# Narrative Trails: A Method for Coherent Storyline Extraction via Maximum Capacity Path Optimization

Fausto German[1], Brian Keith[2], Chris North[1]

[1]Virginia Tech, Blacksburg, Virginia 24061, USA
[2]Universidad Católica del Norte, Av. Angamos 0610, Antofagasta, 1270709, Chile

**Text2Story @ ECIR 2025**

April 10, 2025

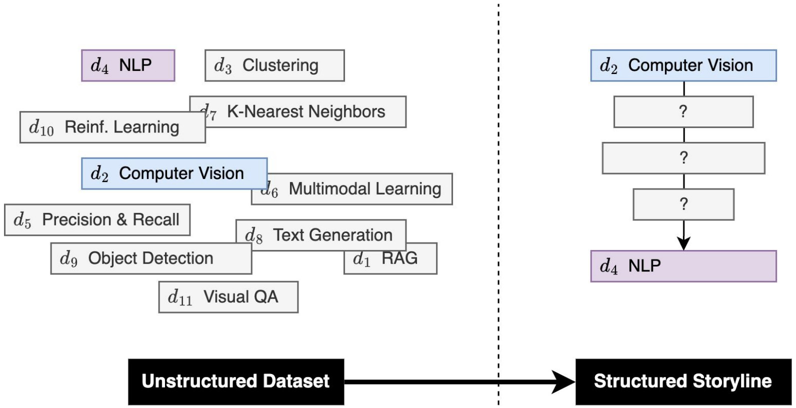
VIRGINIA TECH

Universidad Católica del Norte

# Motivation and Goals

## Finding Structure In Text Corpora

- Humans make sense of complex information through stories
- Text corpora (e.g. news, scientific papers) often contain *latent narratives* embedded within the corpus
- Our goal: Automatically extracting these latent structures




### Ultimate Goal

Automatically and efficiently extract coherent storylines that connects two user-defined endpoints in a large dataset of text documents

# Research Problem

## Challenge

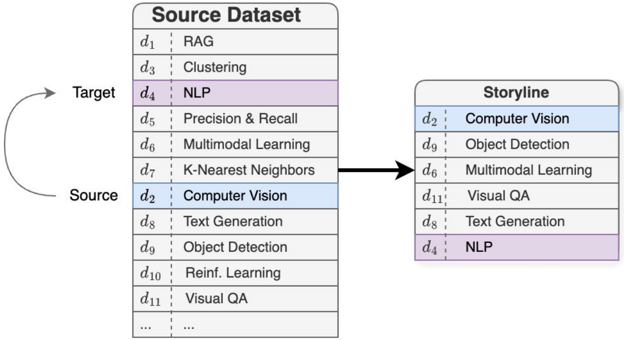
Extracting coherent storylines based on abstract semantics between documents, rather than strict keyword matches.

## Gap

Existing narrative extraction methods often rely on complex word-based heuristics, auxiliary document structures, and linear programming.

## Opportunity

By harnessing the semantic representations from deep learning models directly, we can uncover the latent narratives structures within the embedding space.




## Key Insights

- A storyline is only as coherent as its weakest link.
- Coherent storylines extracted directly from the latent space of deep learning models are maximum-capacity paths.

# Contributions

## Approach

Abstractive approach to narrative extraction, focused primarily on the abstract semantic relationships between documents in a dataset.

## Algorithm

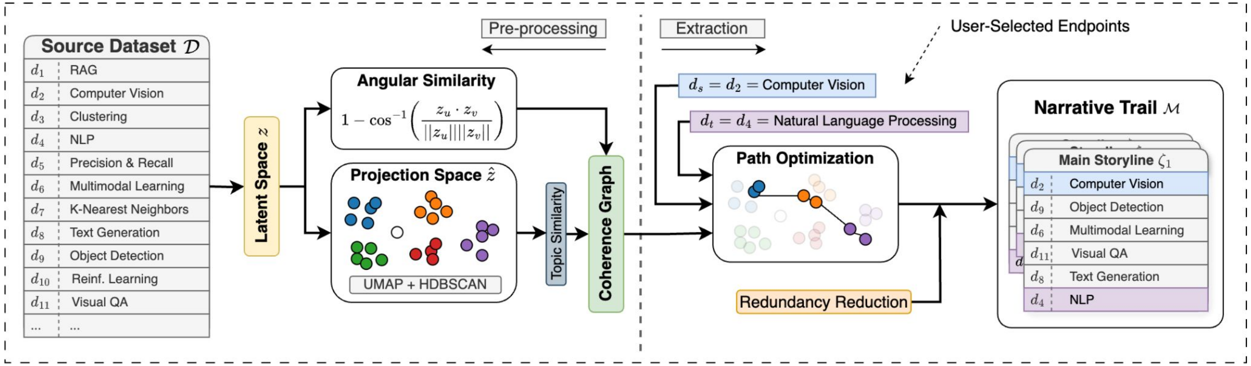
We describe an efficient algorithm based on maximum-capacity path search for coherent storyline extraction from large datasets.

