Amazon Book Sales Data

Team - 50 : Snigdha Raghavan Pradhipa, Vaishnavi Narasimhaiah Sathish,
Samrudhi Ramesh Rao, Kongarasan Sathiya Moorthy

July 20, 2025

In the context of semantic data integration, aligning heterogeneous XML schemas is a critical step toward enabling meaningful data exchange and interoperability across diverse sources. XML schemas define the structural and semantic elements of XML documents through hierarchically organized nodes and paths. However, schemas from different data providers often vary in terminology, structure, or granularity, posing challenges for direct schema alignment.

1 XML Schema Parsing and Matching

The first phase involved parsing the XML schemas using Python's ElementTree library to construct schema trees for both DS and MS. These trees capture the node hierarchy and path context required for matching.

Figure 1: Snapshot of Schema tree 01

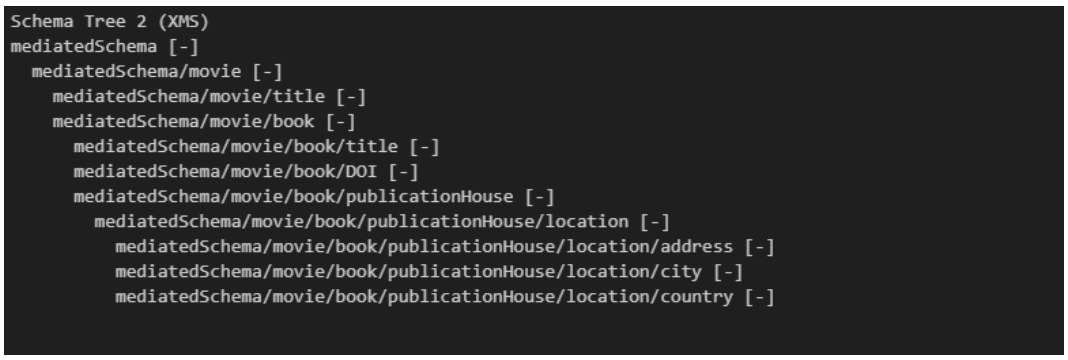


Figure 2: Snapshot of Schema tree 02

2 Node Matching and Path Matching

In the second phase, we implemented multiple string similarity techniques to compute pairwise similarities between nodes and paths from the two schemas:

- **Levenshtein edit distance**, which measures character-level transformations.
- **Jaccard similarity**, which considers overlapping tokens (especially useful for path segments).
- **TF-IDF vectorization with cosine similarity**, capturing weighted textual similarity.
- **FuzzyWuzzy** (token sort ratio), a token-based fuzzy matching method.

2.1 Levenshtein distance / Edit distance

Levenshtein distance, also known as editdistance, is a crucial metric in semantic dataintegration projects. It quantifies the difference between two strings by counting the minimum number of single-character edits—insertions, deletions, or substitutions—required to transform one string into the other.

By leveraging Levenshtein distance, data integration systems can effectively match and merge records from disparate sources, ensuring higher accuracy and consistency in the integrated dataset. This process enhances the overall quality of the data, supporting more reliable analysis and decision-making.

Path:

PATH :
Edit Distance Similarity Matrix

	movie	movie/title	movie/director	movie/director/firstName	movie/director/lastName
mediatedSchema	0.214286	0.214286	0.142857	0.250000	0.260870
mediatedSchema/movie	0.250000	0.250000	0.250000	0.208333	0.260870
mediatedSchema/movie/title	0.192308	0.423077	0.230769	0.269231	0.230769
mediatedSchema/movie/book	0.200000	0.200000	0.240000	0.160000	0.160000
mediatedSchema/movie/book/title	0.161290	0.354839	0.225806	0.258065	0.258065
mediatedSchema/movie/book/DOI	0.172414	0.206897	0.206897	0.172414	0.206897
mediatedSchema/movie/book/publicationHouse	0.119048	0.214286	0.214286	0.238095	0.238095
mediatedSchema/movie/book/publicationHouse/location	0.098039	0.176471	0.215686	0.196078	0.215686
mediatedSchema/movie/book/publicationHouse/location/address	0.084746	0.169492	0.203390	0.220339	0.237288
mediatedSchema/movie/book/publicationHouse/location/city	0.089286	0.160714	0.196429	0.196429	0.214286
mediatedSchema/movie/book/publicationHouse/location/country	0.084746	0.152542	0.203390	0.186441	0.203390

Figure 3: Snapshot of 1st few columns of Levenshtein distance for Paths

PATH :
Edit Distance Similarity Matrix

movie/genre	movie/releaseYear	movie/actors	movie/actors/actor	movie/actors/actor/actorName	...	movie/ratings/rating/score	movie/book
0.214286	0.294118	0.285714	0.222222	0.214286	...	0.192308	0.214286
0.250000	0.250000	0.250000	0.250000	0.285714	...	0.230769	0.250000
0.269231	0.230769	0.230769	0.269231	0.214286	...	0.192308	0.230769
0.200000	0.240000	0.200000	0.240000	0.178571	...	0.192308	0.400000
0.225806	0.193548	0.225806	0.225806	0.225806	...	0.193548	0.322581
0.206897	0.206897	0.206897	0.206897	0.206897	...	0.172414	0.344828
0.190476	0.190476	0.238095	0.214286	0.261905	...	0.214286	0.238095
0.156863	0.215686	0.196078	0.274510	0.215686	...	0.235294	0.196078
0.169492	0.203390	0.203390	0.254237	0.254237	...	0.305085	0.169492
0.142857	0.196429	0.178571	0.250000	0.232143	...	0.285714	0.178571
0.152542	0.203390	0.186441	0.254237	0.254237	...	0.305085	0.169492

Figure 4: Snapshot of 2st set of columns of Levenshtein distance for Paths

PATH :
Edit Distance Similarity Matrix

on	movie/ShootLocation/address	movie/ShootLocation/city	movie/ShootLocation/country	movie/ShootLocation/latitude	movie/ShootLocation/longitude
95	0.185185	0.166667	0.185185	0.142857	0.137931
00	0.185185	0.208333	0.185185	0.214286	0.241379
92	0.185185	0.269231	0.185185	0.285714	0.241379
00	0.148148	0.200000	0.185185	0.142857	0.172414
23	0.225806	0.322581	0.225806	0.290323	0.258065
79	0.172414	0.241379	0.206897	0.172414	0.172414
24	0.261905	0.309524	0.285714	0.333333	0.285714
18	0.274510	0.313725	0.313725	0.313725	0.313725
37	0.389831	0.271186	0.288136	0.338983	0.322034
57	0.267857	0.357143	0.303571	0.321429	0.303571
37	0.271186	0.322034	0.389831	0.322034	0.305085

