primarily handled using `Pandas`, with `Numpy` providing essential support for numerical computations, especially with large data arrays like embeddings. During the exploratory data analysis (EDA) phase, Python libraries such as `Matplotlib` and `Seaborn` were primarily used for generating static visualizations (histograms, bar charts, etc.) to understand data characteristics and trends. Tools for interactive charting were also employed to facilitate a more dynamic exploration of the data during development.

Natural Language Processing and Recommendation Core: The semantic similarity engine relies on the `Sentence-Transformers` library, specifically using the `all-MiniLM-L6-v2` model architecture, to convert project textual information into meaningful dense vector embeddings. The similarity between a user's proposal embedding and the pre-computed project embeddings is calculated using the `cosine_similarity` function from the `Scikit-learn` library. Pre-computed assets like project embeddings and their corresponding identifiers are efficiently stored and loaded using standard Python object serialization techniques.

User Interface and Web Application: The interactive front-end was developed using `Shiny for Python`, enabling the creation of a dynamic web application with reactive components that respond to user inputs. The application integrates specialized Python libraries to render interactive geographical maps for displaying the locations of participating organizations, significantly enhancing data exploration capabilities.

Deployment: The web application has been successfully deployed and is operational on an `AWS EC2` (Amazon Web Services Elastic Compute Cloud) instance, providing accessibility and a stable environment. `Docker` containerization was utilized for the deployment on AWS.

4 Technical Implementation

This section details the core technical methodologies implemented in the project, emphasizing the Natural Language Processing (NLP) techniques for project recommendation and the architecture and deployment of the interactive web application.

4.1 Natural Language Processing for Project Similarity

Textual Data Preparation for NLP

To create a basis for semantic comparison, relevant textual data for each existing Horizon project was consolidated. This involved programmatically combining project titles, their detailed objectives, and associated thematic terms into a unified descriptive string for each project. This aggregated text served as the input for generating its semantic representation.

Generation of Semantic Embeddings

To quantitatively capture the semantic meaning of both the project texts and any user-submitted proposal, a specialized sentence embedding library was utilized. The project specifically employed a pre-trained transformer-based model (identified as `all-MiniLM-L6-v2` from project configuration details) known for its effectiveness in generating high-quality embeddings for sentences and short paragraphs. This model maps a variable-length text document to a fixed-size dense vector in a high-dimensional space, where semantically similar texts are located closer to each other. In an offline preprocessing step, the descriptive text for every Horizon project in the curated dataset was encoded into such a dense vector embedding. These project embeddings were subsequently stored in a numerical array file format, and the corresponding list of project identifiers was preserved using an object serialization library. This allows the application to efficiently load these pre-computed semantic representations at runtime.

Similarity Calculation and Matching Logic

When a user inputs their research proposal text into the web application, the recommendation module's primary matching function is triggered. The input proposal text is first encoded into its own vector representation using the same sentence embedding model. The semantic similarity between the proposal's vector and each stored project's vector is then quantified using the cosine similarity metric. This metric is obtained from a standard machine learning library's pairwise metrics module and is calculated as:

$$\text{similarity}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{||\mathbf{a}|| \, ||\mathbf{b}||}$$

where $\mathbf{a}$ and $\mathbf{b}$ are the vector embeddings of the proposal and a project text, respectively, $\mathbf{a} \cdot \mathbf{b}$ is their dot product, and $||\mathbf{a}||$ and $||\mathbf{b}||$ are their Euclidean norms (magnitudes). The resulting score indicates the cosine of the angle between the two vectors, with higher values signifying greater semantic similarity.

Based on these similarity scores, the projects are ranked, and the top N most similar project identifiers, along with their respective scores, are returned to the application for presentation to the user.

Rationale for Chosen Embedding Model Architecture

The selection of the `all-MiniLM-L6-v2` type model architecture was driven by its recognized ability to provide a strong balance between the quality of semantic representations for similarity tasks and computational efficiency. Such models are relatively compact and fast, making them well-suited for applications requiring the encoding of numerous documents and responsive processing of user queries. Furthermore, their pre-training on extensive and diverse text corpora allows for robust generalization to specialized domains like research project descriptions and proposals.

