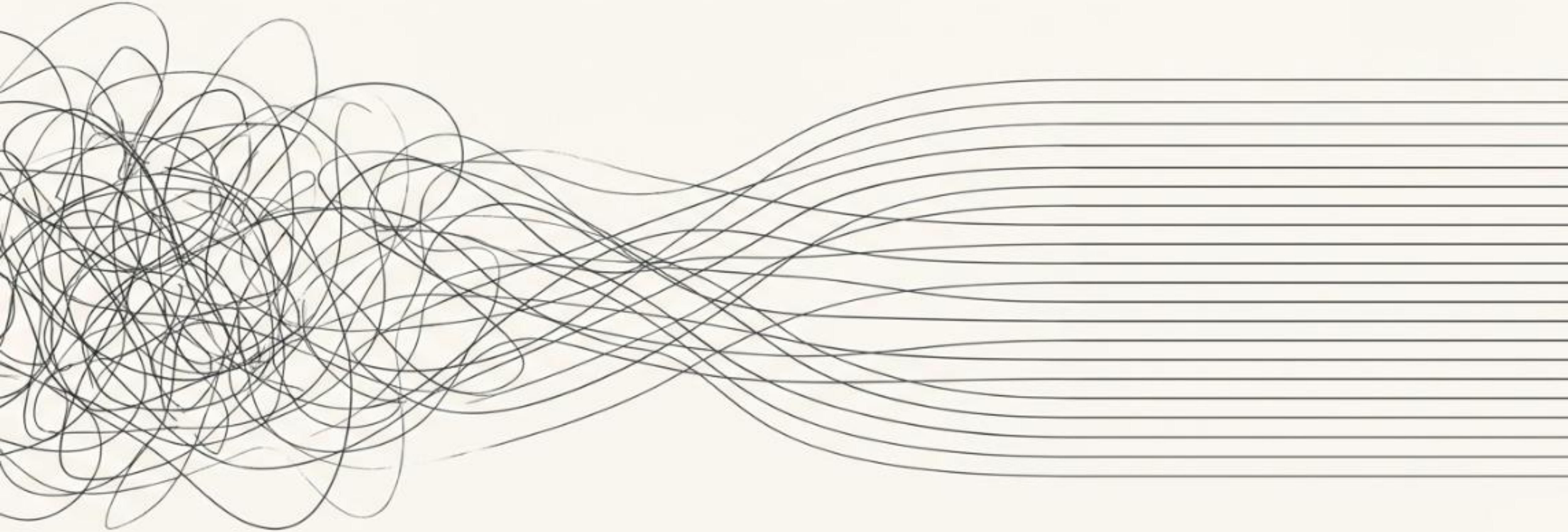
From Raw Text to Rich Insight

Unsupervised Clustering of Technical Documents



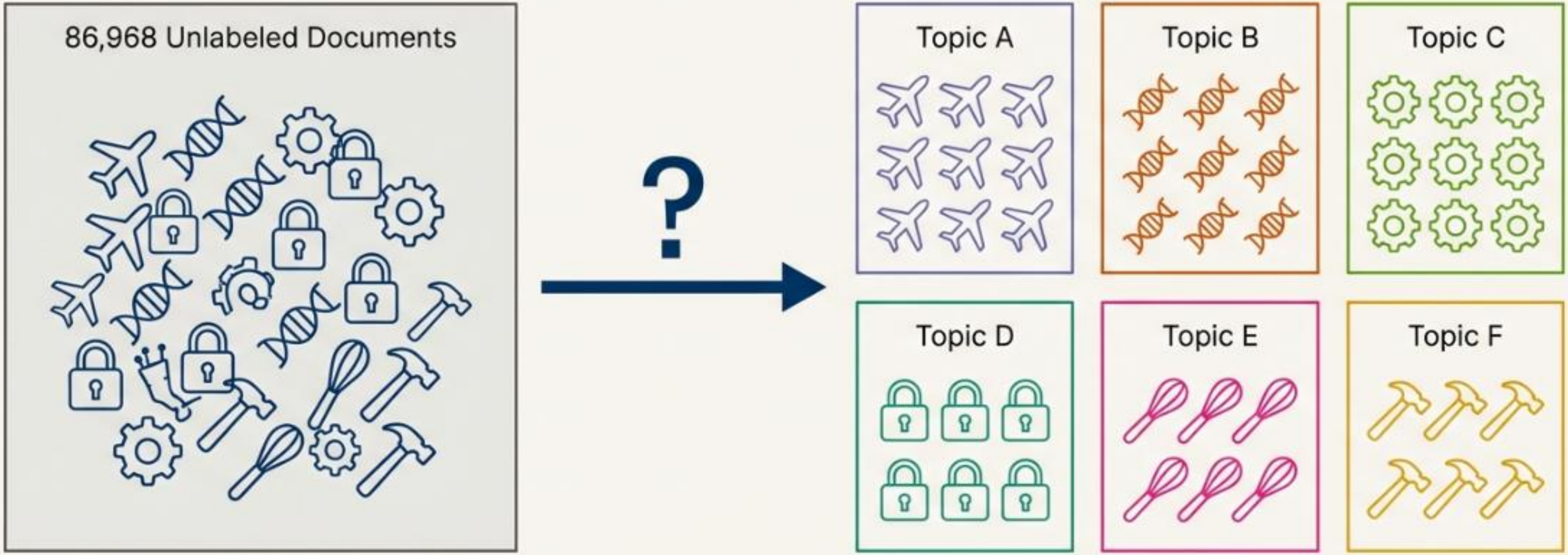
The Mission: Find Structure in Chaos, Automatically

The Challenge

Given a corpus of nearly 87,000 documents from various sources, can we automatically group them into meaningful topics using only the text itself?

The Constraint

We know the documents originate from 6 distinct domains, but this information is held back. The algorithm must work "blind," without access to these ground-truth labels.



