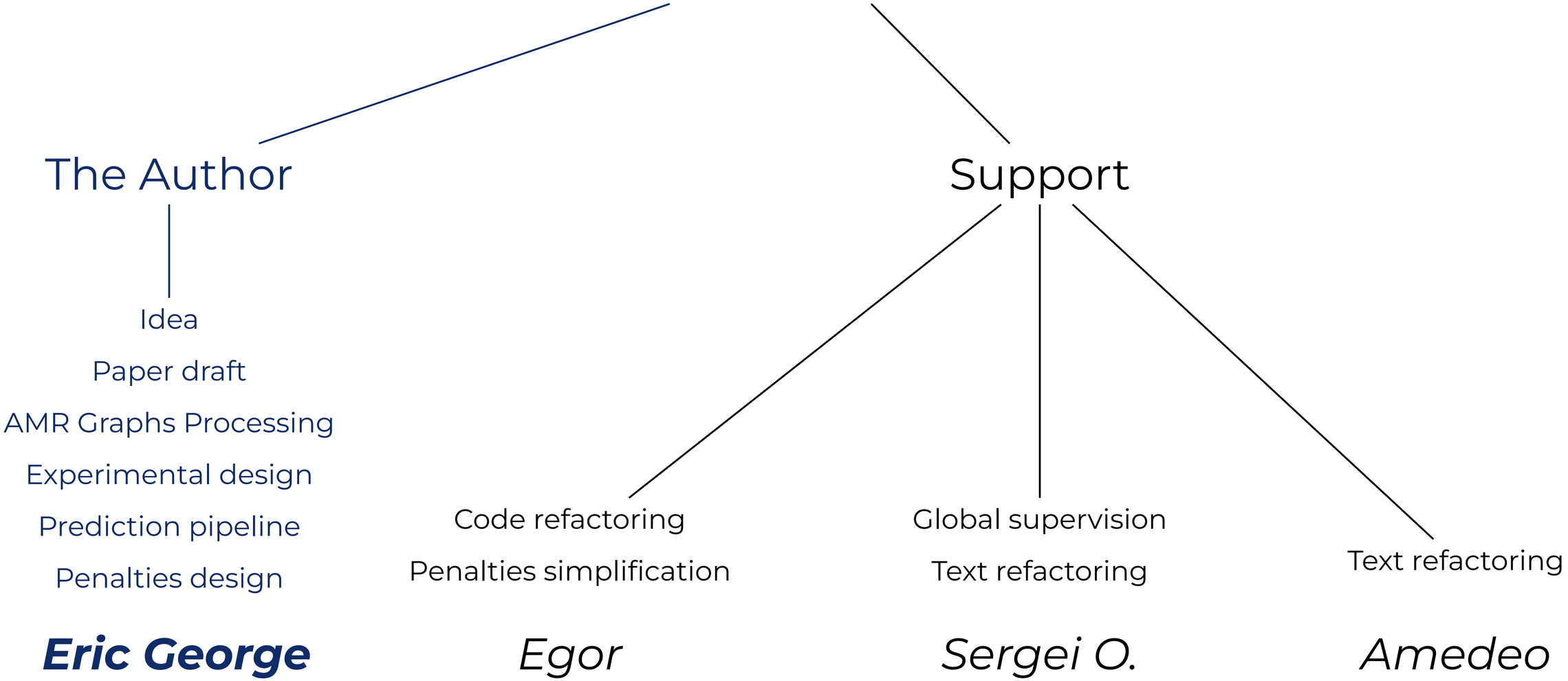
Document Classification via Stable Graph Patterns and Conceptual AMR Graphs

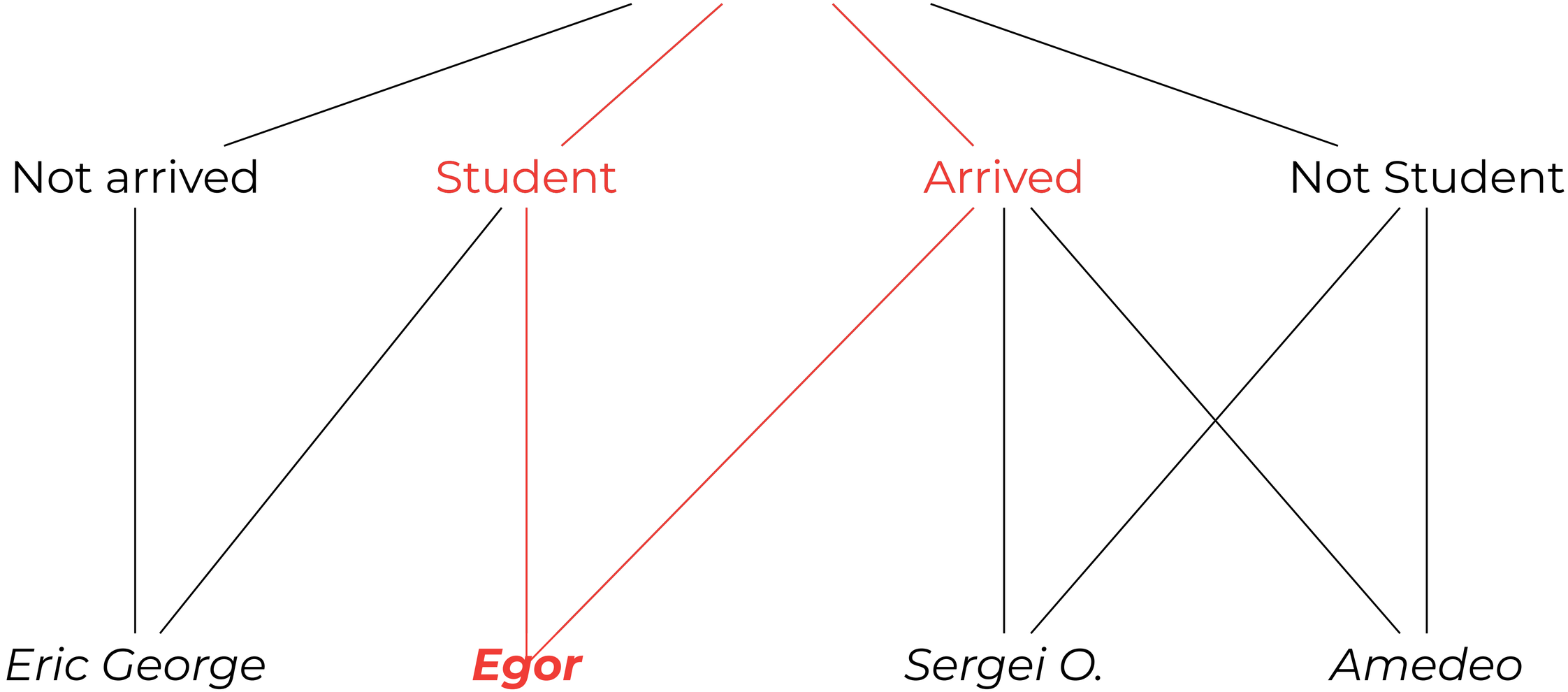
CONCEPTS conference, Cadiz, Spain, September 12, 2024