4.2 Web System Architecture and Deployment

The project provides an interactive interface through a web application developed using a Python-based web framework (`Shiny for Python`), enabling users to directly engage with the NLP-driven recommendation engine.

Application Structure and Workflow: The application is built upon a UI/server architecture:

User Interface (UI): The UI is organized into multiple tabs offering distinct functionalities: proposal input and matching, detailed views of selected projects (including objectives and CORDIS links), participant organization profiles (with details like their roles, countries, and geographical locations visualized on interactive maps using a specialized mapping library), and exploration of funding mechanisms (including a pie chart of matched project topics and details of funding schemes). It incorporates various input elements like text areas and sliders, and output displays such as tables, plots, and text summaries.

Server-Side Logic: The server component manages the application's backend and reactive data flow. *Initialization:* Upon starting, the application loads preprocessed datasets (e.g., the main project data file and detailed organization data file). It instantiates the recommendation module and loads the precomputed project embeddings and their associated identifiers. *Reactive Processing:* User actions, such as submitting a proposal, trigger reactive processes that invoke the recommendation module to fetch similar projects. The results are stored in a reactive data structure. *Dynamic UI Updates:* Output rendering functions observe changes in this reactive data and other user inputs (e.g., selecting a project from a list) to dynamically update the UI elements. A helper function is used to retrieve and format organization data for display, including parsing string-represented lists from the data using an abstract syntax tree evaluation method.

Deployment Model: The application is deployed using `Docker` containerization and hosted on an `AWS EC2` T3-Medium instance provisioned with 16GB of RAM. This cloud infrastructure was selected to effectively support the locally hosted transformer model necessary for the application's NLP inference capabilities. The project's codebase was structured as an installable Python library, which, along with its dependencies, is installed within the `Docker` container. The `Shiny` web application is then executed from this containerized environment, with the appropriate network port exposed for public user access. The NLP component operates by loading pre-computed project embeddings at startup, performing live inference only for encoding user queries, which ensures responsiveness. This containerized cloud deployment strategy provides a consistent, portable, and stable operational environment for the application.

5 Application Functionality and Validation

This section presents the findings from the initial validation of our semantic matching approach and demonstrates the functionality and user experience of the deployed application.

5.1 Initial Validation of the Semantic Matching Approach

Beyond understanding the general characteristics of the dataset through EDA (Section 2.2), it was crucial to perform an initial validation of the core Natural Language Processing (NLP) methodology chosen for matching user proposals to existing projects. This test aimed to assess the feasibility and effectiveness of using sentence embeddings and cosine similarity for this specific task before full-scale implementation in the recommendation engine.

The methodology, detailed in Section 4.1, involves generating semantic vector representations (embeddings) for project texts using a pre-trained sentence transformer model (specifically, one of the `all-MiniLM-L6-v2` architecture type). The similarity between texts is then quantified using the cosine similarity between their respective embeddings.

For this initial validation, a sample query, "war in the balkans," was selected. This query was encoded into a vector using the same sentence embedding model. This query vector was then compared against the pre-computed embeddings of all projects in our curated dataset (which were derived from a combination of project titles, objectives, and associated thematic terms). The cosine similarity score was calculated for each project relative to the query.

Table 1: Top Project Matches for Query "war in the balkans"