Figure 5: Snapshot of 3rd set of columns of Levenshtein distance for Paths

Node:

NODE :
Edit Distance Similarity Matrix

	movie	title	director	firstName	lastName	genre	releaseYear	actors	actor	actorName	...	critic	score
mediatedSchema	0.214286	0.214286	0.285714	0.142857	0.142857	0.142857	0.285714	0.142857	0.142857	0.142857	...	0.214286	0.071429
movie	1.000000	0.200000	0.000000	0.111111	0.125000	0.200000	0.090909	0.000000	0.000000	0.222222	...	0.166667	0.200000
title	0.200000	1.000000	0.250000	0.333333	0.250000	0.200000	0.090909	0.166667	0.200000	0.222222	...	0.333333	0.200000
book	0.200000	0.000000	0.125000	0.000000	0.000000	0.000000	0.000000	0.166667	0.200000	0.111111	...	0.000000	0.200000
DOI	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000
publicationHouse	0.125000	0.187500	0.187500	0.187500	0.187500	0.125000	0.125000	0.250000	0.187500	0.250000	...	0.187500	0.187500
location	0.250000	0.125000	0.125000	0.111111	0.125000	0.000000	0.090909	0.250000	0.375000	0.000000	...	0.250000	0.125000
address	0.142857	0.142857	0.250000	0.000000	0.125000	0.285714	0.181818	0.285714	0.142857	0.222222	...	0.000000	0.285714
city	0.000000	0.400000	0.250000	0.222222	0.125000	0.000000	0.000000	0.166667	0.200000	0.111111	...	0.500000	0.200000
country	0.142857	0.142857	0.125000	0.111111	0.000000	0.285714	0.000000	0.142857	0.142857	0.222222	...	0.285714	0.285714

Figure 6: Snapshot of 1st set of columns of Levenshtein distance for Nodes

NODE :
Edit Distance Similarity Matrix

releaseYear	actors	actor	actorName	...	critic	score	book	DOI	ShootLocation	address	city	country	latitude	longitude
0.285714	0.142857	0.142857	0.142857	...	0.214286	0.071429	0.000000	0.0	0.071429	0.214286	0.142857	0.071429	0.285714	0.214286
0.090909	0.000000	0.000000	0.222222	...	0.166667	0.200000	0.200000	0.0	0.153846	0.142857	0.000000	0.142857	0.250000	0.333333
0.090909	0.166667	0.200000	0.222222	...	0.333333	0.200000	0.000000	0.0	0.153846	0.142857	0.400000	0.142857	0.500000	0.333333
0.000000	0.166667	0.200000	0.111111	...	0.000000	0.200000	1.000000	0.0	0.153846	0.000000	0.000000	0.142857	0.000000	0.111111
0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	1.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.125000	0.250000	0.187500	0.250000	...	0.187500	0.187500	0.187500	0.0	0.250000	0.125000	0.125000	0.125000	0.375000	0.250000
0.090909	0.250000	0.375000	0.000000	...	0.250000	0.125000	0.250000	0.0	0.538462	0.000000	0.250000	0.250000	0.250000	0.333333
0.181818	0.285714	0.142857	0.222222	...	0.000000	0.285714	0.000000	0.0	0.000000	1.000000	0.000000	0.000000	0.125000	0.000000
0.000000	0.166667	0.200000	0.111111	...	0.500000	0.200000	0.000000	0.0	0.153846	0.000000	1.000000	0.428571	0.250000	0.222222
0.000000	0.142857	0.142857	0.222222	...	0.285714	0.285714	0.142857	0.0	0.153846	0.000000	0.428571	1.000000	0.125000	0.222222

Figure 7: Snapshot of 2nd set of columns of Levenshtein distance for Nodes

2.2 Jaccard Similarity

Two sets can be compared for similarity and diversity using the Jaccard Similarity metric. It is frequently employed in recommendation systems text analysis and clustering and is defined as the intersections size divided by the unions size of the sets. From 0 (no similarity) to 1 (identical sets) it goes from there.

Path:

PATH :
Jaccard Similarity Matrix

	movie	movie/title	movie/director	movie/director/firstName	movie/director/lastName	movie/genre	movie/releaseYear	movie/actors
mediatedSchema	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
mediatedSchema/movie	0.500000	0.333333	0.333333	0.250000	0.250000	0.333333	0.333333	0.333333
mediatedSchema/movie/title	0.333333	0.666667	0.250000	0.200000	0.200000	0.250000	0.250000	0.250000
mediatedSchema/movie/book	0.333333	0.250000	0.250000	0.200000	0.200000	0.250000	0.250000	0.250000
mediatedSchema/movie/book/title	0.250000	0.500000	0.200000	0.166667	0.166667	0.200000	0.200000	0.200000
mediatedSchema/movie/book/DOI	0.250000	0.200000	0.200000	0.166667	0.166667	0.200000	0.200000	0.200000
mediatedSchema/movie/book/publicationHouse	0.250000	0.200000	0.200000	0.166667	0.166667	0.200000	0.200000	0.200000
mediatedSchema/movie/book/publicationHouse/location	0.200000	0.166667	0.166667	0.142857	0.142857	0.166667	0.166667	0.166667
mediatedSchema/movie/book/publicationHouse/location/address	0.166667	0.142857	0.142857	0.125000	0.125000	0.142857	0.142857	0.142857
mediatedSchema/movie/book/publicationHouse/location/city	0.166667	0.142857	0.142857	0.125000	0.125000	0.142857	0.142857	0.142857
mediatedSchema/movie/book/publicationHouse/location/country	0.166667	0.142857	0.142857	0.125000	0.125000	0.142857	0.142857	0.142857

