UNIVERSITAT DE BARCELONA

FUNDAMENTAL PRINCIPLES OF DATA SCIENCE MASTER'S THESIS

Exploring Academic Relationships with UMAP: Dimensionality Reduction and Visualization of Topics and Authors in OpenAlex

Author:
Alba GARCIA ROMO

Supervisor:
Dr. Dimitri MARINELLI
Dr. Albert DIAZ-GUILERA

A thesis submitted in partial fulfillment of the requirements for the degree of MSc in Fundamental Principles of Data Science

in the

Facultat de Matemàtiques i Informàtica

June 30, 2025

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica

MSc

Exploring Academic Relationships with UMAP: Dimensionality Reduction and Visualization of Topics and Authors in OpenAlex

by Alba GARCIA ROMO

This thesis applies Uniform Manifold Approximation and Projection (UMAP) to analyse and visualize research works from the OpenAlex database. By using various embedding methods (including transformer-based models and hierarchical topic encodings) the study demonstrates that UMAP projections can effectively capture meaningful structures in the data, revealing relationships among research areas and institutions. Results show that capturing complex topic relationships across multiple domains is a challenging task. Nevertheless, the visualizations reveal significant thematic clusters and author groupings that align with our data analysis. Quantitative evaluation using clustering metrics, such as the silhouette score, confirms the agreement between visual patterns and semantic embeddings. We also show the impact of UMAP hyperparameters on balancing local and global data structure preservation, which influences visualization clarity and interpretability. The resulting interactive, zoomable visual maps provide researchers with a powerful tool to explore and understand the organization of scientific knowledge.

Acknowledgements

Thanks to Albert and Dimitri for their guidance, patience, and for always being available whenever I had a question or needed support. Their feedback helped me stay on track and continuously improve this work.

To my colleagues and friends from the master's program, thank you for the shared days and afternoons at the library, and the constant encouragement. Your support made this journey more manageable and much more enjoyable.

And finally, to my friends and family, thank you for your constant support and your words of motivation during the most challenging moments.

Contents

Chapter 1

Introduction

Making academic knowledge more accessible and understandable is a key challenge in today's research environment. The sheer volume of research published annually makes it hard to stay updated, even within specific fields, requiring creative techniques to organize and access information (Canon, Boyle, and Hepworth, 2022). Additionally, interdisciplinary research combines diverse vocabularies, methods, and publication practices, complicating the mapping of connections across domains (Marrone and Linnenluecke, 2020). Moreover, understanding relationships between research topics and authors is especially important for identifying trends, collaboration networks, and cross-disciplinary work.

This thesis explores these aspects using the OpenAlex dataset, a large, open catalogue of scholarly metadata.

The main goal of this project is to develop an interactive visualization tool that reveals relationships between academic works, both in topical contents and authorship. We use text embedding models to convert research titles and abstracts into numerical representations. These embeddings are then reduced in dimensionality using UMAP to create two-dimensional data maps for visualization and exploration. Embedding models are well suited for capturing semantic similarities, and UMAP is chosen for preserving high-dimensional data structure (McInnes, Healy, and Melville, 2020).

To build the tool, we experiment with different embedding strategies (including hierarchical topic labels and transformer-based models) and UMAP hyperparameters to study their effect on visualization quality and how well they reveal meaningful data groupings.

The analysis uses data from the University of Barcelona and Utrecht University. While most experiments focus on the UB dataset, including a second university adds variety to the experiments and allows us to study relationships between authors from different institutions.

This thesis is structured as follows. Chapter 2 covers the methodology, including dataset, embeddings, UMAP, and visualization tools. Chapter 3 presents topic relationship experiments, evaluating different embedding models qualitatively and quantitatively. Chapter 4 examines UMAP hyperparameters effects. Chapter 5 focuses on author relationship experiments with author embeddings and visualizations. Chapter 6 concludes and suggests future work.

The GitHub repository of the project is available here, and the final interactive visualization can be seen here.

Chapter 2

Methodology and Theoretical Background

As introduced, the final goal of the project is to build a tool that visualizes and highlights different types of relationships between academic papers. This chapter presents both a theoretical and experimental overview of the components used throughout the project pipeline, from data acquisition to interactive visualization.

The process is as follows: first, we select and collect the necessary data from the OpenAlex dataset. Then, we transform the textual features of the works (such as titles and abstracts) into numerical representations using different embedding strategies and models. Next, we reduce the dimensionality of these high-dimensional vectors to a 2D space using UMAP, an algorithm for dimensionality reduction. This step produces a two-dimensional data map, which we then use to create an interactive visualization code from specific Python libraries.

The following sections describe each of these steps in detail. Additionally, we include a brief section to introduce the clustering evaluation metrics that will be used in the experiments to quantify the resulting clusters in the visualizations.

2.1 The OpenAlex Dataset

We begin by presenting the OpenAlex dataset, which is used throughout this project.

OpenAlex is a fully open catalogue of the global research system (OpenAlex Team, 2025d). Launched in 2022, it initially contained metadata for 209 million works (journal articles, books, etc.); 213 million disambiguated authors; 124,000 venues (e.g., journals and online repositories); and 109,000 institutions (Priem, Piwowar, and Orr, 2022). Currently, OpenAlex indexes over 240 million works works, with about 50,000 new records added daily (OpenAlex Team, 2025c).

The dataset is fully and freely available via a web-based GUI, a full data dump, and high-volume REST API. For this project, we primarily used the REST API to retrieve the necessary data.

The OpenAlex dataset is a heterogeneous directed graph, composed of different types of scholarly entities, and the connections between them. The current OpenAlex entities include works, authors, sources, institutions, topics, publishers and founders (more information can be found in OpenAlex Team (2025b)). Figure 2.1 illustrates the connections among them.

We now describe the key entities that are relevant for this project.

Works are scholarly documents like journal articles, books, datasets, and theses. They are central entities because their connections define the scholarly nature of authors, venues, institutions, and topics. Approximately 50,000 new works are added daily from sources like Crossref, PubMed, and arXiv.

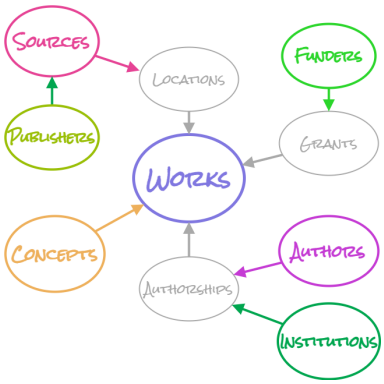


FIGURE 2.1: Sketch of OpenAlex graph model (source: OpenAlex Team (2025b)).

Authors are defined as people who create works. Authors are identified using ORCID IDs when available, though only a small percentage of the author profiles have these identifiers (the percentage is higher for authors of more recent works). To address name disambiguation, OpenAlex uses algorithms that analyse the author's name, their publication record and their citation patterns to reduce misattributions (OpenAlex Team, 2023). Notice that authors are connected to works through the authorship object.

Institutions are organisations to which authors claim affiliations. The identifier for institutions is the ROR ID. Every affiliation is listed by author in order to link institutions to works. These affiliations strings are obtained from both structured sources (eg, PubMed) and unstructured ones (publisher webpages). A two-step algorithm (rules-based and machine-learning-based) is used to extract and normalize affiliation strings. Like authors, institutions are linked to works via the authorship object.

Topics are abstract ideas that works are about. Figure 2.2 shows the Topics hierarchical structure, this structure ensures that each topic maps to exactly one Subfield, Field, and Domain, avoiding ambiguity in classification. Note that there are 4 Domains, 26 Fields, 252 Subfields and 4516 Topics. The full list of topics is available in OpenAlex Team (2025e). See Table 2.1 for an example.

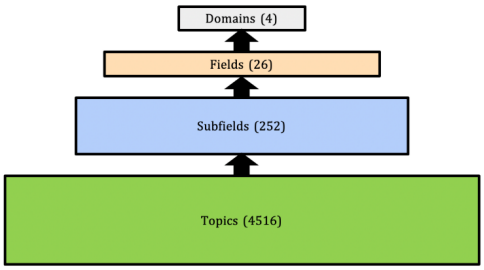


FIGURE 2.2: Topic Structure (source: OpenAlex Team (2025e)).

Works in OpenAlex are tagged with Topics using an algorithm that incorporates both large language models (LLMs) and traditional classifiers, see Figure 2.3. These models consider features such as the title, abstract, source name, and citation context (OpenAlex Team, 2025e). The classification model produces a score for all candidate topics, then the top three topics are assigned to the work, with the highest-scoring one designated as its Primary Topic (OpenAlex Team, 2025a).

Topic:	Natural Language Processing
Subfield:	Artificial Intelligence
Field:	Computer Science
Domain:	Physical Sciences

TABLE 2.1: Example of Topic classification.

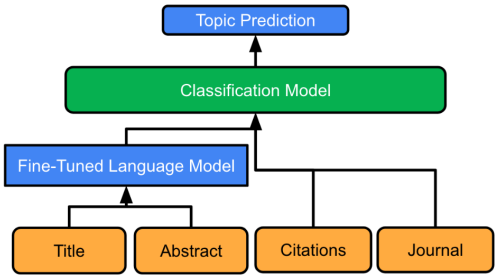


FIGURE 2.3: Outline of Topic Model (source: OpenAlex Team (2025e)).

2.1.1 Data Collection

Now that we understand the basics of the OpenAlex dataset, we describe how it was used in this project.

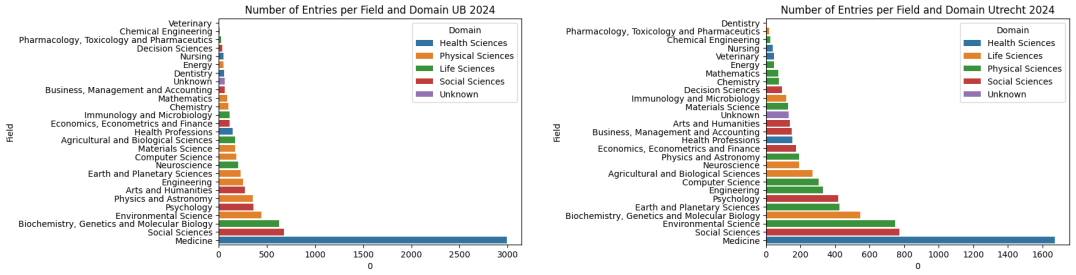
The initial selection criterion was to filter works by the institution of University of Barcelona (UB) and the publication year 2024, generating a dataset of 7,878 works. To introduce more variety into the data, we also used works from Utrecht University published in 2024, with 7,271 entries. Utrecht was selected because it is part of the Charm-EU international university alliance, which includes UB. This second dataset also allows us to explore potential relationships between the two institutions, like collaborations patterns between authors.

Given these filters, the metadata extracted for each work in both datasets includes: title, publication year, list of authors, abstract, and the top three ranked topics. For each topic, we also collected its score, domain, field, and subfield.

2.1.2 Distribution of Fields and Domains

During the exploratory analysis, we examined how works are distributed across domains and fields, based on the field and domain of each work's primary topic. The results revealed a strong imbalance in both datasets, with a dominant concentration in the Health Sciences domain, especially within the field of Medicine (see figure 2.4). This trend was also observed in other universities within the alliance and in other institutions from Barcelona, except for the Universitat Politècnica de Catalunya (UPC), which is primarily focused on engineering disciplines. See Figure A.1 in Annex A for those examples.

This imbalance can be partially explained by the fact that, as of 2022, approximately 23% of all EU scientific publications were in the field of clinical medicine, as reported in Chapter 3 of the Science, Research and Innovation Performance of the EU 2024 report (Directorate-General for Research and Innovation, 2024). One reason for this may be that medical science often produces many short-form publications, such as clinical trials, case reports, and epidemiological studies.



(A) University of Barcelona. (B) Utrecht University.

FIGURE 2.4: Distribution of works per Field and Domain.

Additionally, we have seen that OpenAlex gathers data from highly standardized sources like PubMed. Because of this, it may under-represent research areas that do not have a strong tradition of systematically analysing their own publications (like Humanities). On the other hand, it tends to favour fields like Medicine, where these practices are already well established.

In our experiments, we did not apply any method to address the imbalance across fields, as our goal was to study the topic distribution as it naturally appears in the data. Since both UB and Utrecht exhibit a similar patters, the comparison between the two remains consistent. However, this imbalance implies that fields with fewer publications are less represented in the analysis.

2.1.3 Topic Score Correlations

Another aspect we analysed was the score distribution of the three topics assigned to each work. Our goal was to better understand the behaviour of the OpenAlex Topic Model by exploring the relationship between these three variables. This analysis will also be relevant later, when we study the classification of works by its topics.

To investigate this, we plotted each score against the others and applied a linear regressor to detect possible correlations. The results for the UB 2024 data are shown in Figure 2.5, excluding those scores that were not present in the data. The same analysis is done for Utrecht 2024 data showing the same behaviour, see Figure A.2.

In Figure 2.5 observe that the level of confidence is similar across all three scores. For example, when Score 1 is close to one, Scores 2 and 3 are also close to one. This pattern holds across all scores and shows a strong linear correlation between them.