The Raw Material: A Corpus from Six Distinct Domains

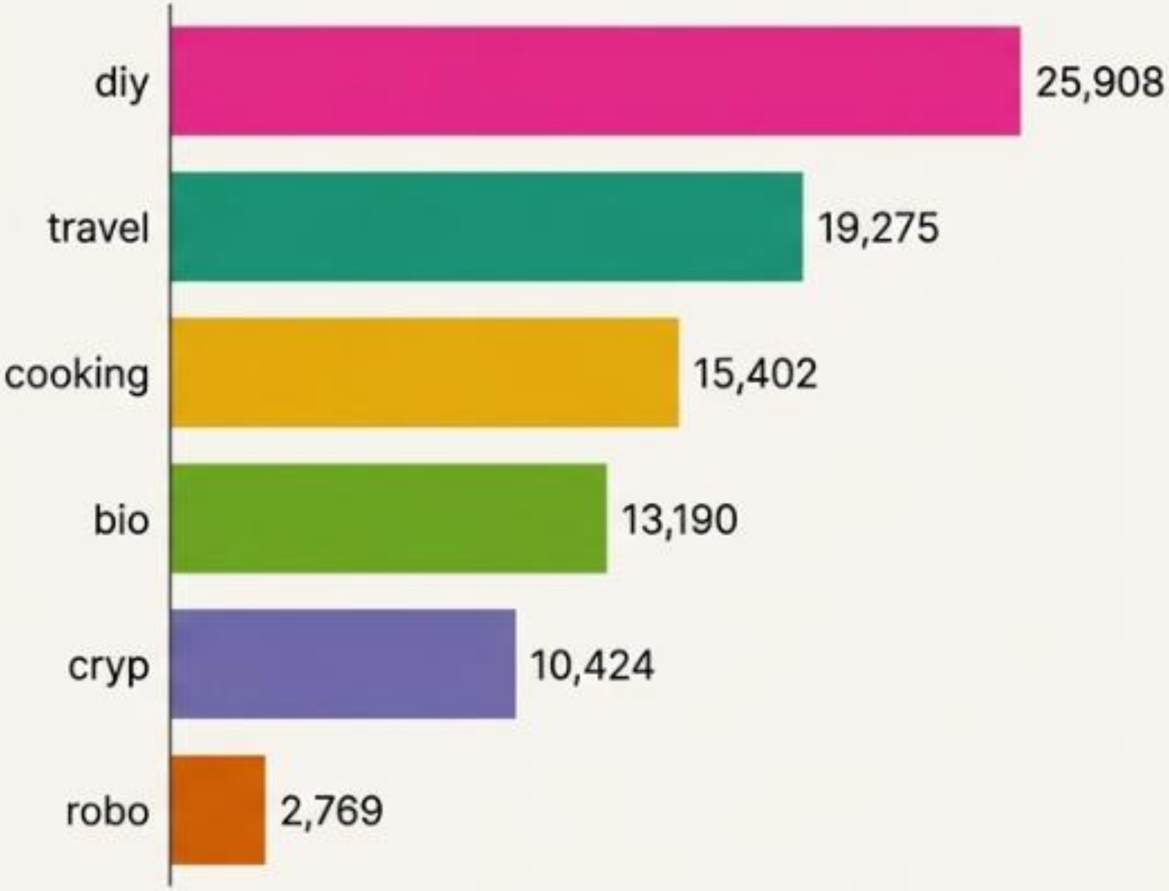
Key Statistics

- **Total Documents:** 87,000
- **Columns:** `title`, `content`, `tags`, `domain`
- **Duplicates & Nulls Removed:** Resulting in 86,968 unique documents.

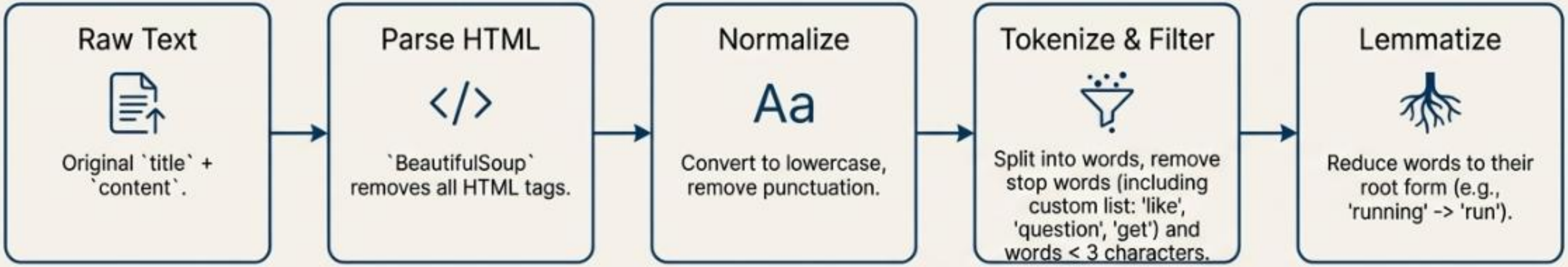
Sample Data

doc_id	clean_title	tags
0	What is the criticality of the ribosome binding...	ribosome binding-sites...
1	How is RNAse contamination in RNA based...	rna biochemistry
2	Are lymphocyte sizes clustered in two groups?	immunology cell-biology...

Source Domain Distribution



Preparing the Evidence: A Rigorous Text Cleaning Pipeline



Raw Text

Original `title` + `content`.

Parse HTML

`BeautifulSoup` removes all HTML tags.

Normalize

Convert to lowercase, remove punctuation.

