

BLESS THIS MESS

**FUNDAMENTALS
OF DATA SCIENCE**

**ROSAMARIA
GRAZIOSI**

**GRUPPO COPIA DI
COPIA DI UNTITLED14**

**IRENE
DI TIMOTEO**

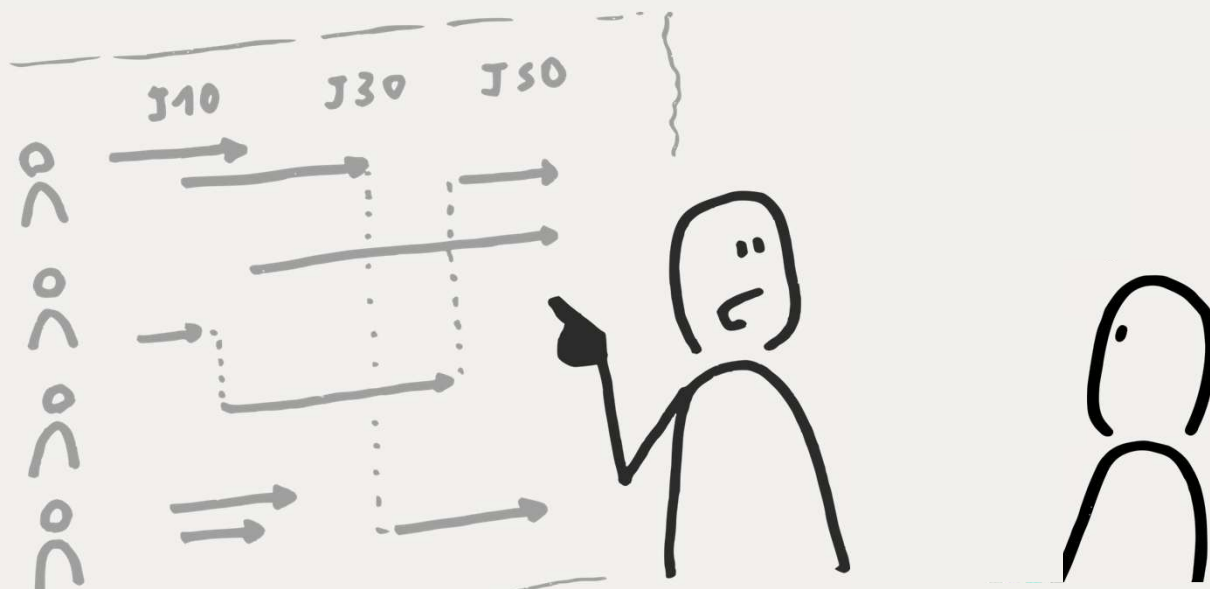
**DATASCIENCE A.Y.
2025/2026**

**JACOPO
CALVANO**

THE PROBLEM

	Copia di Copia di non modificare di Copia di Copia di Copia di Copia di Copia di ...	 me
	Copia di Copia di Progetto_pokemon_finale.ipynb	 me
	Copia di Progetto_pokemon_finale.ipynb	 me
	Copia di Copia di Copia di Progetto_pokemon_finale.ipynb	 me
	Copia di Progetto_pokemon_finale.ipynb 	 me
	Copia di non modificare di Copia di Copia di Copia di Copia di Copia di Copia di ipy...	 me
	lavoro Copia di non modificare di Copia di Copia di Copia di Copia di Copia di Co...	 me
	Copia di Copia di non modificare di Copia di Copia di Copia di Copia di Copia di ...	 me
	Copia di Copia di non modificare di Copia di Copia di Copia di Copia di Copia di ...	 me
	Copia di non modificare di Copia di Copia di Copia di Copia di Copia di Copia... 	 me





FORMALIZING THE CHALLENGE

Given a folder of files

01

Identify relations between files and reorganize them into semantic subfolders

02

Assign relevant titles to files and folders

03

Obtain information from versions and different files

THE DATASET

50 papers from *ArXiv* of 7 subjects, stripped of the title and abstract + 5 versions of each paper

Start with 350 original files

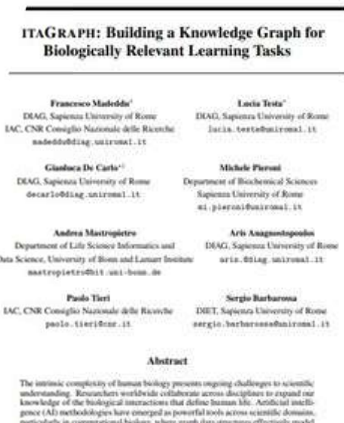
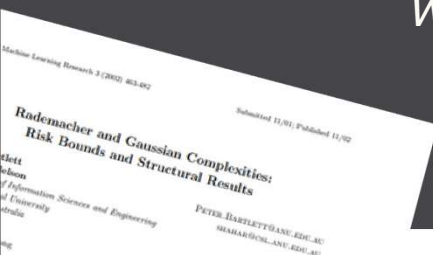