11 rows × 24 columns

Figure 8: Snapshot of 1st few columns of Jaccard Similarity for Paths

PATH :
Jaccard Similarity Matrix

movie/actors/actor/actorName	...	movie/ratings/rating/score	movie/book	movie/book/title	movie/book/DOI	movie/ShootLocation
0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000
0.200000	...	0.200000	0.333333	0.250000	0.250000	0.333333
0.166667	...	0.166667	0.250000	0.500000	0.200000	0.250000
0.166667	...	0.166667	0.666667	0.500000	0.500000	0.250000
0.142857	...	0.142857	0.500000	0.750000	0.400000	0.200000
0.142857	...	0.142857	0.500000	0.400000	0.750000	0.200000
0.142857	...	0.142857	0.500000	0.400000	0.400000	0.200000
0.125000	...	0.125000	0.400000	0.333333	0.333333	0.166667
0.111111	...	0.111111	0.333333	0.285714	0.285714	0.142857
0.111111	...	0.111111	0.333333	0.285714	0.285714	0.142857
0.111111	...	0.111111	0.333333	0.285714	0.285714	0.142857

Figure 9: Snapshot of 2nd set of columns of Jaccard Similarity for Paths

PATH :
Jaccard Similarity Matrix

or/actorName	...	movie/ratings/rating/score	movie/book	movie/book/title	movie/book/DOI	movie/ShootLocation	movie/ShootLocation/address	movie/ShootLocation/city	movie/ShootLocation/country	movie/ShootLocation/latitude	movie/ShootLocation/longitude
0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.200000	...	0.200000	0.333333	0.250000	0.250000	0.333333	0.250000	0.250000	0.250000	0.250000	0.250000
0.166667	...	0.166667	0.250000	0.500000	0.200000	0.250000	0.200000	0.200000	0.200000	0.200000	0.200000
0.166667	...	0.166667	0.666667	0.500000	0.500000	0.250000	0.200000	0.200000	0.200000	0.200000	0.200000
0.142857	...	0.142857	0.500000	0.750000	0.400000	0.200000	0.166667	0.166667	0.166667	0.166667	0.166667
0.142857	...	0.142857	0.500000	0.400000	0.750000	0.200000	0.166667	0.166667	0.166667	0.166667	0.166667
0.142857	...	0.142857	0.500000	0.400000	0.400000	0.200000	0.166667	0.166667	0.166667	0.166667	0.166667
0.125000	...	0.125000	0.400000	0.333333	0.333333	0.166667	0.142857	0.142857	0.142857	0.142857	0.142857
0.111111	...	0.111111	0.333333	0.285714	0.285714	0.142857	0.285714	0.125000	0.125000	0.125000	0.125000
0.111111	...	0.111111	0.333333	0.285714	0.285714	0.142857	0.125000	0.285714	0.125000	0.125000	0.125000
0.111111	...	0.111111	0.333333	0.285714	0.285714	0.142857	0.125000	0.125000	0.285714	0.125000	0.125000

Figure 10: Snapshot of 3rd set of columns of Jaccard Similarity for Paths

Node:

NODE :
Jaccard Similarity Matrix

	movie	title	director	firstName	lastName	genre	releaseYear	actors	actor	actorName	...	critic	score	book	DOI	ShootLocation	address	city	country	latitude	longitude
mediatedSchema	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
movie	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
title	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
book	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DOI	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
publicationHouse	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
location	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
address	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
city	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
country	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0

10 rows × 23 columns

Figure 11: Snapshot of Jaccard Similarity for Nodes

2.3 TF-IDF measure

TF-IDF is used to convert schema element labels into numerical vectors by weighing terms based on their frequency and uniqueness across all labels. Cosine similarity is then applied to compute similarity scores between each pair of elements. This method captures meaningful matches even when labels share partial overlap, making it effective for identifying semantically similar but non-identical terms.

***Matcher type*:** Statistical, vector-based

Path:

	movie	movie/title	movie/director	movie/director/firstName	movie/director/lastName	movie/genre	movie/rele
mediatedSchema	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	
mediatedSchema/movie	0.4400	0.1438	0.1347	0.0880	0.0880	0.1124	
mediatedSchema/movie/title	0.2718	0.8320	0.0832	0.0544	0.0544	0.0695	
mediatedSchema/movie/book	0.3273	0.1070	0.1002	0.0655	0.0655	0.0836	
mediatedSchema/movie/book/title	0.2377	0.7275	0.0728	0.0476	0.0476	0.0607	
mediatedSchema/movie/book/DOI	0.2192	0.0716	0.0671	0.0438	0.0438	0.0560	
mediatedSchema/movie/book/publicationHouse	0.2447	0.0799	0.0749	0.0489	0.0489	0.0625	
ediatedSchema/movie/book/publicationHouse/location	0.1997	0.0653	0.0611	0.0400	0.0400	0.0510	
chema/movie/book/publicationHouse/location/address	0.1654	0.0540	0.0506	0.0331	0.0331	0.0423	
tedSchema/movie/book/publicationHouse/location/city	0.1654	0.0540	0.0506	0.0331	0.0331	0.0423	
chema/movie/book/publicationHouse/location/country	0.1654	0.0540	0.0506	0.0331	0.0331	0.0423	

Figure 12: Snapshot of 1st set of columns of TF-IDF Matcher for Paths

e/genre	movie/releaseYear	movie/actors	movie/actors/actor	movie/actors/actor/actorName	...	movie/ratings/rating/score	movie/book	movie/book/title	movie/boo
0.0000	0.0000	0.0000	0.0000	0.0000	...	0.0000	0.0000	0.0000	0
0.1124	0.1124	0.1438	0.1008	0.0762	...	0.0762	0.1936	0.1196	
0.0695	0.0695	0.0888	0.0623	0.0471	...	0.0471	0.1196	0.6922	0
0.0836	0.0836	0.1070	0.0750	0.0567	...	0.0567	0.7440	0.4597	0
0.0607	0.0607	0.0777	0.0545	0.0412	...	0.0412	0.5403	0.8744	0
0.0560	0.0560	0.0716	0.0502	0.0379	...	0.0379	0.4981	0.3078	0
0.0625	0.0625	0.0799	0.0561	0.0424	...	0.0424	0.5561	0.3436	0
0.0510	0.0510	0.0653	0.0458	0.0346	...	0.0346	0.4539	0.2805	0
0.0423	0.0423	0.0540	0.0379	0.0286	...	0.0286	0.3759	0.2323	0
0.0423	0.0423	0.0540	0.0379	0.0286	...	0.0286	0.3759	0.2323	0
0.0423	0.0423	0.0540	0.0379	0.0286	...	0.0286	0.3759	0.2323	0