By Eric George Parakal, **Egor Dudyrev**, Sergei O. Kuznetsov, Amedeo Napoli

Authors roles



Authors statuses

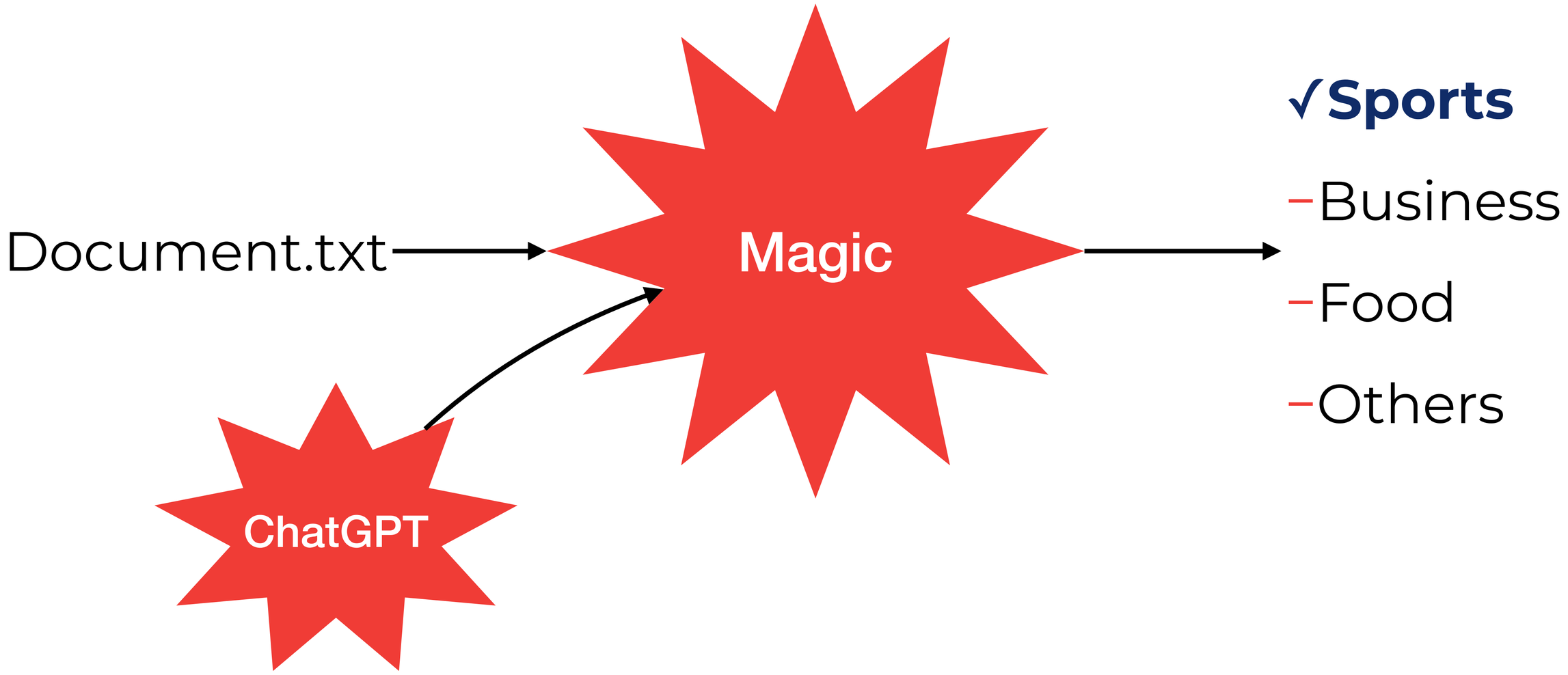


Eric George's thesis

Document Classification via Stable Graph Patterns and Conceptual AMR Graphs

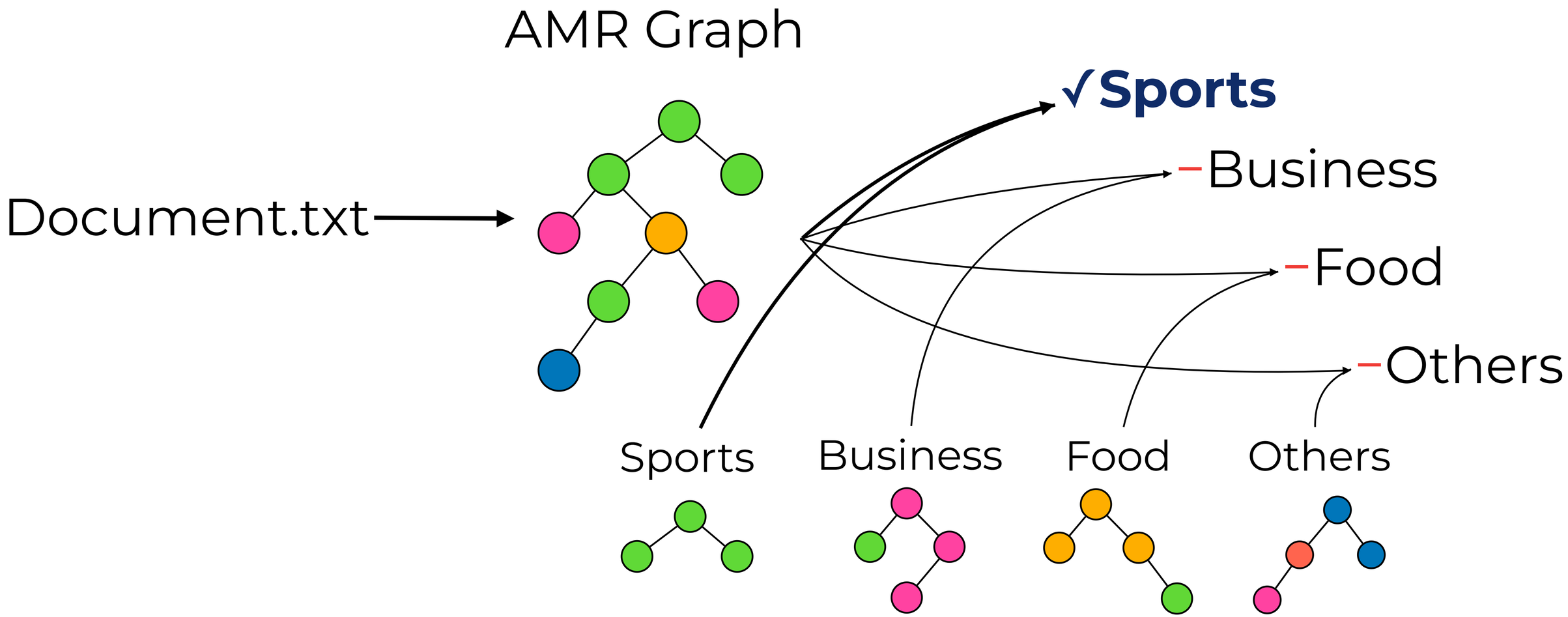
Egor's thesis

How to classify documents



Explainable Document Classification

S.O. Kuznetsov, E.G. Parakal, 2023



Abstract Meaning Representation Graph

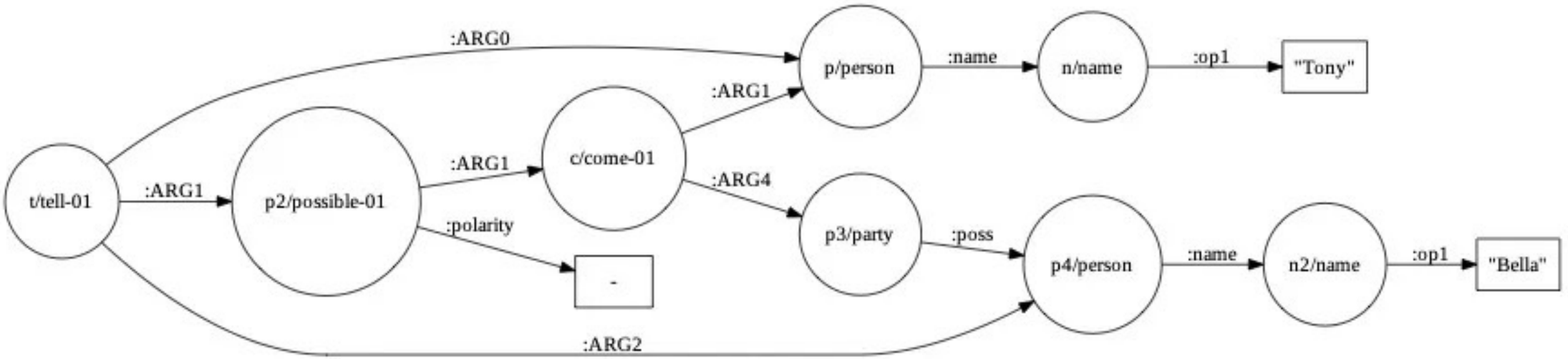


Fig. from "What are abstract meaning representation graphs" @ Medium

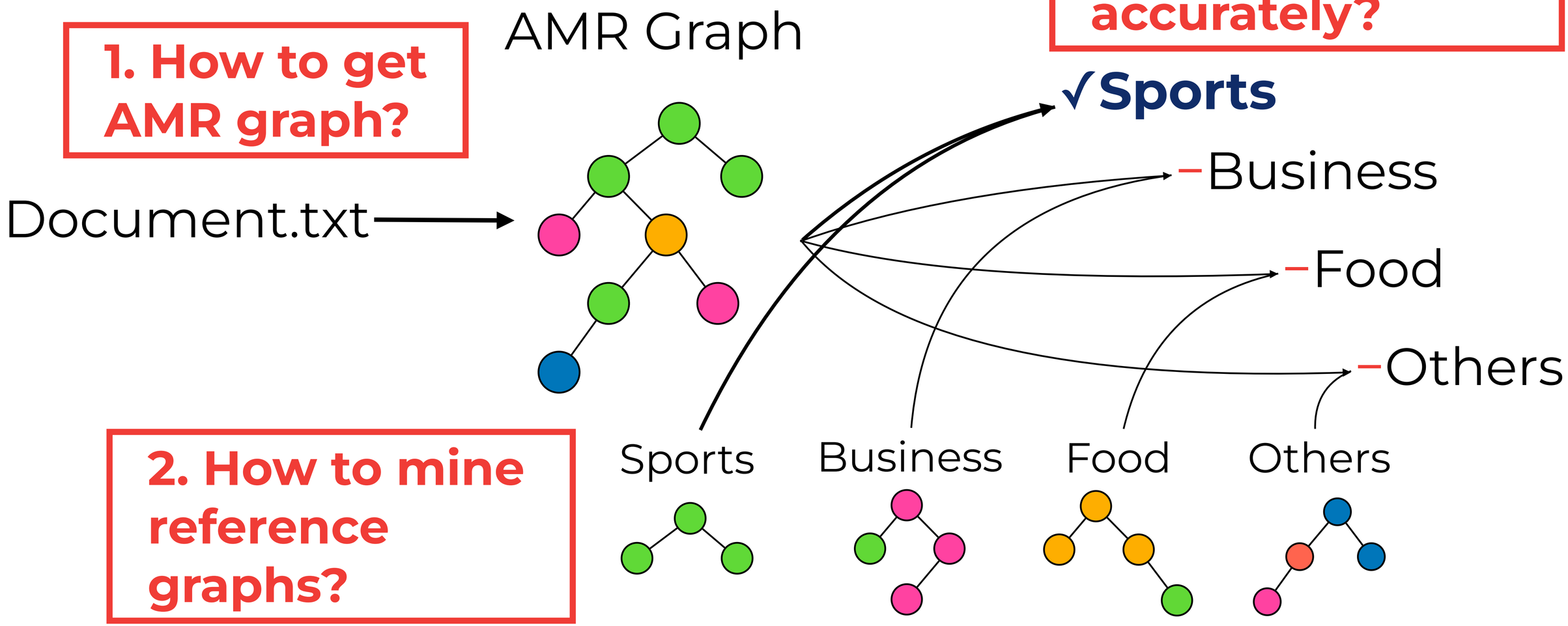
By L. Banarescu et al., LAW VII & ID, 2013

Problems to study

1. How to get AMR graph?

2. How to mine reference graphs?

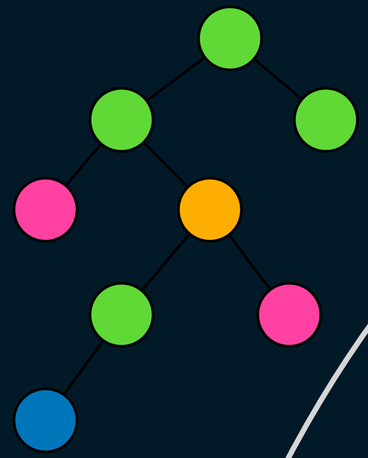
3. How to classify accurately?



Problem 1

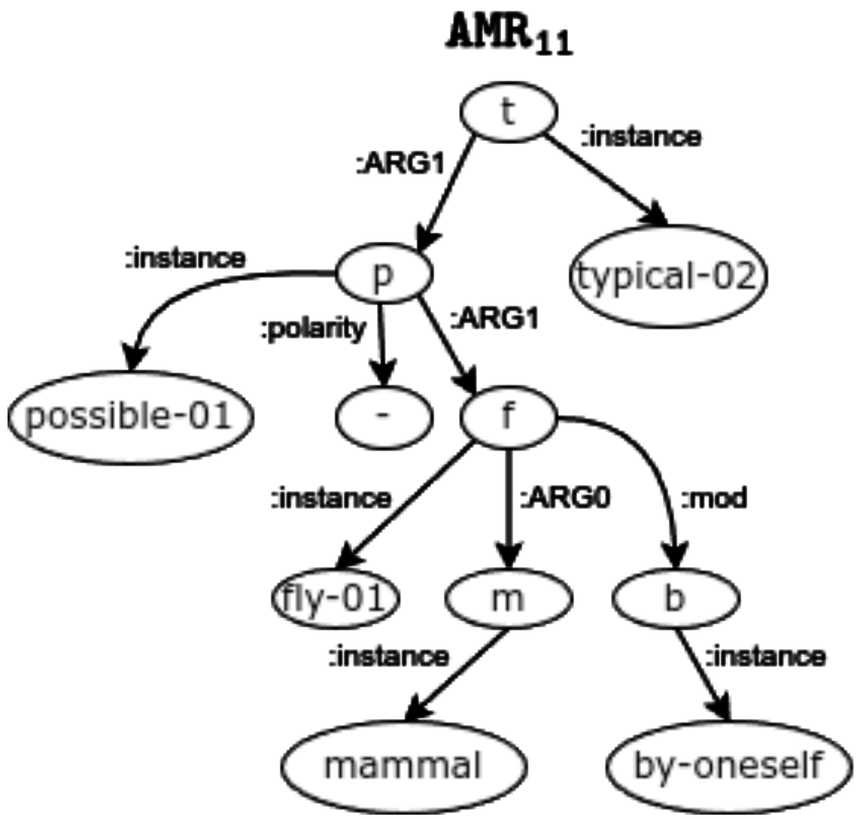
1. How to get AMR graph?

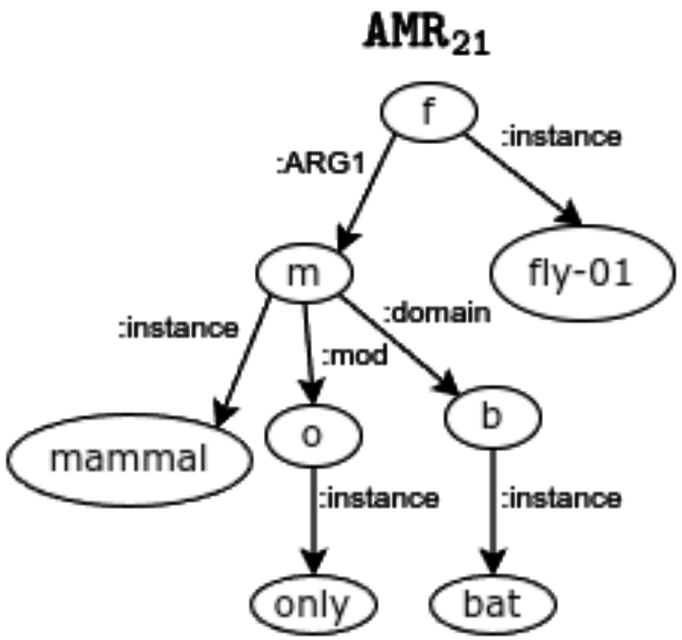