Extract the text, without title and abstract into json file with ids

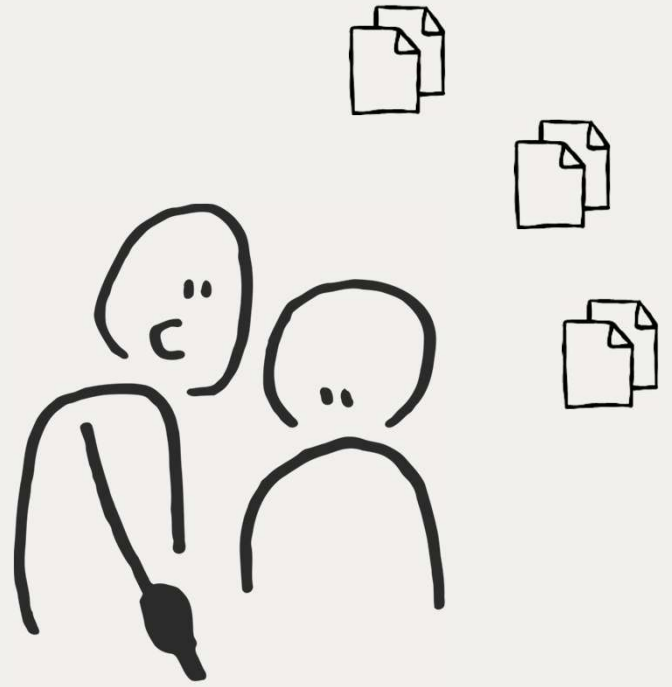
Create 5 versions by selecting random sections from the originals



