POLITECNICO MILANO 1863

MODELLISTICA E CALCOLO SCIENTIFICO
MODELING AND SCIENTIFIC COMPUTING

Graph Neural Network pipeline for unsupervised clinical document analysis

Author: Gabriele Moro

Advisor: Prof. Francesca Ieva, Co-Advisor: Dott. Vittorio Torri

Academic year: 2023-2024

MSc: Mathematical Engineering – Statistical Learning

Motivations

Objective

Given a corpus of clinical documents in textual form, cluster the documents based on their medical similarities.

Application

There is a growing number of **medical data in textual form**. The clusters could help the medical staff to exploit those information, discovering group of patients with similar conditions.

Document clustering

Definition: **Document clustering** is the process of grouping together similar texts.

- Information about clusters enhances the **understanding of the problem**
- **Natural Language Processing (NLP)** task, converting non-mathematical representations such as text and words into a mathematical representation i.e. vectors of real numbers, also called **embeddings**
- Once document embeddings are produced, the clustering phase is performed by **classical clustering algorithms** (e.g. k-means, agglomerative clustering)

Document clustering for medical domain

Clustering medical documents brings an higher level of complexity:

- **Privacy** issues

  → Lack of publicly available corpora to train the models

- **Technical** and rich of **abbreviations** lexicon

  → Large vocabulary size

- Almost always complete **absence of document labeling**

  → Only unsupervised approaches are possible

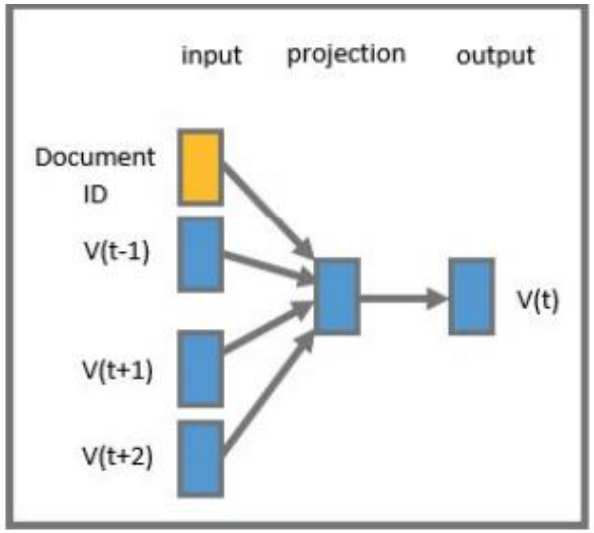
  → Complicated evaluation phase

Existing approaches for document clustering

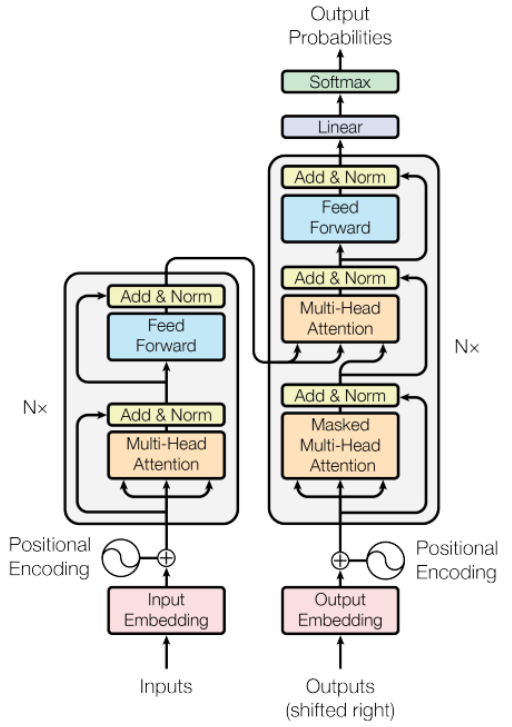
TF-IDF

$$\text{TF-IDF}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \times \frac{|D|}{|\{d \in D : t \in d\}|}.$$

Doc2Vec



Transformers based models



Graph Neural Networks

Definition: A **graph** is defined as the pair **G = (E,V)** where **E** is the set of edges and **V** is the set of vertices.

Graphs applications: Social networks, chemical molecules, and road maps.

Graph Neural Networks (GNNs) are a family of models that exploit the learning schema of the classical Neural Networks by using graphs as input data (e.g. *GAT, GIN, GCN*).

The learning goal can be either at **graph level, node level and edge level.**

The learning schema of GNNs follow a «message passing mechanism», where nodes learn an enhanced representation of themself by aggregating information about neighbouring nodes.

$$h_i^{(k+1)} = f\left(h_i^{(k)}, \left\{h_j^{(k)} : j \in \mathcal{N}(i)\right\}\right). \qquad \text{Generic GNN}$$

$$h_i^{(k+1)} = RELU\left(U^{(k)} Average_{j \in \mathcal{N}(i)} h_j^{(k)}\right). \qquad \text{GCN}$$

GNNs for NLP

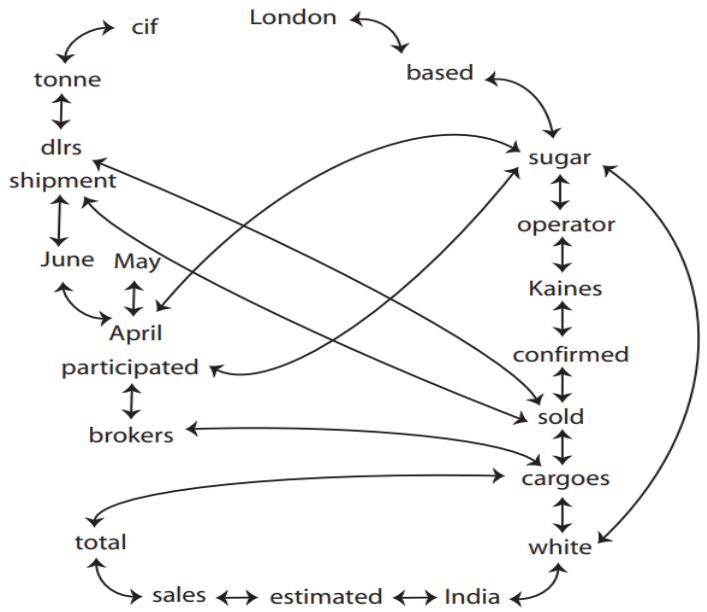
GNNs provided state of the art results over the similar task of **supervised text classification.**

1. **Homogeneous graphs:** One graph for each document, words are represented by nodes and the connections between words are represented by edges. (e.g. TextLevelGCN)

→ **Graph Classification**

2. **Heterogeneous graph:** One graph for the whole corpus, nodes are both words and documents, with different types of connections between the nodes. (e.g. TextGCN)

→ **Node Classification**



Short text example for homogeneous graph structure:
"London-based sugar operator Kaines Ltd confirmed it sold two cargoes of white sugar to India out of an estimated overall sales total of four or five cargoes in which other brokers participated. The sugar, for April/May and April/June shipment, was sold at between 214 and 218 dlrs a tonne cif, it said."

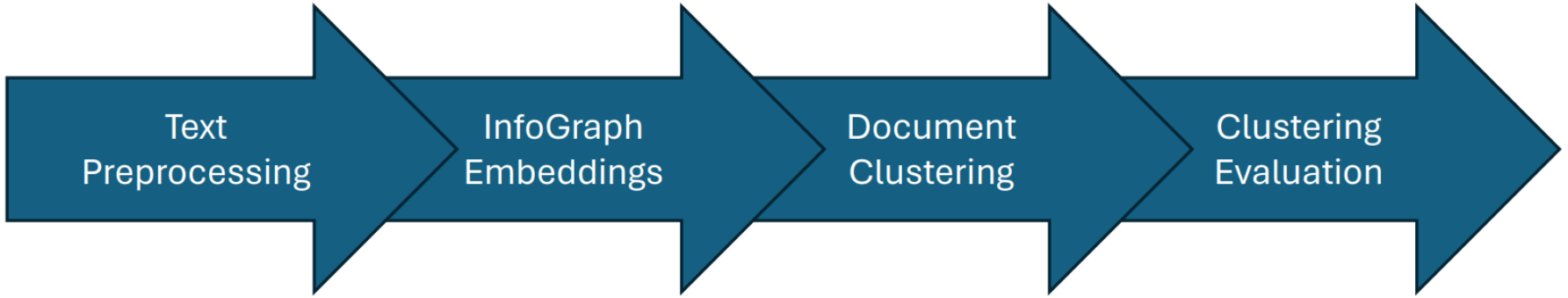
Pipeline for clustering with GNNs

Text preprocessing: Data cleaning, word embeddings, graph construction.