The coefficients from the linear regression models in the second case, all range from $r = 0.95$ to 0.98. These values confirm the strong positive correlation, indicating that all three scores tend to increase or decrease together. This suggest that the OpenAlex Topic Model is consistent when detecting relevant Topics of a work: when it is confident about the primary topic, it typically also assigns high scores to secondary topics, implying that the work is thematically rich and easily to classify. In contrast, when the model is uncertain about the primary topic, the other scores also tend to be low, suggesting that the work is more difficult to classify within the available topic categories.

2.2 Text Representation with Embeddings

After collecting data, we need to find a way of representing each work based on different criteria. In this section, we present the theoretical background of the method

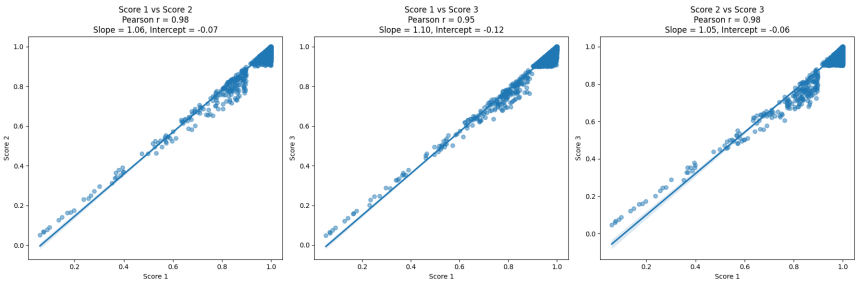


FIGURE 2.5: Correlation between scores in UB 2024 data.

we chose, along with the different models we used.

Our first approach involves building a high-dimensional space where each work is mapped using the Topic, Subfield, Field, and Domain tags as orthogonal dimensions. This represents the "natural" embedding of a work when we are interested in exploring the relationships between works and their assigned topics. This approach is further developed in Section 3.1.

After we study this natural topic space, we will also create representations of the works based on their title and abstracts. The underlying assumption is that papers of the same research area will also have similar title, and abstracts and the other way around; in contrast, works from different areas should have more distinct features. Therefore, we treat the title and abstract as text features that will characterize each work.

Since machine learning algorithms operate on numerical data, we need to convert this textual information into vector representations. This is where text and sentence embeddings come into play. Here we want to give a brief overview of how these models work, but it is not the main focus of this project.

Text embeddings are vector representations that map text into a continuous mathematical space, where semantically similar words or sentences are located near each other. These embeddings are generated by neural network models trained to capture the meaning and context of text.

Transformer-based models, like BERT (Devlin et al., 2019), capture the context of each word by examining each token in the context of every other token. This is done through a self-attention mechanism, which allows the model to weigh the importance of different tokens relative to each other. As the input text passes through successive layers of the transformer, the embeddings are refined and enriched with contextual information. This process results in deeply contextualized vector representations of each token, shaped by the full sequence of input text.

2.2.1 Embedding Models Used in the Experiments

In this section, we describe the different models used in our experiments, along with their main characteristics.

We begin with a model from the Sentence Transformers Python module (Reimers and Gurevych, 2019a). This library offers a wide range of pre-trained models based on the Sentence-BERT (SBERT) model. As described in the original SBERT paper (Reimers and Gurevych, 2019b), the SBERT model is a modification of the pre-trained BERT network that use siamese and triplet network structures to derive semantically meaningful sentence embeddings. This design reduces the computational cost of similarity comparisons, while preserving BERT-level accuracy.

From these pre-trained models models based on SBERT, we selected the model **all-MiniLM-L6-v2**. Trained on a large and diverse dataset of over 1 billion training pairs, according to the documentation (Reimers and Gurevych, 2019a), this model offers strong performance on 14 sentence embedding tasks (relevant to our use case).

Next, we also make use the **Nomic Embed model** (nomic-embed-text-v1), an open-source model with a context length of 8192 tokens. This is also the model used in the example data map from the DataMapPlot examples (McInnes, 2024).

The Nomic model is a modified version of BERT with architectural changes, for example, the activation layer or the batch size. When using this model, it is necessary to include a task-specific instruction prefix in the text input (Nomic AI, 2024). In our case, we used the "clustering" prefix, indicating that the embeddings are intended for grouping similar texts, discovering common topics, or removing semantic duplicates.

Finally, we also use the **SPECTER2 model**, a successor to the original SPECTER model (Cohan et al., 2020). Given the combination of title and abstract of a scientific paper, the model can be used to generate effective embeddings. SPECTER2 has been trained on over 6 million triplets of scientific paper citations, making it particularly well-suited to our task and data (A. Singh et al., 2022).

Overall, our selection of these three models is based on accessibility, performance, and relevance to our dataset. Chapter 3, presents the results produced by these models.

2.3 UMAP Algorithm for Dimensionality Reduction

Once we have the embeddings that represent the works, we need to reduce them to a 2D space so we are able to visualize them. These process is done through dimensionality reduction algorithms. Principal component analysis (PCA), t-distributed stochastic neighbour embedding (t-SNE) and uniform manifold approximation and projection (UMAP) are among these algorithms. This project focuses on UMAP due to the advantages discussed below and builds upon the arXiv example (McInnes, 2024), which also utilized UMAP. In this section, we will give an overview of algorithm without detailing the more theoretical or mathematical aspects, that would go beyond the scope of this thesis.

As we have introduced, UMAP (Uniform Manifold Approximation and Projection) is a flexible, non-linear dimensionality reduction algorithm designed to capture the underlying structure of high-dimensional data. Its goal is to learn the manifold structure of the data and construct a lower-dimensional representation that preserves the essential topological characteristics of that manifold (McInnes, Healy, and Melville, 2020).

Dimensionality reduction techniques generally fall into two categories: global methods like PCA, which attempt to preserve all pairwise distances across the dataset, and local methods, like t-SNE and UMAP, which prioritize preserving the relationships among nearby points, focusing on local neighbourhoods to uncover the data's intrinsic geometry.

In contrast with t-SNE, UMAP offers a better preservation of the data's global structure in the final projection (Coenen and Pearce, 2020). This can be attributed to UMAP's strong theoretical foundations, which allow the algorithm to better strike a balance between emphasizing local versus global structure, which we will explain a bit further. Figure 2.6, from McInnes, Healy, and Melville (2020) shows a comparison between UMAP and t-SNE for different datasets. Note that UMAP successfully

reflects much of the large scale global structure while also preserving the local fine structure, this is also shown in the several examples of Coenen and Pearce (2020).

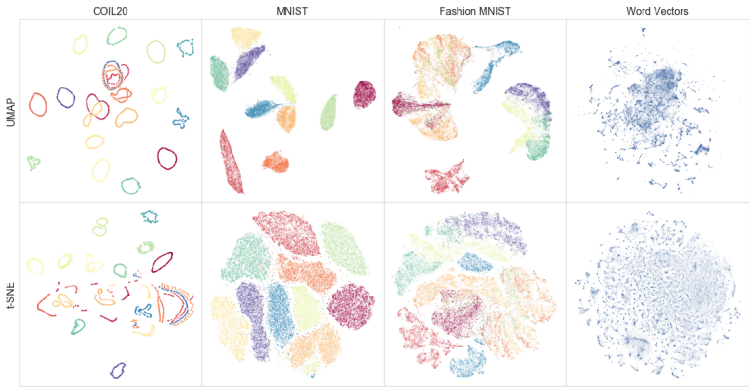


FIGURE 2.6: Comparison of UMAP and t-SNE for several datasets (source: McInnes, Healy, and Melville (2020)).

Another difference from t-SNE is that UMAP is more scalable and faster to compute, which offers a clear advantage when visualizing large datasets. This improvement in performance is analysed in the original UMAP paper (McInnes, Healy, and Melville, 2020) using several datasets.

Let us now introduce how UMAP works, without going into much detail of the theoretical aspects. The algorithm operates in two main phases. First, it builds a weighted n-nearest neighbour graph in the original high-dimensional space, represented in Figure 2.7. This graph encodes the local structure of the data by connecting each point to its nearest n neighbours, where the strength of each connection (the edge weight) reflects the proximity between points. UMAP defines these edge weights using a smooth exponential function that ensures each point is strongly connected to at least its closest neighbour (McInnes, Healy, and Melville, 2020). This weighting function can be interpreted probabilistically, with each edge representing the likelihood that two points are connected.

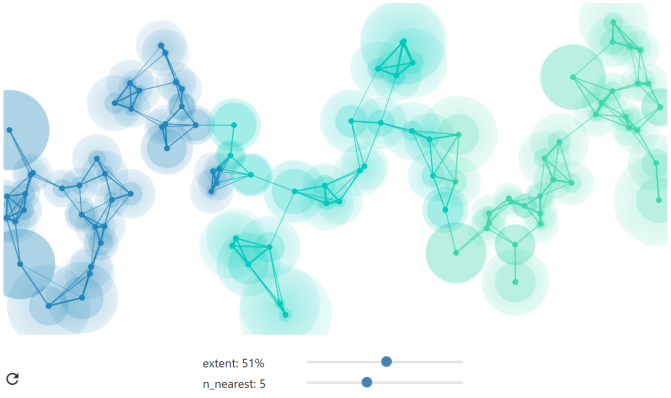


FIGURE 2.7: Representation of the weighted n-nearest neighbour network (source: Coenen and Pearce (2020)).

Given this set of local graphs, we now require a method to combine them into a unified topological representation. Because n-nearest neighbour graphs are inherently asymmetric, UMAP applies a symmetric transformation to the weighted adjacency matrix of G. This step merges local neighbourhood views into a unified topological structure that better captures the global layout of the data.

Once we have this unified graph G, UMAP enters into its second phase. In this step, the algorithm computes a low-dimensional layout of G using a force directed algorithm. This approach uses attractive forces along graph edges to bring connected points closer together and repulsive forces between all pairs of points to prevent overcrowding. The goal is to optimize a non-convex objective function that minimizes the difference between the original graph and the low-dimensional embedding. The result is the low dimensional representation that optimizes this objective function.

2.3.1 UMAP Hyperparameters

Now that we have seen the basic workings of the algorithm, we can study it hyperparameters. In terms of hyperparameters, UMAP provides several variables that are used to fine-tune the lower-dimensional representation. McInnes, Healy, and Melville (2020) describes two key hyperparameters. The first one is the number of neighbours n, already introduced. It defines the local scale at which the manifold is approximated. Conceptually, it sets the size of the local neighbourhood used to estimate the structure of the data manifold. Smaller values of this parameter focus more on capturing fine-grained local structure, potentially at the expense of the global shape, while larger values emphasize broader, large-scale features but may smooth over important local variations. This trade-off means that choosing an appropriate neighbourhood size depends on whether one is more interested in detailed local clustering or the overall topology of the data.

The second one is the `min_dist` variable defined as the desired separation between close points in the embedding space. It controls how tightly UMAP packs points together in the low-dimensional embedding. Unlike n (or `n_neighbors`), which governs the construction of the graph from the high-dimensional space, `min_dist` affects the layout of the embedding itself. Lower values preserve more of the local structure, often resulting in dense clusters, while higher values produce more evenly spaced layouts, which can be beneficial for clarity in visualizations. As such, `min_dist` is often considered an aesthetic parameter.

Lastly, The UMAP library in python also supports a variety of distance metrics through the `metric` parameter (McInnes, Healy, Saul, et al., 2018). While Euclidean distance is the default and works well for many standard use cases, other options such as cosine, Manhattan, or correlation distances can be used depending on the data type and domain. For example, cosine distance is particularly useful when working with high-dimensional sparse vectors such as text embeddings, as it focuses on angular similarity rather than magnitude.

Both the `n_neighbors` and `min_dist` are studied in McInnes, Healy, and Melville (2020) and applied to different datasets. Just for completeness, we have included the example for the MNIST dataset in Figure A.3 in the Appendix A. In our case, we will explore these hyperparameters later when applying the algorithm to our data in Chapter 4.

2.3.2 UMAP Limitations

Finally, we must talk about some of the limitations of the UMAP algorithm. One of the main concerns, as explained in McInnes, Healy, and Melville (2020), lies in interpretability. Like many non-linear methods, UMAP does not offer easily interpretable dimensions in the output space (unlike PCA, where each axis corresponds to a direction of maximum variance in the original data).

Another limitation comes from from its core assumption: that data lies on a well-defined manifold. In datasets with significant noise or very small sample sizes, UMAP may falsely detect structure where there is none. In situations where the dataset has highly variable density across regions, UMAP will attempt to "even out" these differences, which might not be desirable if preserving relative distances is a priority. As the sample size increases, this issue tends to diminish, but it still requires caution in small or noisy datasets.

Additionally, it is important to be cautious when interpreting the geometry of UMAP plots. For example, the size of clusters in the visualization is not meaningful. This is a consequence of the algorithm's focus on preserving local distances when constructing its graph representation. Similarly, the distances between clusters is likely to be meaningless. While it is true that the global positions of clusters are better preserved in UMAP, the distances between them are not meaningful. It still builds the embedding based on local neighbourhoods, which means that the spacing between clusters is largely arbitrary and should not be over-interpreted (Coenen and Pearce, 2020).

2.3.3 Experimental Configuration

In our experiments, we will apply the UMAP algorithm using the `umap` Python library (McInnes, Healy, Saul, et al., 2018). Since we will study the different embedding models first, we will use UMAP with its default configuration, allowing us to obtain a baseline for the embeddings of the data. Later, as mentioned above, we will investigate its hyperparameters in more detail, exploring their effects on the resulting layout and identifying the settings that best suit the characteristics of our dataset.