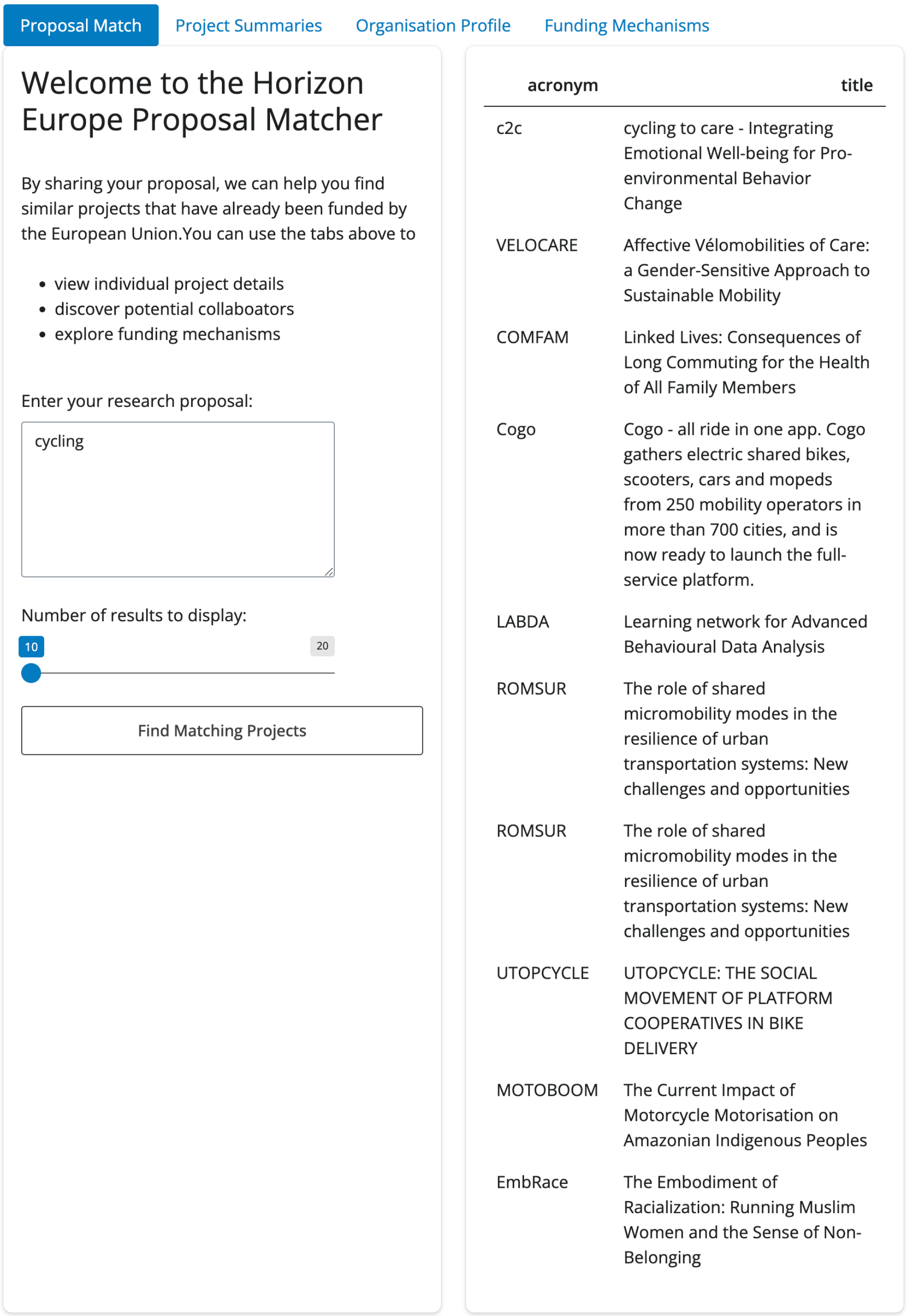
Project ID	Acronym	Title	Similarity
101077076	MACAUTH	Screening Souls, Building Nations. Macedonia(s...	0.524
101054647	CivilWars	The Age of Civil Wars in Europe, c. 1914-1949	0.498
101106810	Expertise	Serving the revolution: educational networks i...	0.489

The results of this test (Table 1) were encouraging. The top-ranked projects demonstrated clear thematic relevance to the "war in the balkans" query. For instance, Project `101077076` (MACAUTH: "Screening Souls, Building Nations. Macedonia(s...)") achieved a similarity score of approximately 0.524. Similarly, Project `101054647` (CivilWars: "The Age of Civil Wars in Europe, c. 1914-1949") also ranked highly with a score of approximately 0.498. Other projects in the top results, such as `101106810` (Expertise, score ≈ 0.489), also related to historical or socio-political studies. The similarity scores, generally in the range of 0.4 to 0.55 for relevant matches, indicated a discernible semantic overlap. This ability to retrieve pertinent projects based on a broad query provided initial confidence in the chosen sentence embedding model and similarity metric, justifying their use in the core recommendation engine. These preliminary findings were instrumental in confirming that the selected NLP approach was a viable and promising direction for the project.

