

# Taxonomy Extraction for Customer Service Knowledge Base Construction

**Bianca Pereira**<sup>1</sup>, Cécile Robin<sup>1</sup>, Tobias Daudert<sup>1</sup>, John P. McCrae<sup>1</sup>, Pranab Mohanty<sup>2</sup> and Paul Buitelaar<sup>1</sup>

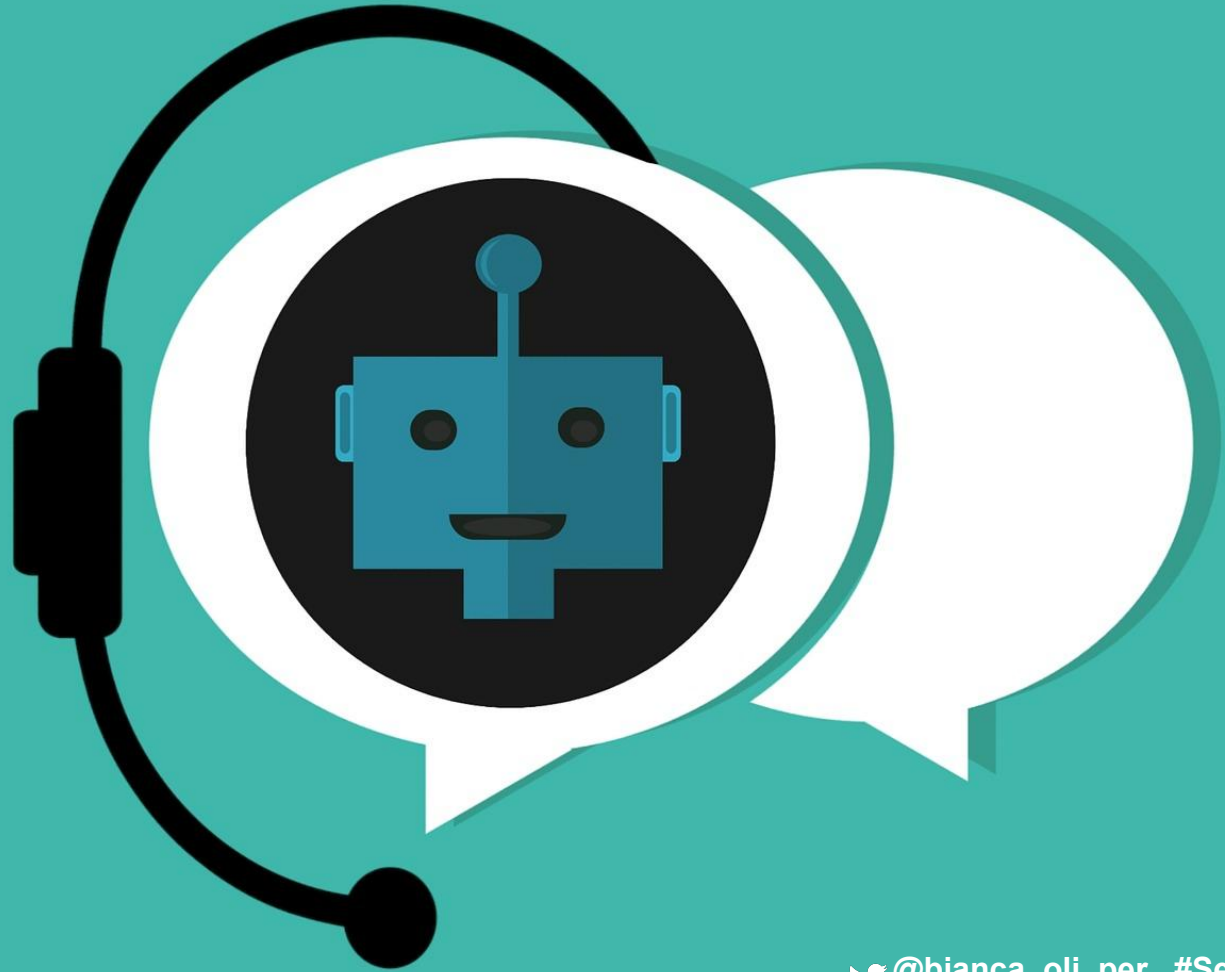
<sup>1</sup> *Insight Centre for Data Analytics. Data Science Institute. National University of Ireland Galway.*

<sup>2</sup> *Fidelity Investments USA*





All our agents are busy at the moment.  
Please, wait in the line..