InfoGraph embeddings: Preprocessed texts in graph form are translated into vectors of real numbers.

Document clustering: The produced document embeddings are clustered.

Clustering evaluation: The results need to be evaluated, without ground-truth labels.

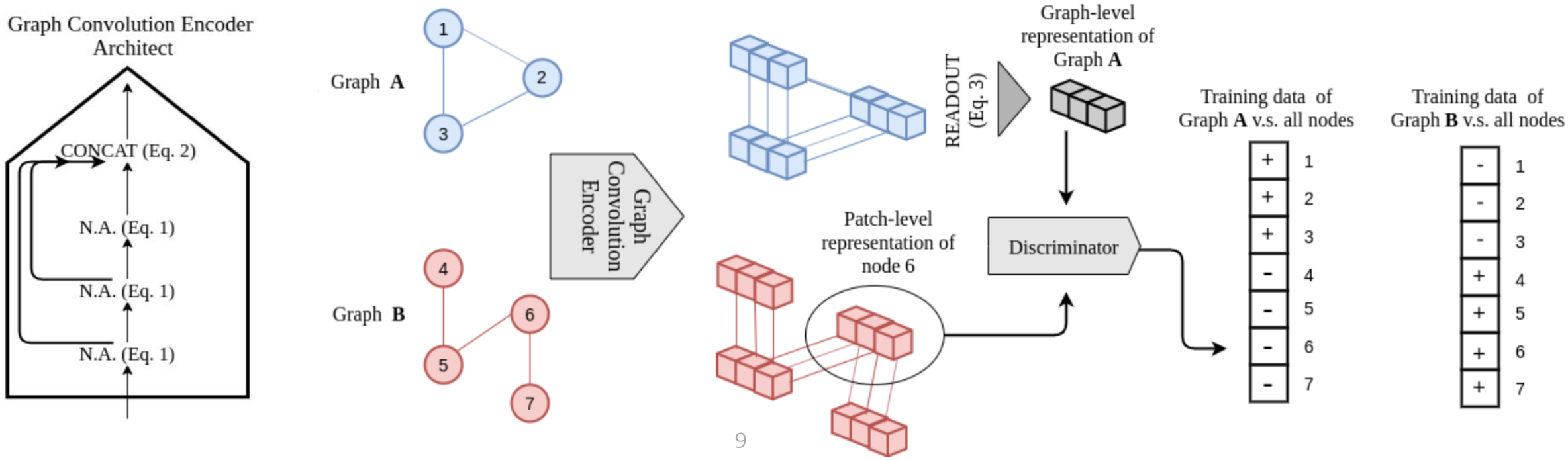


InfoGraph (Sun et al. 2019)

InfoGraph uses **GIN**, the most powerful GNN model for distinguishing different graphs.

A **discriminator** predicts the belonging of a node to a graph.

The parameters of GIN and of the discriminator are trained maximazing the **Mutual Information** between nodes and graphs representations.



Motivations on the choice of InfoGraph

Self-supervised and **unsupervised** model, no need of labels.

One of the few GNN models exploitable due to the labeling absence in clinical documents.

InfoGraph was originally tested on chemical molecules datasets (TUDatasets) on the downstream task of classification.

Once embeddings are produced, instead of training a classifier we deploy a classical clustering technique.

Dataset

The **European Clinical Case Corpus (E3C)** is a freely available multilingual corpus of clinical cases documents. Focus on the Italian subset of it.

«È descritto il caso clinico di una donna di 84 anni, forte fumatrice per circa 30 anni. Non familiarità per malattie renali. Principali rilievi anamnestici: ipertensione…»

Data organization

L1, L2 have CUI clinical annotations

in addition to the text.

	L1	L2	L3
Documents	86	174	10213
Tokens	24319	49900	13601915

Sources of the documents

- AIFA documents discarded
- Journal Publications
- University exam text

Source	Docs	Avg Length	Std Dev	Min	Max
Giornale di Clinica Nefrologica e Dialisi	10	4698.4	1017	2745	6579
Miur	1439	230.3	129.3	38	915
Clinical Management Issues	1	4463	0	4463	4463
Hematology Reports	6	1143.8	328.5	789	1664
Italian Journal of Medicine	534	1170.9	300.3	364	2850
Italian Journal of Emergency Medicine	4	2760.2	2161.8	420	5241
Sapienza Università di Roma	89	598.5	235.6	197	1287
Microbiologia Medica	17	1405.7	1129.1	535	5593
La Pediatria Medica e Chirurgica	6	3104.5	2288.5	428	5842
Pediatric Reports (PAGEPresse)	8	3465.3	1834.3	1334	6096
Pediatric Reports (MDPI)	6	3117.6	314.6	2658	3439
Prof. Dr. Francescopaolo Mattioli	9	6055.1	2817.8	4213	13387
Agenzia Italiana del Farmaco	8084	10373.4	4775.6	2074	58013

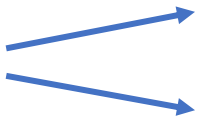
Evaluation Methods

Issues:

- Absence of ground-truth labels
- Impossible to apply Silhouette score or WSS for document clustering evaluation

Proposed evaluation methods:

- UMLS dictionary (L1 CUI annotations)