2.4 Interactive Visualization with `DataMapPlot`

Once we have the 2D dimension representation of the works, the final step of our process is to generate a useful visualization. This visualization needs to support certain features, such as zooming, filtering, and interactivity, so we can adapt it to our needs and extract meaningful insights. For this task, we have chosen the `DataMapPlot` library, which is also the tool used in the example data map from arXiv (McInnes, 2024). Note that all visualization experiments have been tested on Google Chrome and Microsoft Edge and for computer devices, compatibility with other browsers or devices may vary.

`DataMapPlot` is a small library designed to help create aesthetically pleasing data map plots. As described in its documentation Leland McInnes (2023), it can generate both static plots or simple interactive plots, only by passing the data map points and the label clusters of points in the data map.

The main strength of `DataMapPlot` is that it supports the visualization of multiple clustering layers, this reveals hierarchical groupings of data as the user zoom in and out of the map.

In terms of interactivity, the visualizations supports hover tooltips, a search bar, and click behaviour. Tooltips can be fully customized with HTML to display any metadata we want, for example paper titles, authors, topic scores, or links. The search bar allows users to quickly locate specific items, and we can define what happens when a user clicks on a point.

Filtering is another important feature, especially when working with large datasets. Using the filters parameter, we can add filters through and histogram extracted for any categorical variable included in the metadata, for example the year or field of study. This is useful to isolate and explore specific subsets of the data without needing to generate a new map.

Additional customizations are more focused on overall appearance of the visualization. These include both light and dark themes, font control and appearance options for the points in the scatter plot with. These are useful for keeping labels readable and consistent, especially when the map includes many data points. In terms of colour, there are many ways to personalize the style: from adjusting label colour, to shifting the default palette, or applying a custom colour map.

Altogether, these settings make `DataMapPlot` a practical and versatile tool for building clean, interactive, and informative data maps. The combination of interactivity, filtering, and visual customization has been essential for making our topic and author visualizations both engaging and insightful.

2.4.1 Experimental Configuration

Finally, this subsection presents the configuration choices we made for our visualizations and the reasoning behind them. As it will be detailed later, we analyse both topic relationships and author relationships within our data, each of them will present different configurations on the final visualization.

Topic Relationships

For the topic relationship analysis, we include label layers corresponding to the Domain, Subfield, and Field of the Primary Topic. We also add a filter by Primary Field, which helps clarify clusters when applied. Regarding hover information, we display the paper's title along with the names of its three Fields, including Secondary and Tertiary Fields. This setup supports visualizing the correlations between topics that emerged during our analysis, which we will revisit in Section 3.1. For instance, it could be relevant that the Primary Topic of a work falls under Health Sciences, but the Secondary and Tertiary Topics relate to Life Sciences.

The search bar is enabled, allowing users to find works by name, and the on-click feature is also available, linking to the work's details when clicked.

In terms of colour, `DataMapPlot` automatically generates a cyclic HSL palette based on the geometry of the data map. We keep the default palette for scatter plot points since the library detects cluster labels and adjusts colours accordingly. For the hover data, the different Fields are colour-coded using five distinct colours representing the Domains as in Table 2.2.

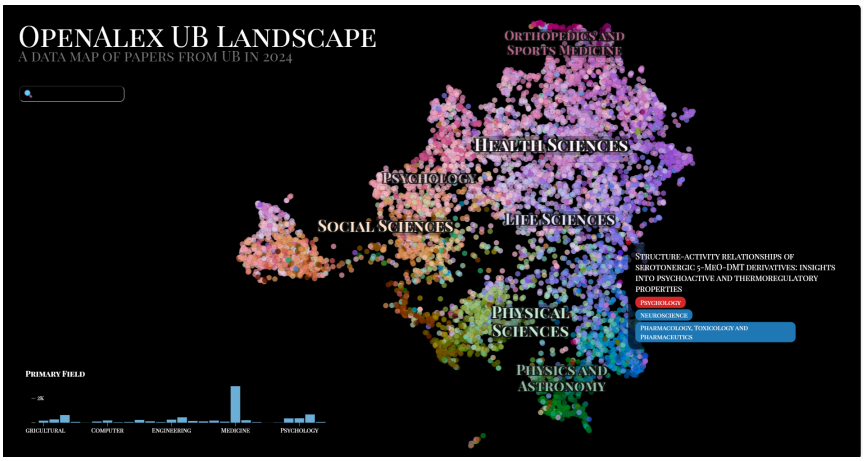
Physical Sciences	#2ca02c, green
Health Sciences	#9467bd, purple
Life Sciences	#1f77b4, blue
Social Sciences	#d62728, red
Unknown	#ff7f0e, orange

TABLE 2.2: Domain Colours.

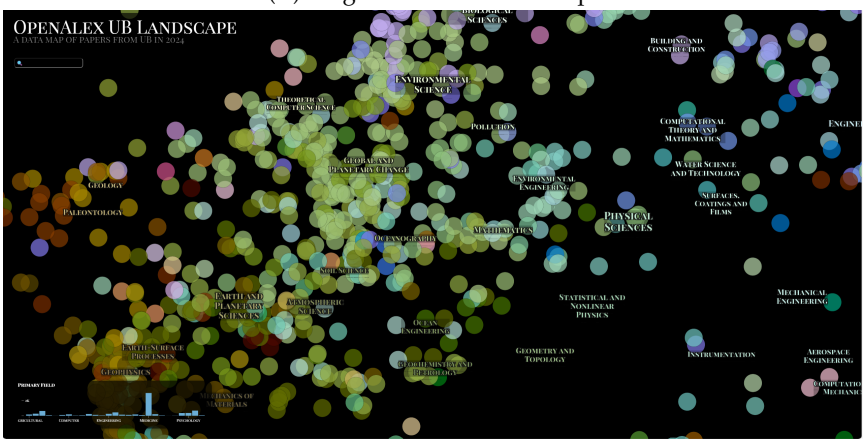
These colours are chosen for their distinctiveness, being opposite each other on the colour wheel, and are used consistently across the project for all visualizations

that filter by Domain. The histogram filter colour, a light blue, is selected to avoid visual conflict with these colours.

Figure 2.8 shows a static preview of the interactive visualization. We will discuss the results in more detail later, including the overall shape and cluster patterns. For now, focus on the different custom visualization options. Observe the topic labels, at the default zoom level, the four Domains are visible (2.8a). As you zoom in, more specific layers appear, revealing the Fields and Subfields (2.8b); in this example, within the Physical Sciences domain. Also observe the tooltip information, with the name of the work and the colour coded fields as we explained. In the image (2.8a), a paper was classified closer to the Life Sciences cluster even when its main topic was from Social Sciences (red). Looking at the metadata in the tooltip, this may be because the other two topics where from the Life Sciences domain (blue). In this case, this extra data gives us more context on the clustering of the works.



(A) Original Zoom and Tooltip.



(B) Fields and Subfields when Zoomed In.

FIGURE 2.8: Preview of the Topics Interactive Map.

Figure A.4 in the Appendix A, shows two examples of using the histogram filter for two different Fields in the visualization.

Author Relationships

For the author visualizations, we use a single label layer representing the Primary Topic Domains of all papers attributed to each author. However, the data points are not coloured based on these labels. Instead, they are coloured by institution, with

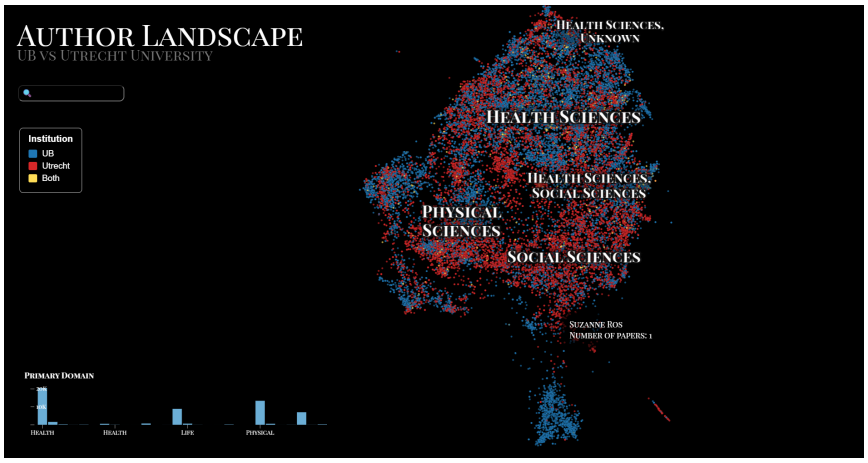


FIGURE 2.9: Preview of the Authors Interactive Map.

a legend added for clarity. This approach helps combine author and topic relationships, making it easier to identify authors working in similar research areas. These configurations are possible thanks to the flexibility of the visualization library. Additionally, we include the same Domain labels as a histogram filter to reinforce the visualization with consistent information. The hover text is customized to display the author's name and their number of publications, and both the search bar and on-click functionalities are enabled using the author name.

Figure 2.9 shows a static preview of the interactive visualization with these customizations. And Figure A.5 in the Appendix A, shows two examples of using the histogram filter for two different Domains in the visualization.

2.5 Clustering Metrics

For each embedding strategy, we also want to be able to interpret the results in a quantitive way, beyond just visual inspection. For this purpose, we rely on established clustering evaluation metrics that provide quantitative insight into how well-defined, compact, and distinct the clusters are. Our intention with these metrics is not to asses the quality of the generated clusters, but to see if they also reflect the behaviour that we see in the qualitative analysis. In this section, we introduce the three metrics used in the experiments: the Silhouette Score (Rousseeuw, 1987), the Davies–Bouldin Index (Davies and Bouldin, 1979), and the Calinski–Harabasz Index (Caliński and and, 1974). A mathematical explanation of the metrics can be found in the Appendix A.

The Silhouette Score is a measure of clustering quality that captures how well a data point is assigned to its cluster relative to others. Its values range from -1 to 1: a score closer to 1 indicates that the point is well-clustered, closer to its own group and far from others, a score near 0 suggests it is located near the boundary between clusters, and a score closer to -1 means it may have been assigned to the wrong cluster.

The Davies-Bouldin Index (DBI) measures clustering quality by assessing both the compactness of clusters and the separation between them. A lower DBI value is preferred, as it implies low internal dispersion and well-separated clusters.

The Calinski-Harabasz Index (or Variance Ratio Criterion) evaluates clustering quality based on the ratio of between-cluster dispersion to within-cluster dispersion. Higher values imply well-defined, compact clusters that are clearly separated.

Chapter 3

Topic Relationships: Evaluation of Embedding Models

In this chapter, we present the experiments carried out to explore topic relationships between works and to comapre different embedding methods. The main goal is to understand how each model behaves and to find which one captures topic-based structure more effectively.

We study two datasets: the works from the University of Barcelona (UB) in 2024, and the works from Utrecht University in the same year. Most of the experiments are done using the UB data, where we compare four different embedding strategies. For the Utrecht dataset, we apply only two of these methods as a complementary analysis.

For the UB dataset, we start by representing the works in a high-dimensional space defined by the hierarchical topic labels from OpenAlex. Then, we move to the text embedding models, applying Sentence Transformers and Nomic model to the work titles. Finally, we include the SPECTER2 model, which uses both the title and the abstract of the works. In contrast, for the Utrecht dataset we apply only the hierarchical topic embedding and the Nomic model.

In all cases for this section, we will generate the 2D map using the default parameters of UMAP (`n_neighbors = 15`, `min_dist = 0.1` and `metric = euclidean`), and evaluate the result both qualitative (by visual inspection) and quantitive.

For the quantitative analysis, we will use three clustering metrics, introduced in section 2.5 of Chpater 2, the Silhouette Score, the Davies-Bouldin Index and the Calinski-Harabasz Index. The three of them are computed using the labels of the Domain of the Primary Topic in all cases.

3.1 Embedding using Hierarchical Topics Vector Basis

We begin with the representation of works by its "natural" embedding (C. K. Singh et al., 2023). This embedding is based on the hierarchical topic tags provided by OpenAlex. For each work, we extract its three Topics along with the associated Subfield, Field, and Domain. We create one-hot encoded vectors for each of these features—Domain, Field, Subfield, and Topic—and concatenate them for each topic. For each work, we then combine the three topic vectors by summing them element-wise. If a feature appears multiple times, such as when several topics share the same Field, this is reflected in the summed vector by values greater than one. Table A.1 in the Appendix A shows an example of this process. Note that all features not shown in the table would have a value of zero.