5.2 Application Functionality and User Experience

The project culminated in an interactive web application, deployed on an AWS EC2 instance, designed to provide researchers with a seamless and insightful interface for exploring Horizon Europe projects. The application's features are organized across several intuitive tabs.

The user's journey typically begins with the **Proposal Matching Interface**. Here, researchers can input their specific research proposal or keywords, such as ”cycling,” into a text area and simply adjust a slider to select the number of top similar projects they wish to retrieve. A simple click on the ”Find Matching Projects” button initiates the backend semantic search, promptly displaying a table summarizing the most relevant existing Horizon projects by acronym, title, and similarity score. Users found this interface intuitive, and the matches presented for queries like ”cycling” quickly provided a list of project titles with several demonstrating clear thematic relevance (e.g., ”cycling to care,” ”UTOPCYCLE,” ”shared micromobility modes”), offering an efficient starting point for exploring existing funded projects, even if some results were more broadly related to general mobility or behavior.



Welcome to the Horizon Europe Proposal Matcher

By sharing your proposal, we can help you find similar projects that have already been funded by the European Union.You can use the tabs above to

- view individual project details
- discover potential collaboators
- explore funding mechanisms

Enter your research proposal:

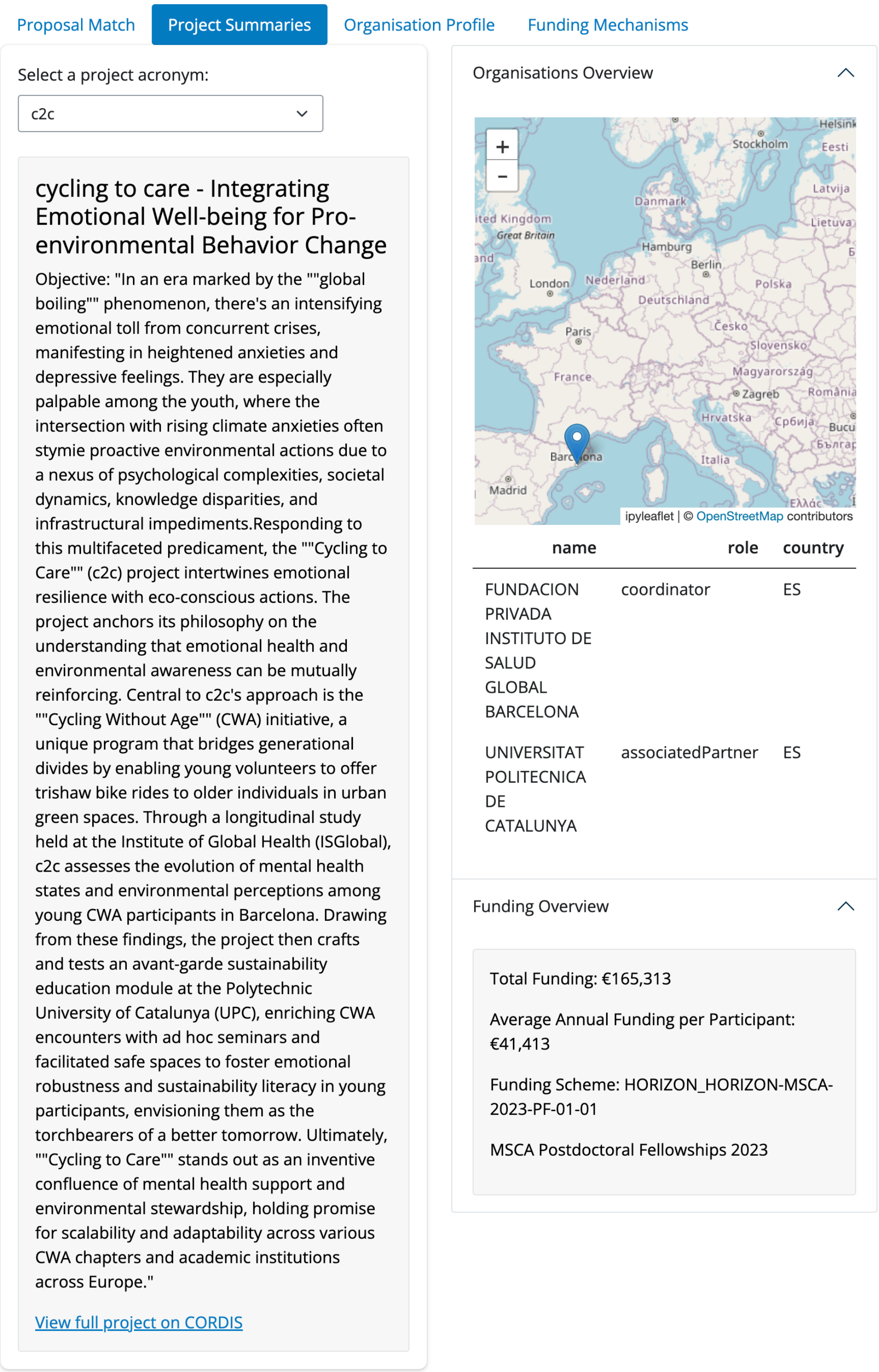
cycling

Number of results to display:

acronym	title
c2c	cycling to care - Integrating Emotional Well-being for Pro-environmental Behavior Change
VELOCARE	Affective Vélomobilities of Care: a Gender-Sensitive Approach to Sustainable Mobility
COMFAM	Linked Lives: Consequences of Long Commuting for the Health of All Family Members
Cogo	Cogo - all ride in one app. Cogo gathers electric shared bikes, scooters, cars and mopeds from 250 mobility operators in more than 700 cities, and is now ready to launch the full-service platform.
LABDA	Learning network for Advanced Behavioural Data Analysis
ROMSUR	The role of shared micromobility modes in the resilience of urban transportation systems: New challenges and opportunities
ROMSUR	The role of shared micromobility modes in the resilience of urban transportation systems: New challenges and opportunities
UTOPCYCLE	UTOPCYCLE: THE SOCIAL MOVEMENT OF PLATFORM COOPERATIVES IN BIKE DELIVERY
MOTOBOOM	The Current Impact of Motorcycle Motorisation on Amazonian Indigenous Peoples
EmbRace	The Embodiment of Racialization: Running Muslim Women and the Sense of Non-Belonging

Figure 7: Proposal Matching Interface.