Tokenize & Filter

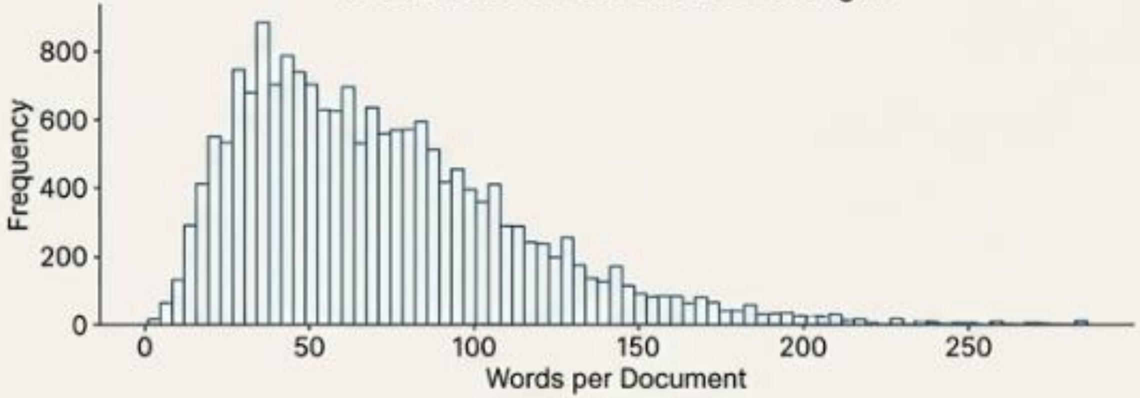
Split into words, remove stop words (including custom list: 'like', 'question', 'get') and words < 3 characters.

Lemmatize

Reduce words to their root form (e.g., 'running' -> 'run').

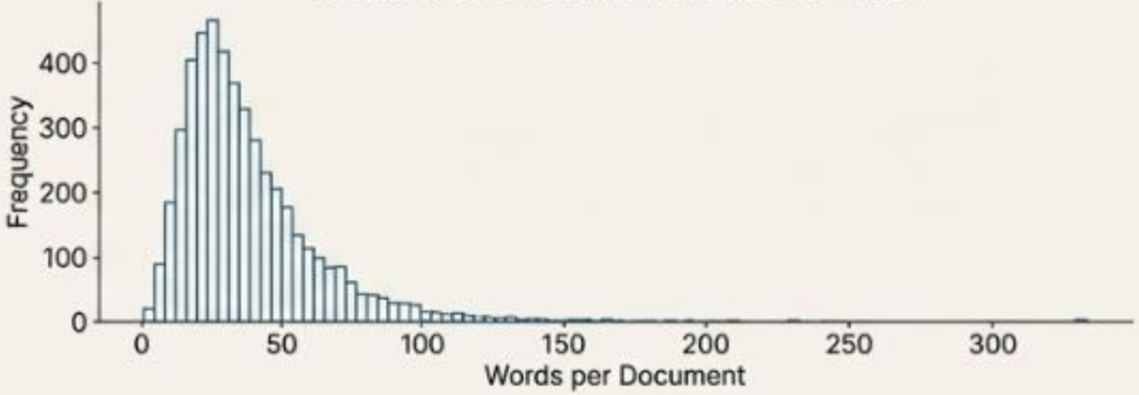
The Impact of Cleaning

Distribution of raw document length



Mean words/doc: 117.0

Distribution of cleaned document length



Mean words/doc: 54.9

Key Takeaway: Cleaning reduces document length by over 50%, focusing the model on the most signal-rich terms.

Capturing Semantic Meaning with SBERT Embeddings

The 'Why'

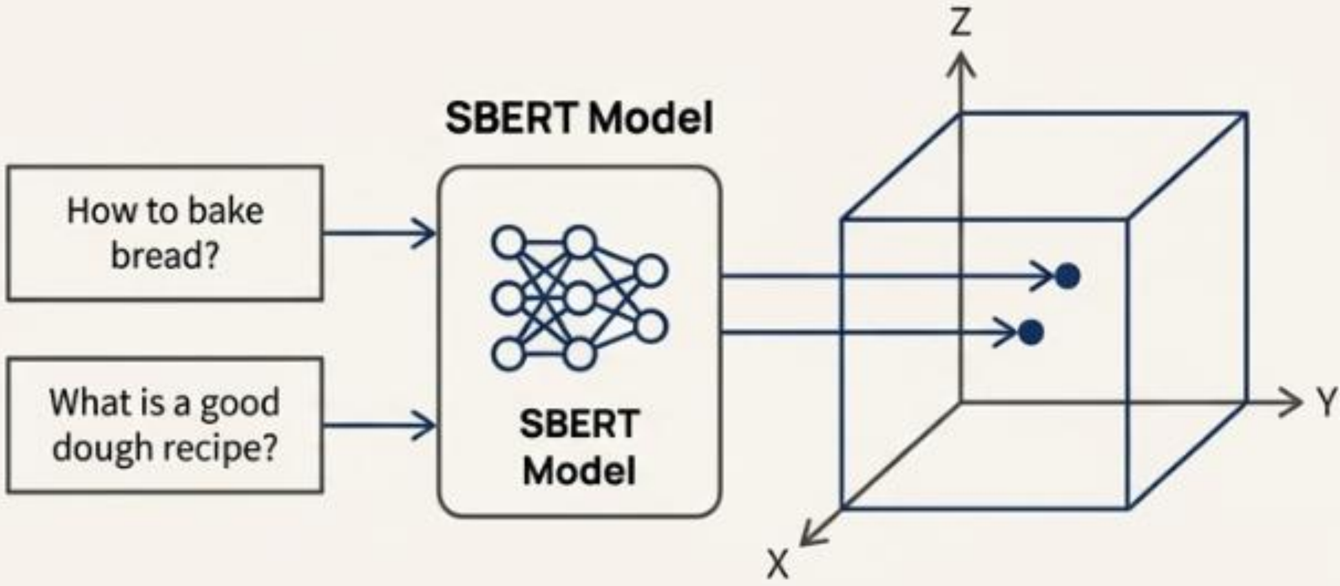
Traditional methods like TF-IDF count words but often miss underlying meaning and context. Sentence-BERT (SBERT) generates dense vector embeddings (768 dimensions) that place semantically similar documents close together in vector space. This is crucial for discovering nuanced topics.

The Model

`all-mpnet-base-v2` - A high-performance model pre-trained on a massive text corpus, ideal for generating general-purpose embeddings.

The Outcome

- Input: 86,968 cleaned text documents.
- Output: A numerical matrix of shape `(86968, 768)`, ready for clustering.



