Article Search Tool

Group member: Zhuoqing Xu, Yushi Zhao, Xinyu Zhang, Jitong Liu

CONTENT

BACKGROUND & PROBLEM DEFINITION

Business Case

The system enables efficient retrieval of relevant documents by leveraging topic modeling and similarity ranking, helping users quickly find information for tasks like research exploration or content discovery.

Background

Quickly finding highly relevant literature is a critical need.

Strategic Objectives

The project aims to deliver:

- Improved Efficiency
- Enhanced Relevance

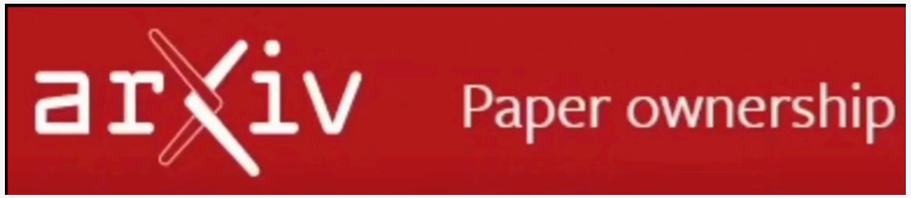
DATA source specification & procurement details

Data source: Hugging Face

- **Data:** **arxiv-abstracts-2021**
- **Link:** https://huggingface.co/datasets/gfissore/arxiv-abstracts-2021
- **Size:** 1.43G, 2M entries
- **Description:** The dataset consists of metadata and abstracts for research papers from the ArXiv repository for the year 2021.

- **Data Fields**

id: ArXiv ID
submitter
authors
title
comments
journal-ref
doi
report-no
abstract
categories