Following the initial matching, users can proceed to the **Detailed Project Summaries** tab for a deeper investigation of specific projects. By selecting a project acronym from a dropdown menu, the user is presented with comprehensive details, including its full objective, a direct link to its CORDIS page, an interactive map visualizing the geographical locations of its collaborating organizations (with coordinators often distinctly marked), a table of these participants, and a concise funding overview. This tab was found to be exceptionally informative by users, with the map being highly effective for understanding collaborative networks, and the combination of objective text, participant details, CORDIS link, and funding information consistently praised as valuable for thorough research into prior funded work.



Select a project acronym:

c2c

cycling to care - Integrating Emotional Well-being for Pro-environmental Behavior Change

Objective: "In an era marked by the ""global boiling"" phenomenon, there's an intensifying emotional toll from concurrent crises, manifesting in heightened anxieties and depressive feelings. They are especially palpable among the youth, where the intersection with rising climate anxieties often stymie proactive environmental actions due to a nexus of psychological complexities, societal dynamics, knowledge disparities, and infrastructural impediments.Responding to this multifaceted predicament, the ""Cycling to Care"" (c2c) project intertwines emotional resilience with eco-conscious actions. The project anchors its philosophy on the understanding that emotional health and environmental awareness can be mutually reinforcing. Central to c2c's approach is the ""Cycling Without Age"" (CWA) initiative, a unique program that bridges generational divides by enabling young volunteers to offer trishaw bike rides to older individuals in urban green spaces. Through a longitudinal study held at the Institute of Global Health (ISGlobal), c2c assesses the evolution of mental health states and environmental perceptions among young CWA participants in Barcelona. Drawing from these findings, the project then crafts and tests an avant-garde sustainability education module at the Polytechnic University of Catalunya (UPC), enriching CWA encounters with ad hoc seminars and facilitated safe spaces to foster emotional robustness and sustainability literacy in young participants, envisioning them as the torchbearers of a better tomorrow. Ultimately, ""Cycling to Care"" stands out as an inventive confluence of mental health support and environmental stewardship, holding promise for scalability and adaptability across various CWA chapters and academic institutions across Europe."