## Extensibility

We provide a repository with details of our algorithm that can be used to reproduce our results or to extend our methods.

GitHub Repository

# Methodology Overview




1. Constructing a projection space from the data
2. Building a coherence graph from the embeddings
3. Finding a path of maximum capacity between two endpoints

# Building the Base Coherence Graph

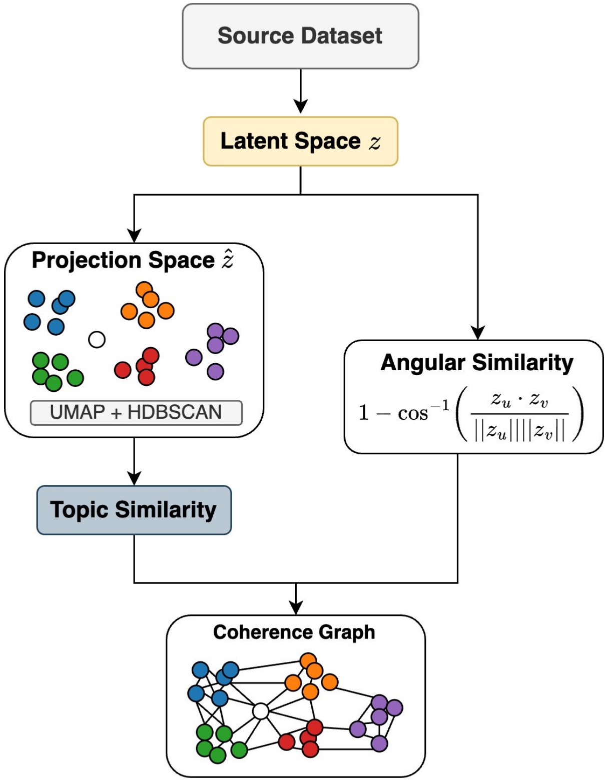
The coherence of a storyline is defined by *both* high content similarity and high topic similarity.

## Projection Space

Used to determine topic similarity as the Jensen-Shannon Divergence between the topic probability distributions of the documents obtained with UMAP and HDBSCAN.

## Base Coherence Graph

Encodes the pairwise coherence between documents using their angular (content) similarity in the embedding space and topic similarity in the projection space.

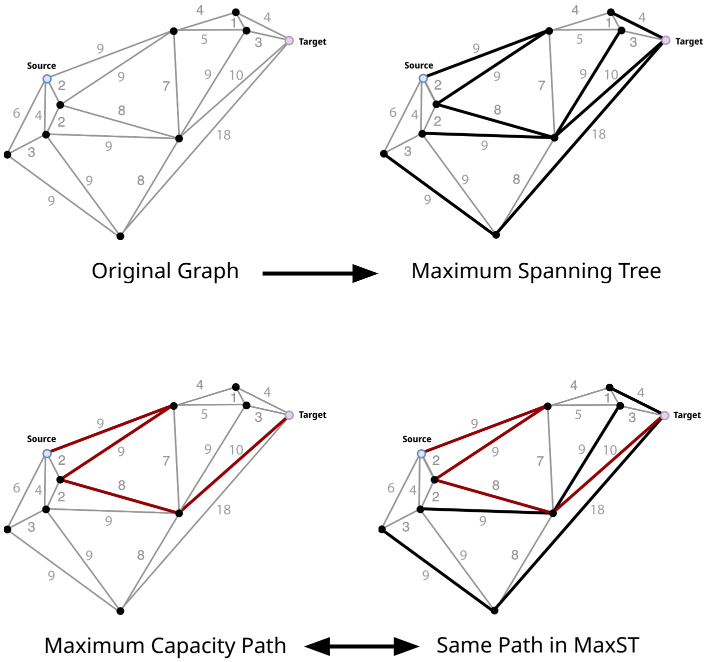





VIRGINIA TECH

# Building the Sparse Coherence Graph

Maximizing the minimum link is equivalent to finding a path of maximum capacity, which is the path between two nodes in an undirected graph's maximum spanning tree.