End with a single json with 1750 texts and ids



THE PIPELINE



Documents
embedding



Pretrained
Siamese NN

Generation of file
and folder titles



Pretrained
LLM

Semantic
clustering



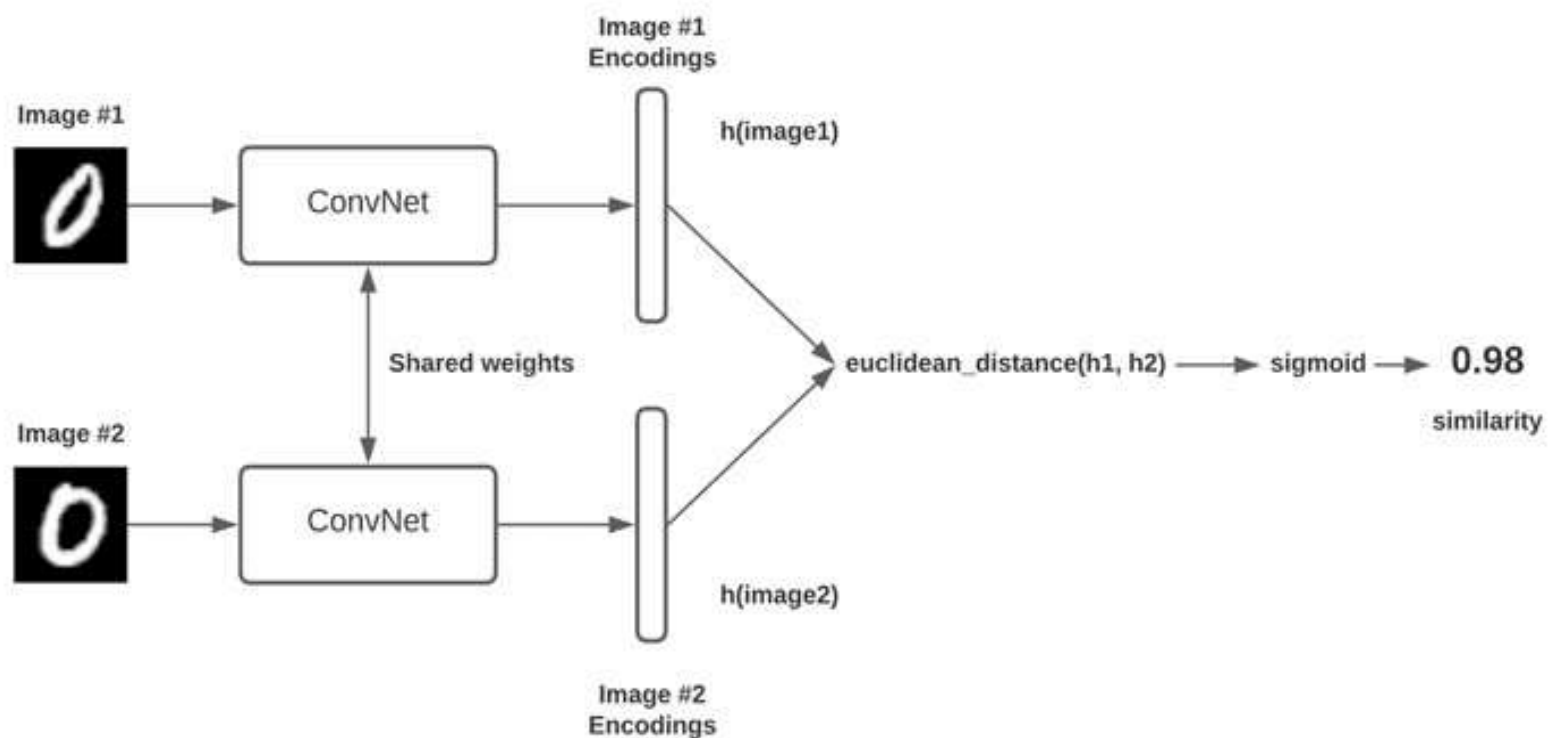
Leiden
algorithm

Information
retrieval

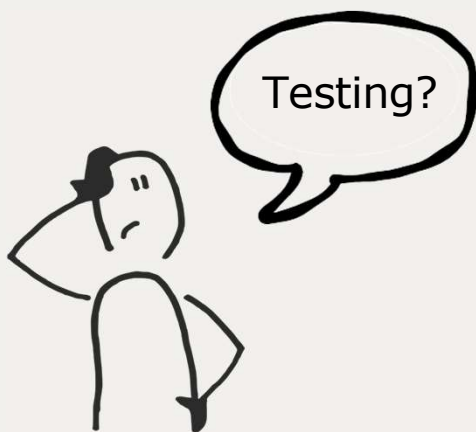


Cosine
similarity

DOCUMENTS EMBEDDING

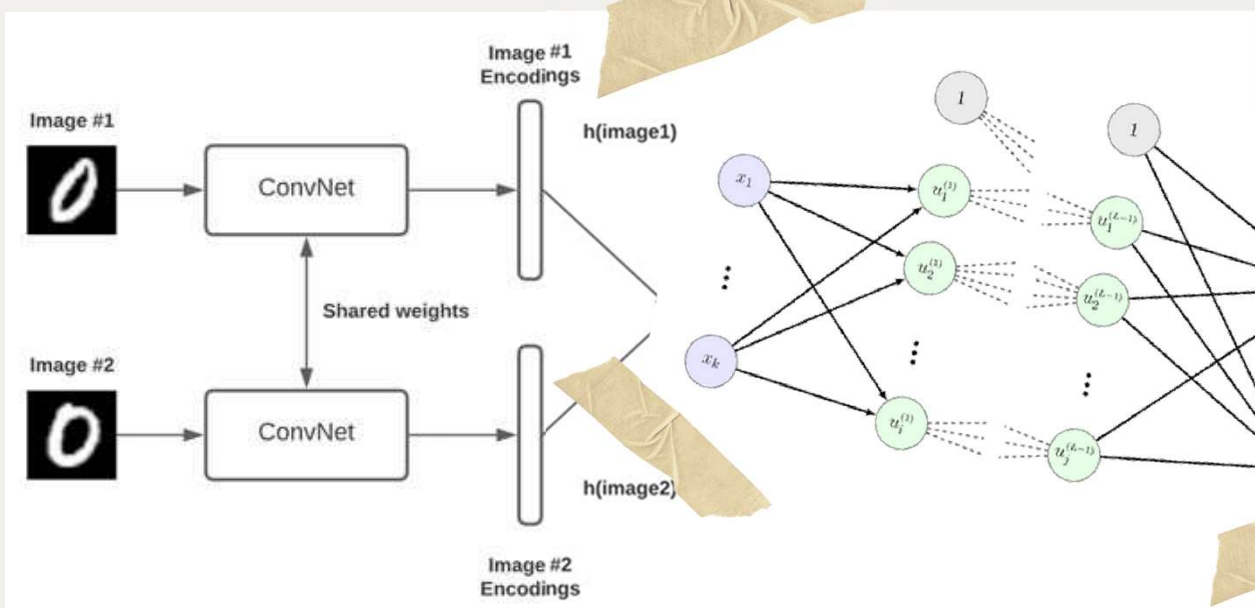


Picture from: <https://pyimagesearch.com/2020/11/30/siamese-networks-with-keras-tensorflow-and-deep-learning/>