# Domain-specific Knowledge



# Relevant vs. Irrelevant topics



Good morning. How can I help you?

Hi, how can I deposit money from my old 401(k) plan into my retirement account?



I see you have a traditional IRA account. You may be interested in changing to a rollover IRA instead...

Good morning. How can I help you?

Hi, how can I deposit money from my old 401(k) plan into my retirement account?



I see you have a traditional IRA account. You may be interested in changing to a rollover IRA instead...

Good morning. How can I help you?

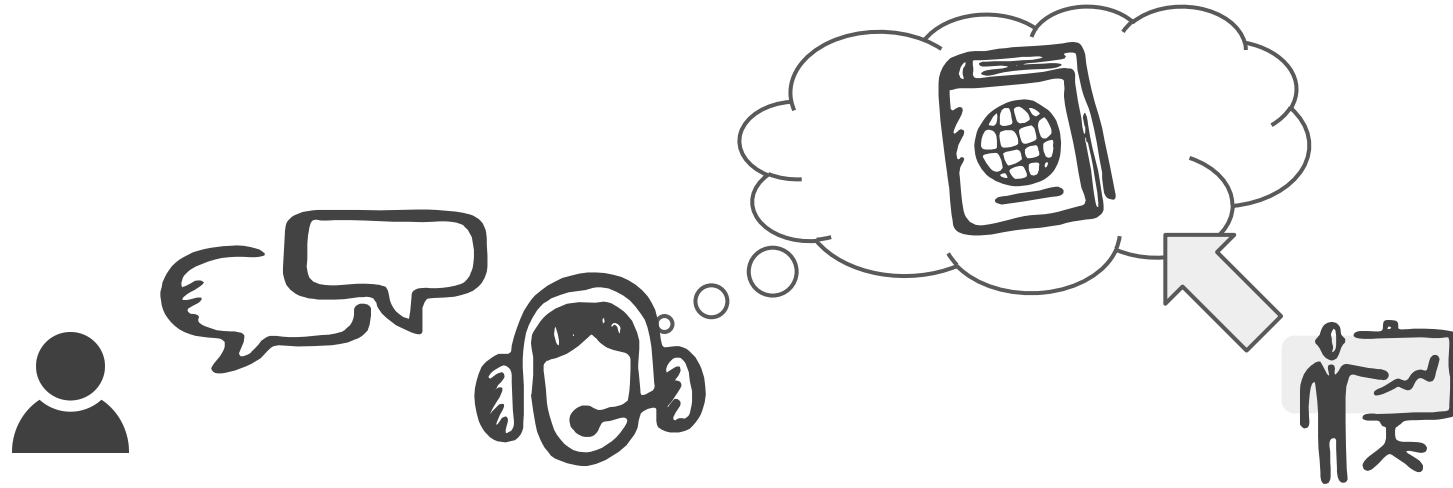
Hi, how can I deposit money from my old 401(k) plan into my retirement account?















I see you have a traditional IRA account. You may be interested in changing to a rollover IRA instead...



# Explainable/Modifiable



# Customer Agent Challenge

	Human	Chatbot [19]	KG-based Chatbot	
			General [7]	Domain [3]
Domain-specific Knowledge				
Relevant vs. Irrelevant Topics				
Explainable/Modifiable				

# Taxonomy Extraction for Customer Service Knowledge Base Construction

**Bianca Pereira**<sup>1</sup>, Cécile Robin<sup>1</sup>, Tobias Daudert<sup>1</sup>, John P. McCrae<sup>1</sup>, Pranab Mohanty<sup>2</sup> and Paul Buitelaar<sup>1</sup>

<sup>1</sup> *Insight Centre for Data Analytics. Data Science Institute. National University of Ireland Galway.*