## Maximum Spanning Tree

Used as an optimization tool to reduce the search space for large graphs.

## Inducing Directionality

After finding the maximum spanning tree, explicit directionality—such as date order, citations, or hyperlinks—can be applied to the edges.

* Graph diagrams modified from Wikipedia's page on "Minimum spanning tree", accessed on April 02, 2025.

# Extracting Storylines

Use Dijkstra's algorithm with a MaxiMin objective on the sparse coherence graph.

## Extracting $k$ Distinct Storylines

Exclude documents already visited in previous storylines and re-execute the algorithm on the remaining documents.

## Redundancy Reduction

- Our objective of maximizing the minimum edge can lead to long storylines.
- We mitigate this issue by finding possible shortcuts in the storyline that maintain similar levels of coherence.

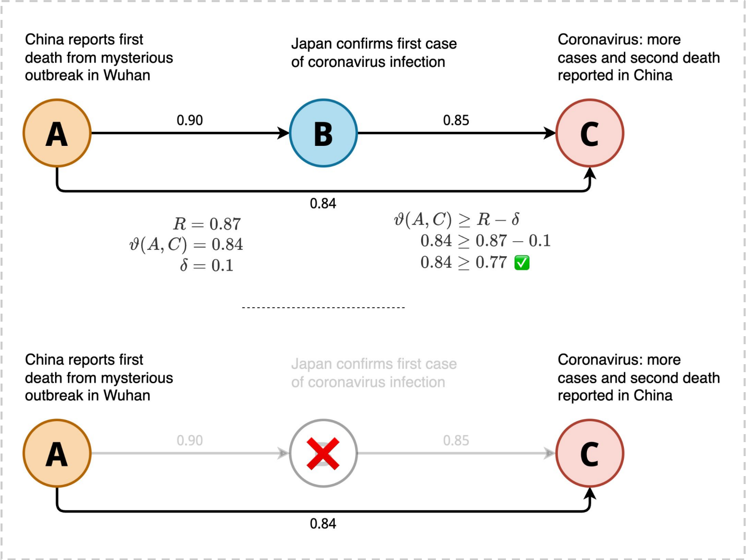



Illustration of the redundancy reduction technique

# Experiments & Evaluations

- (RQ1) How well does Narrative Trails align with human-derived shortest semantic paths?
- (RQ2) How do the storylines extracted by Narrative Trails compare to those extracted by the current state-of-the-art method?

## Evaluation Baselines

- WikiSpeedia
- Narrative Maps
- Shortest Paths
- Random Sampling

## Datasets

- WikiSpeedia: Human-derived shortest-paths from the Wikipedia network
- News Data: Articles about the COVID-19 pandemic and the 2021 Cuban Protests
- AMiner Subset: Research articles related to machine learning and AI
- VisPub: Research articles in the information visualization space

Wikipedia

News Data

AMiner

VIS



Virginia Tech

# Experiments & Evaluations

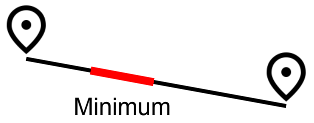
- (RQ1) How well does Narrative Trails align with human-derived shortest semantic paths?
- (RQ2) How do the storylines extracted by Narrative Trails compare to those extracted by the current state-of-the-art method?

## Evaluation Baselines

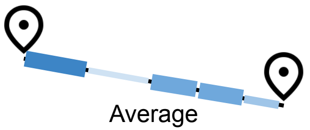
- WikiSpeedia
- Narrative Maps
- Shortest Paths
- Random Sampling