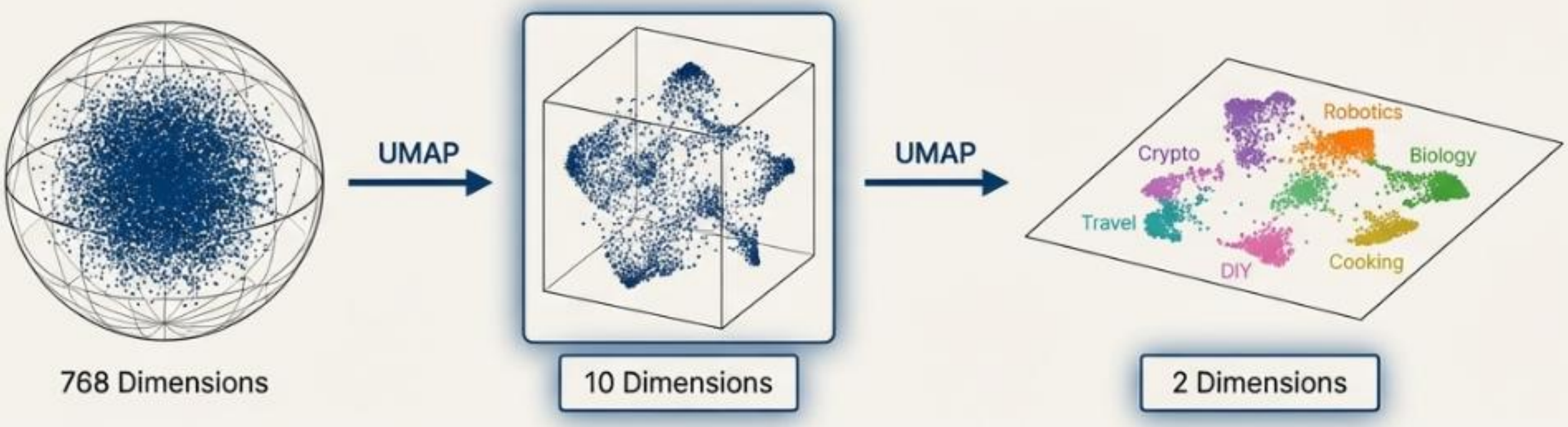
The Cartographer's Lens: Projecting the Data with UMAP

The Problem

Clustering directly in 768 dimensions is computationally expensive and can be affected by the "curse of dimensionality."

The Solution

We use UMAP (Uniform Manifold Approximation and Projection) to reduce the dimensionality of our embeddings. UMAP excels at preserving the global structure and semantic relationships from the original high-dimensional space.



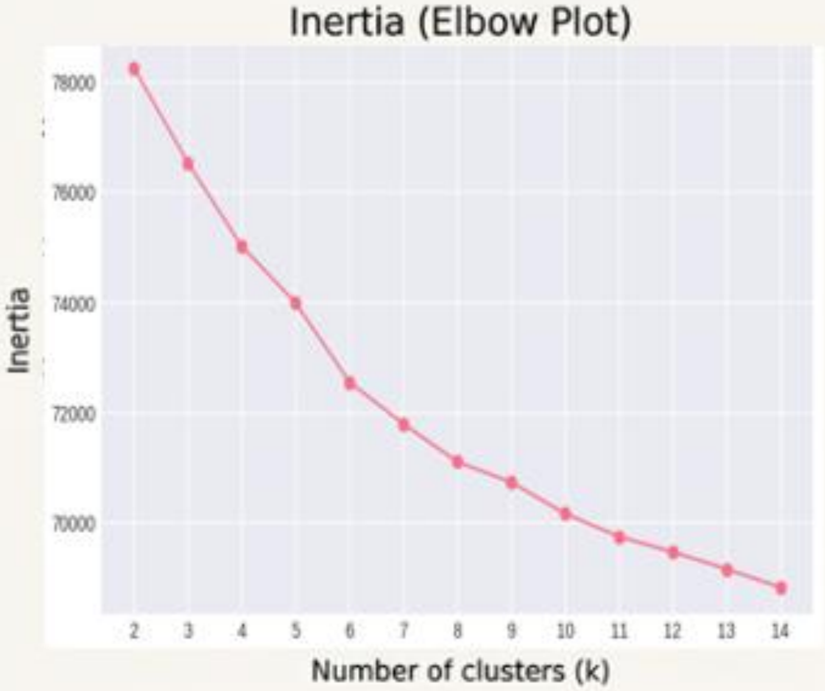
Our Process

- For Clustering: 768 dimensions → **10 dimensions** (`n_neighbors=50`, `min_dist=0.0`)
- For Visualization: 768 dimensions → **2 dimensions** (`n_neighbors=30`, `min_dist=0.1`)

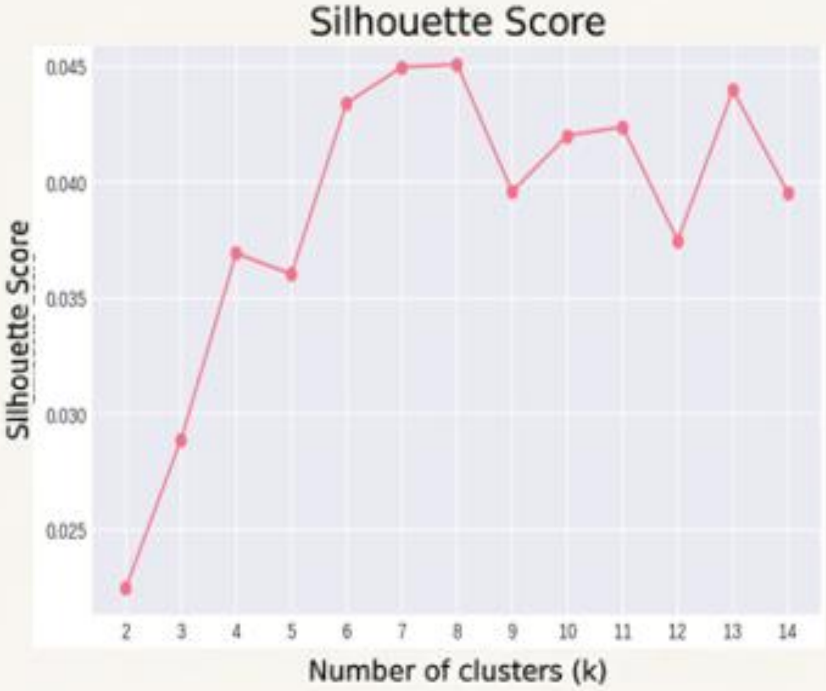
(86,968, 10)

The Core Decision: How Many Clusters Exist in the Data?

