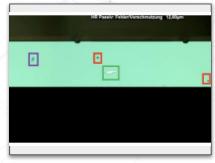
Graph neural networks for information extraction

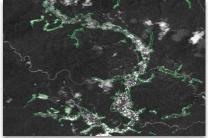
EuroPython 2021 Augusto Stoffel dida Machine Learning



Deep learning landscape







Computer Vision

Example Project

Paragraf 9 [7846; 10313] (Schoenheitsreparaturklausel)

§ 9 Instandhaltung der Mieträume und Schönheitsreparaturen

5 ÜBERTRAGUNGSKLAUSEL

UMFANGSKLAUSEL

Die Durchführung der Schönheitsreparaturen obliegt dem Mieter. Diese umfassen insbesondere das Tapezieren, Anstreichen der Wände und Decken, das Pflegen der Fußböden, das Streichen der Innentüren, der Fenster und Außentüren...

7 FRISTENKLAUSEL

Bei normaler Benutzung sind die Schönheitsreparaturen, ab Vertragsbeginn gerechnet, in Küche Bad und WC alle 3 Jahre für alle übrigen Räume alle 5 Jahre, auszuführen.

8 ENDRENOVIERUNGSKLAUSEL

[...] Demzufolge ist die Mietsache bei Beendigung des Mietverhältnisses unabhängig von der Mietdauer und unabhängig davon, wann zuletzt die vertragsgemäßen Schönheitsreparaturen stattgefunden haben, mit fachmännisch frisch weiß gestrichenen Decken und Wänden sowie im übrigen schadensfrei und gereinigt zurückzugeben.

§ 18 Rückgabe der Mieträume

FRISTENKLAUSEL

Bei Beendigung des Mietverhältnisses bzw. bei vorherigem Auszug hat der Mieter die Mieträume geräumt, in sauberem Zustand und mit allen - auch den von ihm selbst beschafften- Schlüsseln zurückzugeben. Die Mieträume sind in mangelfreiem Zustand,

Natural Language Processing (NLP)

Example Project

Sample GNN use case

_____ 🖗 _____ Dewey Hobbs GmbH

Dewey Hobbs - Hobbsstraße 21 - 12345 Hobbshausen

ABC Chemicals AG Kochstraße 2 5678 Salzstadt Dewey Hobbs GmbH Deweystraße 21 12345 Hobbshausen

Tel.: 0211 12345 67 E-Mail: accounting@hobbs.de Internet: www.deweyhobbs.de

Rechnung

Rechnung Nr. 2020-06-1001 Bestellung-Nr.: KA12345 Kunden-Nr.: 1001
Bitte bei Zahlungen und Schriftverkehr angeben! Datum: 07.06.2020

Pos	Leistung	MwSt.	Einzelpreis	Anzahl	Gesamtpreis
1	Röntgenblitzröhre ZW 367 H 01G 12 120kV /28mm / 20 cm	19 %	11,00 EUR	2	22,00 EUR
2	50kg Seesand reinst	19 %	22,00 EUR	1	22,00 EUR
3	Salzsäure 1 mol / 10l, 20l Standardlösung 7647-01-0	19 %	33,00 EUR	2	66,00 EUR