We also want to note that, in this case, we used a binary variable (1 or 0) to indicate whether a topic is present in a work, instead of using the actual topic scores

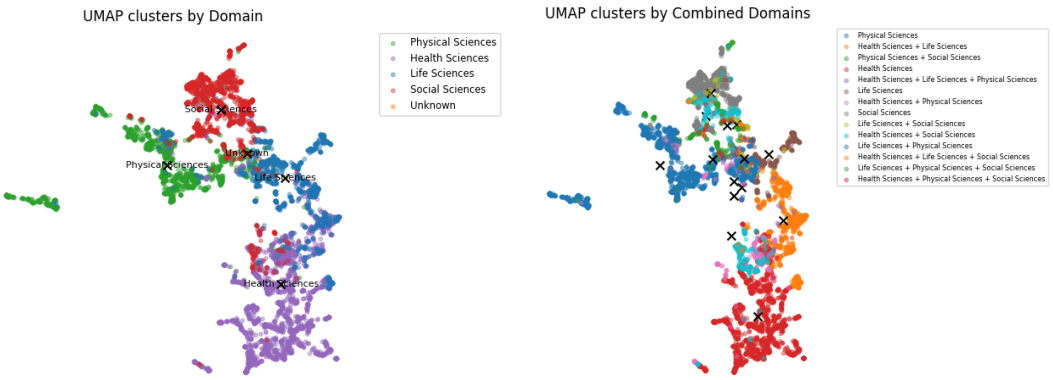
given by OpenAlex. This decision is based on the analysis presented in Section 2.1.3, where we observed a strong linear correlation between the scores of the three topics. In practice, this means that when one topic score is high, the others tend to be high as well, and vice versa. This consistent pattern, suggests that the scores are not providing significantly different information from the presence or absence of a topic. Therefore, for simplicity and to avoid introducing unnecessary complexity, we chose to represent the features as binary variables.

This approach was applied to both datasets introduced earlier: the works from the University of Barcelona in 2024 and those from Utrecht University in the same year. The results are presented in the following sections.

3.1.1 UB Dataset

Using the explained construction, we applied UMAP for dimensionality reduction to the UB data and generated several visualizations[1]. For reference, we include two static visualizations here: Figure 3.1a shows the data map coloured by the Domain of the Primary Topic, while Figure 3.1b uses a colour scheme based on the combination of all three Domains assigned to each work.

In the first figure, we can observe relatively distinct clusters, with Health Sciences overlapping slightly with both Life Sciences and Social Sciences. There is also some overlap between Life Sciences and Physical Sciences. The second figure, however, provides a more detailed view of how cross-domain works are distributed. In the middle section of the map, we see an orange cluster and a light blue cluster representing works that combine Life Sciences and Social Sciences, and Life Sciences with Health Sciences, respectively. These correspond to the overlapping regions seen in the first figure. The light blue dots also appear in the upper section of the map, closer to the Social Sciences cluster, something not visible when only the Primary Domain was considered. Similarly, we now observe pink dots scattered across both the Health and Physical Sciences areas, revealing interactions between these two domains that were previously hidden.



(A) UMAP projection by Primary Domain. (B) UMAP projection by Combined Domains.

FIGURE 3.1: UMAP projections for UB data and Topic embeddings.

As revealed by the visualizations, we find that works are highly dependent on the combination of all three topics, not just the primary one. Moreover, understanding the mix of Topics and Domains present in the dataset is essential for interpreting the results. For this reason, we examine this aspect of the data in more detail. To

[1] An interactive version of the resulting map can be found at the following link.

simplify the analysis, we focus on the four Domains, although a similar study could be conducted using the Fields or Subfields.

Overall, we find that 4,972 works have all three topics assigned to the same Domain, while 2,906 works include at least one topic from a different Domain. This means that approximately 37% of the works can be considered cross-domain. We are also interested in identifying which pairs of Domains are most commonly combined, as this can provide insight into how research areas intersect and which fields tend to collaborate or overlap more frequently. To explore this, we count the occurrences of each pair of different Domains across all works. The results are presented in the co-occurrence matrix shown in Figure 3.2. It is clear that the most common cross-domain pairing is between Health and Life Sciences, more than twice as frequent as other combinations. This may explain why we see more overlap between these two clusters compared to others. We also observe some co-occurrence between Life and Physical Sciences, as well as between Life and Health Sciences, which aligns with the patterns seen in the clusters of Figure 3.1b.

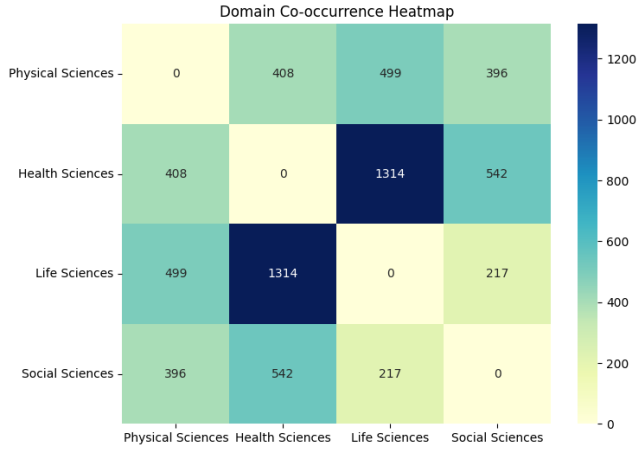


FIGURE 3.2: Domain co-occurrence matrix for UB.

Finally, as part of the quantitative analysis, we applied the clustering metrics introduced earlier in the chapter. The results are shown in Table 3.1. The Silhouette Score of 0.1170 is low, but positive, it suggests that the clusters have some cohesion and separation, though there is also some overlap and the clusters are not very well defined overall. The Davies-Bouldin Index value of 1.0722 indicates moderate cluster separation (lower values, closer to 0, indicate better separation). The Calinski-Harabasz Index, is relatively high at 3619.08, implying that the clustering captures meaningful differences between groups. Overall, these results suggest that the clustering method identifies some relevant structure in the topic embedding space, but the clusters are not sharply defined. Note that this results do align with the previous analysis of the visualization, where we are able to differentiate between the four main groups, but where we also observe overlapping between them.

Clustering Metric	**Value**
Silhouette Score	0.1170
Davies-Bouldin Index	1.0722
Calinski-Harabasz Index	3619.0762

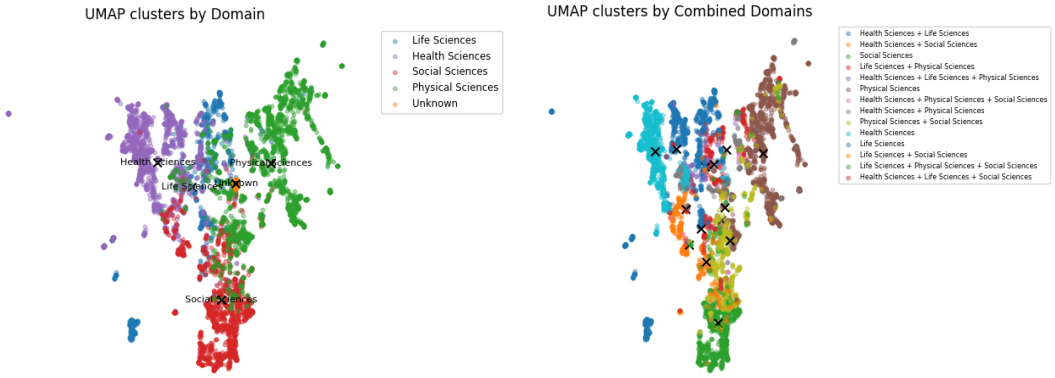
TABLE 3.1: Clustering metrics for Topic embedding of UB data.

In conclusion, the clustering metrics and visualizations together provide a clear understanding of the UB dataset's topic structure. While some distinct clusters are present, there is also considerable overlap, particularly across different domains. This reflects the interdisciplinary nature of the research and the complexity of the topic relationships. These findings offer a solid foundation for the following chapters, where the implications of these connections will be explored further through the use of embedding models.

3.1.2 Utrecht Dataset

In this section, we repeat the analysis performed on the UB data, using the same configuration of the high-dimensional space[2]. Here, we include two static visualizations for reference: Figure 3.3a shows the data map coloured by the Domain of the Primary Topic, while Figure 3.3b presents a colour scheme based on the combination of all three Domains assigned to each work.

As before, in the first visualization, we observe that the five distinct clusters are generally well-defined with some expected overlap. In this case, the central region of the visualization reveals the highest degree of overlap. Specifically, the Life Sciences cluster overlaps much more with the others. Additionally, there is noticeable overlap involving the Social Sciences cluster with both the Health Sciences and Physical Sciences clusters. The second visualization provides a more detailed representation of these relationships. On the left, the combination of Health and Social Sciences, shown in orange, corresponds to the overlap observed between these clusters in the first visualization. Similarly, near the bottom center, a yellow cluster represents the overlap between Social Sciences and Physical Sciences, which is more pronounced in this dataset compared to the UB data. At the center, we still see the Life Sciences cluster, now overlapping with a broader range of cluster combinations.



(A) UMAP projection by Primary Domain. (B) UMAP projection by Combined Domains.

FIGURE 3.3: UMAP projections for Utrecht data and Topic embeddings.

For this dataset, we find that 4,425 works have all assigned topics within the same Domain, while 2,846 works include topics from more than one Domain, meaning that approximately 39% of the works can be considered cross-domain. Referring to the co-occurrence matrix shown in Figure 3.4, which displays the frequency of domain combinations across topic pairs, we can observe the following. While the

[2]As before, an interactive visualization of the resulting map is available at the following link

combination of Health and Life Sciences remains the most frequent, it is less dominant than in the UB data. In contrast, we observe a higher presence of other cross-domain combinations. For instance, between Physical and Social Sciences, which is consistent with the patterns already identified in the combined cluster visualization.

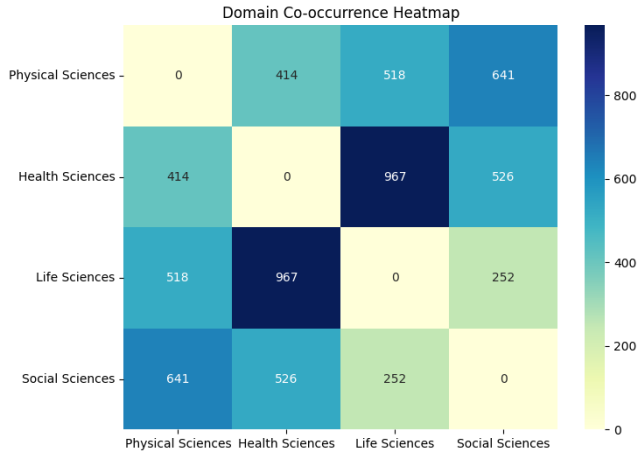


FIGURE 3.4: Domain co-occurrence matrix for Utrecht University.

Finally, as part of the quantitative analysis, we applied the clustering metrics. The results are shown in Table 3.2. The Silhouette Score of 0.1031, suggest a weak but present clustering structure, where data points are, on average, slightly closer to their own cluster than to others. The Davies-Bouldin Index of 1.4738 indicates moderate separation between clusters, and the Calinski-Harabasz Index, with a value of 3056.02, reflects a relatively high ratio of between-cluster dispersion to within-cluster dispersion. Altogether, it implies that this projection captures some structure in the data, with discernible variance between groups. Similar to the UB data, these results suggest that this embedding strategy is able to identify relevant patterns in the topic embedding space, although the clusters are not clearly separated and show considerable overlap.

Additionally, in this case, the clustering metrics perform slightly worse than for the UB dataset, which may suggest that Utrecht University exhibits a higher degree of cross-domain research activity. This observation is consistent with what we know about the data. When comparing the co-occurrence matrices of both institutions, we see that the number of interactions between Domains is slightly higher in the Utrecht dataset than in the UB papers.

Clustering Metric	**Value**
Silhouette Score	0.1031
Davies-Bouldin Index	1.4738
Calinski-Harabasz Index	3056.0227

TABLE 3.2: Clustering metrics for Topic embedding of Utrecht University data.

In conclusion, we obtain results that are very similar to those of the UB data. Based on both the clustering metrics and the visualizations, we observe that while distinct clusters are present, there is also considerable overlap, particularly across different Domains. In this case, cross-domain relationships appear more frequently for certain Domain combinations that were less prominent in the UB dataset. Again,

this reinforces the interdisciplinary nature of the research and highlights the complexity of the relationships between topics.

3.2 Embedding using Transformer-based Models

After analysing the results obtained using the hierarchical topic-based embeddings, we proceeded to explore the transformer-based models introduced in Section 2.2.1. As previously discussed, these models rely on the assumption that works with similar titles and abstracts are likely to be related in terms of content and topic, and vice versa. For the Sentence Transformer and Nomic models, we use only the title as input, while for the SPECTER2 model, we incorporate both the title and abstract.

After obtaining the embeddings for each model, we apply the UMAP transformation (with default parameters) to generate the same interactive visualization as before with `DataMapPlot`. We also replicate the qualitative and quantitative analyses conducted in the previous section. The results derived from these models are presented in the following sections.

3.2.1 UB Dataset

Sentence Transformer Model

We begin by analysing the Sentence Transformer (ST) model `all-MiniLM-L6-v2`, which generates embeddings based only on the title of each work. The titles are passed to the model without any pre-processing, as transformer-based models are trained on naturally language and are designed to handle linguistic variability, including filler words, punctuation, and inconsistent phrasing. Therefore, we do not find it necessary to remove these elements, especially for short texts like titles.

After generating the embeddings, we apply UMAP for dimensionality reduction and examine the resulting visualization and clustering metrics[3]. The static visualization is in Figure 3.5, showing the data map coloured by the Domain of the Primary Topic and it has the centroids of each cluster marked in black.

In this projection, the Health and Life Sciences clusters exhibit the most noticeable overlap, followed by the overlap between the Life and Physical Sciences clusters. Towards the bottom of the visualization, the Social Sciences cluster appears somewhat better defined, although it still shows a degree of overlap with the Physical Sciences cluster.