PyUMLSSimilarity

UMLS API

- Doc2Vec

- ChatGPT semi-automatic labeling for L1

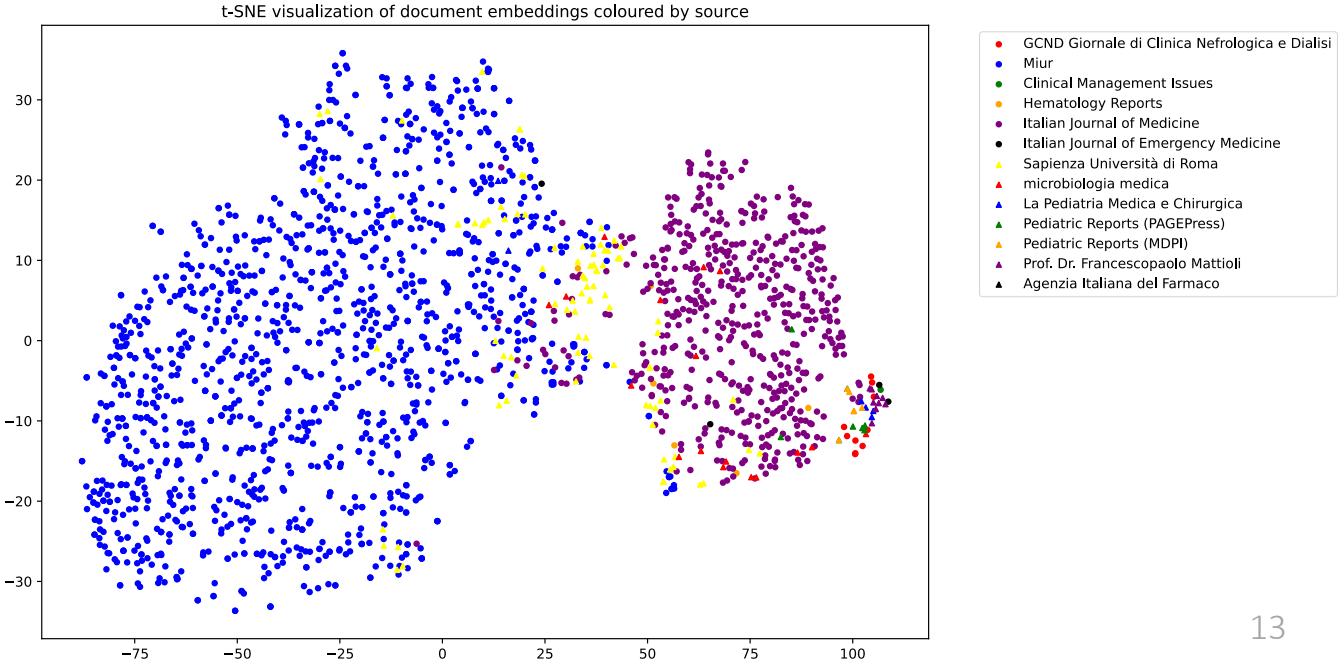
Results

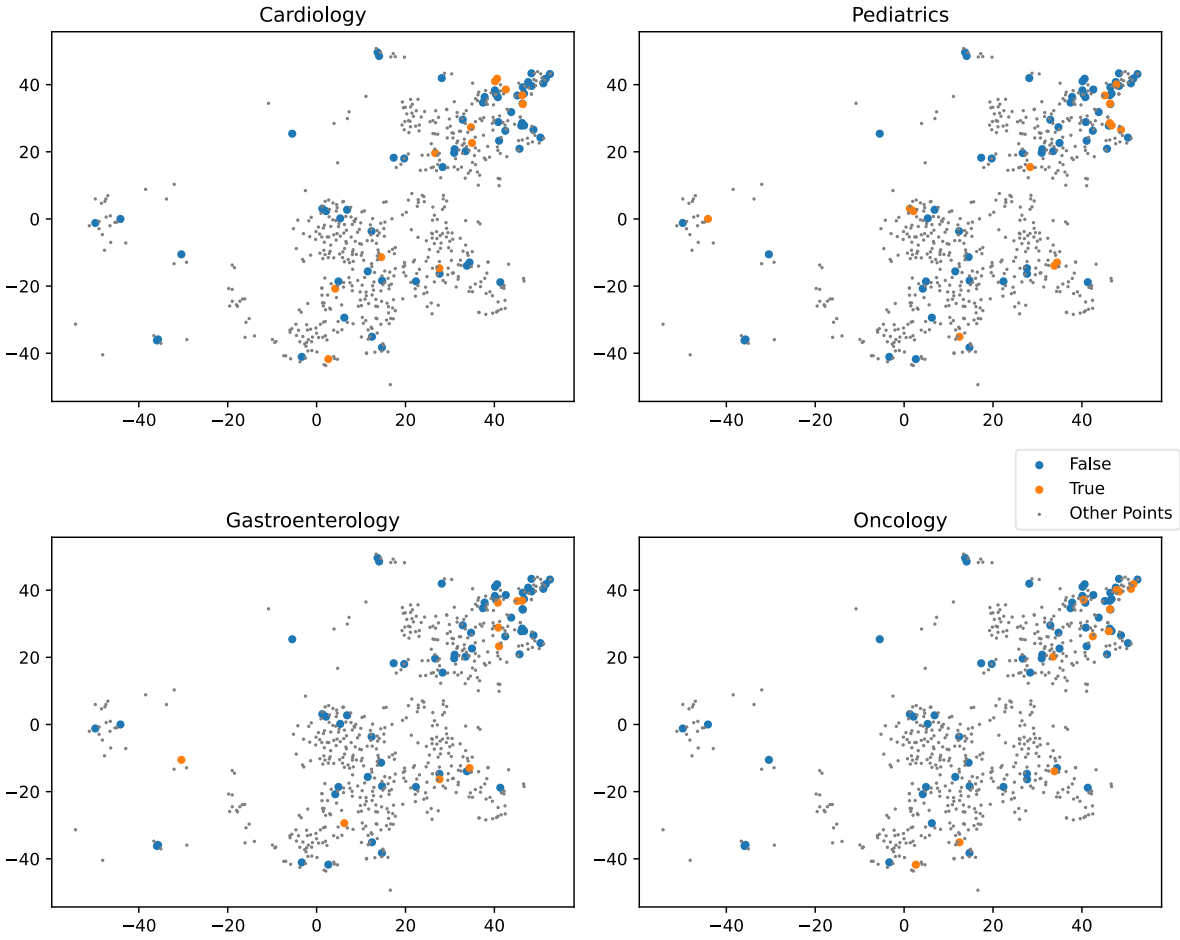
Shortcomings:

- UMLS → computational issues, poor quality of API
- Doc2Vec → no certainties of providing valuable results

Results:

- Embeddings captures information about the source
- The medical meaning is not captured

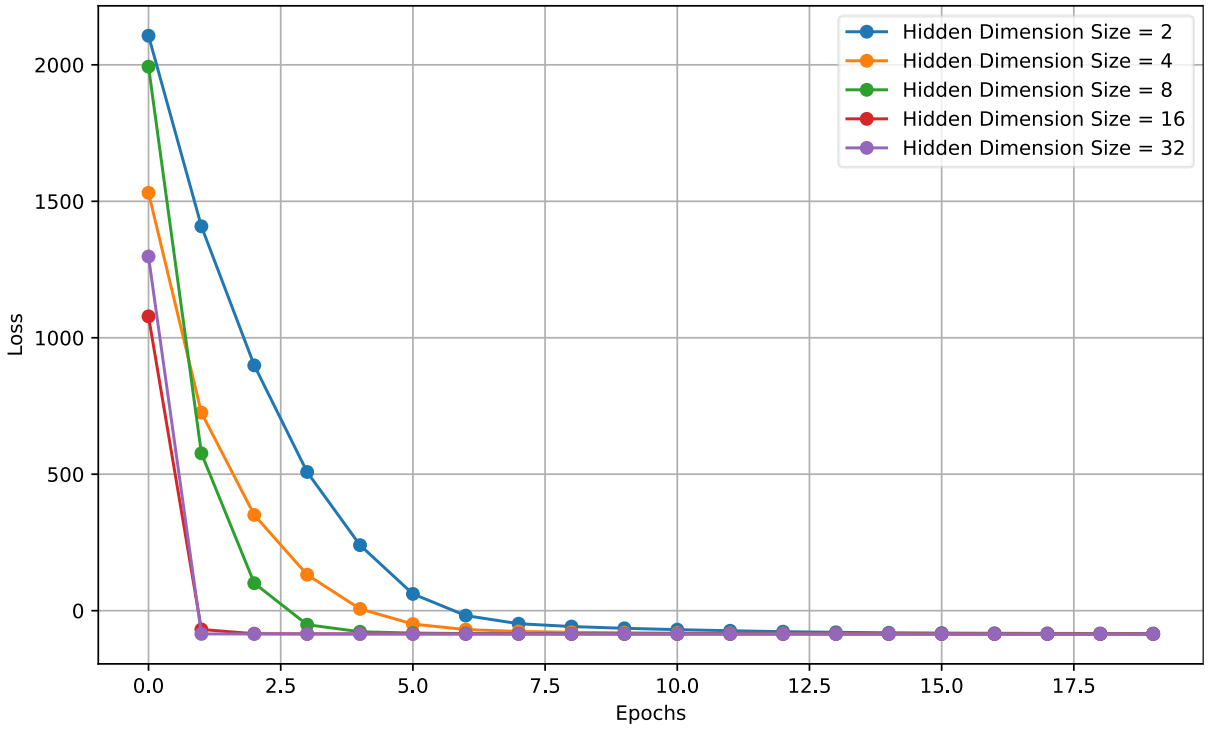




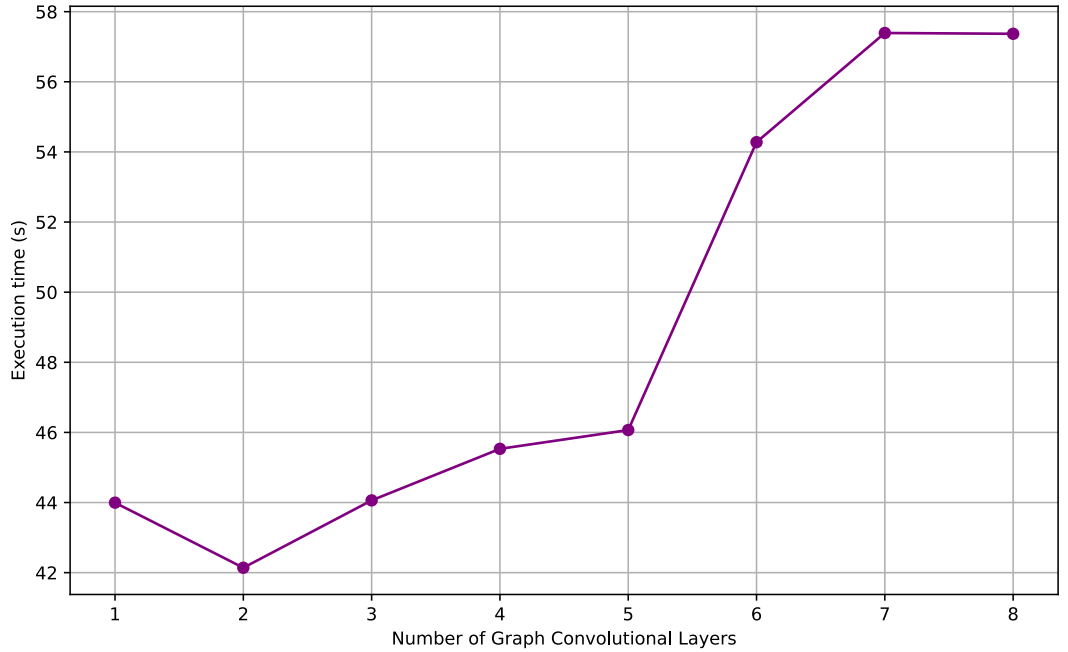
Computational analysis

No GPUs (11th Gen Intel(R) Core(TM) i7-11800H @ 2.30GHz, 32,0 GB ram)