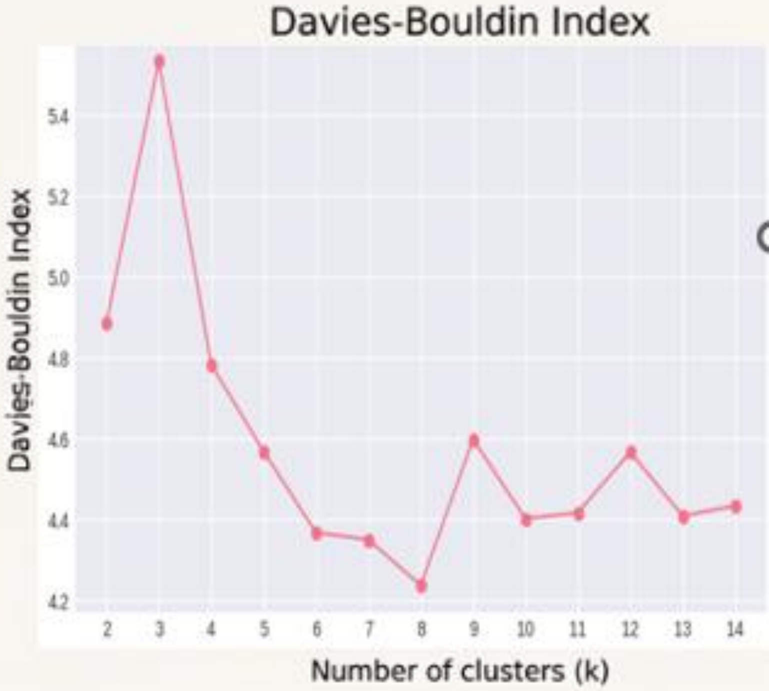
We ran K-Means for `*k*` values from 2 to 14 and evaluated the results using **multiple standard metrics** to find the optimal number of clusters.



Inertia (Elbow Plot)



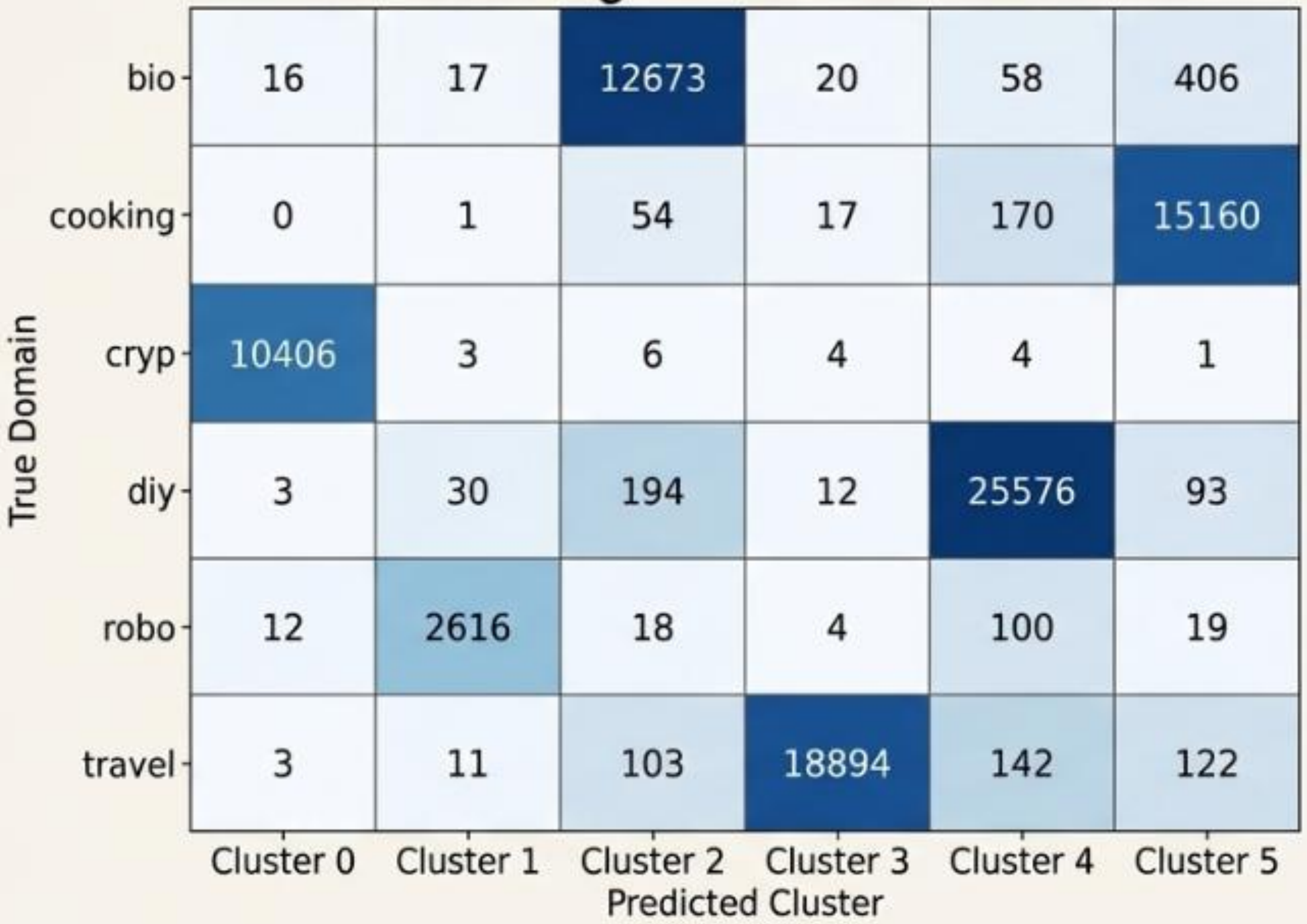
Silhouette Score



Davies-Bouldin Index

The Reveal: A Near-Perfect 98.1% Match to Ground Truth

Cluster Assignment vs. True Domain



True Domain	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
bio	16	17	12673	20	58	406
cooking	0	1	54	17	170	15160
cryp	10406	3	6	4	4	1
diy	3	30	194	12	25576	93
robo	12	2616	18	4	100	19
travel	3	11	103	18894	142	122