The clustering metrics for the Sentence Transformer embeddings of the UB data are shown in Table 3.3. The Silhouette Score of 0.0266 is very close to zero, indicating that the clusters are weakly defined and there is significant overlap between them. The Davies-Bouldin Index of 2.7895 is relatively high, suggesting poor separation and high similarity between clusters. The Calinski-Harabasz Index, although still moderately high at 1554.33, is lower than in previous cases, indicating less distinct group structure. Overall, these results suggest that the clustering structure is weak and not well-separated in the embedding space produced by the Sentence Transformer model. This is consistent with what we see in Figure 3.5 where there are not clear clusters overall.

[3] An interactive version of the resulting map is available at the following link

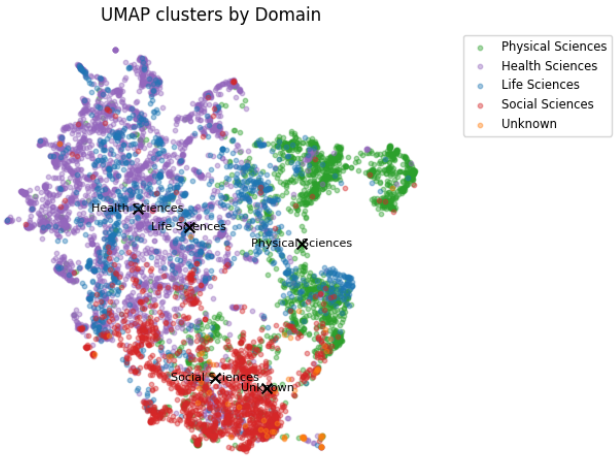


FIGURE 3.5: UMAP projection for ST model and UB 2024 data.

Clustering Metric	Value
Silhouette Score	0.0266
Davies-Bouldin Index	2.7895
Calinski-Harabasz Index	1554.3284

TABLE 3.3: Clustering metrics for ST embedding of UB data.

Nomic Embedding Model

Next, we evaluate the Nomic embedding model, which also uses only the title as input. This model can be configured with a task-specific instruction as prefix in the text input. In our case, we used the "clustering" prefix followed by the work title, indicating that our goal is to group similar title and discover similarities between them. As before, we did not perform any preprocessing of the title texts.

For reference, the static visualization is in Figure 3.6, showing the data map coloured by the Domain of the Primary Topic[4].

In this projection, we also observe that the Health and Life Sciences clusters show the most visible overlap. The Physical Sciences and Social Sciences clusters appear more clearly defined, although they still exhibit a noticeable degree of overlap with the Life Sciences and Health Sciences clusters, respectively. Additionally, we note a small cluster of works on the left side of the visualization, mostly from the Social Sciences domain but also including some from Health and Life Sciences. After exploring the interactive visualization, we find that these works share a common characteristic: they are written in Spanish and have Spanish titles. In this case, the Nomic embedding model appears to capture this linguistic feature, distinguishing these works from the rest.

The clustering metrics for the Nomic embeddings of the UB data are shown in Table 3.4. the results a really similar to the ones for the ST model. The Silhouette Score of 0.0566 is low, indicating weak clustering structure with limited separation between clusters. The Davies-Bouldin Index of 2.3495 suggests moderate to poor separation, with significant similarity between clusters. The Calinski-Harabasz Index, at 1564.62, reflects some variance between clusters but remains relatively modest. Overall, these values point to a weakly defined clustering structure with noticeable overlap between groups, this is also clear with the findings of Figure 3.6.

[4] An interactive version of the resulting map is available at the following link

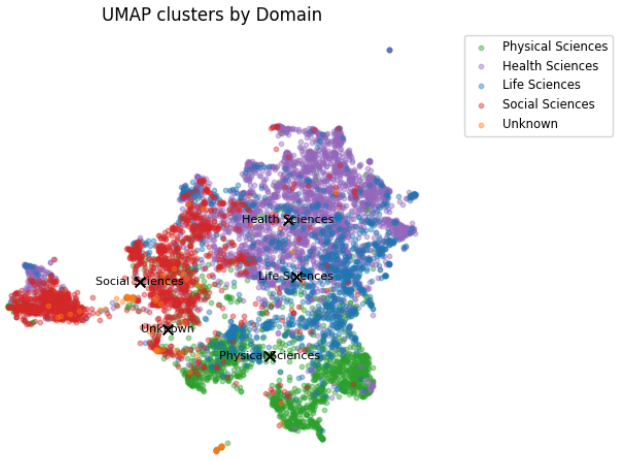


FIGURE 3.6: UMAP projection for Nomic model and UB 2024 data.

Clustering Metric	Value
Silhouette Score	0.0566
Davies-Bouldin Index	2.3495
Calinski-Harabasz Index	1564.6227

TABLE 3.4: Clustering metrics for Nomic embedding of UB data.

SPECTER2 Model

Finally, we analyse the SPECTER2 model, which differs from the previous two by incorporating both the title and abstract of each work. As mentioned earlier, this model is specifically trained on scientific paper data (titles, abstracts, and citation information) which makes it particularly well suited for our context. For this reason, we expect it to better capture the similarities between the papers, more than the other general models. Again, the input texts were passed to the model without any preprocessing.

For reference, the static visualization is in Figure 3.7, showing the data map coloured by the Domain of the Primary Topic[5].

Again, the Health and Life Sciences clusters consistently show the greatest overlap. Also, both the Physical Sciences and Social Sciences clusters appear somewhat better defined in this case, although they continue to exhibit a notable degree of overlap with the Life Sciences and Health Sciences clusters, respectively.

The clustering metrics for the SPECTER2 embeddings of the UB data are presented in Table 3.5. The results are again really similar to the previous models. The Silhouette Score of 0.0775, indicates a modest improvement in cluster cohesion and separation compared to previous models. The Davies-Bouldin Index of 2.1742 suggests moderate separation, with clusters being somewhat more distinct. The Calinski-Harabasz Index, at 1800.49, reflects increased variance between clusters, supporting the presence of a clearer group structure. Overall, these metrics suggest that the SPECTER2 embeddings result in a slightly better-defined clustering compared to the other models. This is consistent with the visualization, where some clusters appear more compact and distinct, although a noticeable degree of overlap remains.

[5]The interactive version of the resulting map is available at the following link

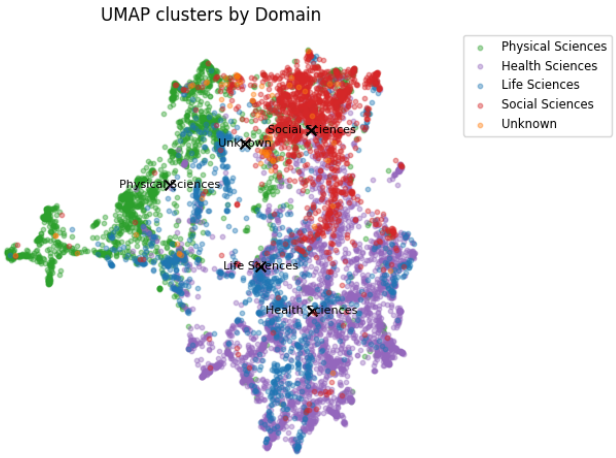


FIGURE 3.7: UMAP projection for SPECTER2 model and UB 2024 data.

Clustering Metric	Value
Silhouette Score	0.0775
Davies-Bouldin Index	2.1742
Calinski-Harabasz Index	1800.4878

TABLE 3.5: Clustering metrics for SPECTER2 embedding of UB data.

3.2.2 Utrecht Dataset

Nomic Embedding Model

Let us now apply Nomic model to the Utrecht dataset. The same configuration is used as in the UB dataset: we embed only the titles of the works without any preprocessing and use the "clustering" prefix provided by the model.

For reference, the static visualization is in Figure 3.8, where the data points are coloured by the Domain of the Primary Topic[6].

In this projection, we observe that the Physical Sciences cluster overlaps with all the other clusters, most notably with Social Sciences, but there is also visible overlap with Life Sciences and Health Sciences. The Life Sciences cluster appears more dispersed and shows clear overlap with the Health Sciences cluster.

The clustering metrics for this case are presented in Table 3.6. The results are similar to those obtained for the UB dataset. The Silhouette Score of 0.0683, suggests moderate cohesion within clusters and some degree of separation. The Davies-Bouldin Index of 1.9653 indicates that the clusters are moderately distinct. The Calinski-Harabasz Index, with a value of 1548.3580, reflects a relatively high variance between clusters. Overall, these metrics suggest that there is some underlying group structure in the embedding, although the clusters are not clearly compact or fully separated, consistent with the visualization.

[6]The interactive version of the resulting data map is available at the following link

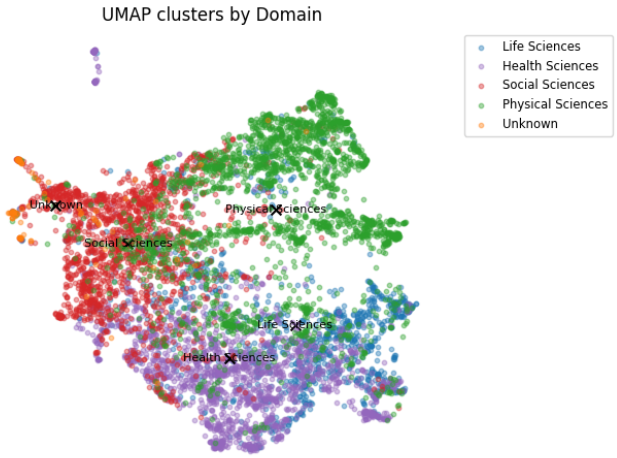


FIGURE 3.8: UMAP projection for Nomic model and Utrecht 2024 data.

Clustering Metric	Value
Silhouette Score	0.0683
Davies-Bouldin Index	1.9653
Calinski-Harabasz Index	1548.3580

TABLE 3.6: Clustering metrics for Nomic embedding of Utrecht data.

3.3 Discussion and Limitations

Now that we have presented the results for the different embedding methods applied to the UB 2024 and Utrecht 2024 datasets, we dedicate this section to a comparative discussion of their outcomes.

First, it is worth noting that the spatial structure produced by the three transformer-based models (Sentence Transformer, Nomic, and SPECTER2) is visually different from the one generated by the topic-based embeddings, for both datasets. In the case of all three embedding models, the resulting UMAP projections exhibit a ring or doughnut-shaped layout (showing that the research areas are not directly connected), contrasting with the more clustered and dense central structure observed in the topic embedding space. Despite this change in geometry, the transformer-based models still manage to capture the most prominent cross-domain relationships identified in the topic-based analysis.

Overall, all results are coherent with the data from the Domain co-occurrence matrices, both for UB data (Figure 3.2) and for Utrecht data (Figure 3.4). For instance, in the case of UB, overlaps between Health and Life Sciences, as well as between Physical and Social Sciences, are consistently reflected across all models. The visualizations are also consistent with the results shown by the clustering metrics, and together, both analyses help us better understand the relationships between the different Domains. This confirms that the visualizations are not only aesthetically coherent but also capture meaningful structural properties of the data, providing confidence in the results obtained through UMAP dimensionality reduction.

In addition, it is also important to consider the type of input each model uses. The topic-based embeddings come from OpenAlex's hierarchical topic classification, which is created using a combination of an LLM and a classifier model. This gives the embeddings a level of interpretability tied to established academic domains. On

the other hand, the transformer-based models generate semantic representations directly from the text based on their specific training. Among these, SPECTER2 stands out because it uses both titles and abstracts, providing richer context compared to models that rely only on titles. These differences in input and model design probably explain some of the variation we see in the clustering results.

One limitation of this analysis is that both the visualizations and clustering metrics rely only on the Domain of the Primary Topic assigned to each work. While this makes interpreting the results easier as a basic reference, it does not fully capture that many works are linked to multiple topics across different domains. It is for this reason that we included the three Fields as tooltip information in the interactive visualizations, giving more context for each work. In the future, it would be useful to explore methods that take all assigned topics into account to better reflect the full complexity of the data.

As future work, clustering metrics could also be applied directly to the original embedding space, not just the 2D projection, to help assess how much information is lost or preserved during dimensionality reduction.

3.4 Conclusion

Let us recap what we have done in this chapter. First, we created a high-dimensional embedding space using the hierarchical structure of the Topics assigned to each work given by OpenAlex. By doing this, we aimed to represent the thematic content of the works in a way that follows the given classification used in the dataset.

Next, we generated data maps of the works by applying UMAP as a dimensionality reduction technique with fixed hyperparameters. We did this both for the topic-based embeddings and for the embeddings produced by three different transformer-based models (Sentence Transformer model, Nomic model and SPECTER2 model). With these data maps, we produce both interactive and static visualizations in order to analyse patterns, clusters, and overlaps between domains.

Looking at the results from the three transformer-based models, we saw that they produced very similar outcomes. All models struggled to form clear clusters, and there was considerable overlap between different domains. Nevertheless, this overlap was consistent with the previous analysis of Domain relationships, showing that domains frequently occurring together in works were also those that overlapped more in the visualizations.

Overall, these results demonstrate the difficulty of capturing complex topic relationships in works that span multiple domains, confirming the interdisciplinary nature of the majority of the research.