Use Siamese NN's embeddings to encode the relations between pairs of files

DOCUMENTS EMBEDDING



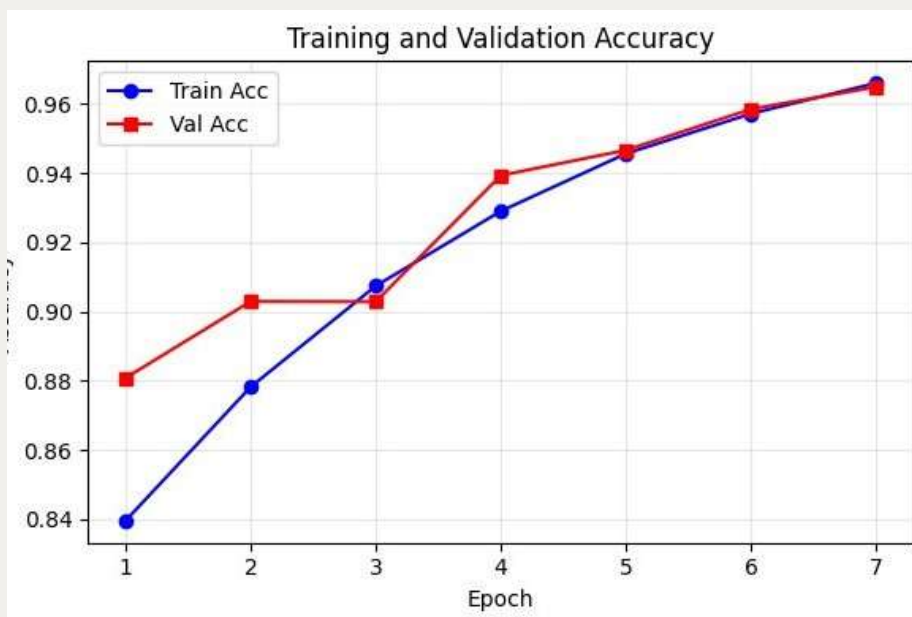
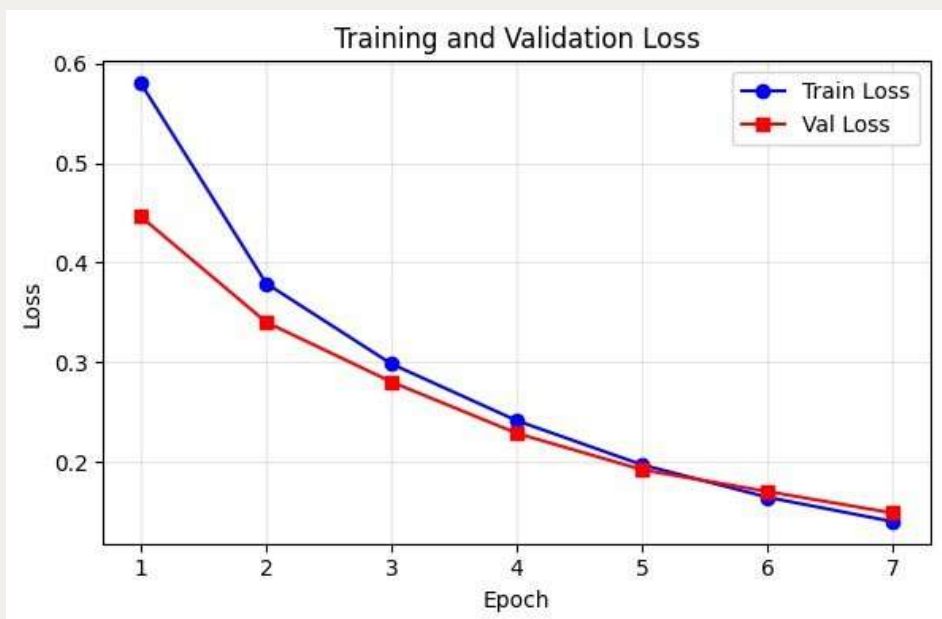
Picture from: <https://pyimagesearch.com/2020/11/30/siamese-networks-with-keras-tensorflow-and-deep-learning/>



Siamese NN + MLP classifier:

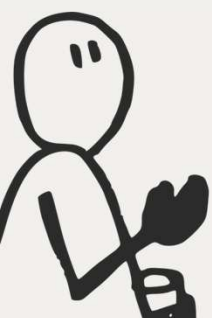
- version
- unrelated
- similar

DOCUMENTS EMBEDDING



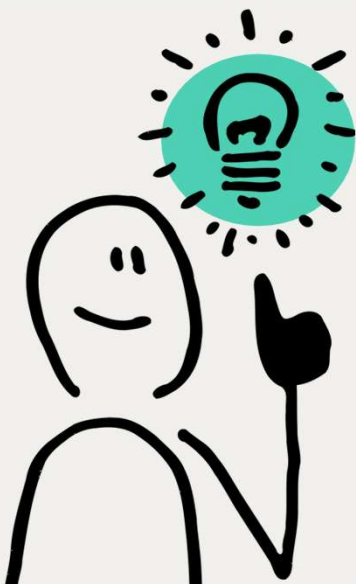
Using
the 3
classes
as a
proxy
we can
evaluate
the
siamese
embeddings