Loss decrease



Training time



Conclusions and future developments

Conclusions

- First attempt of GNNs applied for document clustering
- Sources information captured, not medical ones
- Computationally more efficient than state of the art models

Future developments

- Change the evaluation dataset
- Improve the graph structure for representing documents
- Test other GNNs architectures

Main references

[1] S. Hassan, R. Mihalcea, and C. Banea. Random walk term weighting for improved text classification. *International Journal of Semantic Computing*, 1(04):421–439, 2007.

[2] L. Huang, D. Ma, S. Li, X. Zhang, and H. WANG. Text level graph neural network for text classification, 2019.

[3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[4] F.-Y. Sun, J. Hoffmann, V. Verma, and J. Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. *arXiv preprint arXiv:1908.01000*, 2019.

[5] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

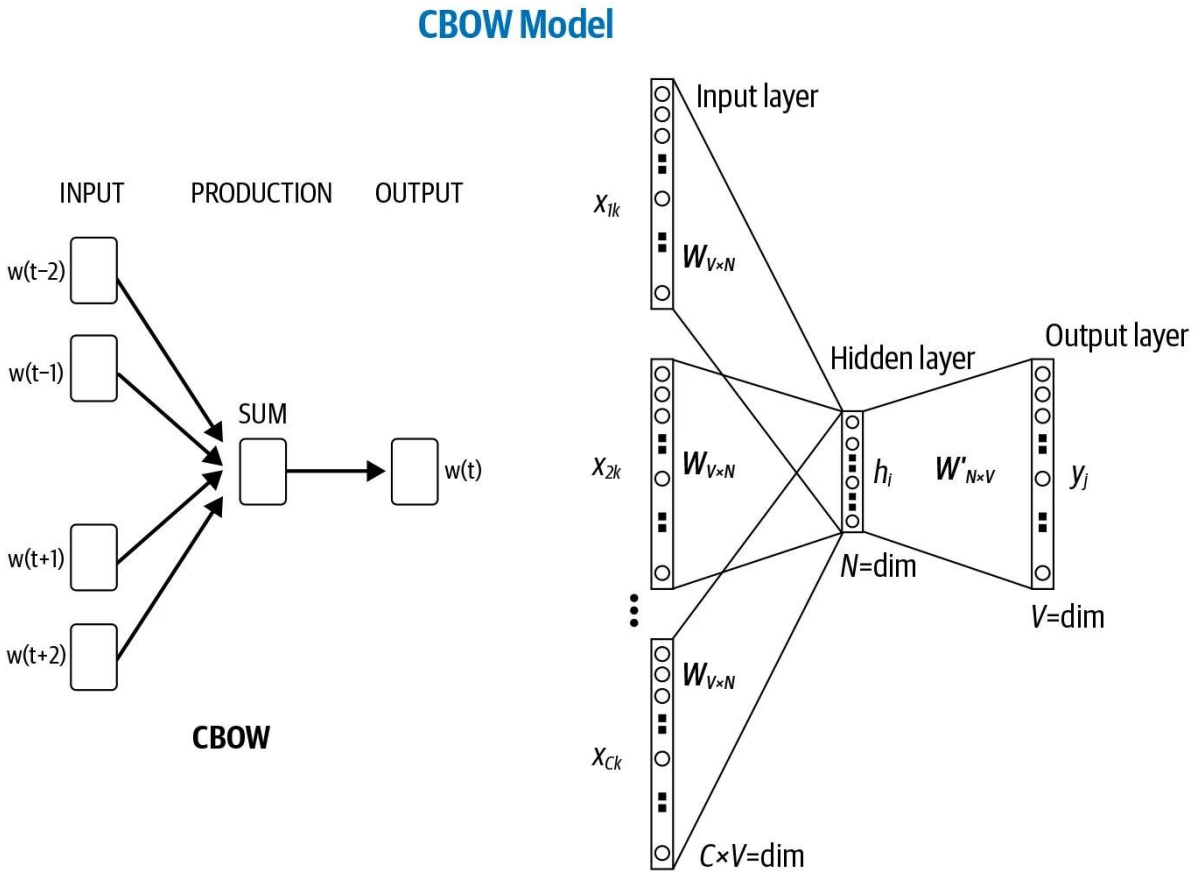
POLITECNICO
MILANO 1863

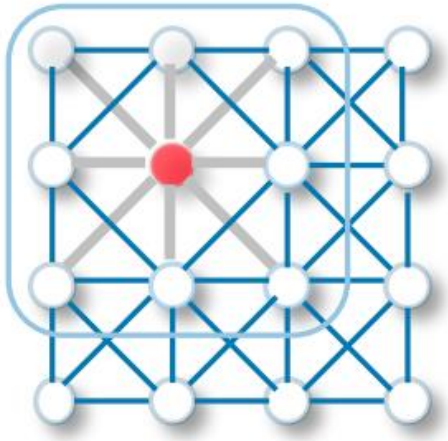
MODELLISTICA E CALCOLO SCIENTIFICO
MOX
MODELING AND SCIENTIFIC COMPUTING

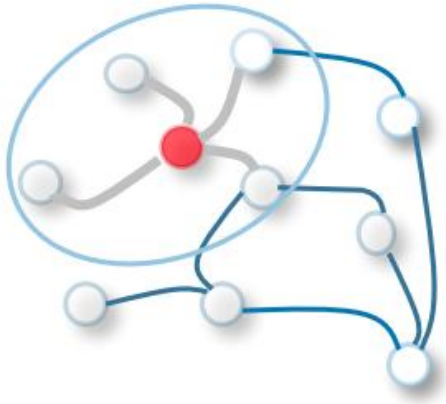
Thank you for your attention!