Sample GNN use case

_____ 🖗 _____ Dewey Hobbs GmbH

Dewey Hobbs - Hobbsstraße 21 - 12345 Hobbshausen

ABC Chemicals AG Kochstraße 2 5678 Salzstadt

Dewey Hobbs GmbH Deweystraße 21 12345 Hobbshausen

Tel.: 0211 12345 67 E-Mail: accounting@hobbs.de Internet: www.deweyhobbs.de

Rechnung

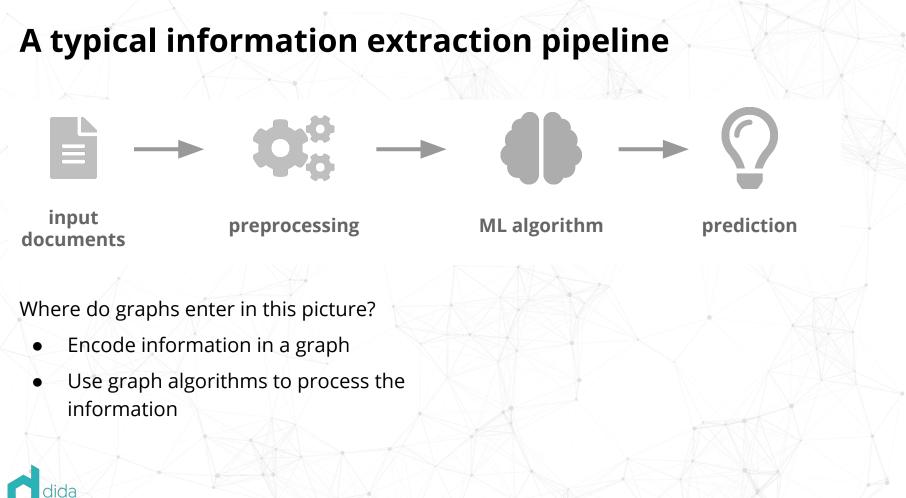
Rechnung Nr. 2020-06-1001

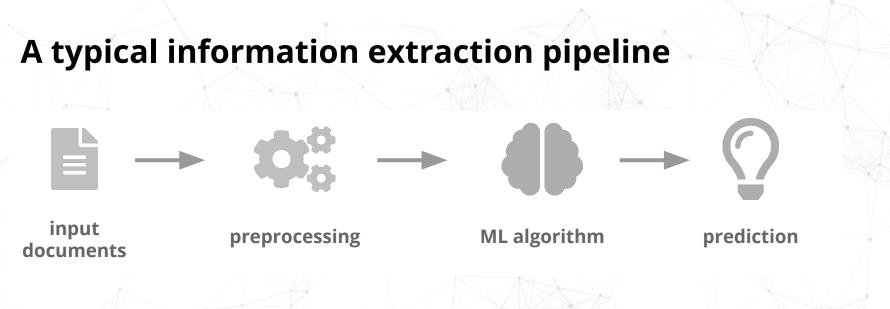
Bestellung-Nr.: KA12345 Kunden-Nr.: 1001

Bitte bei Zahlungen und Schriftverkehr angeben!

Datum: 07.06.2020

Pos	Leistung	MwSt.	Einzelpreis	Anzahl	Gesamtpreis
1	Röntgenblitzröhre ZW 367 H 01G 12 120kV /28mm / 20 cm	19 %	11.00 EUR	Ø	22,00 EUR
2	50kg Seesand reinst	19 %	22.00 EUR	1	22,00 EUR
3	Salzsaure 1 mol / 101 . 201 Standardlösung 7647-01-0	19 %	33.00 EUR	6	66,00 EUR



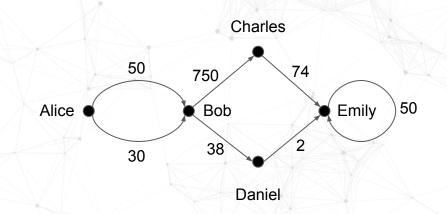


Where do graphs enter in this picture?

- Encode information in a graph
- Use graph algorithms to process the information

Agenda:

- Graphs and GNNs
- Comparison with convolutional networks
- Use case



Variations: simple graph, multigraph, pseudograph; directed or not



B

S

р

q

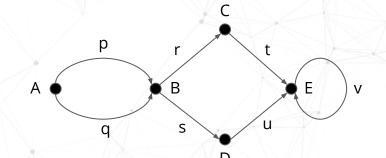
Definition: A *directed pseudograph*, or simply *graph*, is

- a set **1** of vertices
- a set *E* of edges

V

• functions source, target: $\mathcal{E} \to \mathcal{V}$

А

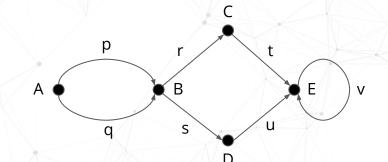


Definition: A *directed pseudograph*, or simply *graph*, is

- a set **1** of vertices
- a set *E* of edges
- functions source, target: $\mathcal{E} \to \mathcal{V}$