DOCUMENTS EMBEDDING



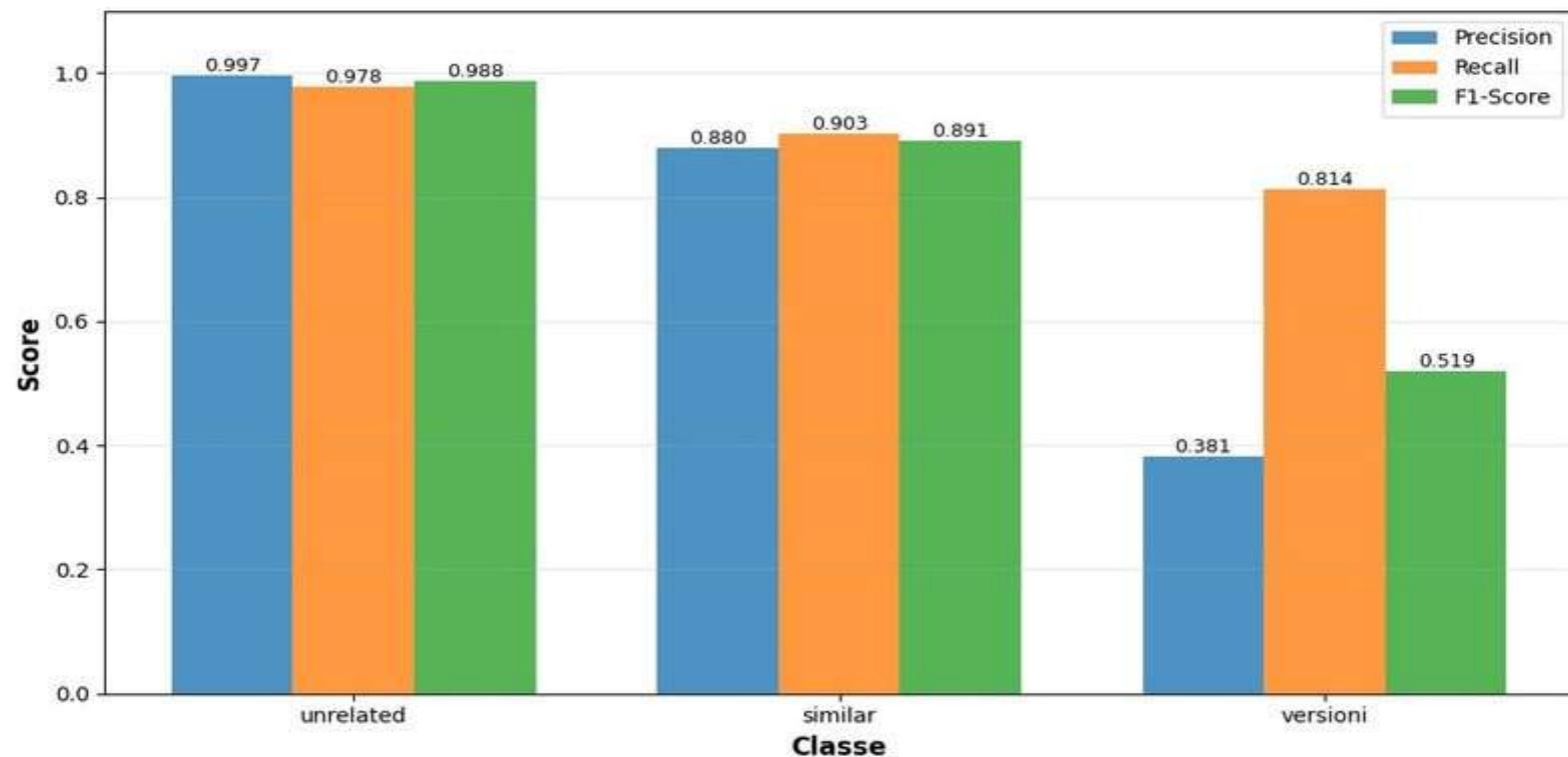
The classes
are unbalanced!