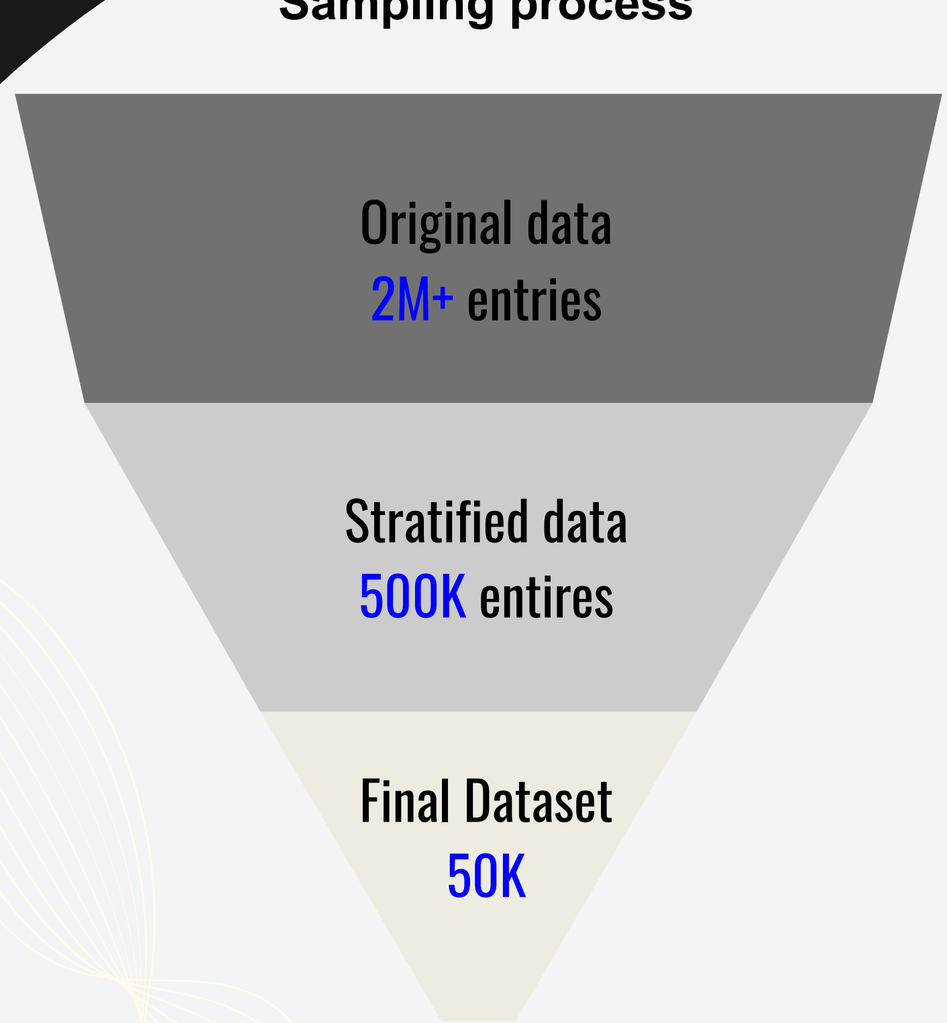
DATA SAMPING& TEXT PREPROCESSING

- **Stratified Sampling:**

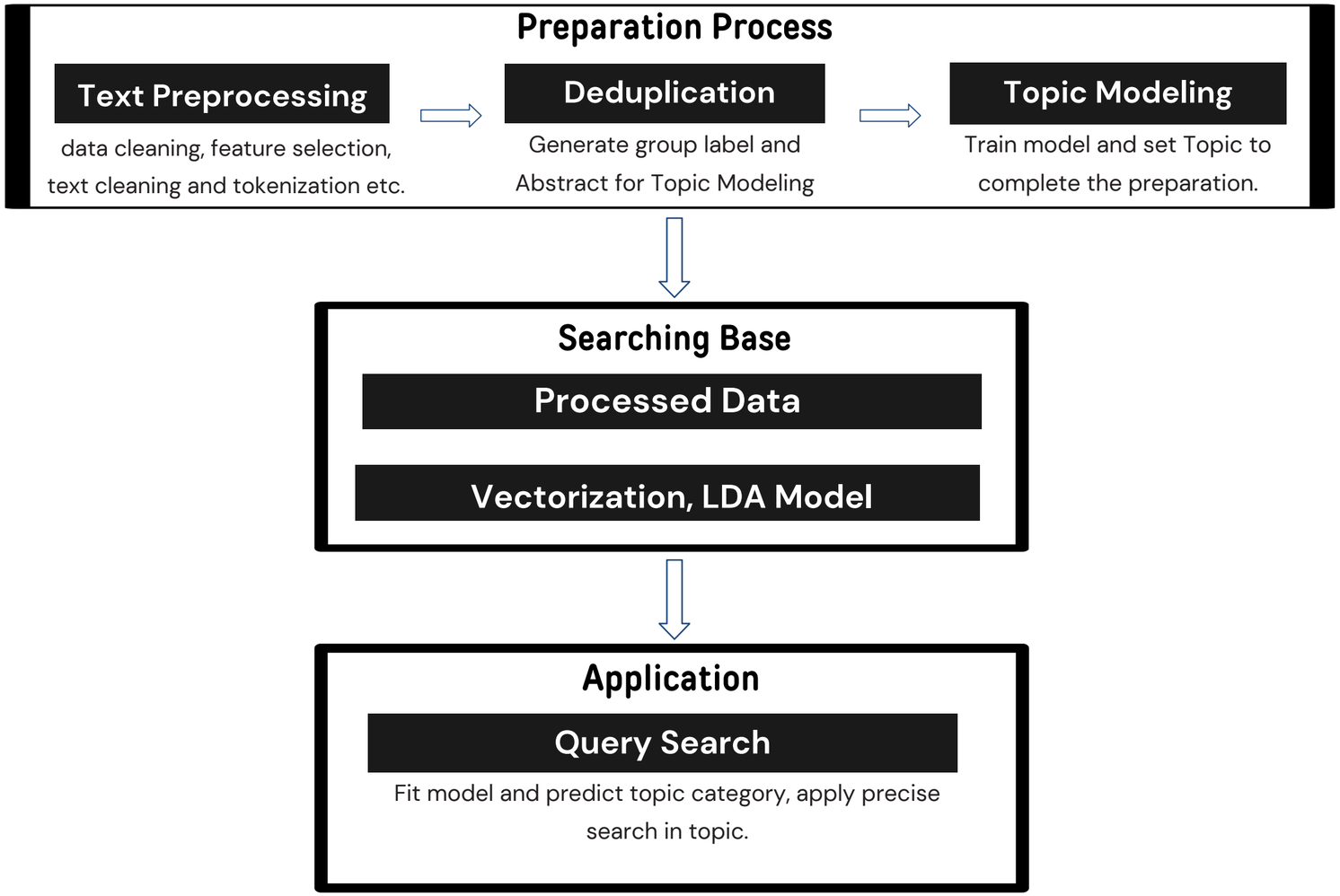
Selected ~500K entries with a balanced distribution across categories. Ensured all categories were proportionally represented.