Example:

- $\Psi = \{A, B, C, ..., E\}$
 - **ℓ** = {p, q, r, ..., v}
- source(p) = A
- target(p) = B
- source(v) = E
- target(v) = E



Definition: A directed pseudograph, or simply graph, is

- a set **1** of vertices
- a set *E* of edges
- functions source, target: $\mathcal{E} \to \mathcal{V}$

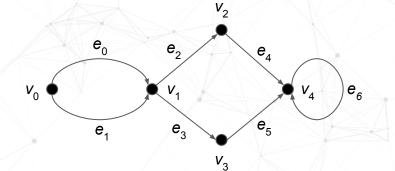
Example:

- **V** = {A, B, C, ..., E}
- *E* = {p, q, r, ..., v}
- source(p) = A
- target(p) = B
- source(v) = E
- target(v) = E

Note: Often there are labels associated to vertices and edges.

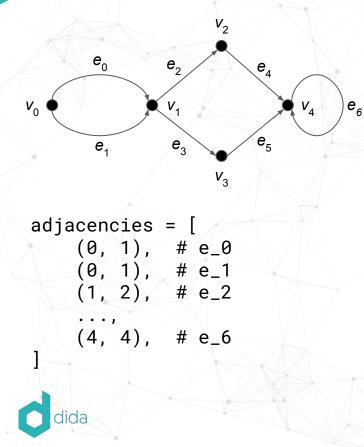
- $L_0: \Psi \rightarrow \{ \text{vertice labels} \}$
- $L_1: \mathcal{E} \to \{\text{edge labels}\}$

Representing graphs in Python

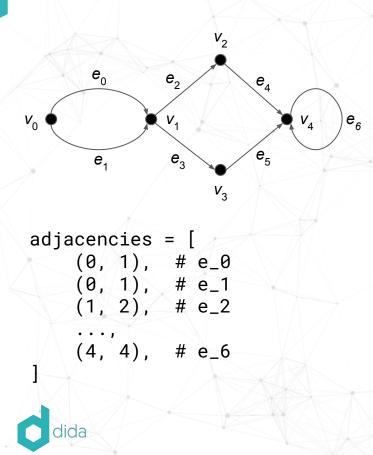




Representing graphs in Python



Representing graphs in Python



A graph with nodes labeled by NodeT and edges labeled by EdgeT can be modeled as follows:

class Graph(Generic[NodeT, EdgeT]):

nodes: List[NodeT]
edges: List[EdgeT]
adjacencies: List[Tuple[int, int]]

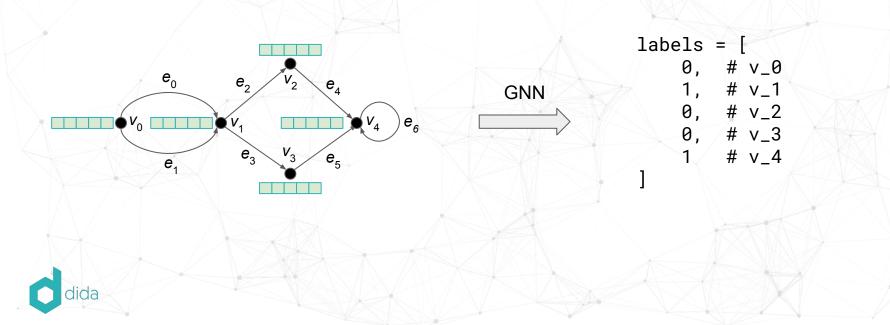
def source(self, k: int) -> int:
 """Source of the kth edge"""
 return self.adjacencies[k][0]

def target(self, k: int) -> int:
 """Target of the kth edge"""
 return self.adjacencies[k][1]