Key Performance Metrics

0.9811

Purity Score

(The percentage of documents correctly assigned to the majority domain within their cluster)

0.9558

Adjusted Rand Index (ARI)

(Measures similarity between true and predicted labels, correcting for chance)

0.9338

Normalized Mutual Information (NMI)

(Measures mutual dependence between the two label sets)

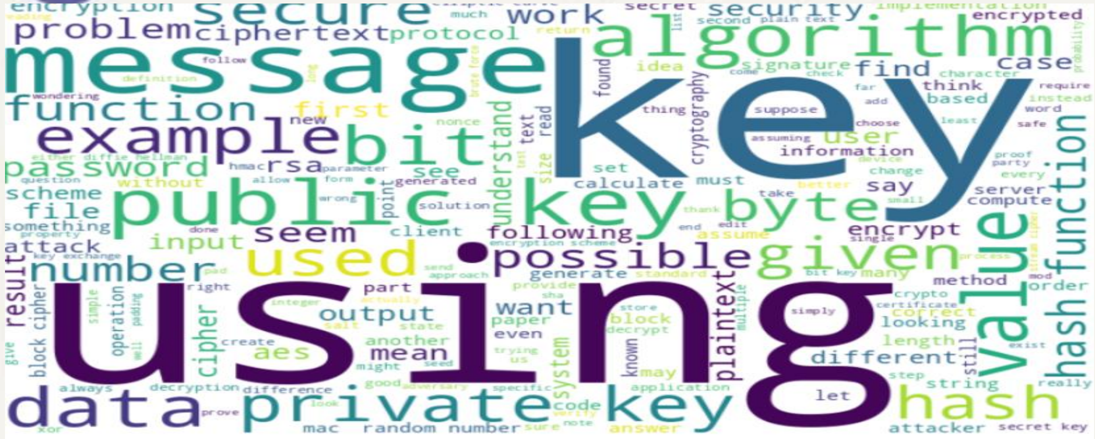
The Verdict: The unsupervised pipeline successfully rediscovered the original thematic structure with extremely high fidelity.

Deconstructing the Clusters: The Six Discovered Worlds

Cluster ID	Documents	Inferred Topic	Majority Domain	Top TF-IDF Terms
● 0	10,440	Cryptography	**cryp** (10406/10440)	key, encryption, hash, message, cipher
● 1	2,678	Robotics	**robo** (2616/2678)	robot, motor, sensor, control, arduino
● 2	13,048	Biology	**bio** (12673/13048)	cell, human, dna, gene, protein
● 3	18,951	Travel	**travel** (18894/18951)	visa, flight, travel, passport, airport
● 4	26,050	DIY / Home Improvement	**diy** (25576/26050)	wall, water, wire, light, house
● 5	15,801	Cooking	**cooking** (15160/15801)	recipe, cooking, chicken, cook, meat

Cluster Profiles: Cryptography & Robotics

● Cluster 0 - Cryptography



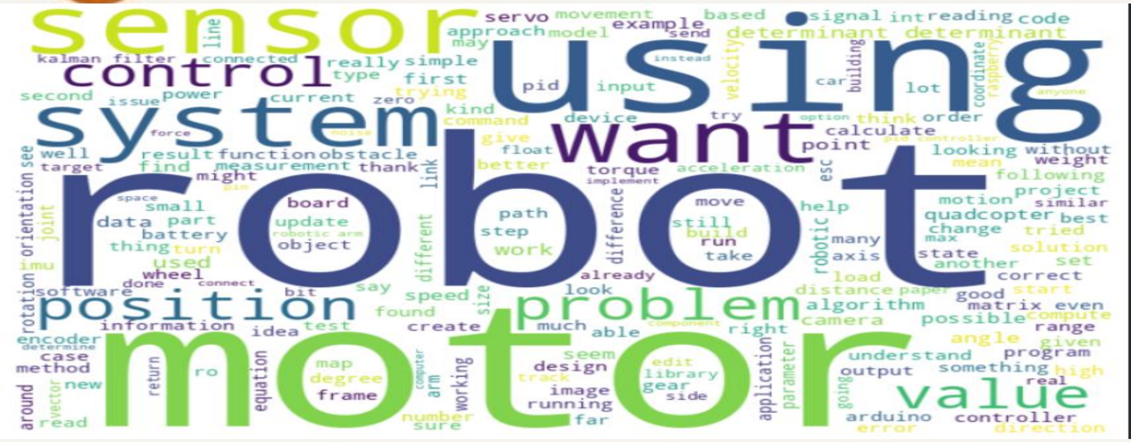
Top Terms

key, encryption, hash, message, bit, cipher, algorithm, rsa, function, aes.

Sample Question

> "What is the difference between a hash and an encryption algorithm?"