Chapter 4

Fine-Tuning UMAP Hyperparameters

After evaluating the different embedding methods, we now turn our attention to fine-tuning the hyperparameters of the UMAP algorithm. In all previous experiments, we used the default settings. In this chapter, we take a closer look at these parameters and explore how they affect the resulting visualizations, considering the possibility that an alternative configuration may better capture the topic relationships observed in the previous chapter.

This analysis will focus only on the UB 2024 dataset and will be applied to both the hierarchical topic embeddings and the SPECTER2 model.

4.1 Number of Neighbours and Min Distance

Let us begin by studying the key parameters of the algorithm, that is `num_neighbors` and `min_dist`. As introduced earlier, `num_neighbors` determines how many neighbouring points are connected when constructing the weighted graph in the high-dimensional space (in the first step of the UMAP algorithm). It controls how much of the local vs. global structure of the graph is captured. In contrast, `min_dist` influences the layout of the resulting embedding, controlling how tightly points are packed together, affecting the visual compactness of clusters.

Figures 4.1 and 4.2 show the results of applying the UMAP algorithm with varying hyperparameters for the UB 2024 dataset with hierarchical topic embedding and SPECTER2 embeddings respectively. As we can see, the `num_neighbors` parameter mostly affects the overall shape of the map. Smaller values, like 5 or 20, produce more fragmented maps with small and dense clusters, while higher values, like 100, lead to smoother and more connected and continuous layouts that capture more global structure. On the other hand, the `min_dist` parameter changes how close points can appear in the 2D space. Low values (0.0125 or 0.05) make the clusters more compact and easier to separate visually, while higher values (like 0.8) spread the points out, which helps avoid clutter but makes the clusters less defined. Observe that these effects appear in both embedding methods, even when they represent different types of information.

In our case, for both embedding methods, we would choose both parameters to be within a medium range, resulting in moderately dense clusters with some degree of overlap and a balanced spread of points. This configuration allows us to preserve the overall shape of the clusters while maintaining interpretability in the visualization.

To complement the visual analysis, we also compute clustering metrics for these projections. The main goal is not to fine-tune the UMAP hyperparameters using

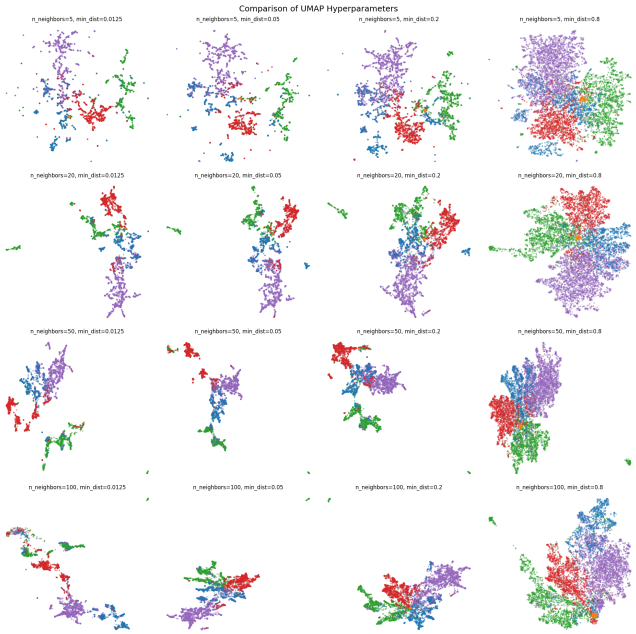


FIGURE 4.1: UMAP projections for UB 2024 with Topic embeddings.

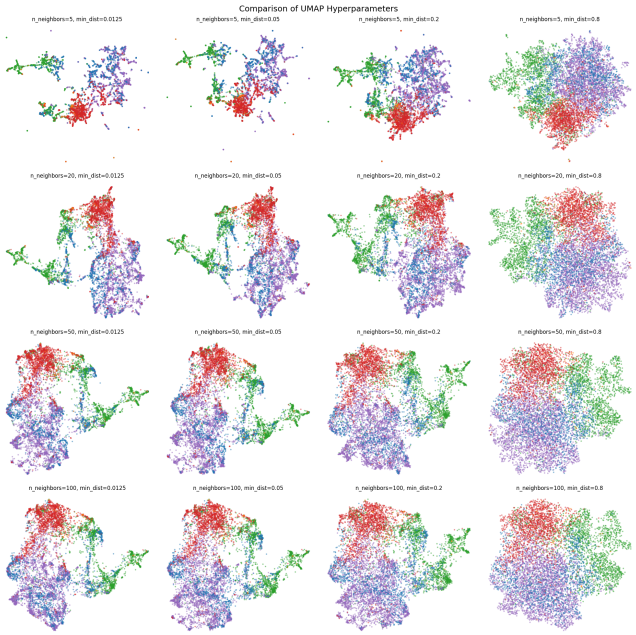


FIGURE 4.2: UMAP projections for UB 2024 with SPECTER2 embeddings.

these metrics, but rather to check whether the patterns seen in the visualizations are supported by quantitative evidence. In other words, we use the metrics to assess whether the clusters that appear more separated and compact in the plots are also evaluated as such numerically.

Tables 4.1 and 4.2 present selected configurations for the UB 2024 dataset using Topic embeddings and SPECTER2 embeddings, respectively. These tables include both the default configuration and other cases that stood out visually. The complete results for all hyperparameters combinations can be found in Appendix A, in Tables A.2 and A.3.

Overall, the metrics align well with the patterns observed in the visualizations. For example, in the Topic embeddings, configurations with `n_neighbors` set to 20 or 50 and smaller `min_dist` values tend to produce more compact and better-separated clusters in the visualizations. These same settings usually result in higher Silhouette and Calinski–Harabasz scores, confirming the visual impressions. A similar trend is observed in the SPECTER2 results, where visually clearer groupings correspond to better metric values.

The default configuration (`n_neighbors = 15`, `min_dist = 0.1`) performs reasonably well in both cases, showing a good balance across all metrics. However, configurations like `n_neighbors = 20` and `min_dist = 0.0125` or `min_dist = 0.0125` sometimes show slight improvements and produce equally interpretable visualizations. Note that these are also the medium-range values as we mentioned before.

n_neighbors	min_dist	**Silhouette**	**Davies-Bouldin**	**Calinski-Harabasz**
15	0.1000	0.1170	1.0722	3619.0722
5	0.0500	0.0399	3.3972	1727.3931
20	0.8000	0.0982	0.9953	3272.9463
50	0.0125	0.2004	1.1493	2248.9446
100	0.0125	0.1267	1.9182	1758.5996

TABLE 4.1: Selected clustering metrics for UMAP hyperparameters configurations on the UB 2024 dataset using Topic embeddings.

n_neighbors	min_dist	**Silhouette**	**Davies-Bouldin**	**Calinski-Harabasz**
15	0.1000	0.0775	2.1742	1800.4878
5	0.8000	0.0597	2.9995	1523.1870
20	0.0125	0.1006	2.0738	1991.3978
50	0.0125	0.0930	2.3555	1815.8835
100	0.8000	0.0597	3.0772	1706.9482

TABLE 4.2: Selected clustering metrics for UMAP hyperparameters configurations on the UB 2024 dataset using SPECTER2 embeddings.

4.2 Distance Metric

Now we study the impact of the distance metric. As mentioned earlier, the default distance metric used by the UMAP Python library is the Euclidean distance. In this section, we explore other available options to see if a different metric can better reflect the structure of our data and lead to improved results. For this experiment, we go back to using the default values for `n_neighbors` and `min_dist`.

Figures 4.3 and 4.4 show the results of applying the UMAP algorithm with different distance metrics to the UB 2024 dataset, using hierarchical topic embeddings and SPECTER2 embeddings respectively. Overall, in both scenarios we find that the visualizations remain relatively stable across the various distance metrics, producing similar shapes overall. Some exceptions are the cosine and correlation distances for the Topic embedding model, where clusters appear more clearly defined and compact; however, this result does not fully align with what we known from interdisciplinary nature of the data (there must be some overlap between different domains). Another exception is the matching distance applied to the SPECTER2 embeddings, which results in scattered points that do not capture any meaningful structure from the embedding space.

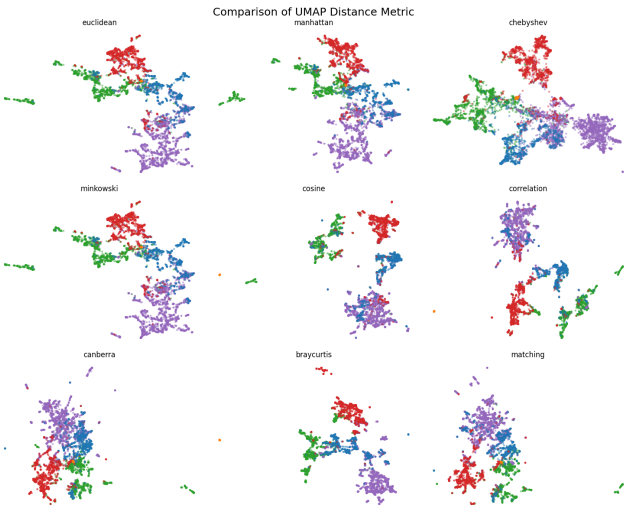


FIGURE 4.3: UMAP projections for UB 2024 with Topic embeddings.

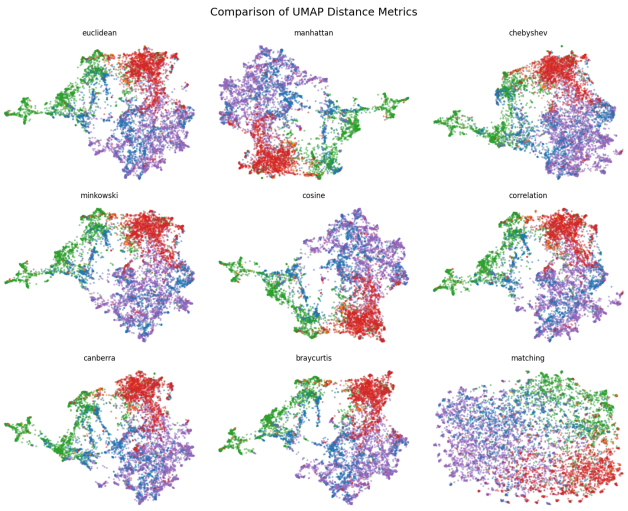


FIGURE 4.4: UMAP projections for UB 2024 with SPECTER2 embeddings.

We also used clustering metrics to quantitatively assess whether the structures observed in the visualizations are supported by numerical evidence.

For the Topic embeddings (Table 4.3), the cosine and correlation distance metrics showed higher Silhouette scores and relatively low Davies-Bouldin indices, which

align with the clear, well-defined clusters seen in the visualizations.

Similarly, for the SPECTER2 embeddings (Table 4.4), the Bray-Curtis and Canberra distances achieved the best clustering metric values, indicating more compact and distinct clusters in their visualizations. However, overall the clustering metrics show less variation across different distance metrics compared to the Topic embeddings. This consistency is also reflected visually, where the SPECTER2 projections maintain a more stable shape regardless of the distance metric used.

Distance Metric	Silhouette	Davies-Bouldin	Calinski-Harabasz
euclidean	0.1170	1.0722	3619.0762
manhattan	0.0633	1.1092	3034.6414
chebyshev	0.2039	0.8897	4323.1968
minkowski	0.1170	1.0722	3619.0762
cosine	0.3482	0.8303	4073.1016
correlation	0.3535	0.9298	3976.5881
canberra	0.0000	1.4135	1861.3761
braycurtis	0.2537	0.9607	3614.1604
matching	0.0571	1.3221	2321.8289

TABLE 4.3: Clustering metrics for different UMAP distance metrics using Topic embeddings.

Distance Metric	Silhouette	Davies-Bouldin	Calinski-Harabasz
euclidean	0.0775	2.1742	1800.4878
manhattan	0.0795	2.2870	1700.1873
chebyshev	0.0859	2.3645	1760.1417
minkowski	0.0860	2.8441	1660.1942
cosine	0.0861	2.2863	1745.8170
correlation	0.0870	2.6180	1670.8789
canberra	0.0964	2.0740	1961.0160
braycurtis	0.0986	2.2034	1729.1510
matching	0.0334	3.1105	1285.3419

TABLE 4.4: Clustering metrics for different UMAP distance metrics using SPECTER2 embeddings.

4.3 Limitations

Let us comment on some limitations of our analysis. Even though it is not possible to test all ranges of hyperparameters due to practical constraints, one limitation of our analysis is that we tested only a limited set of values for `n_neighbors` and `min_dist`. Because of this, there may be better configurations outside the tested values that we did not explore. Additionally, the experiments were conducted solely on the UB 2024 dataset, so the results might not generalize to other datasets or domains without further validation.

4.4 Conclusion

In this chapter, we evaluated different UMAP projections by varying the algorithm's hyperparameters for the UB 2024 dataset using hierarchical topic embeddings and

SPECTER2 embeddings.

Our experiments with different values for `n_neighbors` and `min_dist` showed that these parameters can produce quite different visualizations, so it is important to understand their effects. In our case, mid-range values of `n_neighbors` (around 15 or 20) and low values of `min_dist` (0.0125 to 0.1) produced visualizations that better reflected the behaviour we have seen for cross-domain works, showing the necessary overlap in the clusters. This result was reflected both visually and with the clustering metrics.