Document.txt

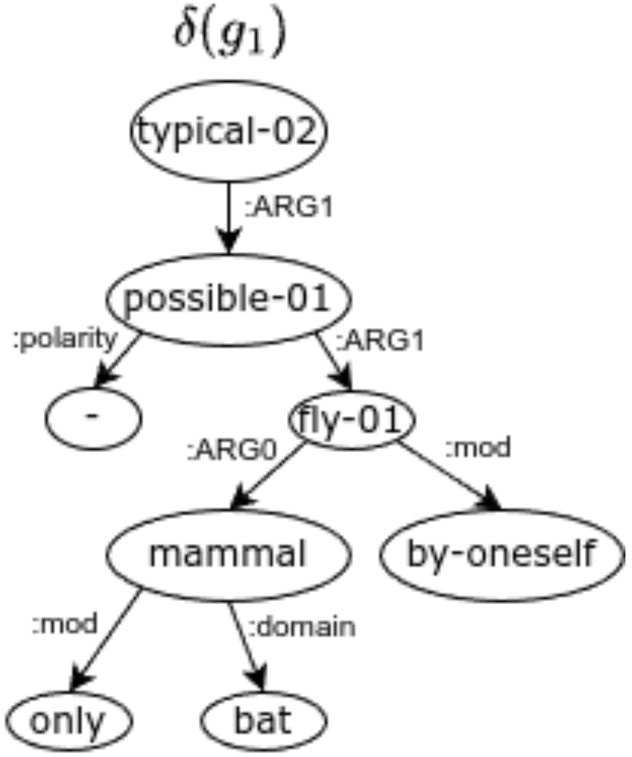


AMR Graph

AMR Graph for a sentence







Doc2Graph algorithm

Algorithm 1 DocToGraph

Input: a document g_i
Output: a graph description $\delta(g_i)$

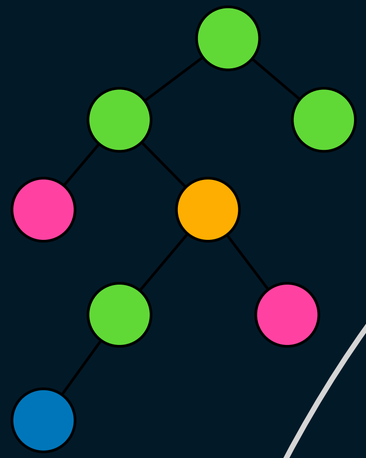
$T_i \leftarrow \text{findSentences}(g_i)$
for all $t_{ji} \in T_i$ **do**
    $\texttt{AMR}_{\texttt{ji}} \leftarrow \text{AMRParser}(t_{ji})$
    $\texttt{MOD}_{\texttt{ji}} \leftarrow \text{ModifyGraph}(\texttt{AMR}_{\texttt{ji}})$
    $\texttt{REF}_{\texttt{ji}} \leftarrow \text{refineGraph}(\texttt{MOD}_{\texttt{ji}})$
end for
$\delta(g_i) \leftarrow \text{mergeGraphs}(\{\texttt{REF}_{\texttt{ji}}\})$
return $\delta(g_i)$

Problem 2

1. How to get AMR graph?

AMR Graph

Document.txt



√ **Sports**

– Business

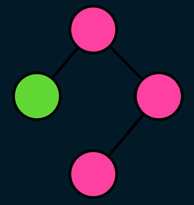
– Food

– Others

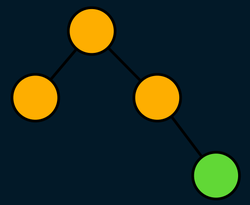
2. How to mine reference graphs?