Figure 13: Snapshot of 2nd set of columns of TF-IDF Matcher for Paths

ie/ShootLocation	movie/ShootLocation/address	movie/ShootLocation/city	movie/ShootLocation/country	movie/ShootLocation/latitude	movie/ShootLocation/longitude
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.1598	0.1008	0.1008	0.1008	0.0940	0.0940
0.0987	0.0622	0.0622	0.0622	0.0581	0.0581
0.1189	0.0750	0.0750	0.0750	0.0700	0.0700
0.0863	0.0544	0.0544	0.0544	0.0508	0.0508
0.0796	0.0502	0.0502	0.0502	0.0468	0.0468
0.0889	0.0560	0.0560	0.0560	0.0523	0.0523
0.0725	0.0457	0.0457	0.0457	0.0427	0.0427
0.0601	0.4730	0.0379	0.0379	0.0353	0.0353
0.0601	0.0379	0.4730	0.0379	0.0353	0.0353
0.0601	0.0379	0.0379	0.4730	0.0353	0.0353

Figure 14: Snapshot of 3rd set of columns of TF-IDF Matcher for Paths

Node:

TF-IDF Node Similarity Matrix:

	movie	title	director	firstName	lastName	genre	releaseYear	actors	actor	actorName	...	critic	score	book	DOI	ShootLocation	addres
mediatedSchema	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.
movie	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.
title	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.
book	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	1.0	0.0	0.0	0.
DOI	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	1.0	0.0	0.
publicationHouse	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.
location	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.
address	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	1.
city	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.
country	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.

10 rows × 23 columns

Figure 15: Snapshot of 1st set of TF-IDF Matcher for Node

TF-IDF Node Similarity Matrix:

ctor	firstName	lastName	genre	releaseYear	actors	actor	actorName	...	critic	score	book	DOI	ShootLocation	address	city	country	latitude	longitude
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0

Figure 16: Snapshot of 2nd set of columns of TF-IDF Matcher for Node

2.4 Fuzzy Matching

FuzzyWuzzy applies Levenshtein distance to sorted tokens of each label, making it robust to word order and minor spelling variations. It returns a similarity score based on how closely the strings match after sorting their tokens alphabetically.

Matcher type: Token-based string similarity

Path:

Fuzzy-Wuzzy Path Similarity Matrix

	movie	movie/title	movie/director	movie/director/firstName	movie/director/lastName	movie/genre
mediatedSchema	0.3158	0.4000	0.4286	0.3684	0.3784	0.3200
mediatedSchema/movie	0.4000	0.3871	0.4118	0.4091	0.4186	0.3226
mediatedSchema/movie/title	0.3226	0.5946	0.4000	0.4400	0.4082	0.3784
mediatedSchema/movie/book	0.3333	0.3333	0.3590	0.3673	0.3750	0.3333
mediatedSchema/movie/book/title	0.2778	0.5238	0.3556	0.4000	0.3704	0.3333
mediatedSchema/movie/book/DOI	0.2941	0.3000	0.3256	0.3396	0.3462	0.3000
mediatedSchema/movie/book/publicationHouse	0.2128	0.3396	0.3571	0.3636	0.3692	0.3019
mediatedSchema/movie/book/publicationHouse/location	0.1786	0.2903	0.3385	0.3200	0.3784	0.2581
mediatedSchema/movie/book/publicationHouse/location/address	0.1562	0.2857	0.3288	0.3373	0.3902	0.2857
mediatedSchema/movie/book/publicationHouse/location/city	0.1639	0.2687	0.3143	0.3500	0.3544	0.2388
mediatedSchema/movie/book/publicationHouse/location/country	0.1562	0.2571	0.3288	0.3133	0.3415	0.2571

Figure 17: Snapshot of 1st set of columns of Fuzzy-Wuzzy Matcher for Path

Fuzzy-Wuzzy Path Similarity Matrix

movie/releaseYear	movie/actors	movie/actors/actor	movie/actors/actor/actorName	...	movie/ratings/rating/score	movie/book	movie/book/title	movie/book/DOI
0.3871	0.3077	0.3125	0.2857	...	0.3500	0.2500	0.3333	0.2143
0.3243	0.3125	0.3158	0.3333	...	0.3478	0.3333	0.3889	0.2941
0.3721	0.3684	0.3636	0.3333	...	0.3846	0.3333	0.5238	0.3000
0.2857	0.3784	0.3721	0.3019	...	0.3137	0.5714	0.4878	0.5128
0.3333	0.3256	0.3673	0.3729	...	0.3509	0.4878	0.6809	0.4889
0.2609	0.3415	0.3404	0.3158	...	0.2909	0.5128	0.4889	0.6512
0.3390	0.3704	0.3667	0.3714	...	0.3824	0.3846	0.4828	0.3929
0.3235	0.3175	0.4058	0.3797	...	0.4156	0.3279	0.4179	0.3385
0.3158	0.3380	0.3896	0.4138	...	0.4471	0.2899	0.4000	0.3014
0.3014	0.2941	0.3784	0.4048	...	0.4390	0.3030	0.3889	0.3143
0.3158	0.3099	0.3896	0.4138	...	0.4706	0.2899	0.3733	0.3014

Figure 18: Snapshot of 1st set of columns of Fuzzy-Wuzzy Matcher for Path

Fuzzy-Wuzzy Path Similarity Matrix

e/ShootLocation	movie/ShootLocation/address	movie/ShootLocation/city	movie/ShootLocation/country	movie/ShootLocation/latitude	movie/ShootLocation/longitude
0.3636	0.2927	0.3158	0.2927	0.3333	0.2791
0.3590	0.3404	0.3636	0.3404	0.3750	0.4082
0.3556	0.3396	0.4000	0.3396	0.4444	0.4000
0.3636	0.3077	0.3265	0.3462	0.3396	0.3704
0.4000	0.3793	0.4000	0.3793	0.4407	0.4000
0.3333	0.3214	0.3396	0.3214	0.3158	0.3448
0.4590	0.4348	0.4242	0.4638	0.4571	0.4789
0.4571	0.4103	0.4533	0.4615	0.4810	0.4500
0.4103	0.5581	0.4096	0.4419	0.4828	0.4545
0.4267	0.4096	0.5250	0.4819	0.4762	0.4706
0.4103	0.4186	0.4819	0.5581	0.4598	0.4545