View full project on CORDIS

Organisations Overview

name	role	country
FUNDACION PRIVADA INSTITUTO DE SALUD GLOBAL BARCELONA	coordinator	ES
UNIVERSITAT POLITECNICA DE CATALUNYA	associatedPartner	ES

Funding Overview

Total Funding: €165,313

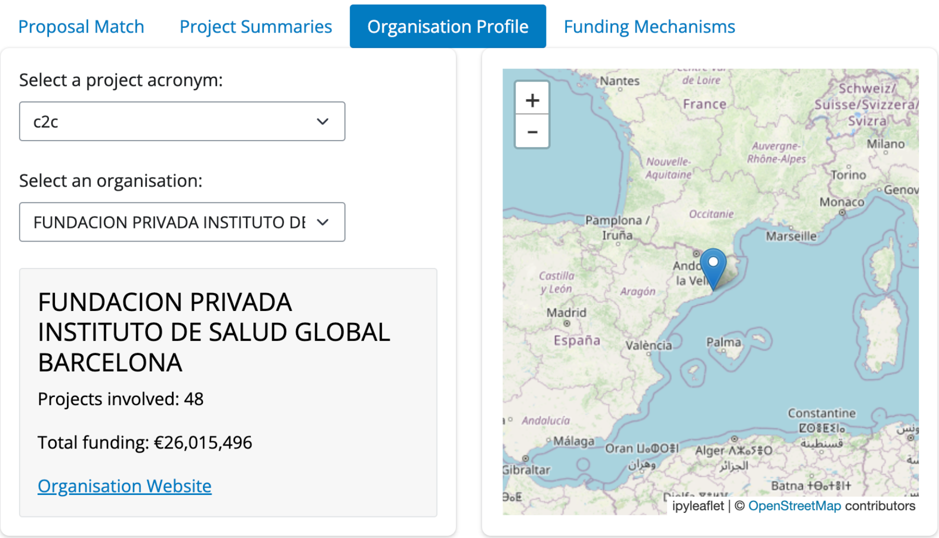
Average Annual Funding per Participant: €41,413

Funding Scheme: HORIZON_HORIZON-MSCA-2023-PF-01-01

MSCA Postdoctoral Fellowships 2023

Figure 8: Detailed Project Summaries Interface.

The application also offers an **Organisation Profile Exploration** tab to facilitate a more granular analysis of participating entities. This feature allows users to first select a project and then a specific organization from that project's participants to view its dedicated profile, which includes its name, the total number of Horizon projects it has participated in, its aggregated funding, a website link if available, and a map pinpointing its location. Users found this tab very useful for quickly assessing an organization's experience and involvement within the Horizon Europe programme, with the clear display of key metrics and direct links aiding in the identification of potential collaborators.



Select a project acronym:

c2c

Select an organisation:

FUNDACION PRIVADA INSTITUTO DI

FUNDACION PRIVADA INSTITUTO DE SALUD GLOBAL BARCELONA

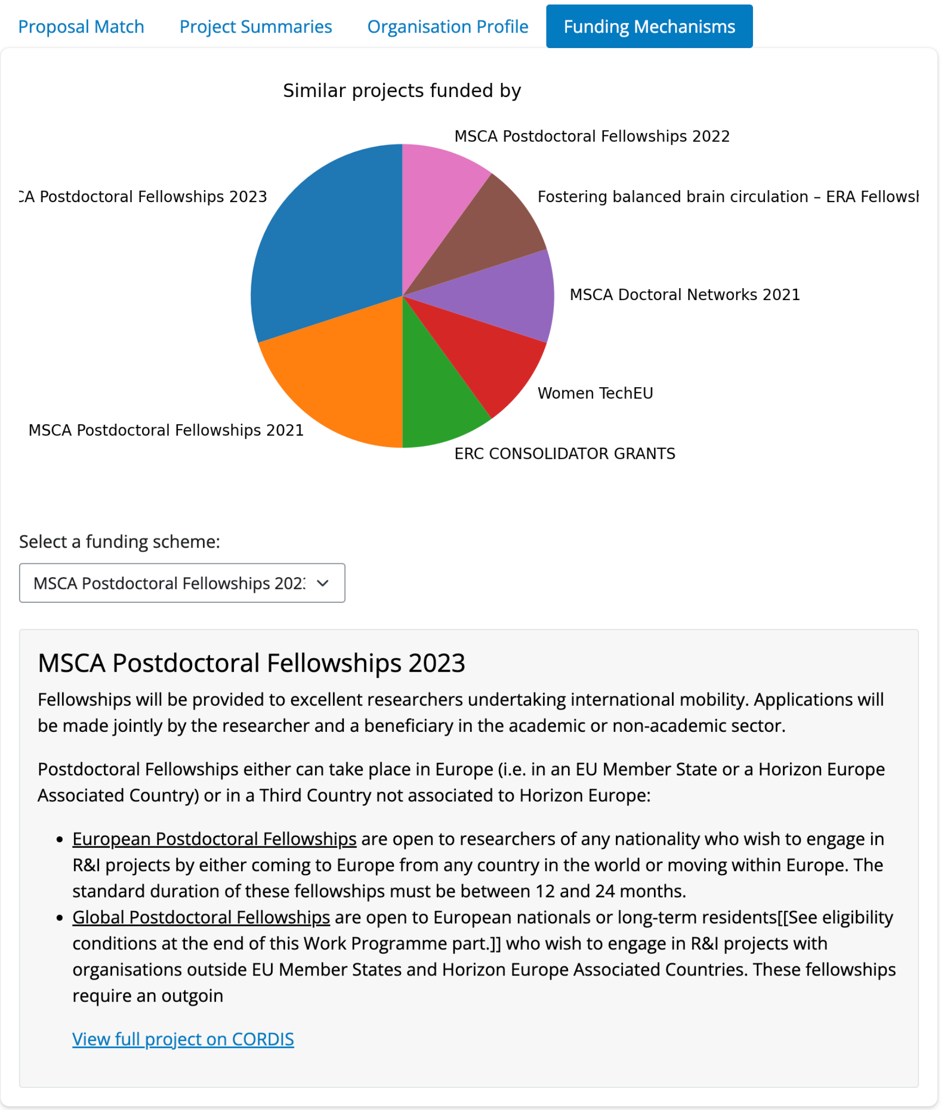
Projects involved: 48

Total funding: €26,015,496

Organisation Website

Figure 9: Organisation Profile Interface.

Finally, the **Funding Mechanism Insights** tab provides context on the funding environment relevant to the user's research idea. This section presents a pie chart visualizing the distribution of funding areas (title_topic) among the projects matched to the user's proposal, offering an immediate overview of aligned program areas. Users can select a specific funding scheme (e.g., "MSCA Postdoctoral Fellowships") from a dropdown to read detailed descriptions of its objectives and eligibility criteria. Crucially, direct hyperlinks to the CORDIS website for these funding schemes are also provided, where researchers can find further details on how to apply. This feature was considered highly advantageous by users for strategic proposal planning and for deeper investigation into potential funding avenues.


Proposal Match
Project Summaries
Organisation Profile
Funding Mechanisms
Similar projects funded by
MSCA Postdoctoral Fellowships 2022
Fostering balanced brain circulation – ERA Fellowsl
MSCA Doctoral Networks 2021
Women TechEU
ERC CONSOLIDATOR GRANTS
MSCA Postdoctoral Fellowships 2021
:A Postdoctoral Fellowships 2023

Select a funding scheme:

MSCA Postdoctoral Fellowships 202:

MSCA Postdoctoral Fellowships 2023

Fellowships will be provided to excellent researchers undertaking international mobility. Applications will be made jointly by the researcher and a beneficiary in the academic or non-academic sector.