DOCUMENTS EMBEDDING



Precision
Recall
F1-Score

Performance per Classe

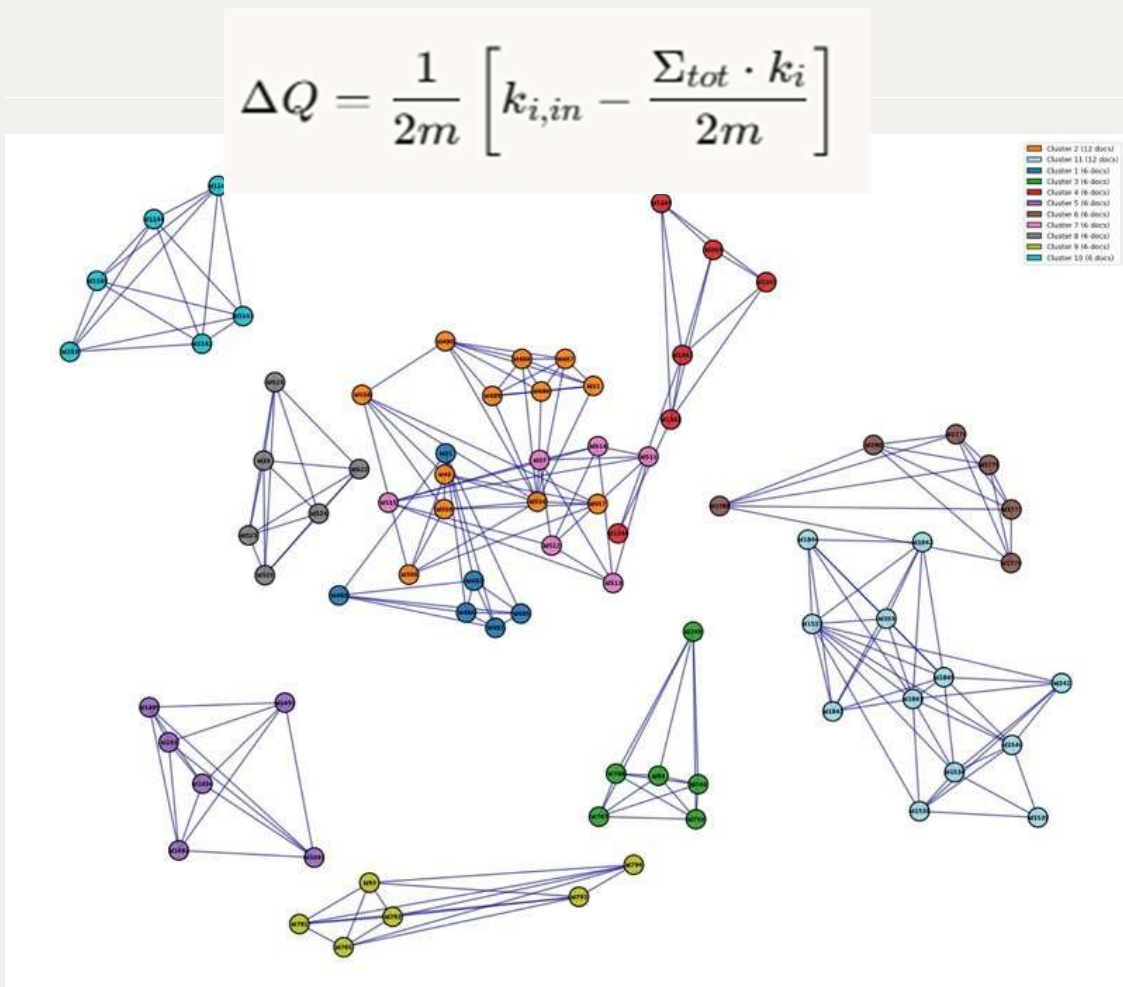


SEMANTIC CLUSTERING

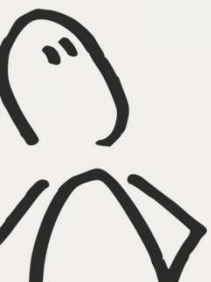
Louvain algorithm

An Heuristic to detect communities in large networks by maxizing modularity. It optimizes modularity greedily and locally using this gain measure

$$\Delta Q = \frac{1}{2m} \left[k_{i,in} - \frac{\Sigma_{tot} \cdot k_i}{2m} \right]$$



Example on
78 documents
11 clusters



SEMANTIC CLUSTERING

● Silhouette (cosine)

Mean: 0.6056 ± 0.0787

Range: [0.4389, 0.7961]

● Davies–Bouldin

Mean: 1.0423 ± 0.1682

Range: [0.6468, 1.4853]