Sports

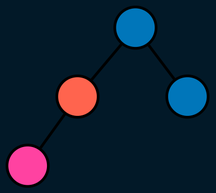
Business

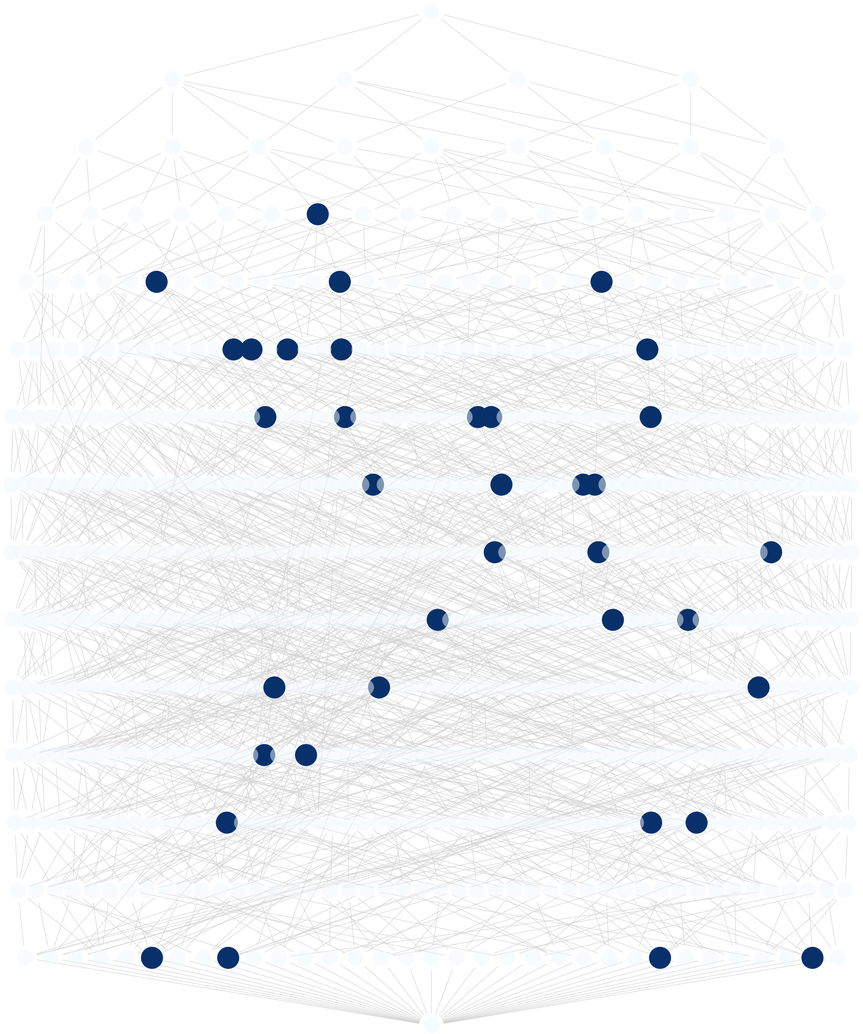


Food



Others





Use pattern structures

Attributes M $\Longrightarrow$ Description space $\underline{\mathbb{D}} = (2^M, \subseteq)$

$\ldots$ $\Longrightarrow$ Description space $\underline{\mathbb{D}} = (\mathbb{D}, \sqsubseteq)$

Select only interesting patterns

Frequent Concepts/Iceberg Lattice

Stable Concept Mining

Δ Stability

FCA

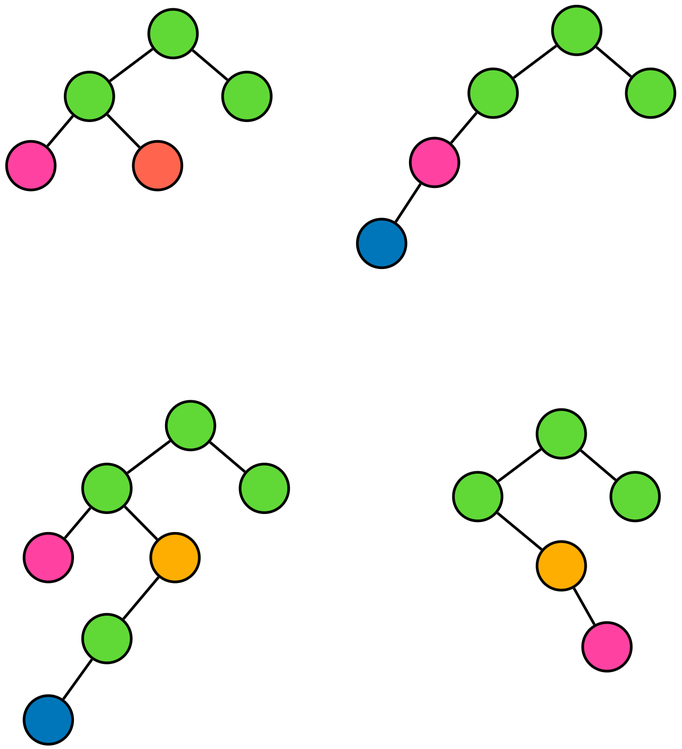
$$\Delta(B) = \mathrm{supp}(B) - \max_{\substack{B_2 \subseteq M \\ B \subset B_2}} \mathrm{supp}(B_2)$$