Postdoctoral Fellowships either can take place in Europe (i.e. in an EU Member State or a Horizon Europe Associated Country) or in a Third Country not associated to Horizon Europe:

- European Postdoctoral Fellowships are open to researchers of any nationality who wish to engage in R&I projects by either coming to Europe from any country in the world or moving within Europe. The standard duration of these fellowships must be between 12 and 24 months.
- Global Postdoctoral Fellowships are open to European nationals or long-term residents[[See eligibility conditions at the end of this Work Programme part.]] who wish to engage in R&I projects with organisations outside EU Member States and Horizon Europe Associated Countries. These fellowships require an outgoin

View full project on CORDIS

Figure 10: Funding Mechanism Insights Interface.

In summary, the deployed web application enables users to perform semantic matching of their research ideas against Horizon Europe projects and explore relevant project details, participating organizations, and funding mechanisms. This directly addresses the challenge of navigating the programme's vast dataset, thereby aiding researchers in identifying key opportunities and formulating stronger, more competitive proposals.

6 Conclusion and Future Work

6.1 Conclusion

This project successfully developed an interactive web application that leverages Natural Language Processing to help researchers navigate the Horizon Europe CORDIS dataset. By providing semantic matching of research proposals to existing projects and offering detailed insights into project specifics, participating organizations, and funding mechanisms, the tool addresses the challenge of information discovery in a complex funding landscape. The Exploratory Data Analysis confirmed the richness of the dataset, and initial validation tests supported the chosen NLP approach. The deployed application serves as a practical tool for researchers aiming to formulate more targeted and competitive proposals.

6.2 Future Work

While the current application provides significant utility, several avenues for future enhancement have been identified:

- **Automated Data Preprocessing Pipeline:** To improve the robustness and scalability of the data backend, future work could involve automating the entire data ingestion and preprocessing pipeline, for example, using workflow management tools like Kedro. This would streamline the process of updating the application with new CORDIS data releases.
- **Cloud-Based Data Hosting and Advanced Analytics:** Utilizing cloud resources (e.g., for hosting data sources and potentially for more advanced analytics or model re-training) could further enhance performance, scalability, and data management practices.
- **Enhanced User Feedback Mechanisms:** Incorporating user feedback on the relevance of recommendations could be used to fine-tune the matching algorithm over time.
- **Expanded Data Integration:** Integrating additional relevant datasets, such as project publications or outputs, could further enrich the insights provided to users.
- **More Methodical Evaluation:** A more systematic evaluation of the recommender's performance, potentially involving user studies or comparison against other matching algorithms, could provide deeper insights and guide further improvements.

7 Appendix

This appendix provides supplementary information related to the project, including links to the source code repository, the deployed application, and details about the development environment and key dependencies.

7.1 GitHub Repository and Deployed Application

- **GitHub Repository:** The complete source code for this project is publicly available at: `https://github.com/David-TMNg/Modern_Data_Analytics.git`
- **Deployed Web Application:** The interactive web application can be accessed at: `http://13.42.51.136/`

7.2 Environment Setup and Key Dependencies

The project was developed and tested within a Python environment.

- **Python Version:** 3.9.11
- **Key Dependencies:** The major dependencies include (a more comprehensive list can be found in the `requirements.txt` file in the GitHub repository):
  - `pandas==2.2.3`
  - `numpy==2.2.5`
  - `scikit-learn==1.6.1`
  - `sentence-transformers==4.1.0`
  - `torch==2.7.0`
  - `transformers==4.51.3`
  - `shiny==1.4.0`
  - `shinywidgets==0.5.2`
  - `ipyleaflet==0.19.2`

- `matplotlib==3.10.3`
- `seaborn==0.13.2`
- `uvicorn==0.34.2`
- `anywidget==0.9.18`
- `htmltools==0.6.0`
- `ipywidgets==8.1.7`
- `Jinja2==3.1.6`
- `plotly==6.0.1`
- `requests==2.32.3`
- `starlette==0.46.2`
- `websockets==15.0.1`