● Conductance

Mean: 0.2075 ± 0.780

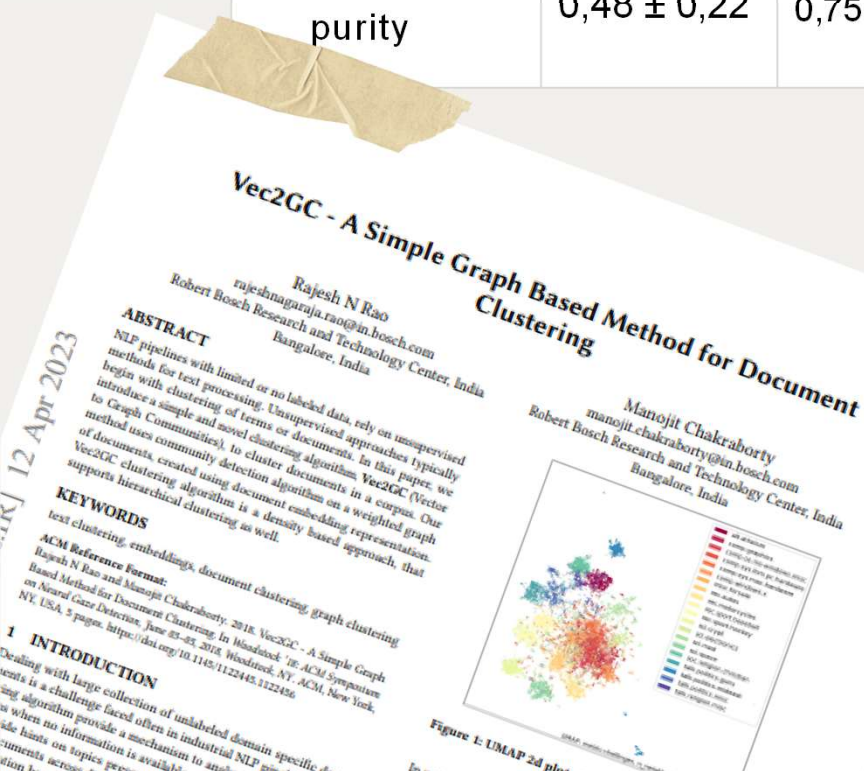
Range: [0.0000, 0.4327]



SEMANTIC CLUSTERING

purity

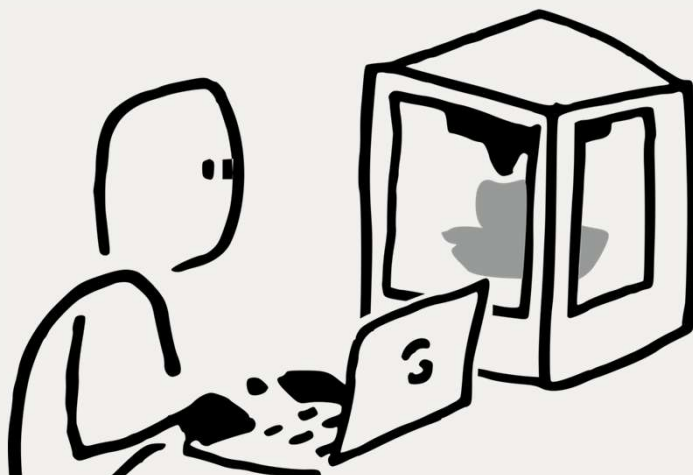
	Our Result #100 folder	Vec2GC DBPedia DS	K-Medoids DBPedia DS
≥ 50% purity	0,93 ± 0,11	0,94	0,80
≥ 70% purity	0,52 ± 0,23	0,88	0,54
≥ 90% purity	0,48 ± 0,22	0,75	0,32



GENERATING TITLES FOR FILES AND FOLDERS



```
messages = [  
    {"role": "system", "content": "Create  
concise titles. Output ONLY the title."},  
  
    {"role": "user", "content": f"Create a title  
(max 10 words):\n\n{snippet}\n\nTitle:"}  
]
```



INFORMATION RETRIEVAL

```
ID: id1
Titolo Generato: Microfluidic Platforms for Studying Angiogenesis in iPSC-ECs
Articoli Simili Trovati:
- Rank 1 (Score: 0.6360) | ID: id1 (COERENZA INTERNA)
- Rank 2 (Score: 0.6360) | ID: id331
- Rank 3 (Score: 0.6360) | ID: id335
- Rank 4 (Score: 0.5455) | ID: id332
- Rank 5 (Score: 0.5245) | ID: id333
- Rank 6 (Score: 0.5163) | ID: id334
```

```
Titolo: Modeling iPSC-derived Endothelial Cell Transition in Tumor Angiogenesis using Petri Nets
Articoli Simili Trovati (incluso se stesso):
- Rank 1 (Score: 0.7636) | ID: id332
- Rank 2 (Score: 0.7259) | ID: id334
- Rank 3 (Score: 0.6618) | ID: id331
- Rank 4 (Score: 0.6618) | ID: id335
- Rank 5 (Score: 0.6618) | ID: id1 (COERENZA INTERNA)
- Rank 6 (Score: 0.6486) | ID: id333
```

```
ID: id1
Titolo: Modeling iPSC-derived Endothelial Cell Transition in Tumor Angiogenesis using Petri Nets
Articoli Simili Trovati (Abstract vs. Testo Completo):
- Rank 1 (Score: 0.8218) | ID: id1 (COERENZA INTERNA)
- Rank 2 (Score: 0.4943) | ID: id17
- Rank 3 (Score: 0.4048) | ID: id24
- Rank 4 (Score: 0.3888) | ID: id212
- Rank 5 (Score: 0.3509) | ID: id19
- Rank 6 (Score: 0.3317) | ID: id172
```