Figure 19: Snapshot of 1st set of columns of Fuzzy-Wuzzy Matcher for Path

Node:

Fuzzy-Wuzzy Node Similarity Matrix

	movie	title	director	firstName	lastName	genre	releaseYear	actors	actor	actorName	...	critic	score	book	DOI	ShootLocati
mediatedSchema	0.3158	0.3158	0.3636	0.2609	0.2727	0.2105	0.4000	0.2000	0.2105	0.2609	...	0.3000	0.2105	0.0000	0.0	0.22
movie	1.0000	0.4000	0.3077	0.2857	0.3077	0.2000	0.1250	0.1818	0.2000	0.2857	...	0.1818	0.4000	0.2222	0.0	0.22
title	0.4000	1.0000	0.3077	0.4286	0.3077	0.2000	0.2500	0.1818	0.2000	0.2857	...	0.3636	0.2000	0.0000	0.0	0.22
book	0.2222	0.0000	0.1667	0.0000	0.0000	0.0000	0.0000	0.2000	0.2222	0.1538	...	0.0000	0.2222	1.0000	0.0	0.23
DOI	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	...	0.0000	0.0000	0.0000	1.0	0.00
publicationHouse	0.1905	0.2857	0.3333	0.2400	0.3333	0.1905	0.2963	0.3636	0.2857	0.3200	...	0.2727	0.2857	0.3000	0.0	0.41
location	0.3077	0.3077	0.3750	0.1176	0.3750	0.1538	0.2105	0.4286	0.4615	0.3529	...	0.4286	0.3077	0.3333	0.0	0.66
address	0.1667	0.1667	0.4000	0.2500	0.2667	0.3333	0.3333	0.4615	0.3333	0.3750	...	0.1538	0.3333	0.0000	0.0	0.10
city	0.2222	0.4444	0.3333	0.3077	0.1667	0.0000	0.0000	0.4000	0.4444	0.3077	...	0.6000	0.2222	0.0000	0.0	0.23
country	0.1667	0.1667	0.4000	0.1250	0.1333	0.3333	0.1111	0.4615	0.5000	0.3750	...	0.3077	0.5000	0.1818	0.0	0.30

Figure 20: Snapshot of 1st set of columns of Fuzzy-Wuzzy Matcher for Node

Fuzzy-Wuzzy Node Similarity Matrix

lame	lastName	genre	releaseYear	actors	actor	actorName	...	critic	score	book	DOI	ShootLocation	address	city	country	latitude	longitude
2609	0.2727	0.2105	0.4000	0.2000	0.2105	0.2609	...	0.3000	0.2105	0.0000	0.0	0.2222	0.2857	0.2222	0.0952	0.3636	0.3478
.2857	0.3077	0.2000	0.1250	0.1818	0.2000	0.2857	...	0.1818	0.4000	0.2222	0.0	0.2222	0.1667	0.2222	0.1667	0.3077	0.4286
4286	0.3077	0.2000	0.2500	0.1818	0.2000	0.2857	...	0.3636	0.2000	0.0000	0.0	0.2222	0.1667	0.4444	0.1667	0.6154	0.4286
0000	0.0000	0.0000	0.0000	0.2000	0.2222	0.1538	...	0.0000	0.2222	1.0000	0.0	0.2353	0.0000	0.0000	0.1818	0.0000	0.1538
0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	...	0.0000	0.0000	0.0000	1.0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2400	0.3333	0.1905	0.2963	0.3636	0.2857	0.3200	...	0.2727	0.2857	0.3000	0.0	0.4138	0.1739	0.2000	0.2609	0.5000	0.4000
).1176	0.3750	0.1538	0.2105	0.4286	0.4615	0.3529	...	0.4286	0.3077	0.3333	0.0	0.6667	0.1333	0.3333	0.4000	0.5000	0.3529
2500	0.2667	0.3333	0.3333	0.4615	0.3333	0.3750	...	0.1538	0.3333	0.0000	0.0	0.1000	1.0000	0.0000	0.1429	0.4000	0.2500
.3077	0.1667	0.0000	0.0000	0.4000	0.4444	0.3077	...	0.6000	0.2222	0.0000	0.0	0.2353	0.0000	1.0000	0.5455	0.3333	0.3077
.1250	0.1333	0.3333	0.1111	0.4615	0.5000	0.3750	...	0.3077	0.5000	0.1818	0.0	0.3000	0.1429	0.5455	1.0000	0.1333	0.3750

Figure 21: Snapshot of 1st set of columns of Fuzzy-Wuzzy Matcher for Node

2.5 Applying Threshold and Evaluating Performence

To convert continuous similarity scores into binary alignment decisions, we applied a thresholding step. For each similarity matrix (TF-IDF, FuzzyWuzzy, Jaccard, Levenshtein), a similarity score above a defined threshold (e.g., 0.5 or 0.75) was

considered a positive match, while scores below the threshold were discarded. This allowed us to extract alignments from noisy similarity data and control the strictness of matching. Separate thresholds were applied for:

Node similarity: 0.6

Path similarity: 0.75

The threshold value significantly influences precision and recall — higher thresholds yield fewer but more confident matches, while lower thresholds increase recall at the cost of precision.