- **Final Sampling:**

Randomly sampled 10% of the stratified dataset.
Final dataset: ~50K entries, optimized for analysis and model training.



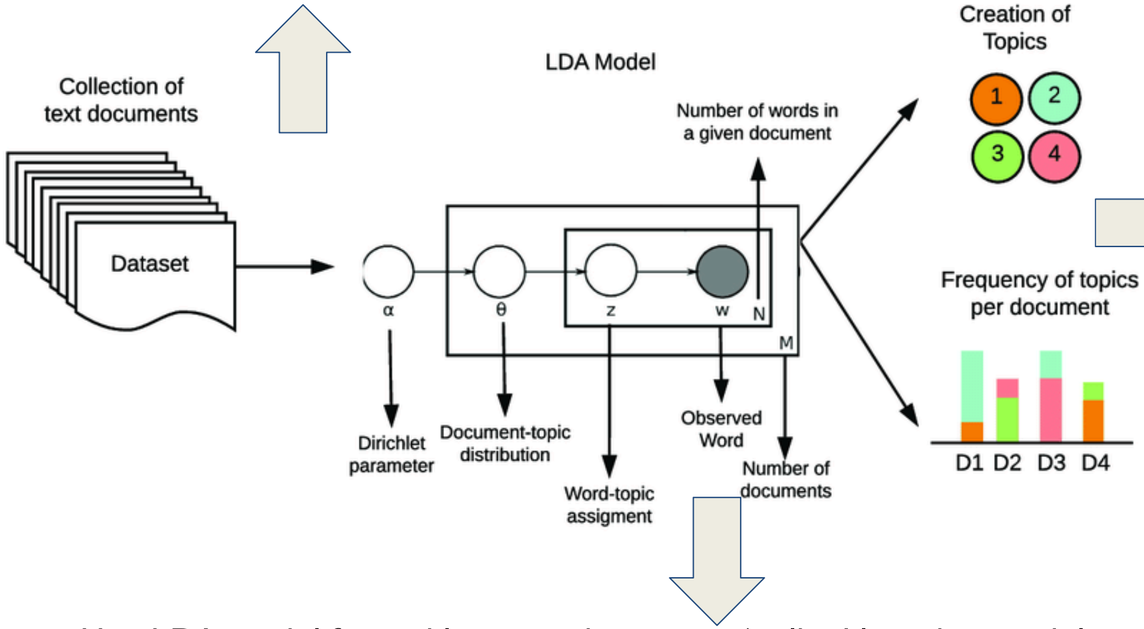
Project Workflow



LDA Process

Vectorization: Convert text data into a sparse matrix.