Regarding the distance metrics, we found that their impact on the final visualization varies depending on the embedding model used. For example, the SPECTER2 model appeared more stable overall, while the Topic embeddings showed more variation, as reflected in both the resulting shapes and the clustering metrics. Another example, cosine and correlation distances produced more compact, non-overlapping clusters with the Topic embeddings, but did not have the same effect with the SPECTER2 embeddings.

Chapter 5

Author Relationships

In this chapter, we move from exploring topic relationships to studying author relationships across different institutions. The main goal stays the same: to create a visualization tool that highlights the most relevant patterns and helps us gain insights from the data.

For this part of the project, we use only the Nomic embedding model, applying it to both the UB 2024 and Utrecht 2024 datasets. The next sections present the experiments carried out to explore the author relationships between these two institutions.

5.1 Author Considerations and Embeddings

Let us begin by explaining how we process the works to extract author-related data. As described earlier, authors are linked to each work through the "authorship" object provided by OpenAlex. To retrieve this information, we read from that entity and extract the name and ID of each author associated with every work. This allows us to generate a list of authors for each work.

Next, we apply a filtering step by removing works that list more than 20 authors. We do this because these works can introduce noise into the analysis and may represent large-scale collaborations that are not the focus of this study. Such collaborations could be studied separately, but here we are more interested in the general author relationships within the datasets. Figure 5.1 shows the distribution of the number of authors per work in the UB 2024 dataset. As the figure illustrates, most works have between 1 and 15 authors, with a noticeable drop beyond that range. It is also worth noting that OpenAlex limits the maximum number of authors per work to 100, which is reflected in the tail of the distribution.

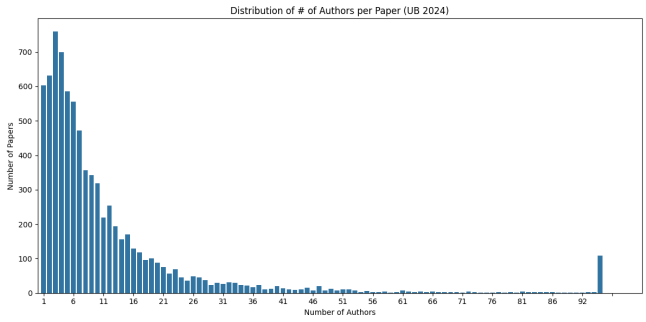


FIGURE 5.1: Distribution of number of authors per work.

After filtering, we restructure the dataset using the `explode` method. This operation creates one row for each author–work pair, effectively flattening the list of authors into individual records. With this format, we group the data by author ID

and collect all the works that each author has contributed to. For each author, we compute a mean embedding vector by averaging the embeddings of all their works. This averaged vector serves as a representation of the author in the high dimensional space. Additionally, for each author, we also collect the Domain of the Primary Topic of all their works, these we will use later in the visualization as metadata.

Finally, we apply the same UMAP reducer that was trained on the work embeddings to project the author embeddings into the same two-dimensional space.

5.2 Joint Visualization for UB and Utrecht Datasets

To construct the visualization tool for the authors, we follow the steps described previously. In this case, we use the Nomic embeddings for the works and recompute the UMAP reducer using the combined set of works from both institutions. This ensures that the reducer captures the geometry of the shared high-dimensional space. The author embeddings and their corresponding 2D projections are then computed using this joint UMAP reducer, as described earlier. Finally, we generate the interactive visualization using `DataMapPlot`, following the configuration detailed in Section 2.4.1.

To analyse the visualization, we will return to Figure 2.9[1]. We observe several patterns. Firstly, the author points appear to be distributed relatively evenly across all domains. There are no clearly defined clusters of blue (UB) or red (Utrecht) authors, except for a small blue island at the bottom of the map. This matches the projection of the UB 2024 data using the Nomic model, where we previously identified a similar cluster corresponding to papers written in Spanish, now fully blue, as they are exclusively from UB. Even though the clusters are not sharply defined, we notice that the top region of the visualization contains more UB authors, while the lower part includes a higher concentration of Utrecht authors. This aligns with our earlier observations about the Health Sciences domain, which is more heavily represented in the UB dataset than in Utrecht's.

5.3 Discussion and Limitations

To further discuss our results, we need to consider several limitations. First, the method assumes that averaging the embeddings of all works gives a meaningful representation of an author's research profile. While this may work well for authors who focus on a single research area, it may not reflect the profiles of those who publish in multiple, unrelated fields. A more refined approach, such as assigning weights to works or filtering them based on publication year, could provide more accurate author embeddings, especially if future experiments include data from other years.

Another point to consider is the combination of a linear operation, like averaging, with a non-linear method such as UMAP. Averaging assumes that the embedding space behaves in a mostly linear way, and that the meaning of a group of texts can be captured by their average. This works reasonably well in high-dimensional spaces created by models like Nomic or SPECTER2, which are designed to place similar texts close together. However, since UMAP is a non-linear algorithm that focuses on preserving the local structure of the data, the compatibility of the two

[1] An interactive version of the resulting map can be found at the following link.

methods may not be guaranteed. Although this approach still helps to reveal useful patterns, especially when looking at a large number of authors, the interaction between linear averaging and non-linear dimensionality reduction may limit how accurately relationships are represented. Future work could explore whether other non-linear aggregation methods work better alongside UMAP.

5.4 Conclusion

The author visualizations presented in this chapter offer an exploratory approach to mapping author relationships through aggregated work embeddings. By averaging the embeddings of each author's publications, we generate a single vector representation that allows authors to be positioned in the same semantic space as individual works. This approach provides a useful way to observe topical similarities between researchers and to explore patterns across institutions.

While there are limitations to this approach, the resulting visualization serves as a practical exploratory tool. It helps us examine how authors are distributed across research areas and highlights potential collaboration patterns between institutions. Overall, we have achieved our goal of building a visualization that effectively captures and reveals meaningful author relationships within the data.

Chapter 6

Conclusion

This thesis has explored the challenge of visualizing the structure of academic knowledge by leveraging open scholarly metadata and modern machine learning techniques. Using the OpenAlex dataset, we developed an interactive visualization tool that captures relationships between academic works both in terms of topical content and authorship.

The experiments in Chapter 3 showed that embedding models—especially those based on transformer architectures and hierarchical topic labels—provide consistent and meaningful semantic representations of research works. When these embeddings are combined with the UMAP dimensionality reduction algorithm, it becomes possible to generate two-dimensional maps that reveal clear clusters of related topics. We found that the visualisations created with these models reflected the relationships between topics previously identified. For example, in the UB 2024 dataset, all embedding models showed noticeable overlap between Health Sciences and Life Sciences, which matched what we had observed in earlier analyses. We also applied clustering metrics to support these visual insights with quantitative results, showing a good match between visual cluster overlap and numerical evaluation. One interesting pattern we noticed was the frequent appearance of a doughnut-like shape in the maps, suggesting that some research areas are not directly connected to others.

In Chapter 4, we explored the impact of UMAP hyperparameters and observed how these settings influence the layout and interpretability of the visualizations. For our data, we saw that mid-range values often produces more coherent and readable representations, where the spacing between points and the compactness of clusters was better balances, making the visualisations easier to interpret.

In Chapter 5, we focused on the analysis of author relationships. We constructed author embeddings by aggregating the embeddings of their associated works and found that the resulting maps meaningfully captured the relationships between researchers. These visualizations offer a valuable resource for identifying authors working in similar research areas, serving as a useful tool for exploring potential collaborations or understanding the structure of research communities.

Even with these positive results, we must have in mind the following limitations. Firstly, the inherent limitations of UMAP, which can be reflected in our visualizations. For instance, while the overall structure of clusters is clear and consistent with the semantic content of the works, we must be cautious when interpreting the distances between clusters or their specific layout in the 2D space. In several maps, clusters that appear far apart may in fact share related topics, and the size of clusters is not always proportional to the number of works they contain. These effects are expected, given UMAP's focus on preserving local relationships rather than global geometry.

Another limitation comes from the way we used the Primary Topic and Domain labels to colour the visualizations. As mentioned earlier, this can sometimes lead to a simplified view of the data, especially for works that belong to more than one topic or do not clearly fit into a single domain. We tried to address this by including extra metadata in the visualization, helping to better explain the context of each work within its cluster. Still, it is important to keep in mind that the topic labels from OpenAlex are assigned automatically, so some misclassification might be present, which could affect how certain clusters or topic relationships appear in the final maps. Nevertheless, these labels are assumed to be accurate for the purposes of this project.

Finally, we want to discuss several ways this project could be improved in the future. One important step would be to turn the current prototype[1] into a more complete and optimized website. Instead of using a static dataset, the tool could connect directly to the OpenAlex API and allow live queries that generate visualizations on the spot. It would also be helpful to make sure the tool works well across different browsers and devices.

Additionally, a stability analysis of the current results could be carried out by running the embedding and visualization processes multiple times with different random seeds. This would help assess the robustness of the patterns observed and increase confidence in the consistency of the visualizations.

Another area worth exploring is the way OpenAlex assigns topics to research works. Gaining a better understanding of this process could help identify possible errors or inconsistencies in the topic classification. In addition, the visualizations could be improved by going beyond the primary topic. Including secondary and tertiary topics would provide a more complete view of each work and help produce more meaningful results, especially in the context of interdisciplinary research.

Finally, another interesting direction for future work would be to analyse how researchers move between different areas of knowledge over time. By looking at author trajectories, it would be possible to explore whether authors stay within the same research area or shift focus and collaborate across disciplines, as shown in previous studies like C. K. Singh et al. (2023). This could reveal important patterns in the evolution of scientific careers and the development of interdisciplinary work.

[1] Here is the link the the prototype website.

Appendix A

Additional Information

A.1 Mathematical Details for Clustering Metrics

The Silhouette Score for a point i is defined as:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where $a(i)$ is the average distance from point i to other points within the same cluster, and $b(i)$ is the smallest average distance from point i to points in any other cluster.

The Davies-Bouldin Index is defined as:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \frac{R_{ii} + R_{jj}}{R_{ij}}$$

where k is the number of clusters, R_{ii} and R_{jj} denote the compactness of clusters i and j, respectively, and R_{ij} denotes the distance between clusters i and j.

The Calinski-Harabasz Index is calculated as:

$$CH = \frac{B/(K-1)}{W/(N-K)}$$

where B is the between-cluster sum of squares, W is the within-cluster sum of squares, N is the total number of data points, and K is the number of clusters.

The between-cluster sum of squares is defined as:

$$B = \sum_{k=1}^{K} n_k \|C_k - C\|^2$$

and the within-cluster sum of squares is defined as:

$$W = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \|X_{ik} - C_k\|^2$$

where n_k is the number of observations in cluster k, C_k is the centroid of cluster k, C is the global centroid, and X_{ik} is the i-th point in cluster k.

A.2 Extra Figures and Tables

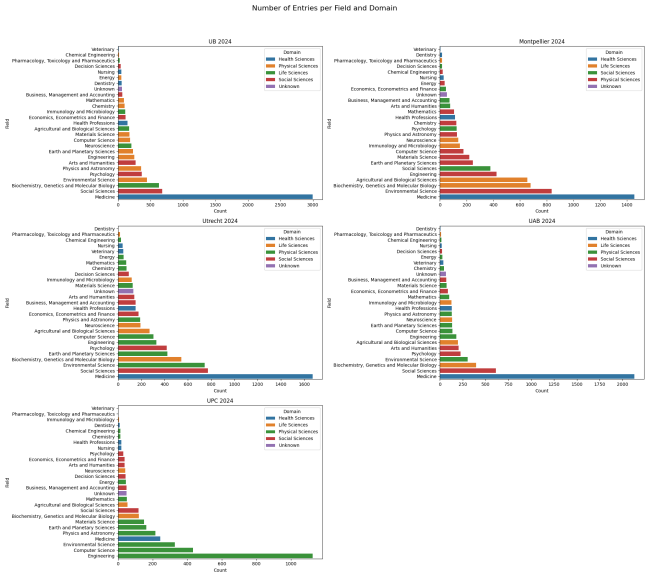


FIGURE A.1: Distribution of works per Field and Domain for different universities.

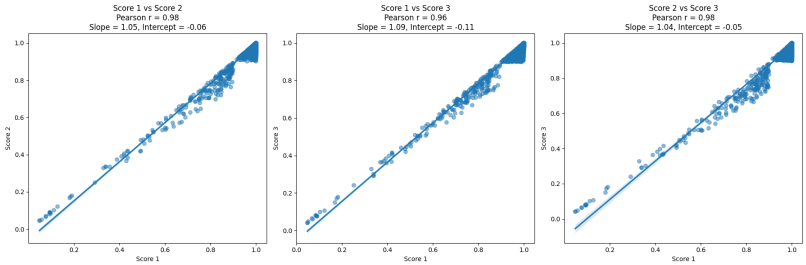


FIGURE A.2: Correlation between scores in Utrecht 2024 data.