● Cluster 1 - Robotics



Top Terms

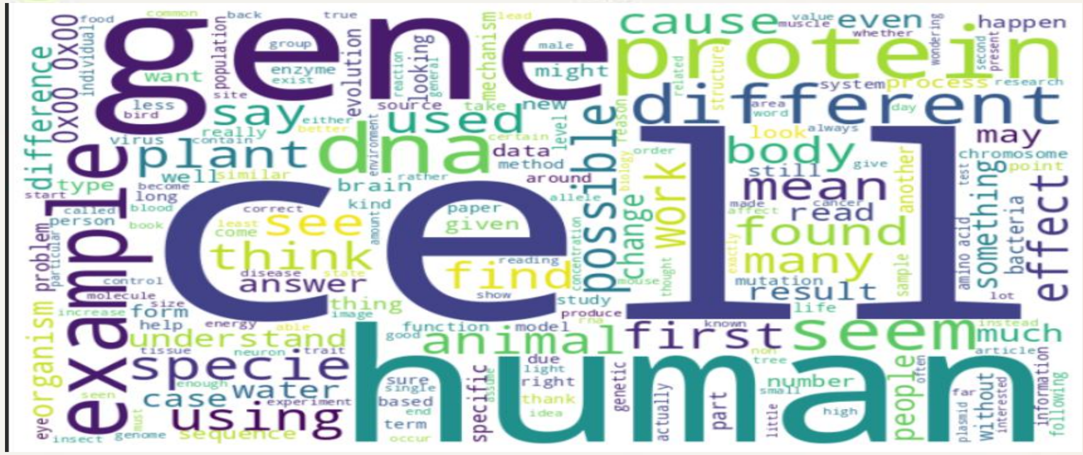
robot, motor, sensor, using, control, arduino, servo, position, controller, camera.

Sample Question

> "How to control a servo motor using Arduino and a distance sensor?"

Cluster Profiles: Biology & Travel

Cluster 2 - Biology



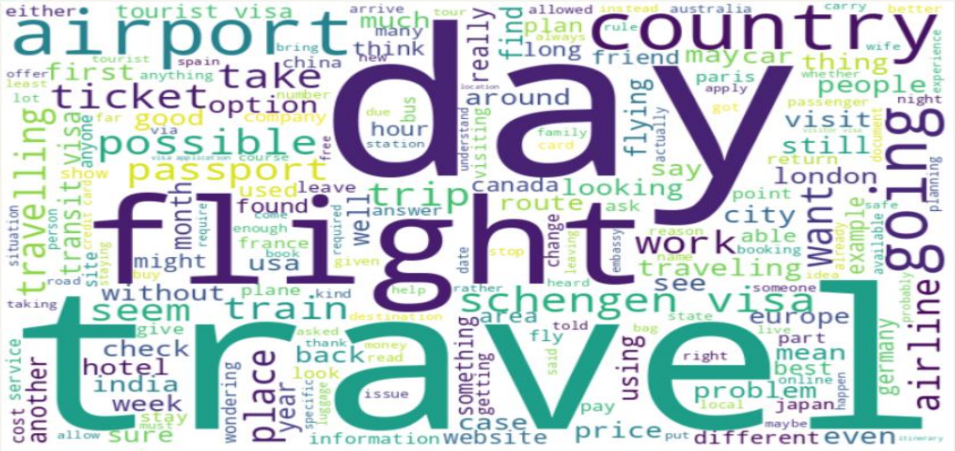
Top Terms

cell, human, dna, gene, protein, specie, plant, body, animal, blood.

Sample Question

> “What is the criticality of the ribosome binding site in gene expression?”

Cluster 3 - Travel



Top Terms

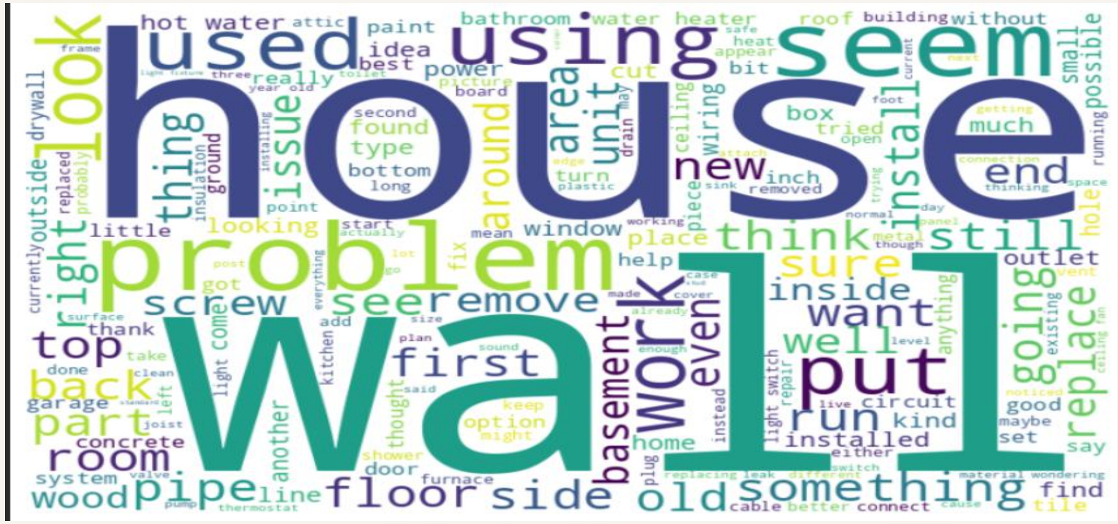
visa, flight, travel, passport, airport, day, ticket, country, schengen, visit.

Sample Question

> “India to Jamaica via UK and USA: do I need transit visas?”