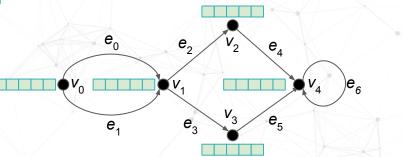
Graph neural networks

Let's consider the *node classification* problem:

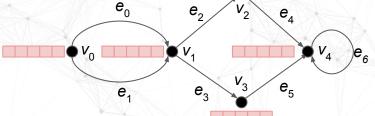
- Input: a graph G of type Graph [Tensor, EnumT]
- Output: a labeling of its vertices, $L: \Psi \to \{0, 1\}$



Graph neural networks



GNN layer

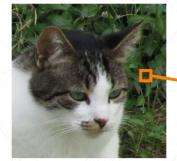


For instance: $y_v = \sigma \left(\sum_{\substack{e \\ w \to v}} \frac{1}{\deg_{t(e)}^+(v)} W_{t(e)} x_w \right)$ where

- y_v is the output feature vector of node v
- x_{w} is the input feature vector of node w
- t(e) is the type of the edge $e: v \rightarrow w$
- W_t is a learned weight matrix corresponding to edge type t
- $\deg_t^+(v)$ is the number of incoming edges of type *t* for node *v*
- σ is an activation function

Schlichtkrull et at, "Modeling Relational Data with Graph Convolutional Networks", arxiv:1703.06103

Graph (convolutional) networks generalize CNNs from computer vision





 $y_{i,j} = \sigma \left(\sum_{\substack{k=-1,0,1 \ l = 1,0,1}} W_{k,l} x_{i+k,j+l} \right)$

where

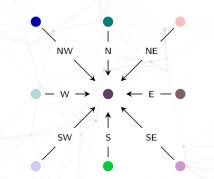
Original image

A 3 \times 3 patch

- *y_{i,j}* is the output feature vector of pixel (*i*, *j*)
 x_{i,j} is the input feature vector of pixel (*i*, *j*)
 W_{k,l} are learned weight matrices
- σ is an activation function

Graph (convolutional) networks generalize CNNs from computer vision





 $y_{i,j} = \sigma \left(\sum_{\substack{k=-1,0,1 \ l= 1,0,1}} W_{k,l} x_{i+k,j+l} \right)$

where

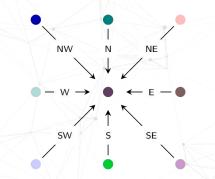
Original image

A graph representing the 3×3 patch

- $y_{i,i}$ is the output feature vector of pixel (i, j)
- x_{i,j} is the input feature vector of pixel (i, j)
 W_{k,l} are learned weight matrices
- σ is an activation function

Graph (convolutional) networks generalize CNNs from computer vision





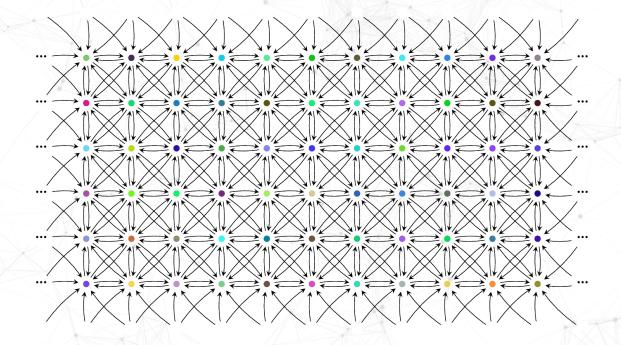
 $y_p = \sigma \left(W_0 x_p + \sum_{q \to p} W_t x_q \right)$

where

Original image

A graph representing the 3×3 patch

- x_p is the input feature vector of pixel p
- y'_p is the output feature vector of pixel p
- $t \in \{N, NW, W, ...\}$ is a cardinal direction
- W_t are learned weight matrices
- σ is an activation function