The performance of similarity and matching algorithms in schema integration tasks is assessed using three key metrics: **precision, recall, and F1-score.**

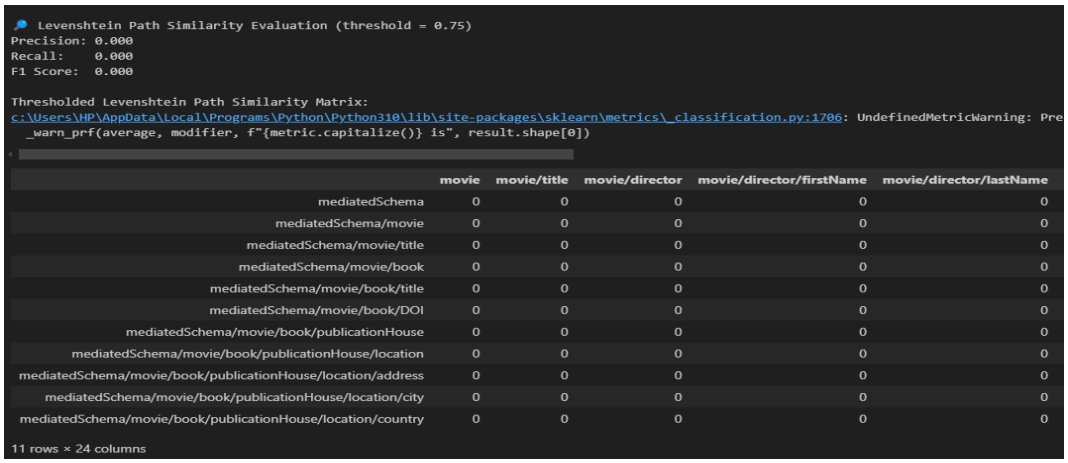
Precision Precision is an indicator of prediction reliability, as it quantifies the percentage of predicted matches that are actually accurate. A method with high precision suggests that it generates few false positives, ensuring that most of the matches it proposes are correct.

Recall In contrast, recall quantifies the percentage of true correct matches that the method successfully detects. It reflects the method's capacity to capture the entire collection of relevant correspondences. A technique with high recall guarantees wider coverage and reduces the number of false negatives.

F1-score The F1-score offers a balanced assessment of the method's overall efficacy by combining precision and recall into a single value, calculated as their harmonic mean. In practice, a good matching strategy aims to achieve both high precision and high recall. However, in many real-world scenarios, there is often a trade-off between the two. Thus, assessing all three metrics together helps to understand the strengths and weaknesses of a method under various thresholds or similarity measures.

The Evaluation Scores are:

- **Levenshtein Path Similarity:** Threshold = **0.75**



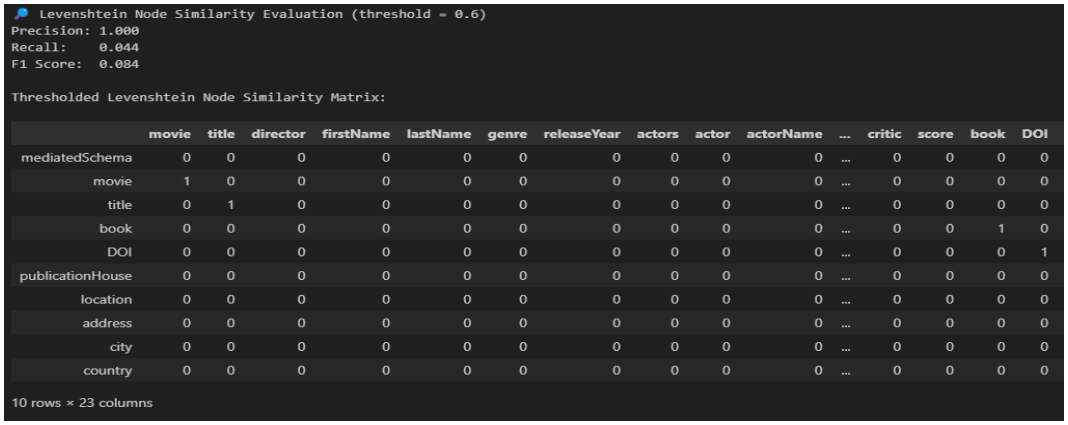
```
🔎 Levenshtein Path Similarity Evaluation (threshold = 0.75)
Precision: 0.000
Recall:    0.000
F1 Score:  0.000

Thresholded Levenshtein Path Similarity Matrix:
c:\Users\HP\AppData\Local\Programs\Python\Python310\lib\site-packages\sklearn\metrics\_classification.py:1706: UndefinedMetricWarning: Pre
  _warn_prf(average, modifier, f"{metric.capitalize()} is", result.shape[0])
```

	movie	movie/title	movie/director	movie/director/firstName	movie/director/lastName
mediatedSchema	0	0	0	0	0
mediatedSchema/movie	0	0	0	0	0
mediatedSchema/movie/title	0	0	0	0	0
mediatedSchema/movie/book	0	0	0	0	0
mediatedSchema/movie/book/title	0	0	0	0	0
mediatedSchema/movie/book/DOI	0	0	0	0	0
mediatedSchema/movie/book/publicationHouse	0	0	0	0	0
mediatedSchema/movie/book/publicationHouse/location	0	0	0	0	0
mediatedSchema/movie/book/publicationHouse/location/address	0	0	0	0	0
mediatedSchema/movie/book/publicationHouse/location/city	0	0	0	0	0
mediatedSchema/movie/book/publicationHouse/location/country	0	0	0	0	0

11 rows × 24 columns

Figure 22: Snapshot of Path Scores for Levenshtein Similarity

- **Levenshtein Node Similarity:** Threshold = **0.6**



Levenshtein Node Similarity Evaluation (threshold = 0.6)
Precision: 1.000
Recall: 0.044
F1 Score: 0.084