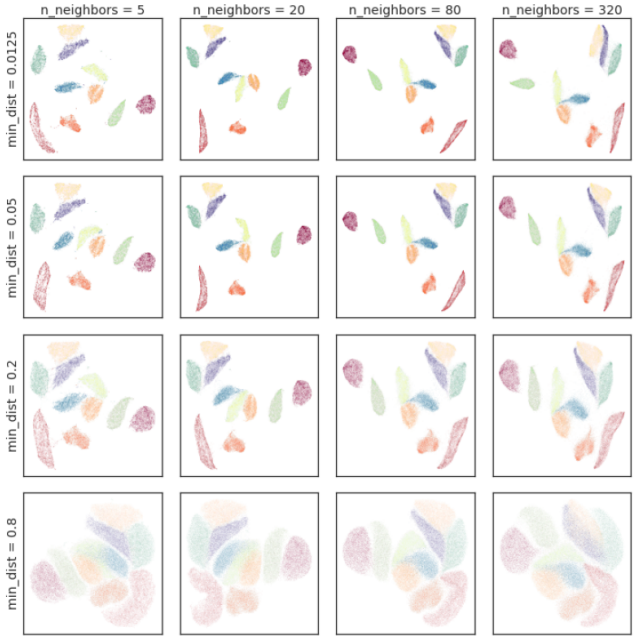
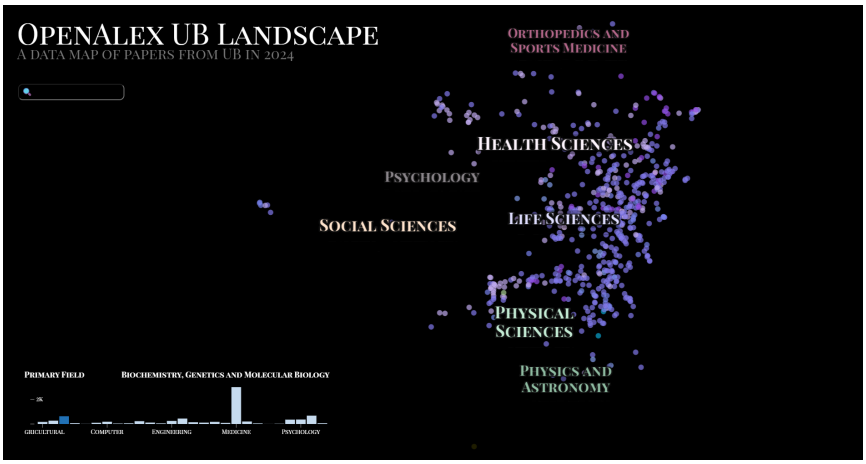
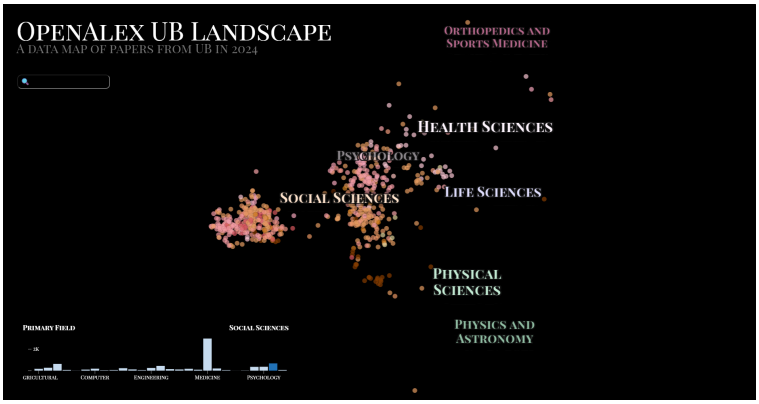


FIGURE A.3: Variation of UMAP hyperparameters for the MNIST dataset (source: McInnes, Healy, and Melville (2020)).



(A) Filter in Life Sciences domain.

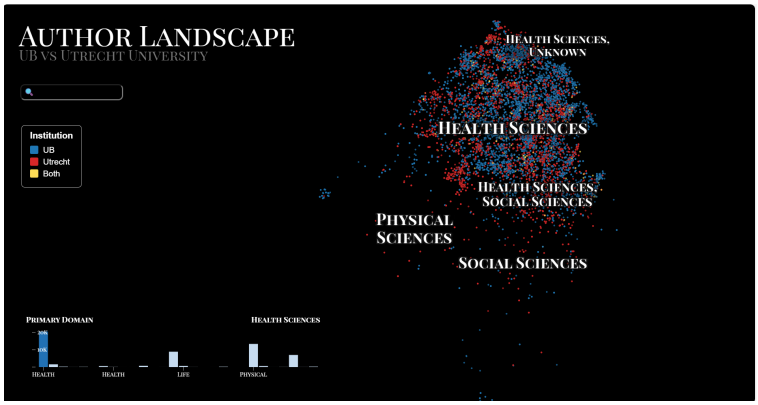


(B) Filter in the Social Sciences domain.

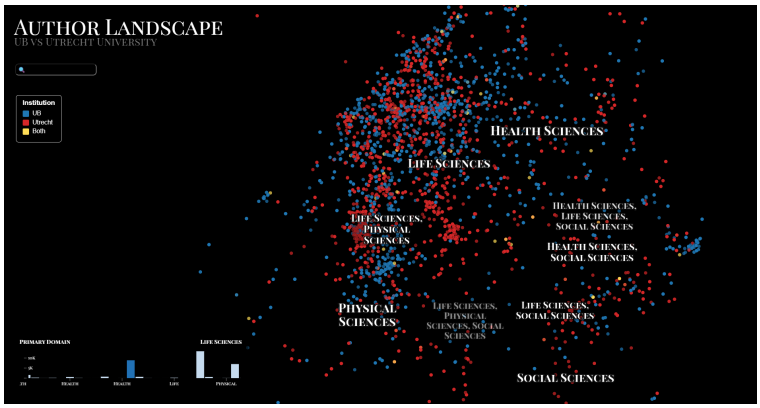
FIGURE A.4: Preview of the filter options in the Topics interactive map.

Original Work Information	**Non-Zero Features**
Topic 1: Dark Matter and... Topic 2: Computational Physics... Topic 3: Particle physics...	TOPIC_Dark Matter and...: 1.0 TOPIC_Computational Physics...: 1.0 TOPIC_Particle physics...: 1.0
Subfield 1: Nuclear and High Energy... Subfield 3: Nuclear and High Energy... Subfield 2: Artificial Intelligence	SUBFIELD_Nuclear and High...: 2.0 SUBFIELD_Artificial Intelligence: 1.0
Field 1: Physics and Astronomy Field 3: Physics and Astronomy Field 2: Computer Science	FIELD_Physics and Astronomy: 2.0 FIELD_Computer Science: 1.0
Domain 1: Physical Sciences Domain 2: Physical Sciences Domain 3: Physical Sciences	DOMAIN_Physical Sciences: 3.0

TABLE A.1: Example of topic-Based vector representation.



(A) Filter in Health Sciences domain.



(B) Filter in the Life Sciences domain.

FIGURE A.5: Preview of the filter options in the Authors interactive map.

n_neighbors	min_dist	Silhouette	Davies-Bouldin	Calinski-Harabasz
5	0.0125	0.07891	1.7102	2341.3286
5	0.0500	0.03987	3.3972	1727.3931
5	0.2000	0.06752	2.3806	2314.7666
5	0.8000	0.03190	3.1356	2074.8865
15	0.1000	0.1170	1.0722	3619.0722
20	0.0125	0.0885	1.1750	2653.2361
20	0.0500	0.0615	1.2574	2577.3235
20	0.2000	0.0647	2.1258	2497.3599
20	0.8000	0.0982	0.9953	3272.9463
50	0.0125	0.2004	1.1493	2248.9446
50	0.0500	0.1831	1.5606	2064.7981
50	0.2000	0.1835	1.0898	2052.6782
50	0.8000	0.0570	1.2492	2253.7119
100	0.0125	0.1267	1.9182	1758.5996
100	0.0500	0.0302	1.2546	2195.1226
100	0.2000	0.0581	1.5290	2181.9959
100	0.8000	0.0743	1.7080	1855.4984

TABLE A.2: Full clustering metrics for different UMAP hyperparameters for UB 2024 with Topic embeddings.

n_neighbors	min_dist	Silhouette	Davies-Bouldin	Calinski-Harabasz
5	0.0125	0.0797	2.4456	1607.7668
5	0.0500	0.0717	2.4899	1535.7249
5	0.2000	0.0817	2.3897	1654.4471
5	0.8000	0.0597	2.9995	1523.1870
15	0.1000	0.0775	2.1742	1800.4878
20	0.0125	0.1006	2.0738	1991.3978
20	0.0500	0.0868	2.1022	1949.5598
20	0.2000	0.0702	2.2720	1866.5370
20	0.8000	0.0570	2.3989	1789.4564
50	0.0125	0.0930	2.3555	1815.8835
50	0.0500	0.0899	2.3704	1785.9542
50	0.2000	0.0805	2.7701	1774.2121
50	0.8000	0.0614	2.4521	1793.7520
100	0.0125	0.0949	2.7127	1776.7771
100	0.0500	0.0886	2.4001	1800.5967
100	0.2000	0.0794	2.4182	1767.1643
100	0.8000	0.0597	3.0772	1706.9482

TABLE A.3: Full clustering metrics for different UMAP hyperparameters for UB 2024 with SPECTER2 embeddings.

Bibliography

Caliński, T. and J Harabasz and (1974). "A dendrite method for cluster analysis". In: *Communications in Statistics* 3.1, pp. 1–27. DOI: 10.1080/03610927408827101. eprint: https://www.tandfonline.com/doi/pdf/10.1080/03610927408827101. URL: https://www.tandfonline.com/doi/abs/10.1080/03610927408827101.

Canon, Chelsea, Douglas Iain Ross Boyle, and KJ Hepworth (2022). "Ethical and Effective Visualization of Knowledge Networks". In: *Digit. Humanit. Q.* 16. URL: https://api.semanticscholar.org/CorpusID:252311345.

Coenen, Andy and Adam Pearce (2020). *Understanding UMAP*. URL: https://pair-code.github.io/understanding-umap.

Cohan, Arman et al. (2020). "SPECTER: Document-level Representation Learning using Citation-informed Transformers". In: *ACL*.

Davies, David L. and Donald W. Bouldin (1979). "A Cluster Separation Measure". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.2, pp. 224–227. DOI: 10.1109/TPAMI.1979.4766909.

Devlin, Jacob et al. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv: 1810.04805 [cs.CL]. URL: https://arxiv.org/abs/1810.04805.

Directorate-General for Research and Innovation (2024). *Science, Research and Innovation Performance of the EU 2024*. URL: https://research-and-innovation.ec.europa.eu/knowledge-publications-tools-and-data/publications/all-publications/science-research-and-innovation-performance-eu-2024-report_en.

Leland McInnes (2023). *DataMapPlot Documentation*. URL: https://datamapplot.readthedocs.io/en/latest/index.html.

Marrone, Mauricio and Martina K. Linnenluecke (2020). "Interdisciplinary Research Maps: A New Technique for Visualizing Research Topics". In: *PLOS ONE* 15.11, e0242283. DOI: 10.1371/journal.pone.0242283. URL: https://doi.org/10.1371/journal.pone.0242283.

McInnes, Leland (2024). *DataMapPlot Examples: arXiv*. URL: https://lmcinnes.github.io/datamapplot_examples/arXiv/.

McInnes, Leland, John Healy, and James Melville (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv: 1802.03426 [stat.ML]. URL: https://arxiv.org/abs/1802.03426.

McInnes, Leland, John Healy, Nathaniel Saul, et al. (2018). "UMAP: Uniform Manifold Approximation and Projection". In: *The Journal of Open Source Software* 3.29, p. 861. URL: https://umap-learn.readthedocs.io/en/latest/index.html.

Nomic AI (2024). *nomic-embed-text-v1*. https://huggingface.co/nomic-ai/nomic-embed-text-v1.

OpenAlex Team (2023). *Author Disambiguation*. URL: https://help.openalex.org/hc/en-us/articles/24347048891543-Author-disambiguation.

— (2025a). *OpenAlex API Documentation: Topics*. URL: https://docs.openalex.org/api-entities/topics.

OpenAlex Team (2025b). *OpenAlex API: Entities Overview*. https://docs.openalex.org/api-entities/entities-overview.

— (2025c). *OpenAlex API: Works*. URL: https://docs.openalex.org/api-entities/works.

— (2025d). *OpenAlex Documentation*. URL: https://docs.openalex.org/.

— (2025e). *OpenAlex Topic Classification Whitepaper*. https://docs.google.com/document/d/1bDopkhuGieQ4F8gGNj7sEc8WSE8mvLZS/edit.

Priem, Jason, Heather Piwowar, and Richard Orr (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. arXiv: 2205.01833 [cs.DL]. URL: https://arxiv.org/abs/2205.01833.

Reimers, Nils and Iryna Gurevych (2019a). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. https://www.sbert.net/.

— (Nov. 2019b). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. URL: https://arxiv.org/abs/1908.10084.

Rousseeuw, Peter J. (1987). "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20, pp. 53–65. ISSN: 0377-0427. DOI: https://doi.org/10.1016/0377-0427(87)90125-7. URL: https://www.sciencedirect.com/science/article/pii/0377042787901257.

Singh, Amanpreet et al. (2022). "SciRepEval: A Multi-Format Benchmark for Scientific Document Representations". In: *Conference on Empirical Methods in Natural Language Processing*. URL: https://api.semanticscholar.org/CorpusID:254018137.

Singh, Chakresh Kumar et al. (2023). *Charting mobility patterns in the scientific knowledge landscape*. arXiv: 2302.13054 [physics.soc-ph]. URL: https://arxiv.org/abs/2302.13054.