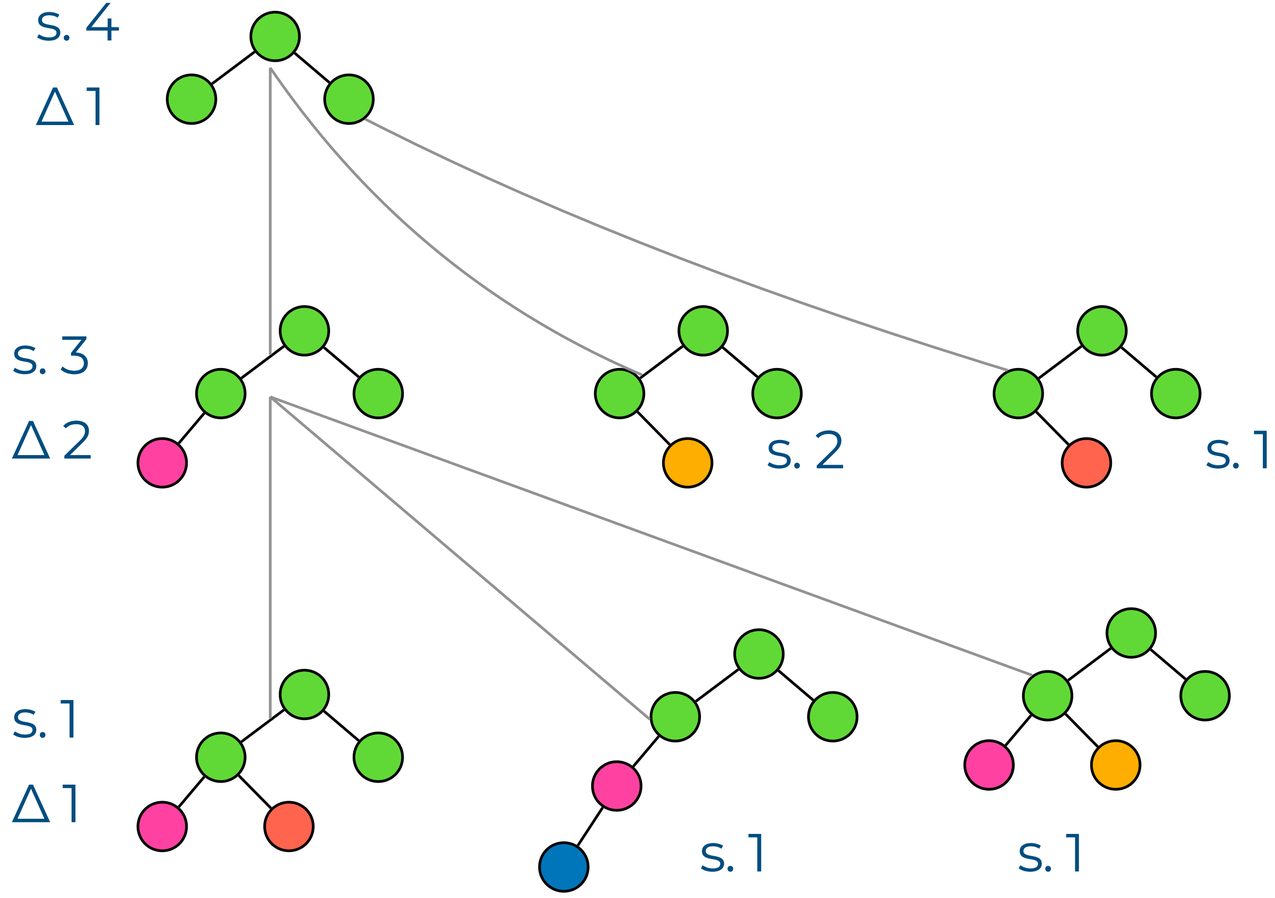
Pattern Structures

$$\Delta(D) = \mathrm{supp}(D) - \max_{\substack{D_2 \in \mathbb{D} \\ D \sqsubset D_2}} \mathrm{supp}(D_2)$$

A. Buzmakov et al., ICFCA, 2014

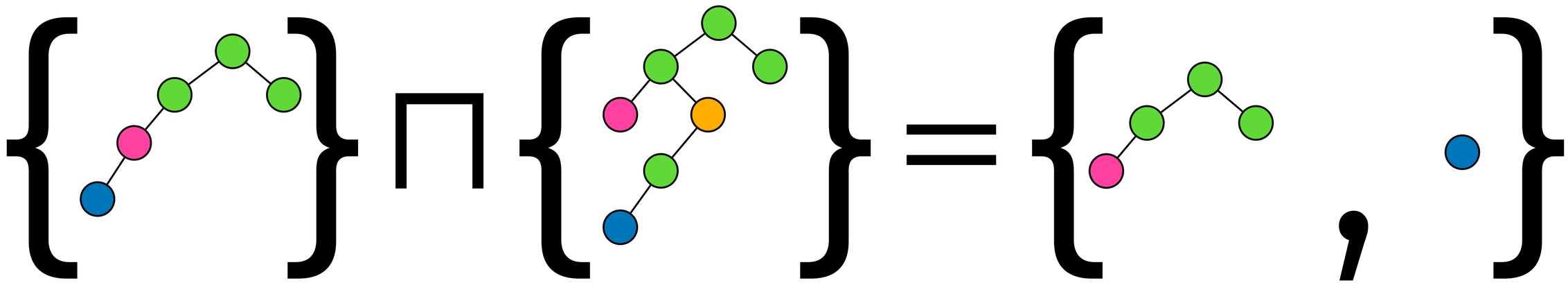
Stability of graphs





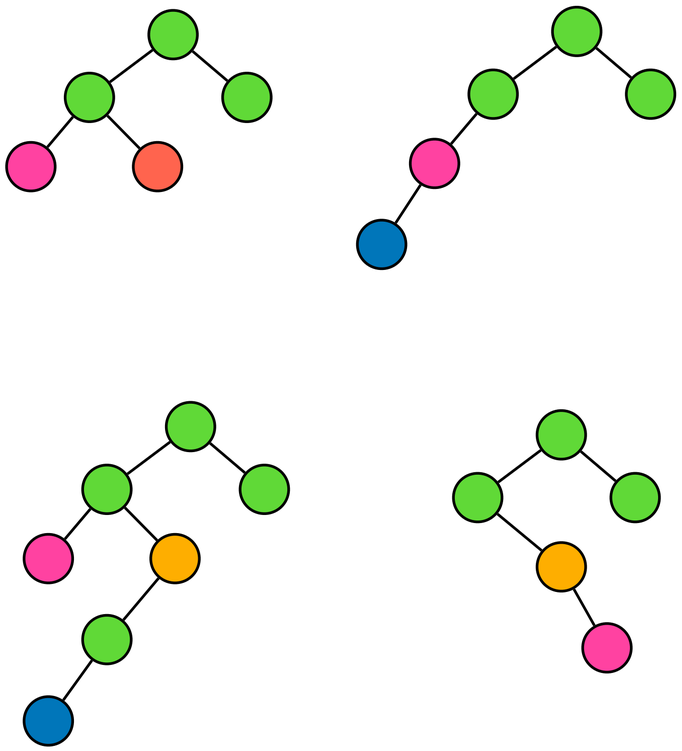
A small problem

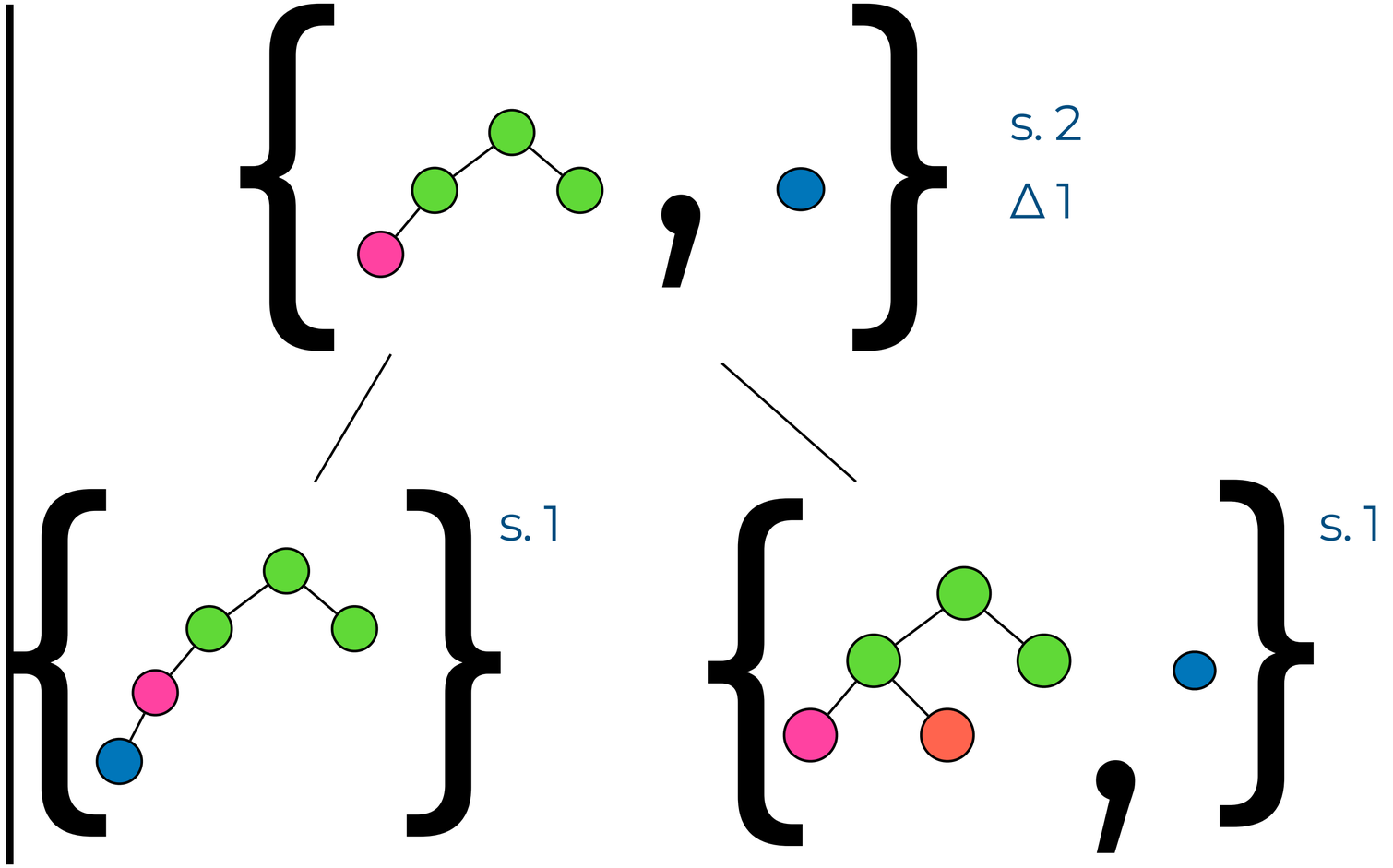
Graphs do not form a lattice



Antichains of graphs do

Stability of sets of graphs





Projection monotonicity

Sub-attributes $P \subseteq M$

$$\Delta\big(B \mid P\big) = \mathrm{supp}(B \cap P) - \max_{\substack{B_2 \subseteq P \\ B \subset B_2}} \mathrm{supp}(B_2)$$