Cluster Profiles: Home Improvement & Cooking

● Cluster 4 - DIY / Home Improvement



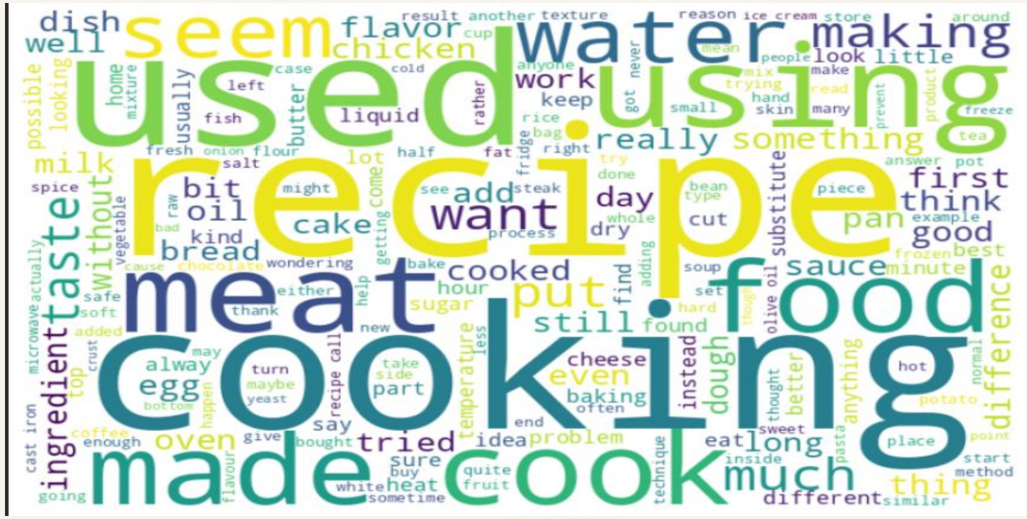
Top Terms

wall, water, wire, light, house, switch, floor, door, pipe, outlet.

Sample Question

> "How can I ground a fluorescent light that I'm attaching a wall outlet plug to?"

● Cluster 5 - Cooking



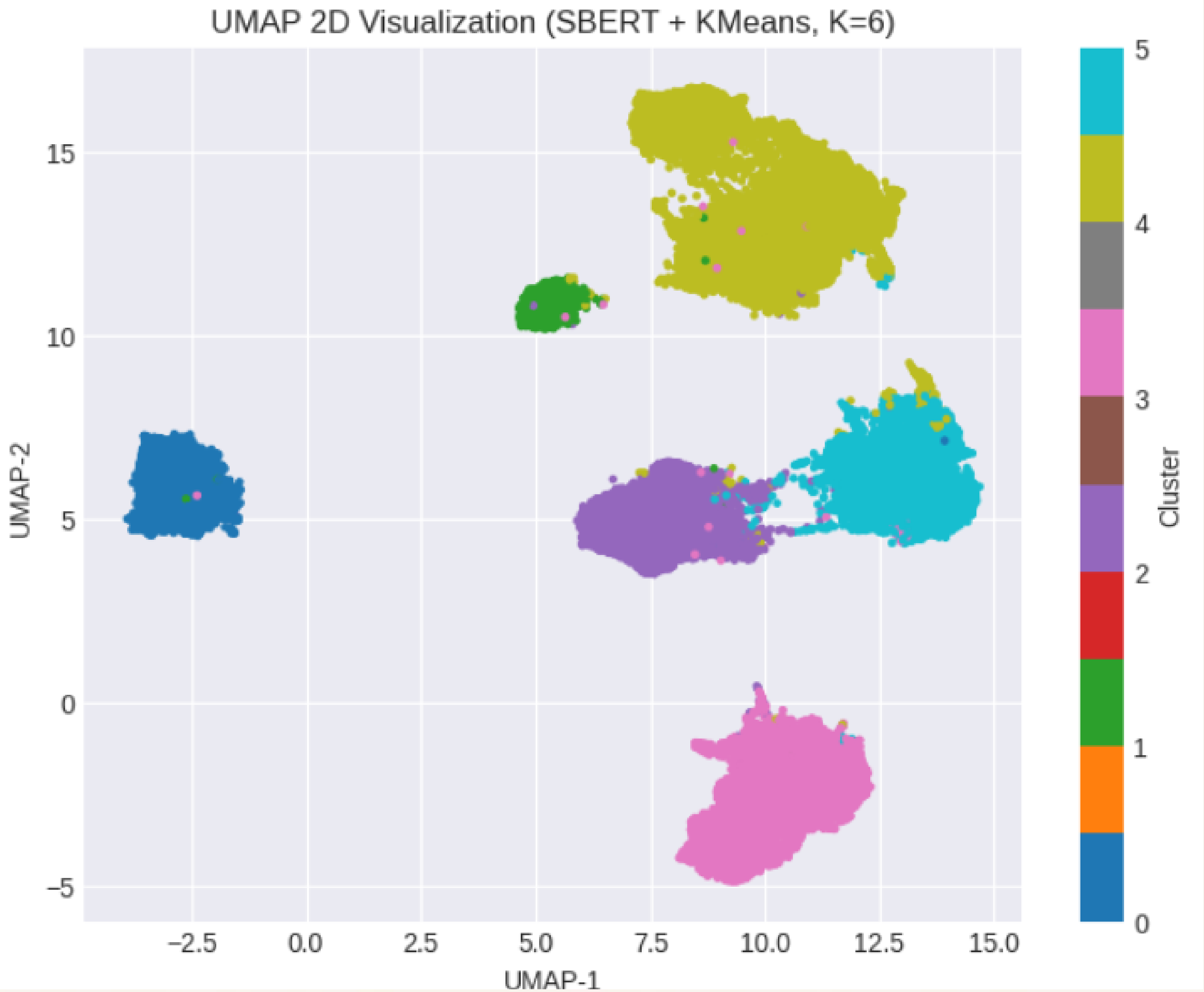
Top Terms

recipe, cooking, chicken, cook, meat, egg, oil, pan, sauce, food.

Sample Question

> "How long should I mature my mincemeat before making Mince Pies?"

The Final Picture: A Visual Map of Discovered Knowledge



The final 2D projection clearly shows six distinct and well-separated clusters. This visual confirms the quantitative metrics, illustrating the success of the SBERT and UMAP pipeline in structuring the document corpus.

From Chaos to Clarity: Key Takeaways

A modern, unsupervised NLP pipeline can discover latent thematic structures in large text corpora with remarkable precision, providing a powerful tool for automated data organization

- **SBERT + UMAP is a Potent Combination**

  State-of-the-art sentence embeddings effectively capture semantic meaning, which UMAP can then project into a cluster-friendly space.

- **Unsupervised Results Can Rival Supervised Accuracy**

  Achieving over 98% purity demonstrates that for datasets with strong thematic separation, unsupervised methods can approach the performance of labeled approaches.

- **A Scalable & Adaptable Framework**

  This methodology is not domain-specific and can be applied to various text organization tasks, such as customer feedback analysis, scientific literature review, or document tagging.