gabriele.moro@mail.polimi.it

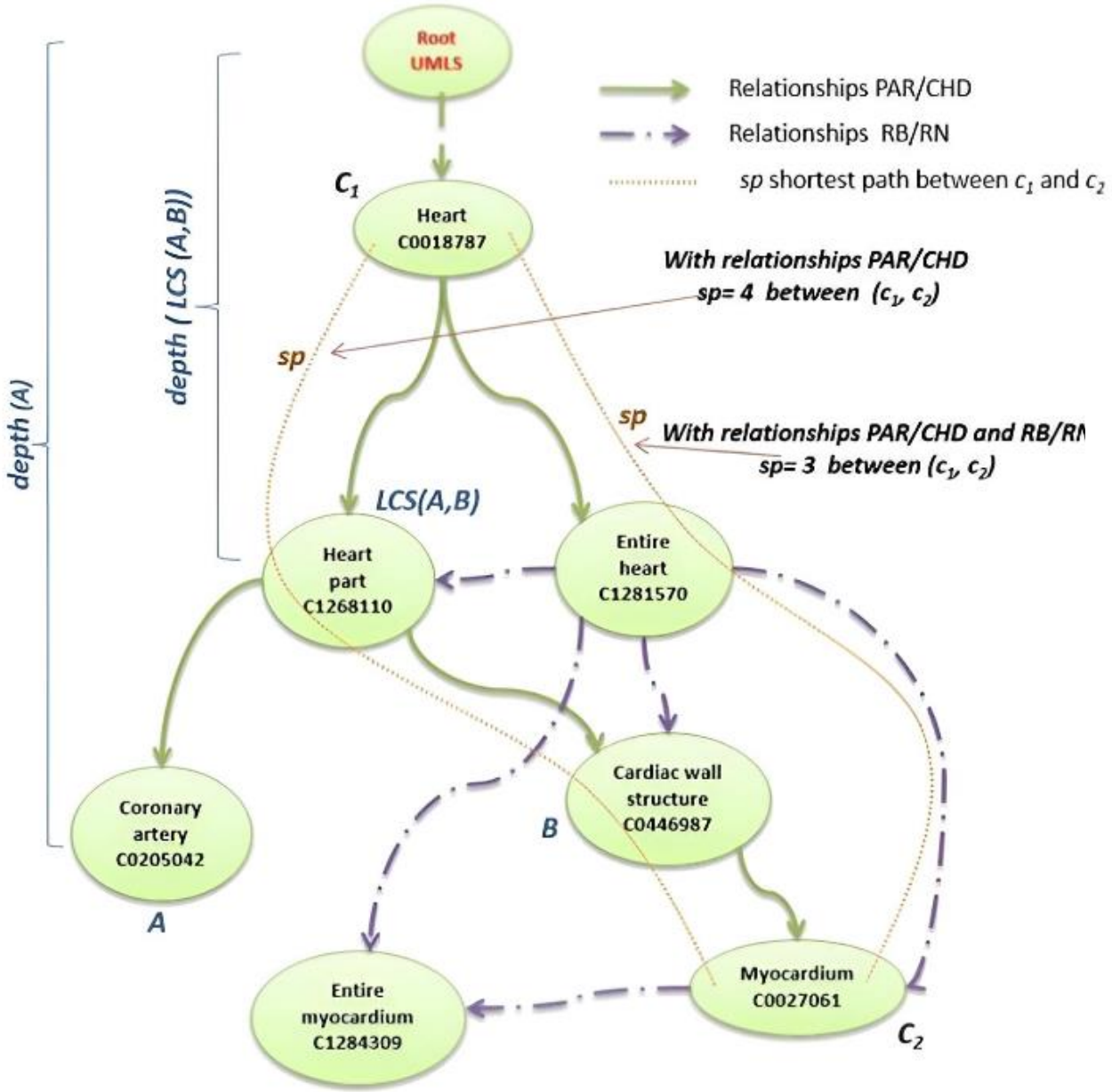
Additional slides

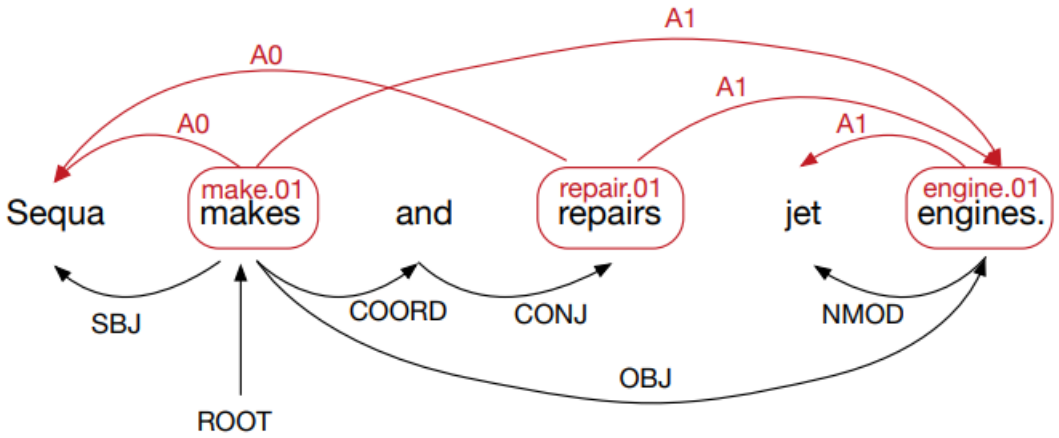




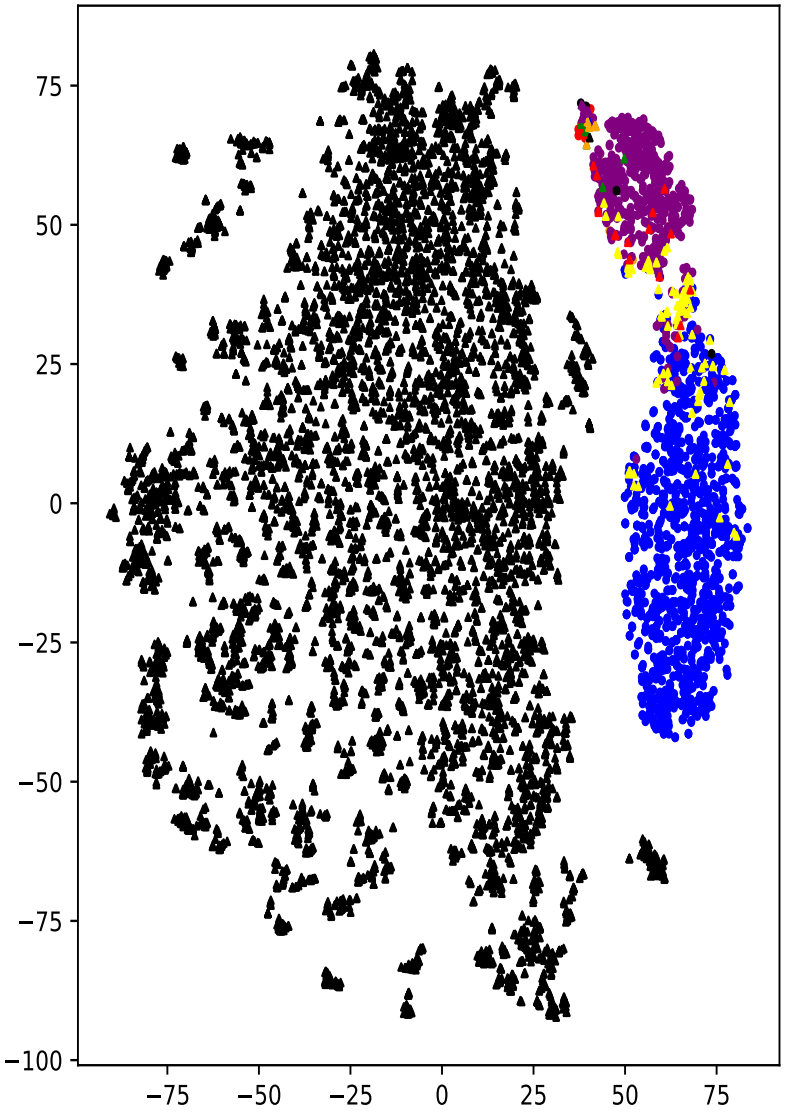


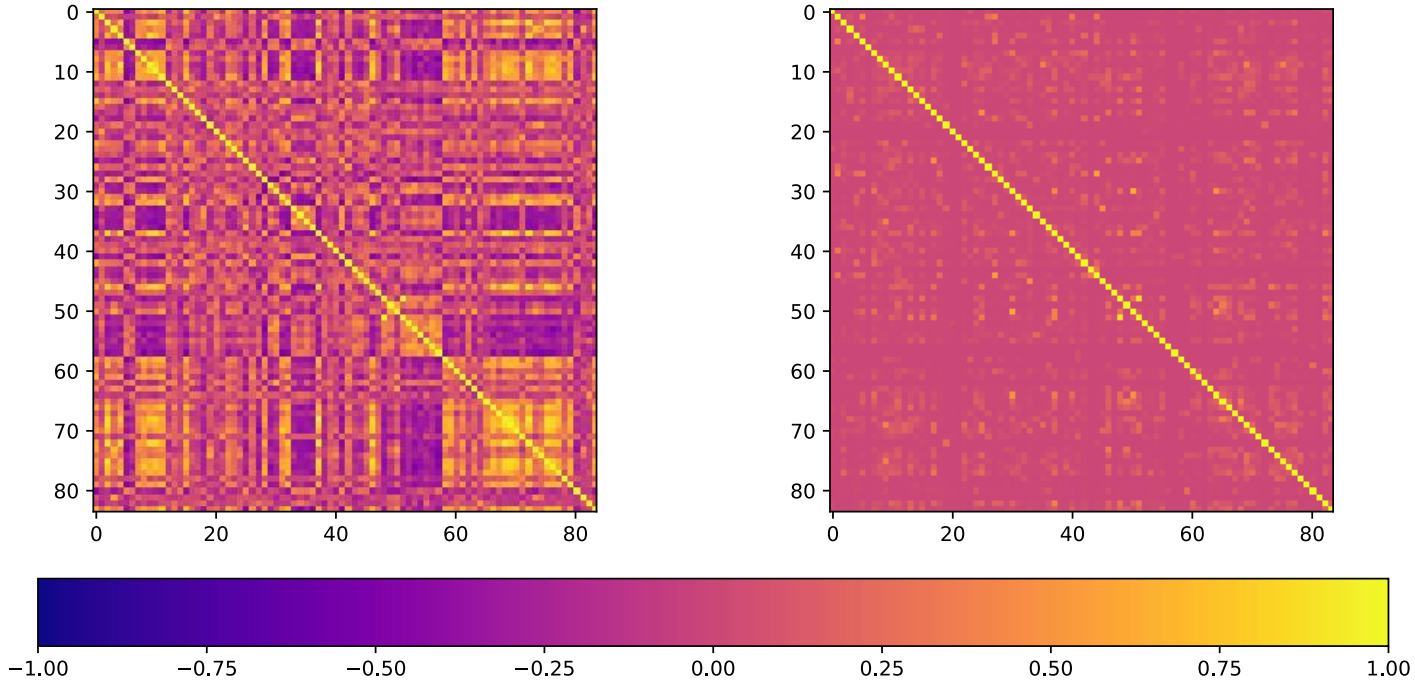
Additional slides



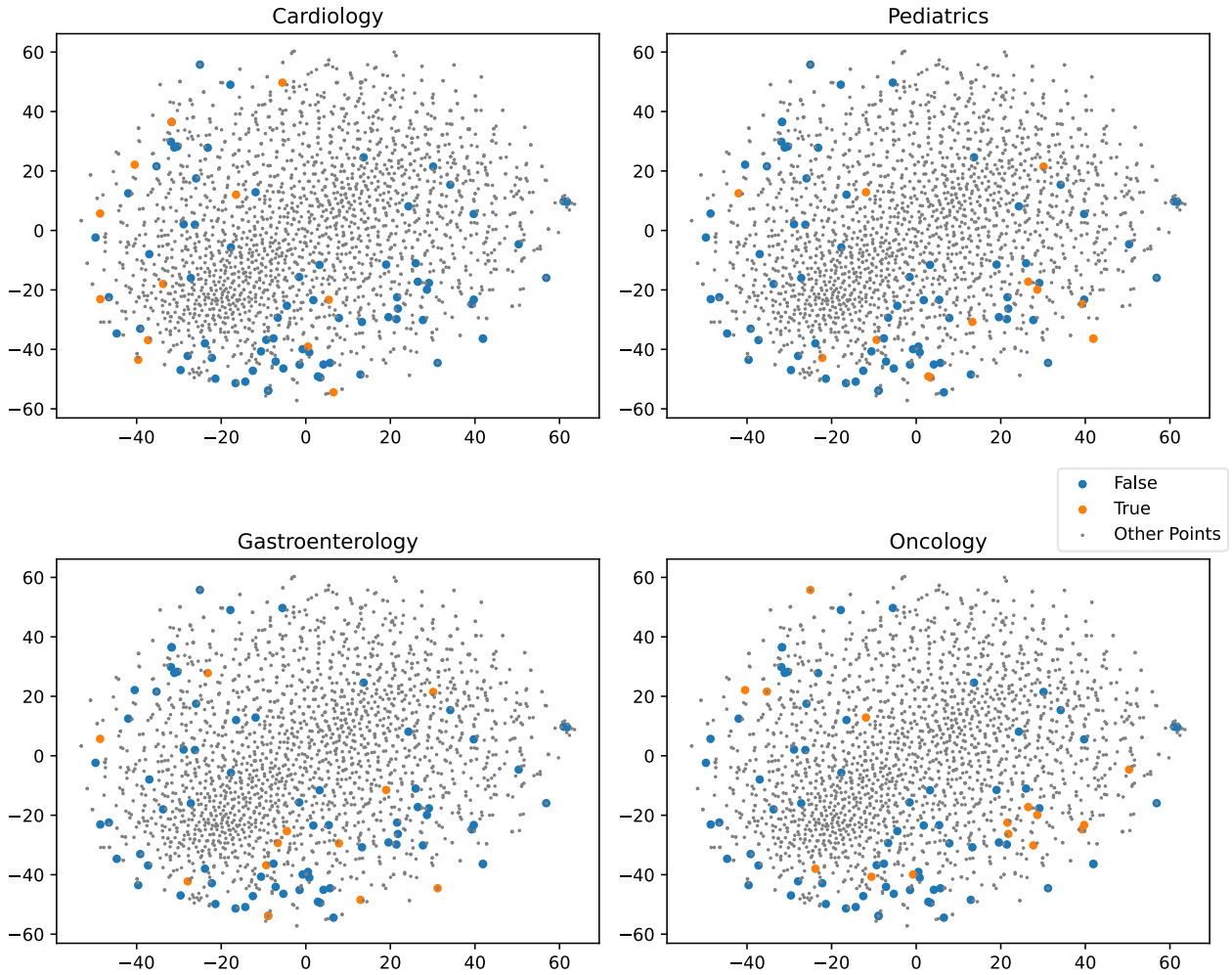


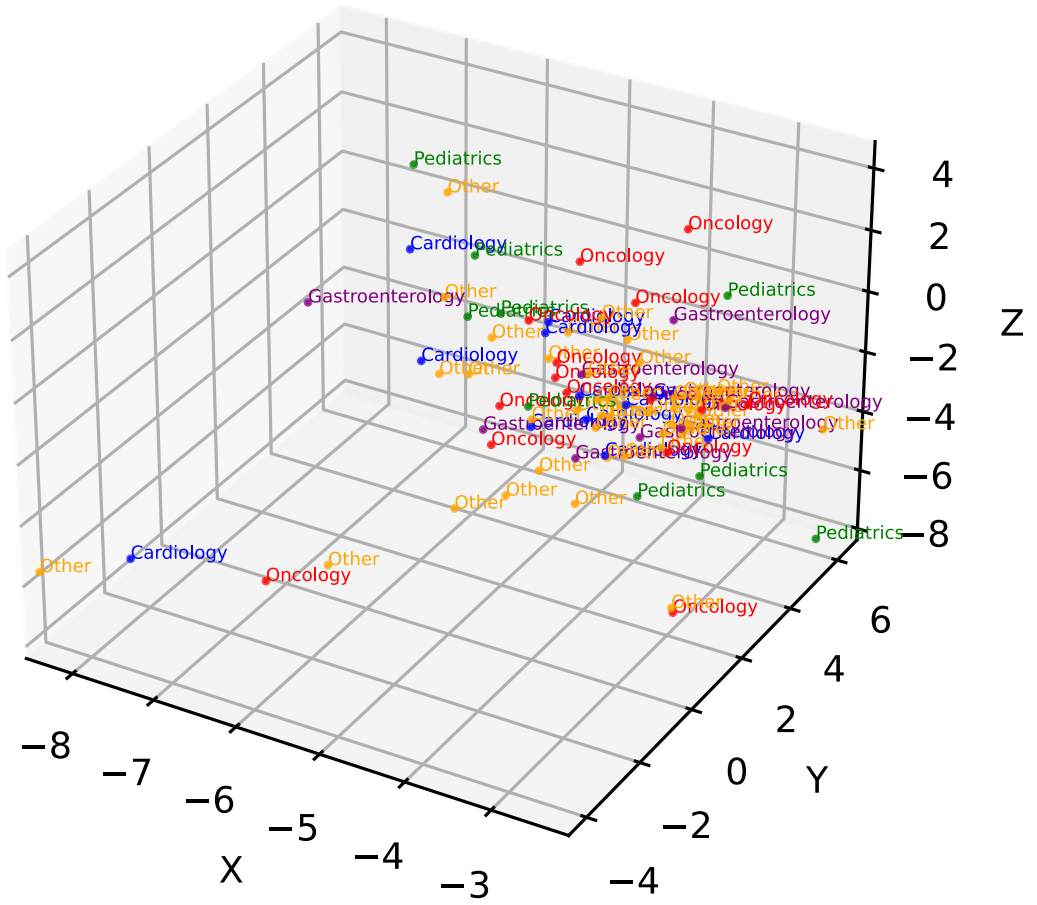
Additional slides



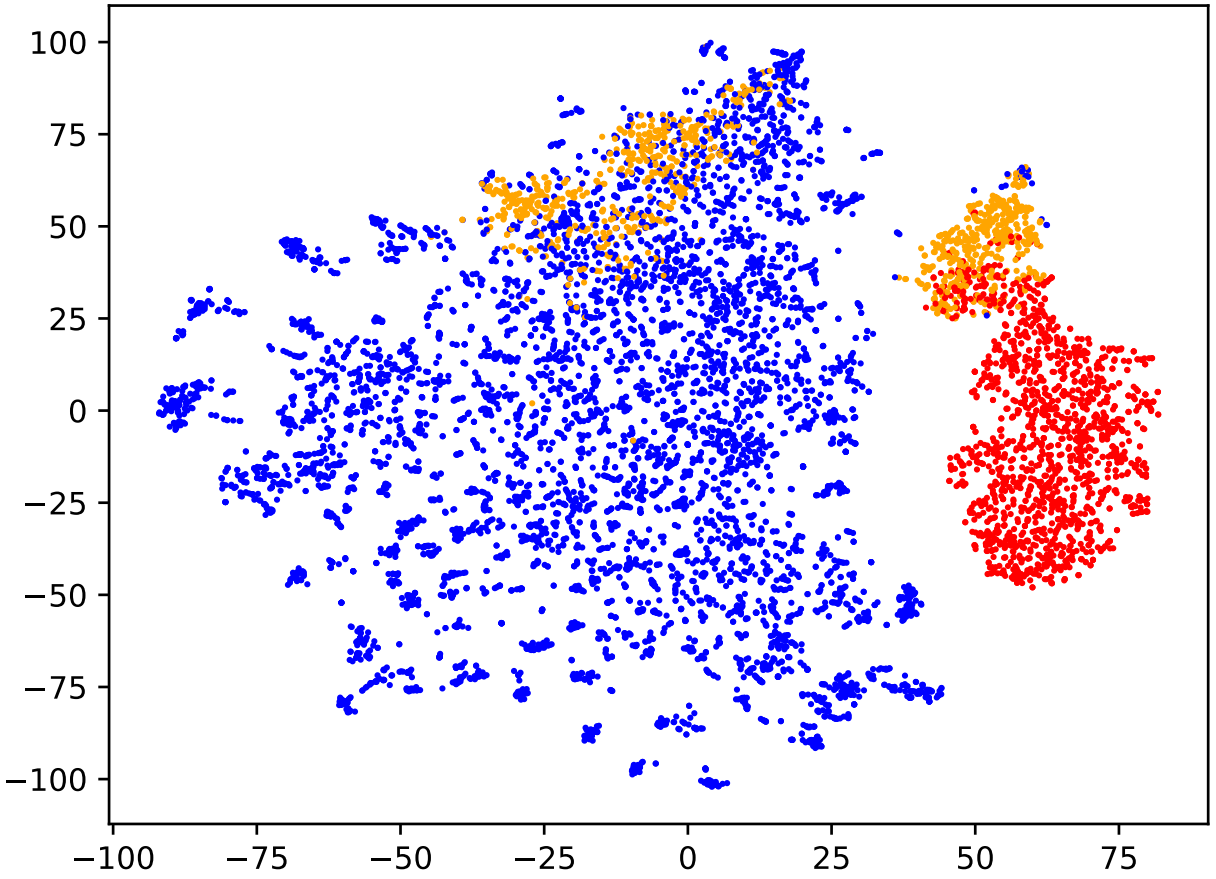


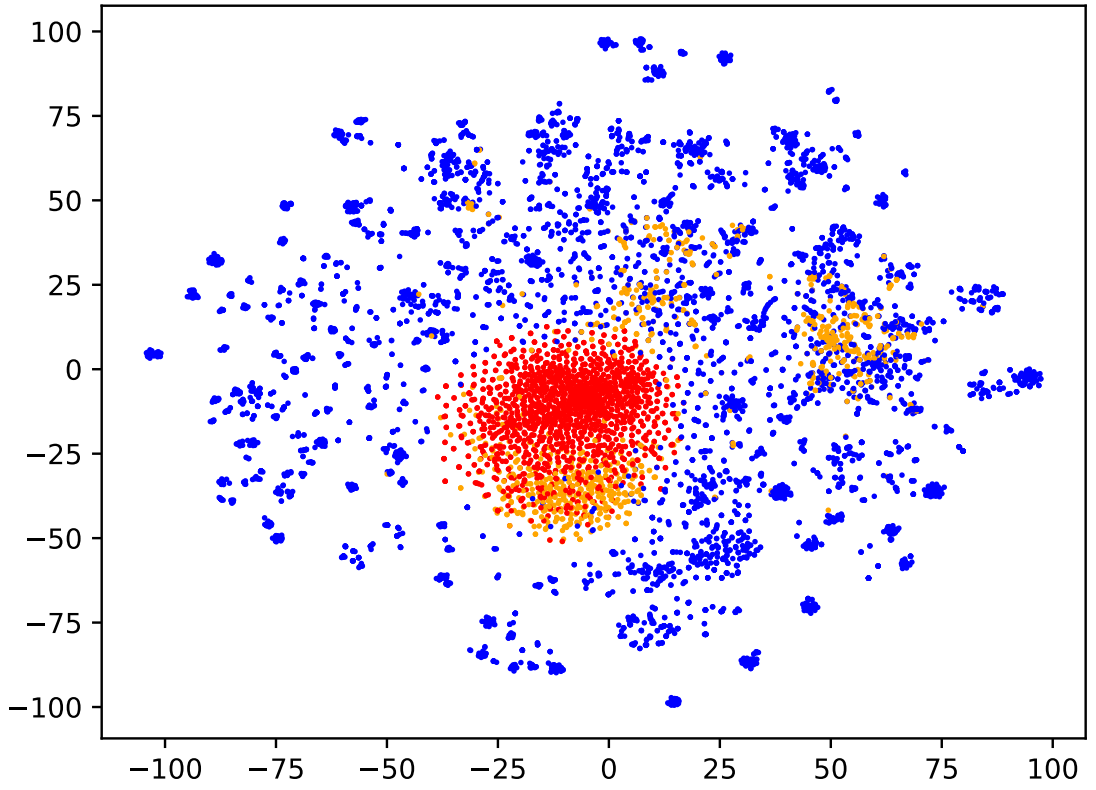
Additional slides



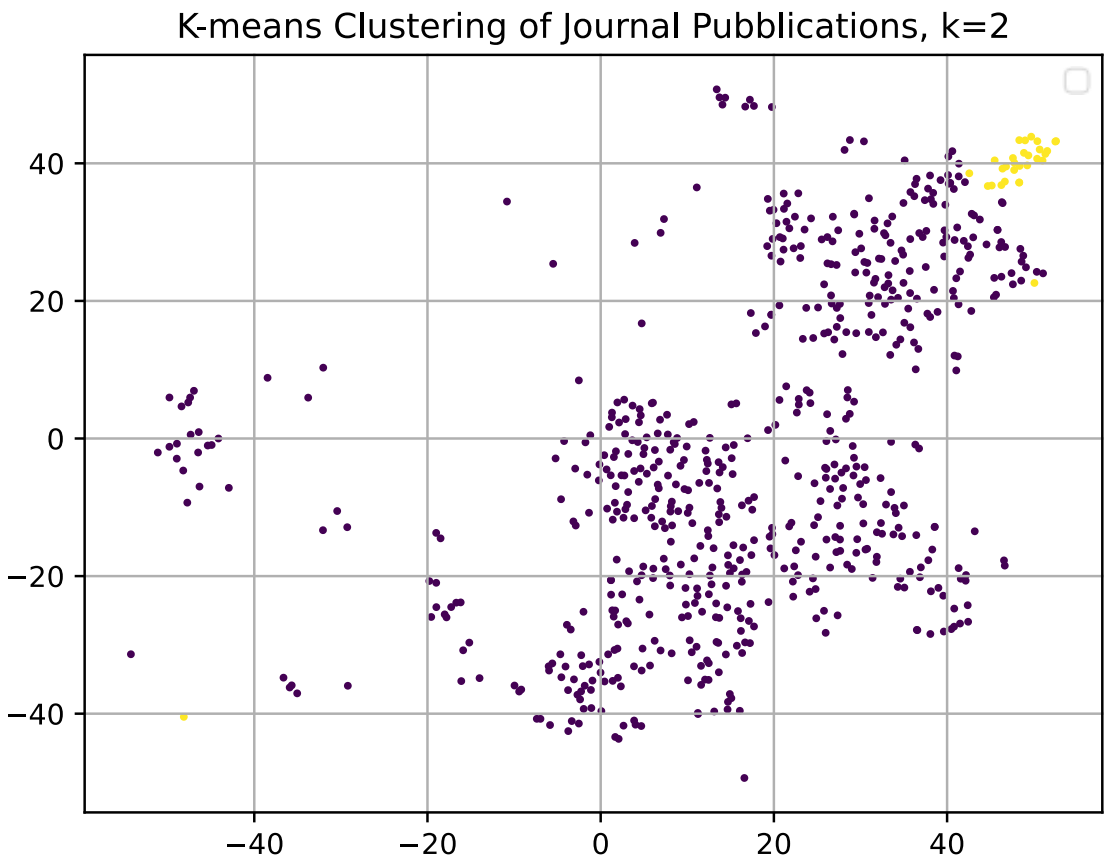


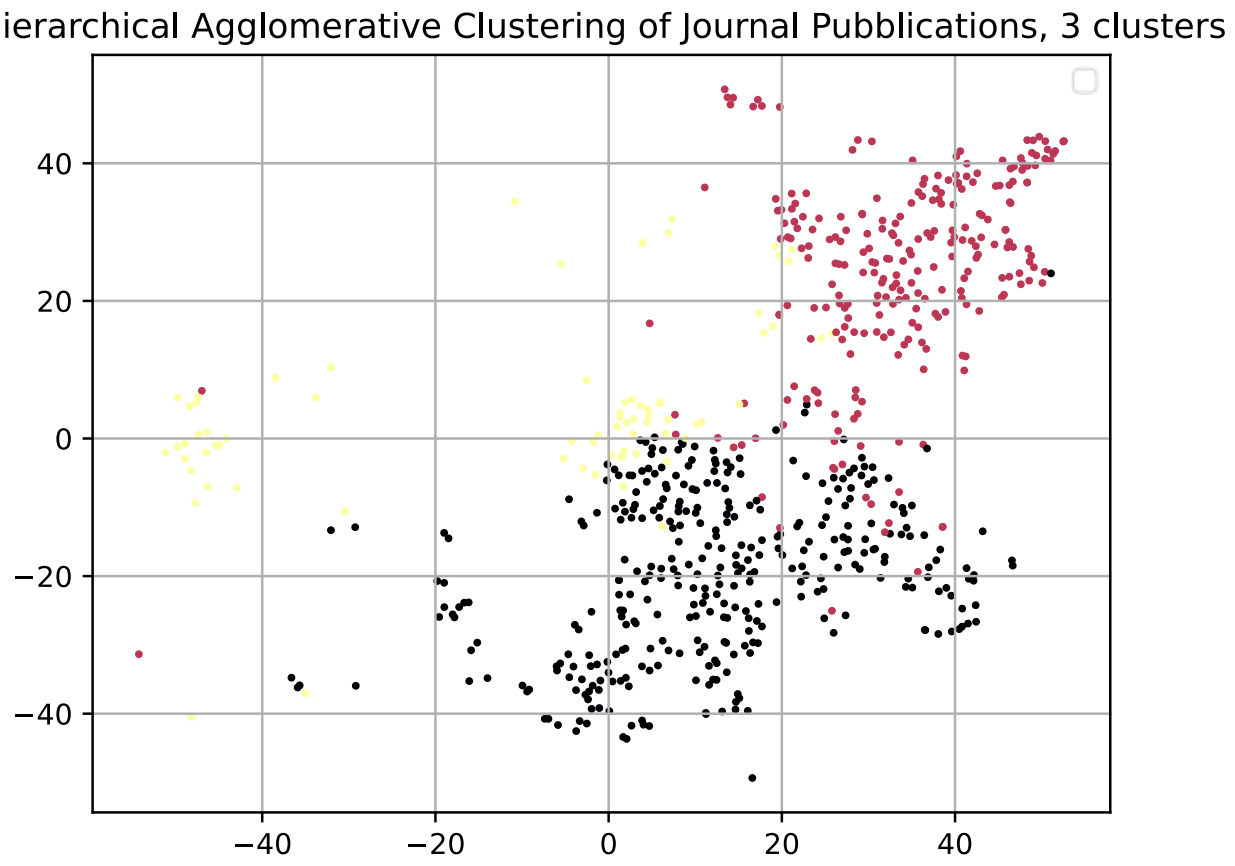
Additional slides





Additional slides





Additional slides

K-means, k = 2, Cardiology

	Cluster 1	Cluster 2
False (Layer1)	49	10
True (Layer1)	10	2
Layer 3	583	18

Hierarchical Agglomerative Clustering, k =3, Cardiology

	Cluster 1	Cluster 2	Cluster 3
False (Layer1)	16	35	8
True (Layer1)	4	8	0
Layer 3	325	212	64

Additional slides

"È descritto il caso clinico di una donna di 84 anni, forte fumatrice per circa 30 anni. Non familiarità per malattie renali. Principali rilievi anamnestici: ipertensione arteriosa da almeno 