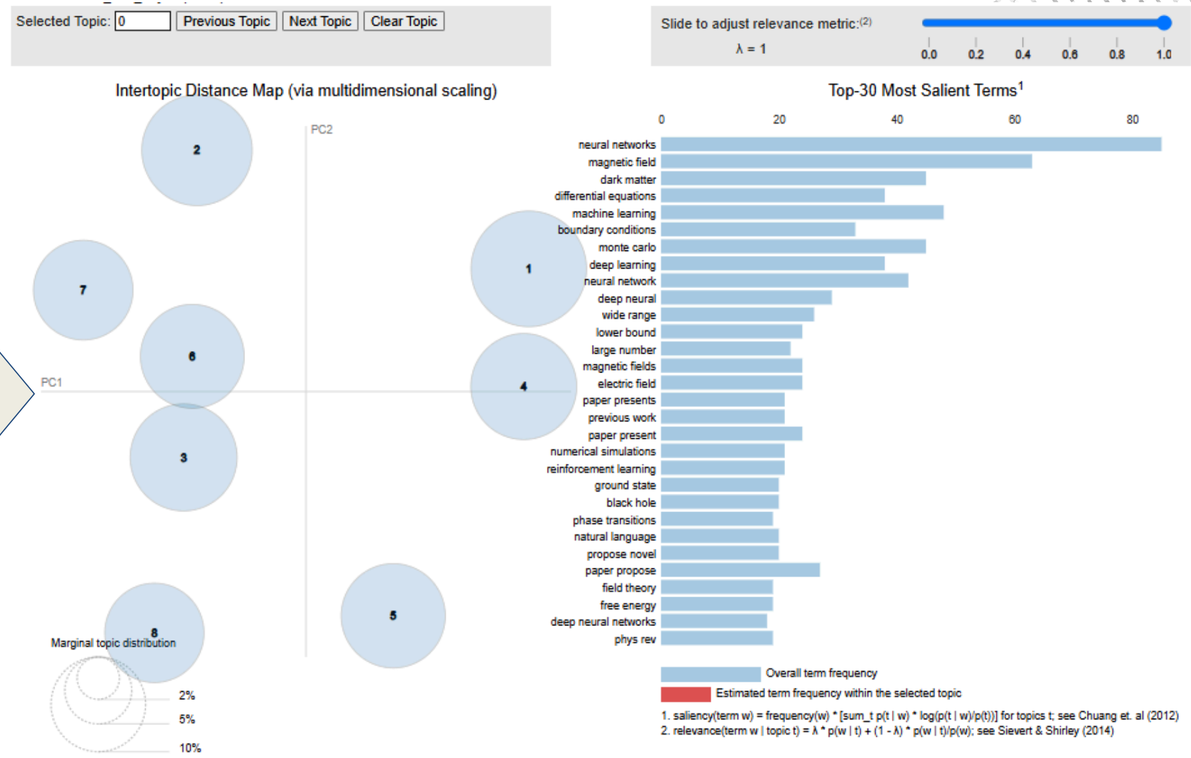
	achieve high	additive noise	address problem	adversarial attacks	agreement experimental	aim paper	algorithm using	analysis shows	analytical expressions	analytical results	...
0	0	0	0	0	0	0	0	0	0	0	...
1	0	0	0	0	0	0	0	0	0	0	...
2	0	0	0	0	0	0	0	0	0	0	...
3	0	0	0	0	0	0	0	0	0	0	...
4	0	0	0	0	0	0	0	0	0	0	...
...	...	...	...	...	...	...	...	...	...	...	...



Use LDA model from sklearn package, use ‘online’ learning model and some other parameter to tune the model.

```
LatentDirichletAllocation
LatentDirichletAllocation(learning_method='online', max_iter=50, n_components=8,
                          random_state=40, topic_word_prior=0.01)
```

Get the visualization plot of topic modeling result and save the model.