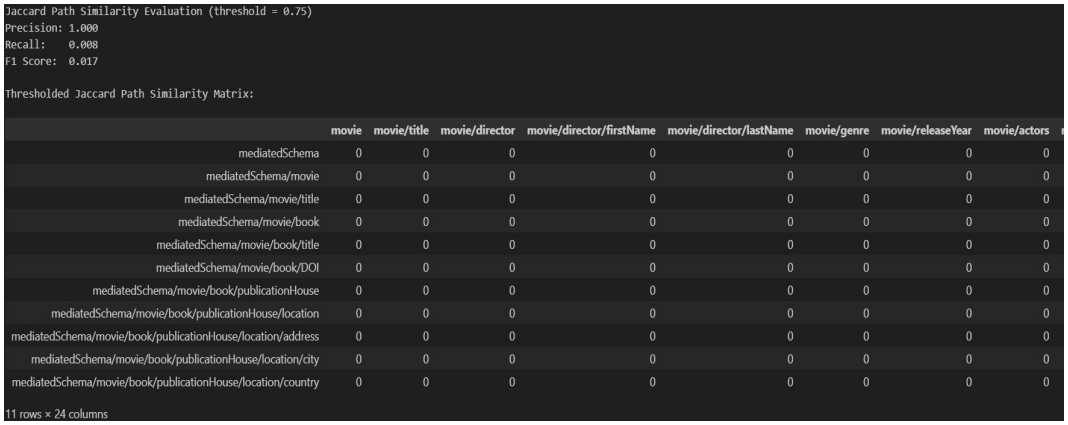
Thresholded Levenshtein Node Similarity Matrix:

	movie	title	director	firstName	lastName	genre	releaseYear	actors	actor	actorName	...	critic	score	book	DOI
mediatedSchema	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
movie	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0
title	0	1	0	0	0	0	0	0	0	0	...	0	0	0	0
book	0	0	0	0	0	0	0	0	0	0	...	0	0	1	0
DOI	0	0	0	0	0	0	0	0	0	0	...	0	0	0	1
publicationHouse	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
location	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
address	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
city	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
country	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0

10 rows × 23 columns

Figure 23: Snapshot of Node Scores for Levenshtein Similarity

- **Jaccard Path Similarity:** Threshold = **0.75**



Jaccard Path Similarity Evaluation (threshold = 0.75)
Precision: 1.000
Recall: 0.008
F1 Score: 0.017

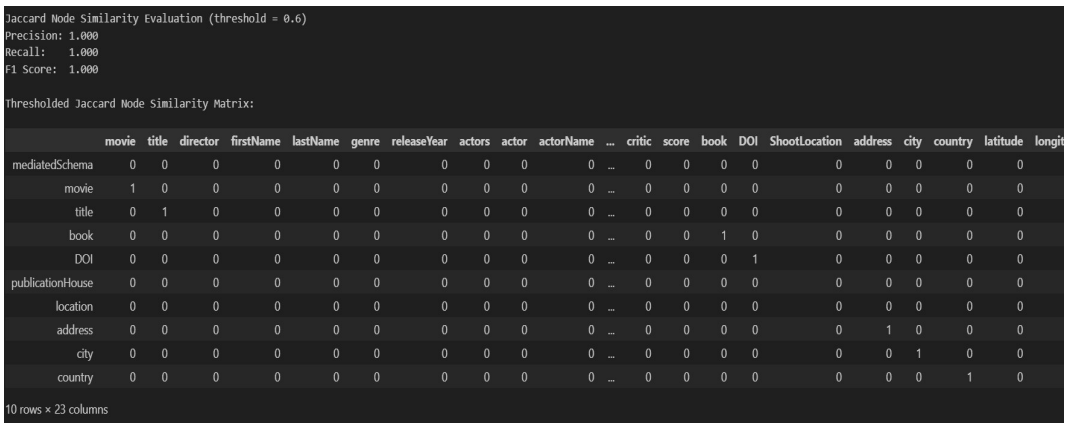
Thresholded Jaccard Path Similarity Matrix:

	movie	movie/title	movie/director	movie/director/firstName	movie/director/lastName	movie/genre	movie/releaseYear	movie/actors
mediatedSchema	0	0	0	0	0	0	0	0
mediatedSchema/movie	0	0	0	0	0	0	0	0
mediatedSchema/movie/title	0	0	0	0	0	0	0	0
mediatedSchema/movie/book	0	0	0	0	0	0	0	0
mediatedSchema/movie/book/title	0	0	0	0	0	0	0	0
mediatedSchema/movie/book/DOI	0	0	0	0	0	0	0	0
mediatedSchema/movie/book/publicationHouse	0	0	0	0	0	0	0	0
mediatedSchema/movie/book/publicationHouse/location	0	0	0	0	0	0	0	0
mediatedSchema/movie/book/publicationHouse/location/address	0	0	0	0	0	0	0	0
mediatedSchema/movie/book/publicationHouse/location/city	0	0	0	0	0	0	0	0
mediatedSchema/movie/book/publicationHouse/location/country	0	0	0	0	0	0	0	0

11 rows × 24 columns

Figure 24: Snapshot of Path Scores for Jaccard Similarity

- **Jaccard Node Similarity:** Threshold = **0.6**



Jaccard Node Similarity Evaluation (threshold = 0.6)
Precision: 1.000
Recall: 1.000
F1 Score: 1.000

Thresholded Jaccard Node Similarity Matrix:

	movie	title	director	firstName	lastName	genre	releaseYear	actors	actor	actorName	...	critic	score	book	DOI	ShootLocation	address	city	country	latitude	longit
mediatedSchema	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	
movie	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	
title	0	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	
book	0	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	
DOI	0	0	0	0	0	0	0	0	0	0	...	0	0	0	1	0	0	0	0	0	
publicationHouse	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	
location	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	
address	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	1	0	0	0	
city	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	1	0	0	
country	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1	0	

10 rows × 23 columns

Figure 25: Snapshot of Node Scores for Jaccard Similarity

- **TF-IDF Path Similarity:** Threshold = **0.75**

```
🔎 TF-IDF Path Similarity Evaluation (threshold = 0.75)
Precision: 1.000
Recall:    0.013
F1 Score:  0.025

Thresholded TF-IDF Path Similarity Matrix:
```

	movie	movie/title	movie/director	movie/director/firstName	movie/director/lastName	movie/genre
mediatedSchema	0	0	0	0	0	0
mediatedSchema/movie	0	0	0	0	0	0
mediatedSchema/movie/title	0	1	0	0	0	0
mediatedSchema/movie/book	0	0	0	0	0	0
mediatedSchema/movie/book/title	0	0	0	0	0	0
mediatedSchema/movie/book/DOI	0	0	0	0	0	0
mediatedSchema/movie/book/publicationHouse	0	0	0	0	0	0
mediatedSchema/movie/book/publicationHouse/location	0	0	0	0	0	0
mediatedSchema/movie/book/publicationHouse/location/address	0	0	0	0	0	0
mediatedSchema/movie/book/publicationHouse/location/city	0	0	0	0	0	0
mediatedSchema/movie/book/publicationHouse/location/country	0	0	0	0	0	0