20 anni controllata dalla terapia farmacologica (calcio antagonisti, diuretici), colelitiasi, ipotiroidismo in terapia sostitutiva ormonale. Non disturbi dell'udito. All'età di 70 anni quadrantectomia mammaria destra per carcinoma. All'età di 80 anni la paziente effettuava due ricoveri durante i quali venivano evidenziati addensamenti polmonari multipli, bilaterali e veniva posta diagnosi di "Alveolite allergica estrinseca". In tale occasione era esclusa patologia infettiva e prescritto steroide a seguito del quale la paziente presentava un miglioramento del quadro clinico; non dati disponibili sulla funzione renale. Nel 2005, all'età di 82 anni, la paziente era ricoverata per insufficienza renale acuta rapidamente progressiva (creatininemia 1,5→6 mg/dL) e insufficienza respiratoria. Dal punto di vista sistemico la paziente presentava dolori osteoarticolari diffusi, febbricola e tosse secca con un solo episodio di emottisi. Alla radiografia del torace venivano evidenziati addensamenti polmonari bilaterali. La TC del torace mostrava aspetto a vetro smerigliato del parenchima polmonare, edema alveolare, addensamenti parenchimali multipli. Un ecocardiogramma risultava nella norma per l'età. La paziente effettuava una broncoscopia che mostrava un albero tracheobronchiale nella norma; l'esame citologico del liquido alveolare mostrava la presenza di numerosi macrofagi contenenti pigmento emosiderinico. Gli esami colturali (germi comuni, BK) risultavano negativi. La paziente eseguiva agobiopsia renale che evidenziava un quadro istologico di glomerulonefrite crescentica necrotizzante con depositi lineari di IgG nelle membrane basali glomerulari compatibile con S. di Goodpasture. Presenza nel siero di anticorpi antimembrana basale glomerulare (anti MBG); negativa la ricerca di anticorpi anti citoplasma dei neutrofili (ANCA). La paziente praticava terapia steroidea dapprima ev poi per os, boli mensili di ciclofosfamide e plasmaferesi con risoluzione del quadro clinico e negativizzazione degli anti MBG. La paziente assumeva terapia immunosoppressiva con acido micofenolico e cortisone per oltre un anno. Nel corso di questo periodo anti MBG e ANCA sempre negativi. La creatinina oscillava tra 1,5-2 mg/dL con proteinuria sempre < 1 g die. All'età di 84 anni dopo circa 6 mesi dalla sospensione della terapia immunosoppressiva recidiva di malattia con grave riacutizzazione dell'insufficienza renale, con necessità di emodialisi e della patologia polmonare. Il quadro laboratoristico evidenziava negatività degli anti MBG e comparsa di positività dei p-ANCA pur se a titolo non elevato (IFI 1:80; EIA 9,6). L'esame TC torace mostrava presenza di limitato ispessimento di natura reticolare e reticolo-nodulare della trama interstiziale. Non eseguita broncoscopia per impossibilità da parte della paziente a sostenere l'esame. Negativi gli esami colturali (escreatocoltura, urinocoltura). Nell'ipotesi di una riattivazione della malattia autoimmunitaria e con l'evidenza di esami colturali negativi la paziente praticava nuovamente boli di cortisone ev ma l'evoluzione clinica era complicata da polmonite da K. pneumoniae che determinava rapidamente il decesso della paziente."