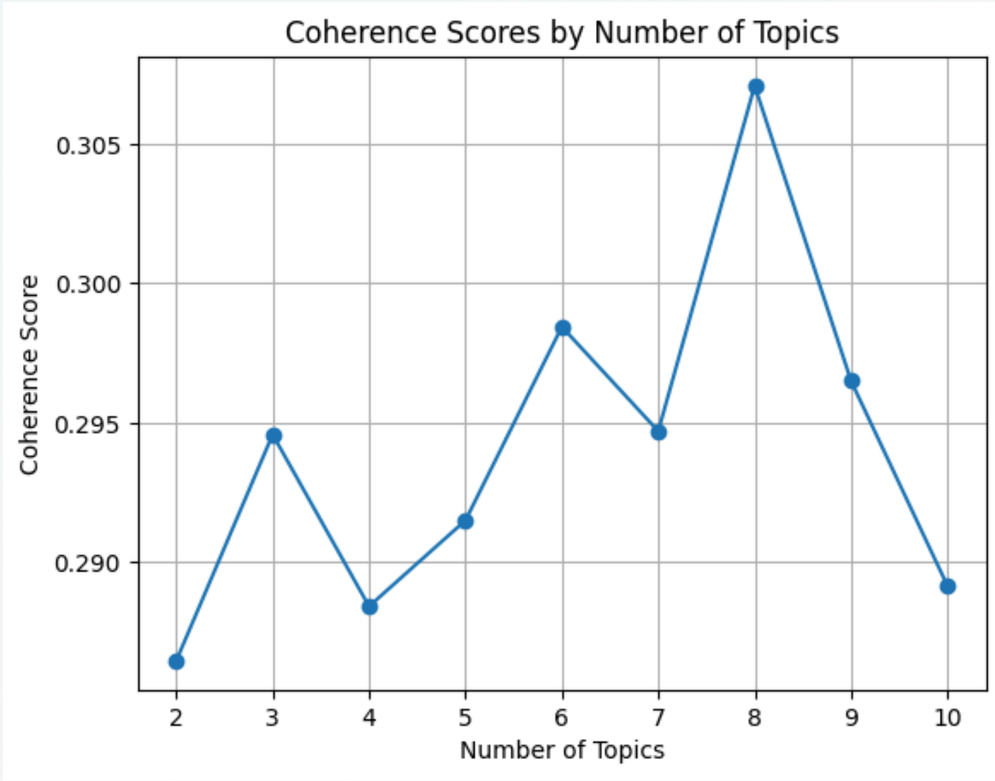
LDA Coherence

Coherence Score

- **Definition:** Measures the interpretability and semantic similarity of topics.
- **Higher Score:** Topics are more meaningful.
- **Lower Score:** Topics are less coherent or noisy.

Results

- **Optimal Topics:** 8 topics (coherence score = **0.3071**).
- **Trend:** Coherence improved up to 8 topics, then declined, indicating over-segmentation.



Query Search Process

Input

Query

Get the query Q from user

Fit the vector matrix constructed in LDA Process

Predict dominant topic

Get topic for Query Q

Use pre-trained LDA model to predict the belong topic of Q

Filter the topic in original dataset

Similarity Search

Articles Discovery

Use Cosine Similarity to search within the topic group and find top 10 article(may less).

```
input_text = "Machine learning\
              models in health care"
```

	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	selected topic
0	0.03125	0.03125	0.031253	0.03125	0.781247	0.03125	0.03125	0.03125	4

	authors	title	abstract	similarity
	Boris Rubinstein	A fast noise filtering algorithm for time seri...	Recent research demonstrate that prediction ...	0.801784

DEMONSTRATION AND KEY RESULT

Query Search Engine Result

- **Accuracy:** Query is matched to top 10 (may less) relevant articles with their abstract.
- **Interpretability:** Each step in the workflow ensures logical filtering and ranking of results.
- **Efficiency :** The workflow combines pre-trained topic modeling and similarity metrics to speed up search process.

Result for Example Query Searching:

	authors	title	abstract	similarity
156069	Boris Rubinstein	A fast noise filtering algorithm for time seri...	Recent research demonstrate that prediction ...	0.801784
478299	Lars Eidnes, Arild N{\o}kland	Shifting Mean Activation Towards Zero with Bip...	We propose a simple extension to the ReLU-fa...	0.717137
84044	Marc Riera, Jose-Maria Arnau, Antonio Gonzalez	CREW: Computation Reuse and Efficient Weight S...	Deep Neural Networks (DNNs) have achieved tr...	0.623610
167237	Friedemann Zenke and Emre O. Neftci	Brain-Inspired Learning on Neuromorphic Substr...	Neuromorphic hardware strives to emulate bra...	0.606977
167494	Yann Ollivier, Corentin Tallec, Guillaume Char...	Training recurrent networks online without bac...	We introduce the "NoBackTrack" algorithm to ...	0.585540

THANK'S FOR WATCHING