Figure 26: Snapshot of Path Scores for TF-IDF Matcher

- **TF-IDF Node Similarity:** Threshold = **0.6**

```
🔎 TF-IDF Node Similarity Evaluation (threshold = 0.6)
Precision: 1.000
Recall:    1.000
F1 Score:  1.000

Thresholded TF-IDF Node Similarity Matrix:
```

	movie	title	director	firstName	lastName	genre	releaseYear	actors	actor	actorName	...	critic	score	book	DOI	ShootLocation	addres
mediatedSchema	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
movie	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
title	0	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
book	0	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	
DOI	0	0	0	0	0	0	0	0	0	0	...	0	0	0	1	0	
publicationHouse	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
location	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
address	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
city	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
country	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	

Figure 27: Snapshot of Node Scores for TF-IDF Matcher

- **Fuzzy Path Similarity:** Threshold = **0.75**

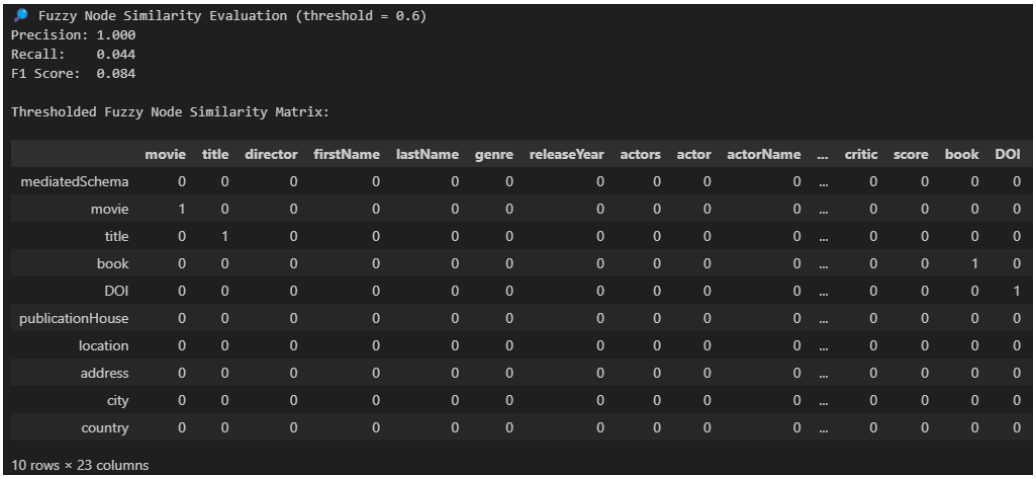
```
🔎 Fuzzy Path Similarity Evaluation (threshold = 0.75)
Precision: 0.000
Recall:    0.000
F1 Score:  0.000

Thresholded Fuzzy Path Similarity Matrix:
c:\Users\HP\AppData\Local\Programs\Python\Python310\lib\site-packages\sklearn\metrics\_classification.py:1706: UndefinedMetricWarning: Pre
  _warn_prf(average, modifier, f"{metric.capitalize()} is", result.shape[0])
```

	movie	movie/title	movie/director	movie/director/firstName	movie/director/lastName
mediatedSchema	0	0	0	0	0
mediatedSchema/movie	0	0	0	0	0
mediatedSchema/movie/title	0	0	0	0	0
mediatedSchema/movie/book	0	0	0	0	0
mediatedSchema/movie/book/title	0	0	0	0	0
mediatedSchema/movie/book/DOI	0	0	0	0	0
mediatedSchema/movie/book/publicationHouse	0	0	0	0	0
mediatedSchema/movie/book/publicationHouse/location	0	0	0	0	0
mediatedSchema/movie/book/publicationHouse/location/address	0	0	0	0	0
mediatedSchema/movie/book/publicationHouse/location/city	0	0	0	0	0
mediatedSchema/movie/book/publicationHouse/location/country	0	0	0	0	0

11 rows × 24 columns

Figure 28: Snapshot of Path Scores for Fuzzy Similarity

- **Fuzzy Node Similarity:** Threshold = **0.6**



🔑 Fuzzy Node Similarity Evaluation (threshold = 0.6)
Precision: 1.000
Recall: 0.044
F1 Score: 0.084

Thresholded Fuzzy Node Similarity Matrix:

	movie	title	director	firstName	lastName	genre	releaseYear	actors	actor	actorName	...	critic	score	book	DOI
mediatedSchema	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
movie	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0
title	0	1	0	0	0	0	0	0	0	0	...	0	0	0	0
book	0	0	0	0	0	0	0	0	0	0	...	0	0	1	0
DOI	0	0	0	0	0	0	0	0	0	0	...	0	0	0	1
publicationHouse	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
location	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
address	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
city	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0
country	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0

10 rows × 23 columns

Figure 29: Snapshot of Node Scores for fuzzy Similarity

3 Contributions

Snigdha Raghavan Pradhipa

- Schema Tree Extraction
- Path & Node Similarity Computation (Levenshtein, Jaccard, TF-IDF, Fuzzy Matching)
- Similarity Matrix Generation
- Threshold-Based Binary Matching
- Precision, Recall, F1 Evaluation and Comparison

Vaishnavi Narasimhaiah Sathish

- Schema Tree Extraction
- Path & Node Similarity Computation (Levenshtein, Jaccard, TF-IDF, Fuzzy Matching)
- Similarity Matrix Generation
- Threshold-Based Binary Matching
- Precision, Recall, F1 Evaluation and Comparison

Samrudhi Ramesh Rao

- Schema Tree Extraction

- Path & Node Similarity Computation (Levenshtein, Jaccard, TF-IDF, Fuzzy Matching)
- Similarity Matrix Generation
- Threshold-Based Binary Matching
- Precision, Recall, F1 Evaluation and Comparison

Kongarasan Sathiya Moorthy

- Schema Tree Extraction
- Path & Node Similarity Computation (Levenshtein, Jaccard, TF-IDF, Fuzzy Matching)
- Similarity Matrix Generation
- Threshold-Based Binary Matching
- Precision, Recall, F1 Evaluation and Comparison