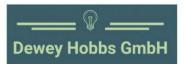


A graph representing the entire image



Idea: Model the document structure as a graph

- each word is a node in the graph
- neighbouring words are connected by edges



Dewey Hobbs - Hobbsstraße 21 - 12345 Hobbshausen

ABC Chemicals AG Kochstraße 2 5678 Salzstadt



Tel.: 0211 12345 67 E-Mail: accounting@hobbs.de Internet: www.deweyhobbs.de

Rechnung

Rechnung Nr. 2020-06-1001	Bestellung-Nr. KA123	45 Kunden-Nr.: 1	.001
Bitte bei Zahlungen und Schriftverkehr ang	eben!	Da	atum: 07.06.2020
Des Loistung		Anaphi	Cosomtorois

Pos	Leistung	MwSt.	Einzelpreis	Anzahl	Gesamtpreis
1	Röntgenblitzröhre ZW 367 H 01G 12 120kV /28mm / 20 cm	19 %	11.00 EUR	6	22,00 EUR
2	50kg Seesand reinst	19 %	22,00 EUR	1	22,00 EUR
3	Salzsaure 1 mol / 101, 201 Standardlösung 7647-01-0	19 %	33.00 EUR	0	66,00 EUR

Prediction Result (example)



Our Order Date

1/2/20

Revision

1

ACME US&A

Company Making Everything Carolina . 12345 .UNITED STATES Tel: +1-123-234-3456 Fed ID 11-123456 Tax Exempt 123345-6789 Fax: VAT Reg. No .:

Seller We-Supply NC Suppliers Corporation P.O. Box 123456 Atlanta GA 12345-98765 UNITED STATES 111-222-3334 Phone: Fax:

Ship Via Road

Terms Of Delivery Free on board

999 Supreme Industrial Road , Anderson , South

WESUPPLY **Delivery Address** ACME US&A 999 Supreme Industrial Road Anderson South Carolina 12345 UNITED STATES

Order

Supplier No

Payment Terms 1 Day Net Delivery Date 3/31/19

For orders placed by AMCE US&A or AMCE US&B, the AMCE US&A / AMCE US&B Standard Conditions of Contract for the Purchase and I or Hire of Goods and Services (US&A/US&B) apply to this order. For orders placed by AMCE US&A USA Corporation, the AMCE US&A USA Corporation Standard Conditions of

Nodes:

Purchase Order

A12345

Each word is a node in the graph

ACME US&A Order 999 Supreme Industrial Road , Anderson , South Carolina . Purchase Order Company Our Order Date 12345 UNITED STATES Making 1/2/20 A12345 Tel: +1-123-234-3456 Fed ID 11-123456 Everything Fax: Tax Exempt 123345-6789 Supplier No Revision VAT Reg. No .: WESUPPLY 1 Seller Delivery Address We-Supply ACME US&A 999 Supreme Industrial Road NC Suppliers Corporation Anderson P.O. Box 123456 Atlanta South Carolina GA 12345-98765 12345 UNITED STATES UNITED STATES 111-222-3334 Phone Payment Terms Ship Via 1 Day Net Road Delivery Date Terms Of Delivery Free on board 3/31/19

А

Fax:

For orders placed by AMCE US&A or AMCE US&B, the AMCE US&A / AMCE US&B Standard Conditions of Contract for the Purchase and I or Hire of Goods and Services (US&A/US&B) apply to this order. For orders placed by AMCE US&A USA Corporation, the AMCE US&A USA Corporation Standard Conditions of

Nodes:

Each word is a node in the graph

Edges:

Neighboring words are connected

Our Order Date

1/2/20

Revision

1

Purchase Order

A12345

ACME US&A

Company Making Everything

Tel: +1-123-234thing Fax: Tax Exempt VAT Reg. No.:

We-Supply NC Suppliers Corporation P.O. Box 123456 Atlanta GA 12345-98765 UNITED STATES Phone: 111-222-3334 Fax:

Ship Via Road

Seller

Terms Of Delivery Free on board 999 Supreme Industrial Road , Anderson , South Carolina . 12345 .UNITED STATES Tel: +1-123-234-3456 Fed ID 11-123456 Fax: Tax Exempt 123345-6789 VAT Reg. No.:

> Delivery Address ACME US&A 999 Supreme Industrial Road Anderson South Carolina 12345 UNITED STATES

Order

Supplier No

WESUPPLY

Payment Terms 1 Day Net Delivery Date 3/31/19

For orders placed by AMCE US&A or AMCE US&B, the AMCE US&A / AMCE US&B Standard Conditions of Contract for the Purchase and / or Hire of Goods and Services (US&A/US&B) apply to this order. For orders placed by AMCE US&A USA Corporation, the AMCE US&A USA Corporation Standard Conditions of

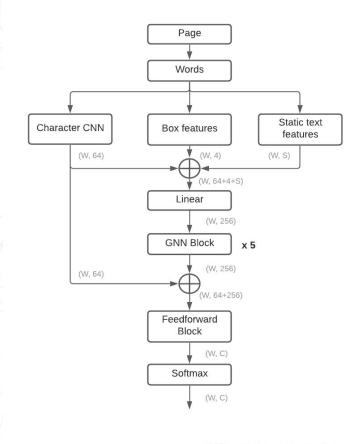
Nodes:

• Each word is a node in the graph

Edges:

Neighboring words are connected

Use case: a possible model architecture



Legend

W: Number of words

S: Number of static text features

C: Number of classes

•: Concatenate tensors

Use case: results

dida

Harling Tet: +1:12:32:34:34:56 / Fed ID 12:1:23:2456 VeryThing Fac: Tax: Exempt 12:33:45-6789 VAT Reg. No.: To: Ship to : To: Value ACME US&A DEPT AT 12:34:56 999 Supreme Industrial Road ATLANTA GA 12:345-45678 ANDERSON, SC United States United States				To : We- DEP ATL	Fax: Tax Exempt 12334 VAT Reg. No.: Supply T AT 123456 ANTA GA 12345-45676 ed States		999 S AND	EUS&A Supreme Industrial Ro ERSON, SC d States	ad				
Phone ()	- Fax () -						N	 Phone ()	- Fax ()	2			-
PO Date	Ship Via	FOB	Planner Cor	nfirming to	Terms	-		PO Date	Ship Via	FOB	Planner	Confirming to	Terms
5/2019	Your Truck	OUR PLANT	DGJ		0.5% 10, NET 30		/	02/15/2019	Your Truck	OUR PLANT	DGJ		0.5% 10, NET 30
Facilit	/ Part / Rev / Description / Deta		ntity Promised Delivery 3.4567 02/18/2019	Vendor Unit Cost 11.1111	Extended Cost			Item Facilit	y / Part / Rev / Description De		Z3.4567 02	elivery Vendor Unit Co 18/2019	
efault GALV, 12GA	.015	127-3/8" Rev 000 U/M EA Order Quantity: 11.0000	2.3456 02/18/2019	222.2222				Purchasing **SHEET TC WIDTH +0/ LENGTH +0	Vendor U/M	8 127-3/85 ⁻² Key 000 U/M EA Order Quantity: 11.0000	(2.345 6 * ()2)	222.2 18/2019 ⁻ 11.2	
fault ALV, 12GA Purchasing SHEET TC WIDTH +0/	GALV, 12GA X 48-7/16" X 12GA48716138GR50 (48-7/16" X 138-3/4" SS Category : INV LERENCE** .015 f0625	Rev 000 U/M EA Order Quantity: 12.0000	2 1111 02/15/2019	333.4444	1,222.30			Default GALV, 12GA Purchasing	BALV. 12GA X 48-7/15* 12GA 287/16138GR50 X 48-7/16* X 138-3/4* 59 Category : INV >LERENCE** -015 >/0625	Rev 000 U/M EA Order Quantity: 12.0000		333.4	

26

Implementations, literature

- PyTorch Geometric
 - "GNN Cheatsheet" in the docs contains an interesting list of papers
- <u>Spektral</u> for TensorFlow
- <u>Deep Graph Library</u> (framework agnostic)
- <u>https://en.wikipedia.org/wiki/Graph_neural_network</u> exists since 5 July 2021