## Evaluation Metrics

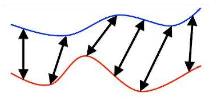
Minimum Storyline Coherence




Reliability: Geometric mean of coherence weights




Dynamic Time Warping Distance & Similarity



# Results

## Alignment with Human-Derived Paths

| Method | Min. Coherence | | | Reliability | | | DTW Similarity | | | nDTW Distance | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $k=1$ | $k=2$ | $k=3$ | $k=1$ | $k=2$ | $k=3$ | $k=1$ | $k=2$ | $k=3$ | $k=1$ | $k=2$ | $k=3$ |
| Wikispeedia | 0.419 | — | — | 0.609 | — | — | — | — | — | — | — | — |
| Random Points | 0.320 | 0.321 | 0.322 | 0.454 | 0.455 | 0.456 | 0.347 | 0.347 | 0.347 | 2.200 | 2.201 | 2.200 |
| Shortest Path | 0.558 | 0.560 | 0.563 | 0.614 | 0.615 | 0.620 | 0.742 | 0.742 | 0.746 | **0.967** | **0.978** | **0.971** |
| Narrative Trails | **0.709** | **0.704** | **0.704** | **0.776** | **0.769** | **0.767** | **0.788** | **0.785** | **0.787** | 1.029 | 1.049 | 1.063 |
| Redundancy Reduced | 0.668 | 0.667 | 0.669 | 0.760 | 0.756 | 0.755 | 0.769$^\dagger$ | 0.768 | <u>0.771</u>$^\dagger$ | 1.055 | 1.076 | 1.088 |
| Narrative Trails (CC) | 0.640 | 0.631 | 0.630 | 0.753 | 0.748 | 0.746 | <u>0.777</u> | <u>0.778</u> | 0.766$^\dagger$ | 1.029 | 1.049 | 1.093 |
| Redundancy Reduced (CC) | 0.630 | 0.625 | 0.624 | 0.737 | 0.735 | 0.734 | 0.759 | 0.761$^\dagger$ | 0.751 | 1.065 | 1.079 | 1.117 |

- Consistently produces storylines with higher minimum coherence, reliability, and Dynamic Time Warping Similarity.
- Shortest Path outperforms Narrative Trails in Dynamic Time Warping Distance due to the inherent nature of the user's task in the WikiSpeedia game.



VIRGINIA TECH

# Results

## Comparison with Narrative Maps

| Method | Min. Coherence | | | Reliability | | | DTW Similarity | | | nDTW Distance | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | News | VisPub | AMnr. | News | VisPub | AMnr. | News | VisPub | AMnr. | News | VisPub | AMnr. |
| Narrative Maps | 0.499 | 0.554 | 0.502 | 0.702 | 0.677 | 0.629 | — | — | — | — | — | — |
| Random Sample | 0.343 | 0.412 | 0.357 | 0.492 | 0.577 | 0.512 | 0.621 | 0.337 | 0.278 | 2.466 | 1.397 | 1.427 |
| Shortest Path | 0.557 | 0.743 | 0.635 | 0.593 | 0.753 | 0.644 | 0.363 | 0.461 | 0.188 | 1.001 | 0.991 | 1.108 |
| Narrative Trails | **0.689** | **0.784** | **0.736** | **0.786** | **0.800** | **0.764** | **0.872** | **0.616** | **0.556** | **0.762** | **0.915**† | **0.962** |
| Redundancy Reduced | 0.638 | 0.756 | 0.691 | 0.739 | 0.777 | 0.724 | 0.845 | 0.570† | 0.455 | 0.825 | 0.946† | 1.025† |

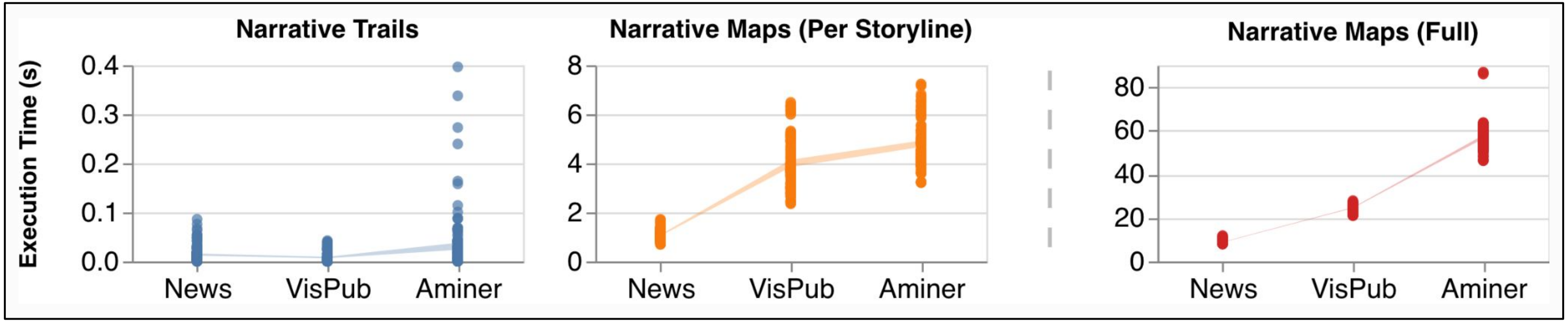
- Narrative Trails outperforms all baselines, including the Narrative Maps state-of-the-art method, in all evaluation metrics.