Attributes M

$$\Delta\big(B\big) = \mathrm{supp}(B) - \max_{\substack{B_2 \subseteq M \\ B \subset B_2}} \mathrm{supp}(B_2)$$

$$\Delta(B \mid P) \geq \Delta(B)$$

$$\Delta(B \mid P) < \Delta_{\min} \implies \Delta(B) < \Delta_{\min}$$

 A. Buzmakov et al., ECML PKDD, 2015

SOFIA algorithm

	m1	m2	m3	m4	…	m_n
g1	X	X	X			
g2	X	X	X	X		
g3	X	X	X			
g4	X	X	X	X		
g5	X	X	X			
g6	X	X	X	X		
g7	X	X	X			
g8	X	X				X

1. Start with $M_0 = \{\}, L_0 = \{\varnothing\}$
2. For $i = 1, \ldots, n$:
   1. $L_i = L_{i-1} \cup L_{i-1} \times \{m_i\}$
   2. $\Delta_i(B) = \text{supp}(B) - \max_{m \in M_i \setminus B} \text{supp}(B \cup \{m\})$
   3. $L_i = \{B \in L_i \mid \Delta_i(B) \geq \Delta_{\min}\}$

+ an optimisation of ΔStability computation

 A. Buzmakov et al., ECML PKDD, 2015

Chain of projections

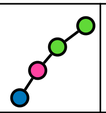
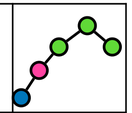
$$\psi_0 : D \mapsto \top \ < \ \psi_1 \ < \ \psi_2 < \ldots < \psi_n : D \mapsto D$$

Every projection ψ is a mapping $\psi : \mathbb{D} \to \mathbb{D}$ on the partial order $(\mathbb{D}, \sqsubseteq)$, which is a kernel (interior) operator, i.e. ψ is:

- Monotone $(x \sqsubseteq y) \mapsto (\psi(x) \sqsubseteq \psi(y))$
- Contractive $(\psi(x) \sqsubseteq x)$, and
- Idempotent $\left(\psi(\psi(x)) = \psi(x)\right)$

 A. Buzmakov et al., ECML PKDD, 2015

gSOFIA algorithm

					...	
g1	X	X	X			
g2	X	X	X	X		
g3	X	X	X			
g4	X	X	X	X		
g5	X	X	X			
g6	X	X	X	X		
g7	X	X	X			
g8	X	X				X

1. Start with $M_0 = \{\}, L_0 = \{\ \top\ \}$
2. For $i = 1, \ldots, n$:
   1. $L_i = L_{i-1} \ \cup \ L_{i-1} \times \{m\}$
   2. $\Delta_i(D) = \mathrm{supp}(D) - \max_{m \in M_i \setminus B} \mathrm{supp}(D \sqcup \{m\})$
   3. $L_i = \{D \in L_i \mid \Delta_i(D) \geq \Delta_{\min}\}$

+ an optimisation of ΔStability computation

+ an optimisation of attribute iteration

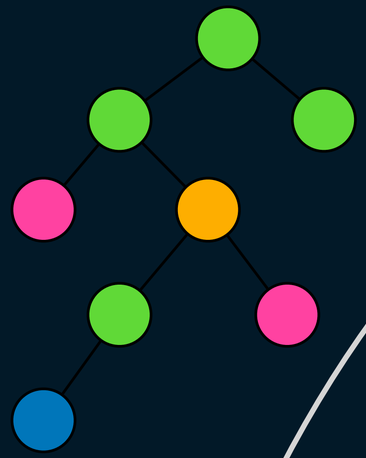
A. Buzmakov et al., ICDM, 2017

Problem 3

1. How to get AMR graph?

Document.txt

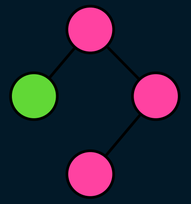
AMR Graph



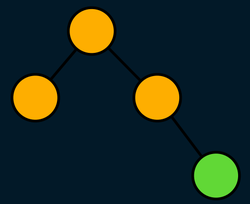
2. How to mine reference graphs?

Sports

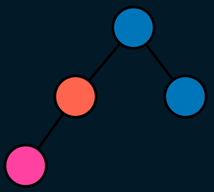
Business



Food



Others



3. How to classify accurately?

√Sports

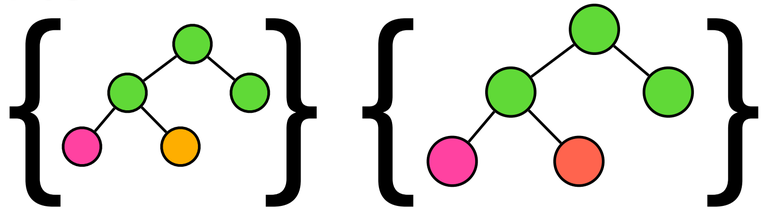
–Business

–Food

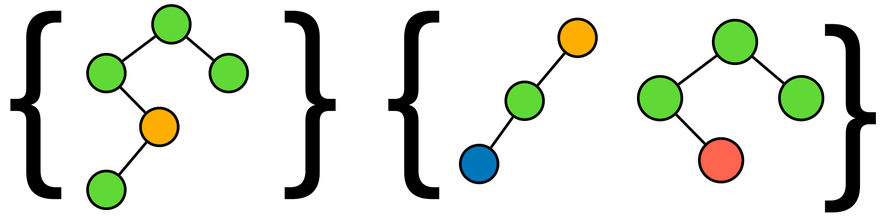
–Others

How to make a prediction?

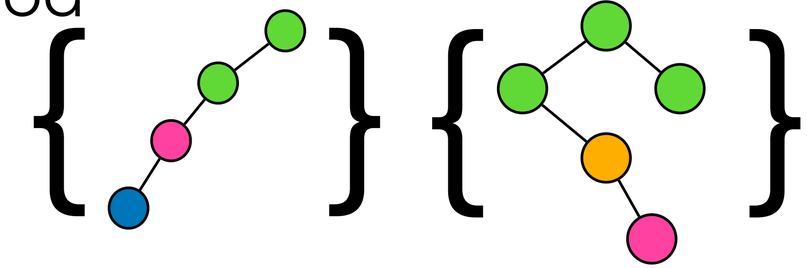
Sports



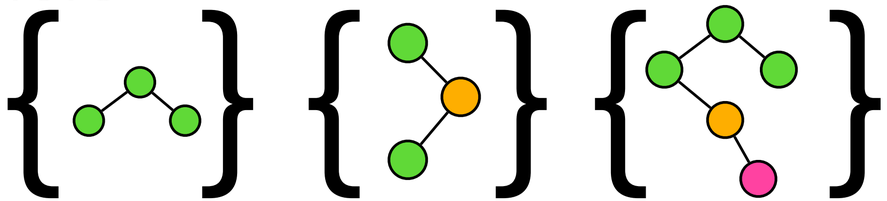
Business



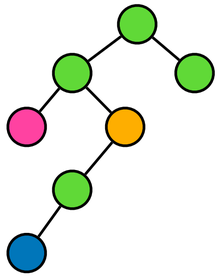
Food



Others

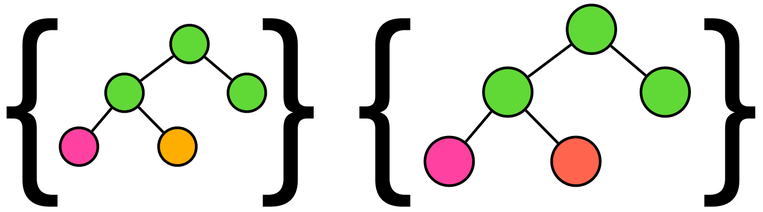


For a graph



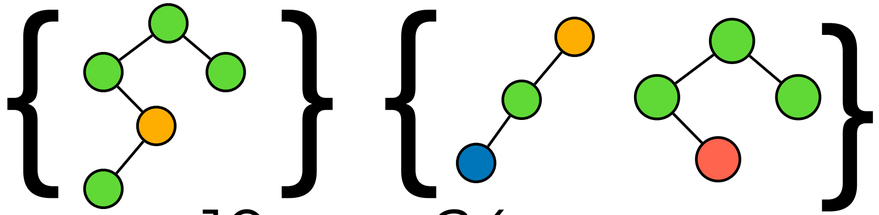
Step 0. Compute a score for every description