Top 6 on 3 different query levels:

- generated title
- original title
- abstract



INFORMATION RETRIEVAL

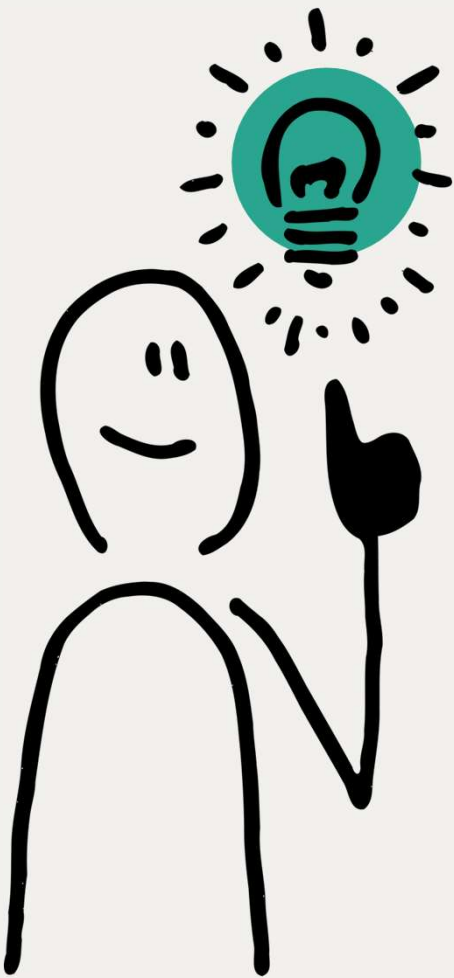
A metric to evaluate the generated titles

Query	Score	%
Gen. title	3.6672	61.1
Orig. title	3.7449	62.4
Abstract	2.3697	39.5

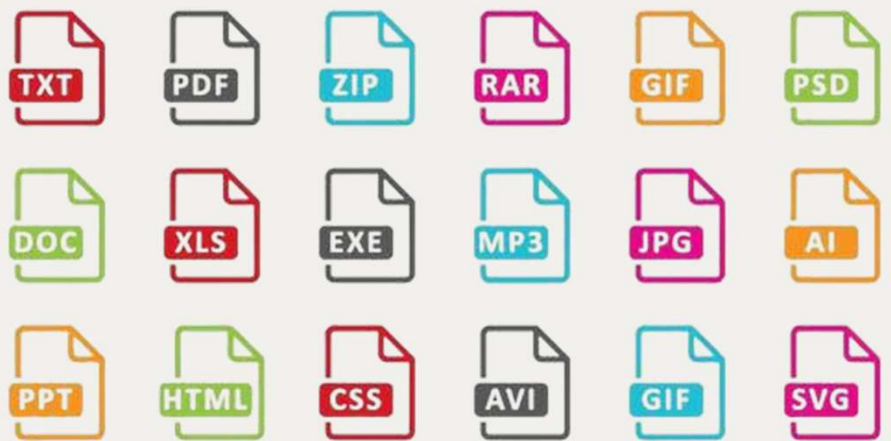
$$\text{score} = \sum_{d \in D} \mathbf{1}_{\text{correct}(d)} + 0.5 \sum_{d \in D} \mathbf{1}_{\text{similar}(d)}$$



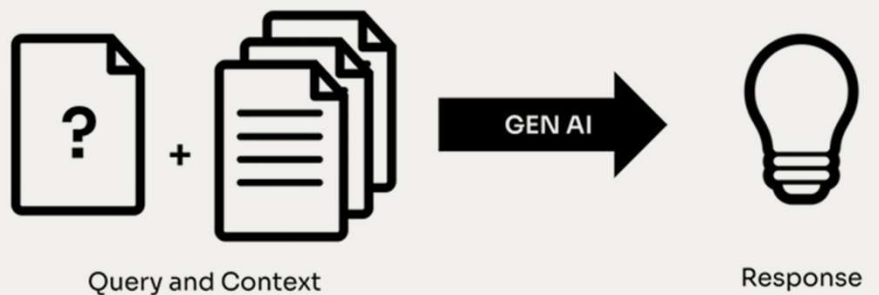
FUTURE



1. extend to other tipe of file



2. RAG



THANK YOU!