VIRGINIA TECH

# Results

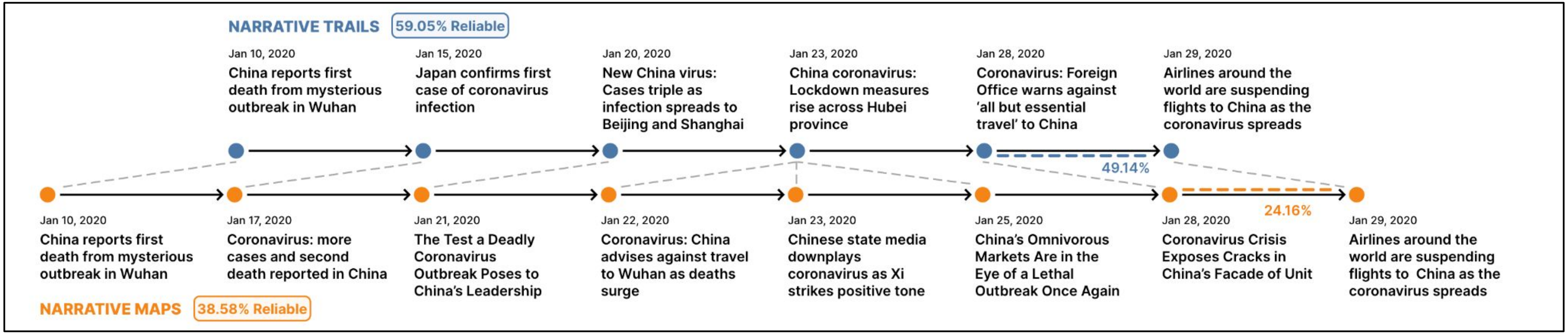
## Comparison with Narrative Maps




- Narrative Trails outperforms all baselines, including the Narrative Maps state-of-the-art method, in all evaluation metrics.
- Narrative Trails is also orders of magnitude faster than the Narrative Maps algorithm at extracting storylines.

# Example Storyline

Storylines about the COVID-19 pandemic's impact on global flights in January 2020, extracted from a collection of news articles using Narrative Trails (blue) and Narrative Maps (orange).

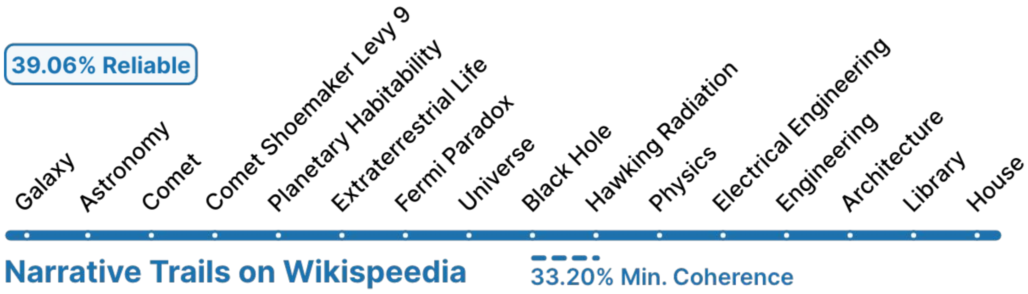


# Limitations & Future Work

1. The differences between the task of the WikiSpeedia game and Narrative Trails may make the evaluations difficult to interpret.
2. The evaluations do not include other narrative extraction methods such as Connect-the-Dots and newsLens due to limited code availability.
3. Features such as Coverage and semantic interaction were disabled in Narrative Maps, limiting its performance.
4. <u>Future Directions</u>
    a. Deep-learning-based search agents with controllable parameters such as length.
    b. Efficient narrative extraction on more complex datasets and tasks, such as multimodal storylines.

# Conclusions

- Abstractive approach to narrative extraction based on latent semantics rather than keyword-based heuristics.
- Successfully extract coherent storylines from large datasets with varying topics, tasks, and graph structures.
- Our efficient and abstractive approach opens the doors to deep-learning-based storyline extraction on tasks beyond text.




GitHub Repository

Source code is publicly available on GitHub

VIRGINIA TECH

# Authors

Fausto German[1], Brian Keith[2], Chris North[1]

[1]Virginia Tech, Blacksburg, Virginia 24061, USA
[2]Universidad Católica del Norte, Av. Angamos 0610, Antofagasta, 1270709, Chile

**Personal Website**

**Text2Story @ ECIR 2025**

April 10, 2025

# Thank You!

Universidad Católica del Norte