Sports



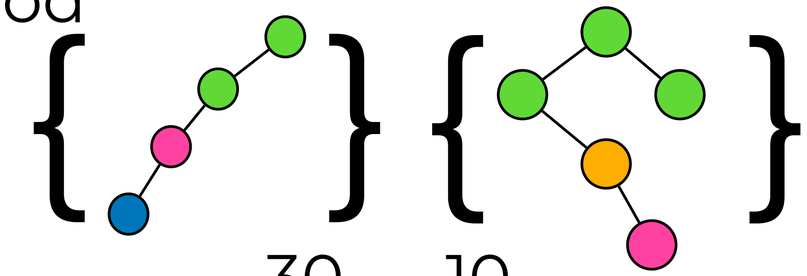
20 10

Business



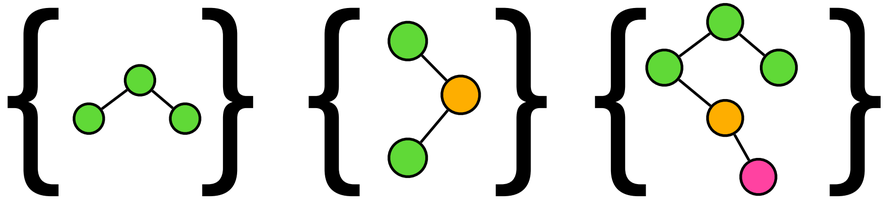
10 24

Food



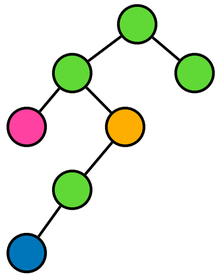
30 10

Others



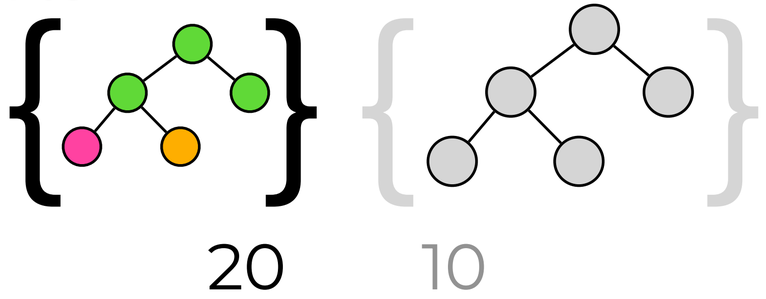
20 6 3

For a graph



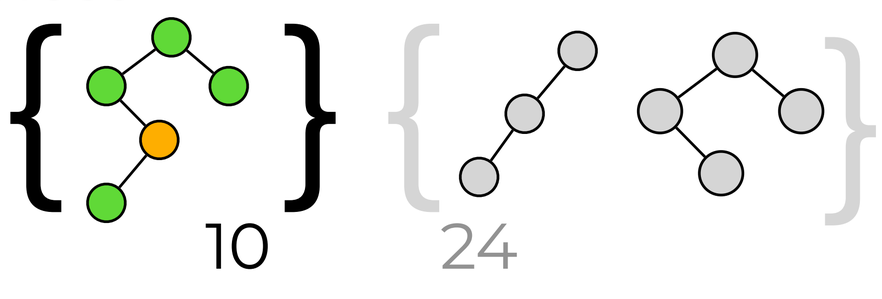
Step 1. Select applicable descriptions

Sports



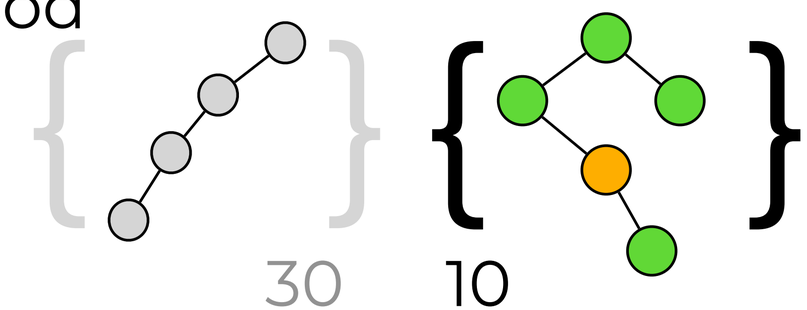
20 10

Business



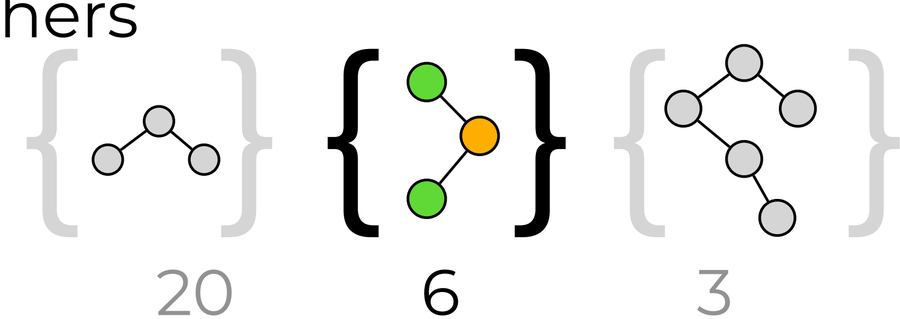
10 24

Food



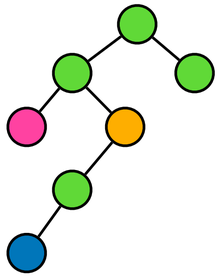
30 10

Others



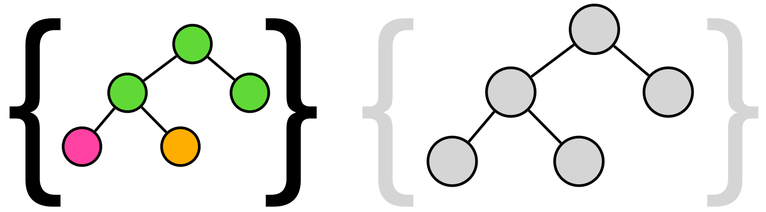
20 6 3

For a graph



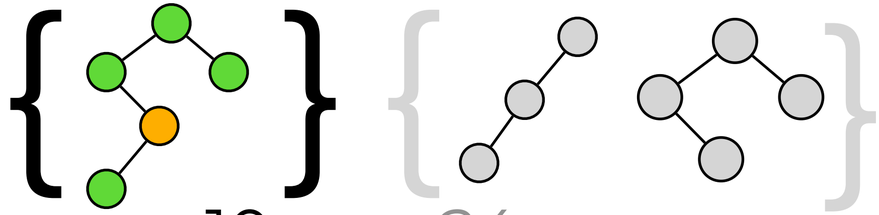
Step 2. Compute integral score per class

Sports $10 = 20/2$



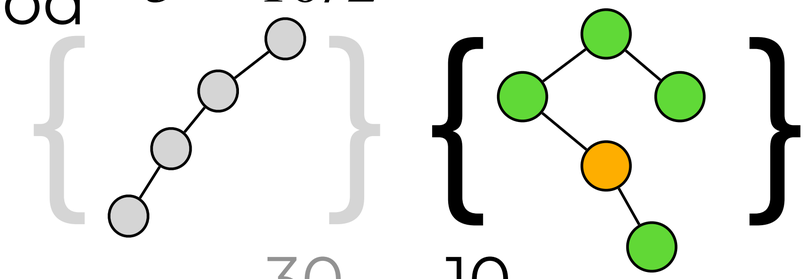
20 10

Business $10 = 20/2$



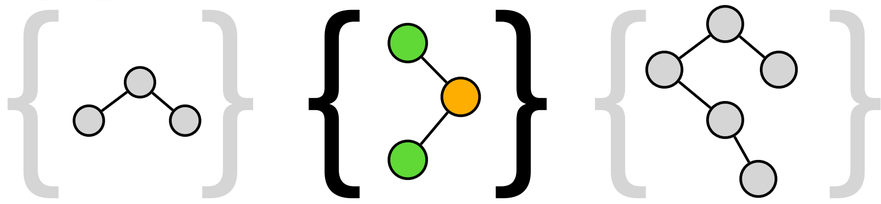
10 24

Food $5 = 10/2$



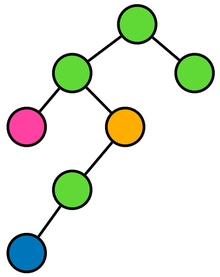
30 10

Others $2 = 6/3$



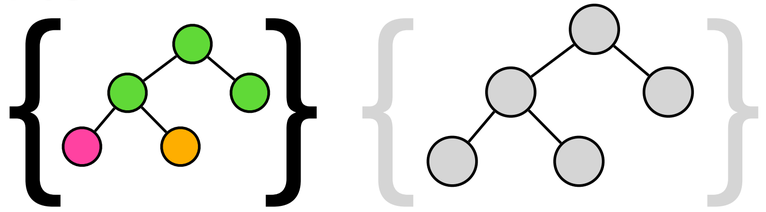
20 6 3

For a graph



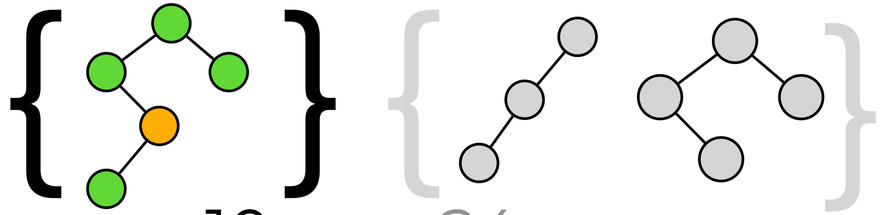
Step 3. Select the class with the maximal score

Sports $10 = 20/2$



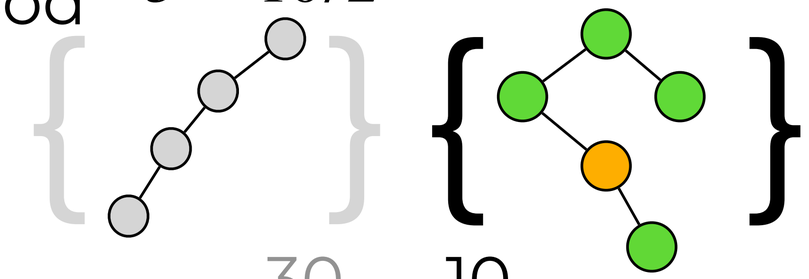
20 10

Business $10 = 20/2$



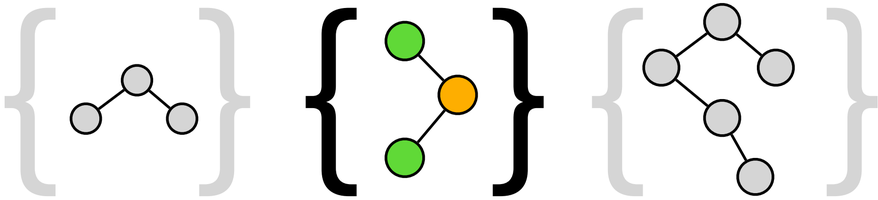
10 24

Food $5 = 10/2$



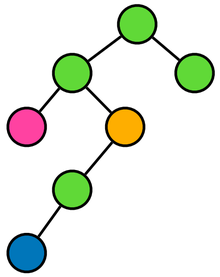
30 10

Others $2 = 6/3$



20 6 3

For a graph

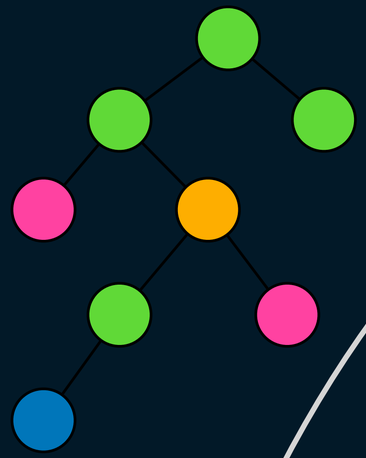


Experiments

1. How to get AMR graph?

Document.txt

AMR Graph



2. How to mine reference graphs?

3. How to classify accurately?

✓ **Sports**

– Business

– Food

– Others

Datasets

10 newsgroups data

1000 documents

10 classes

BBC Sport data

737 documents

5 classes

Results

Dataset	Logistic regression	GCN	SVM
10 newsgroups	**0.9502**	0.8478	0.8018
BBC sport	0.9732	0.8906	**0.9733**

Dataset	$penalty_{\text{baseline}}$	$penalty_1$	$penalty_2$	$penalty_3$	$penalty_4$	$penalty_5$	$penalty_6$
10 newsgroups	0.6425	0.7303	0.7004	0.7004	0.2096	0.6946	**0.7398**
BBC sport	0.7884	**0.9503**	0.8855	0.8858	0.6829	0.8732	**0.9503**

Dataset	$penalty_{\text{baseline}}$	$penalty_1$	$penalty_2$	$penalty_3$	$penalty_4$	$penalty_5$	$penalty_6$
10 newsgroups	0.6714	**0.7957**	0.7735	0.7750	0.3244	0.7876	0.7739
BBC sport	0.7414	0.9227	0.9046	0.8941	0.5437	0.9084	**0.9300**

Dataset	$penalty_{\text{baseline}}$	$penalty_1$	$penalty_2$	$penalty_3$	$penalty_4$	$penalty_5$	$penalty_6$
10 newsgroups	0.7412	0.8246	0.8017	0.7767	0.6266	**0.8287**	0.8021
BBC sport	0.7013	0.9300	0.9300	0.9187	**0.9436**	0.8639	0.9300

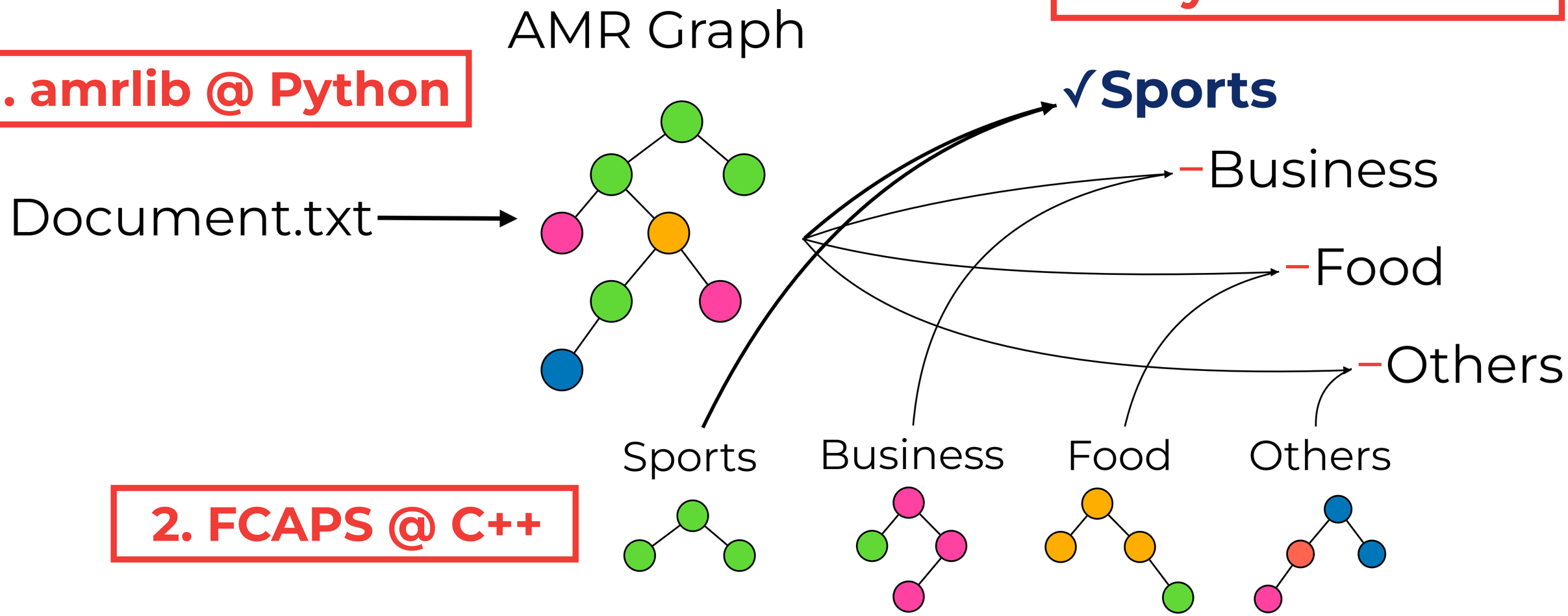
	SotA	**Our**
10 newsgroups	0.95	0.83
BBC Sport	0.97	0.95

Software

1. amrlib @ Python

2. FCAPS @ C++

3. Python



Software 2025

Caspailleur

Characteristic Attribute Sets -pailleur

Ready to use

GitHub PyPI

E.g.: Bob Ross Paintings (TD DM IDMC) via Google Colab

Next release: v0.1.4 “Human API” (this Sunday)

Paspailleur

Pattern Structures -pailleur

In active development

GitHub PyPI

E.g.: Mining stable patterns in complex data (TD DM IDMC) via Google Colab

Next release: v0.1.0 “Pattern Keys and Human API” (October ?)

Expailleur

Examples -pailleur

Examples of using Caspailleur and Paspailleur

GitHub

Document Classification via Stable Graph Patterns and Conceptual AMR Graphs

CONCEPTS conference, Cadiz, Spain, September 12, 2024

By Eric George Parakal, **Egor Dudyrev**, Sergei O. Kuznetsov, Amedeo Napoli