<sup>2</sup> *Fidelity Investments USA*



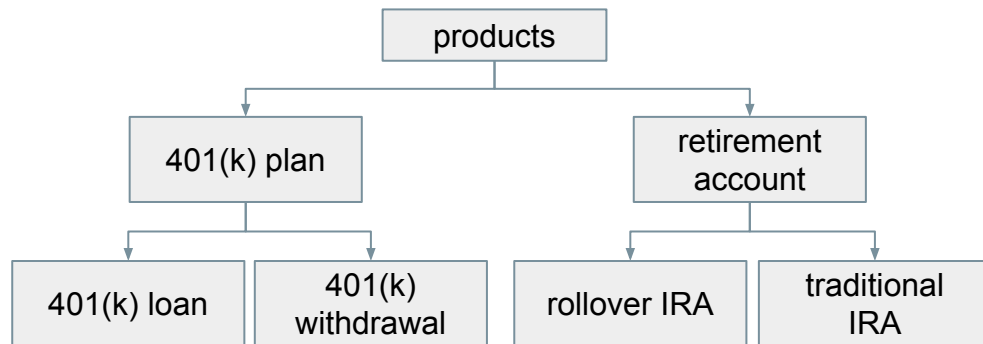
# Objective

Good morning. How can I help you?

Hi, how can I deposit money from my old 401(k) plan into my retirement account?



I see you have a traditional IRA account. You may be interested in changing to a rollover IRA instead...



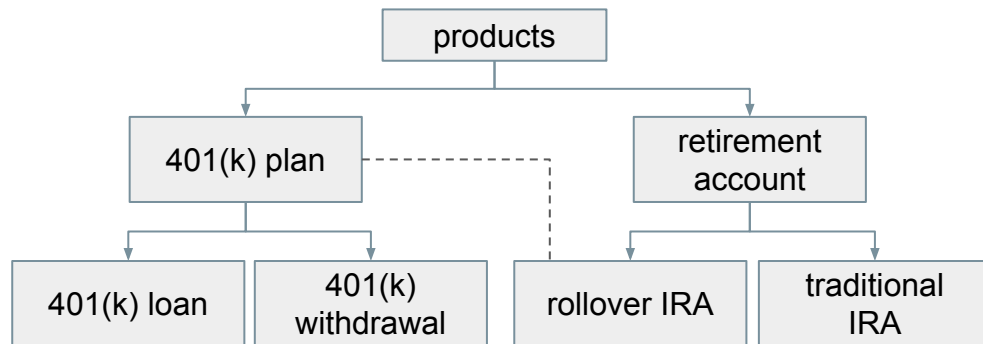
# Objective

Good morning. How can I help you?

Hi, how can I deposit money from my old 401(k) plan into my retirement account?



I see you have a traditional IRA account. You may be interested in changing to a rollover IRA instead...



# Objective

Domain-specific  
Knowledge



Relevant vs. Irrelevant  
Topics

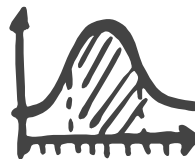
Explainable/Modifiable

# Objective

Domain-specific  
Knowledge



Relevant vs. Irrelevant  
Topics



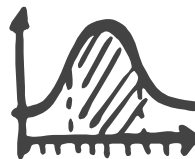
Explainable/Modifiable

# Objective

Domain-specific  
Knowledge



Relevant vs. Irrelevant  
Topics



Explainable/Modifiable

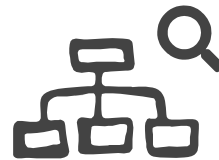




# Objective



Domain-specific

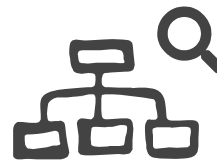
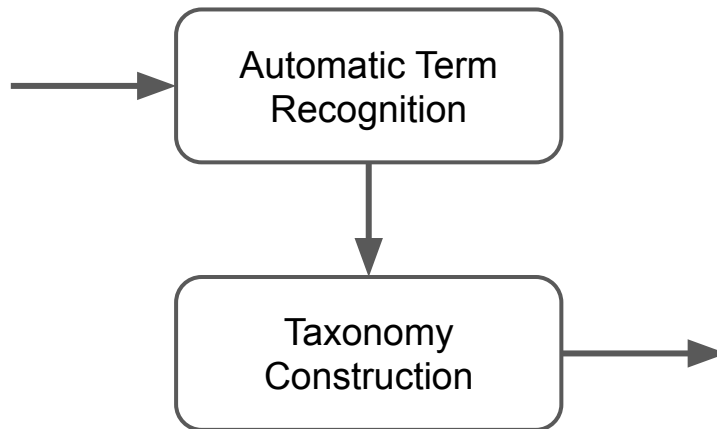


Auditable by the  
organisation

# Objective

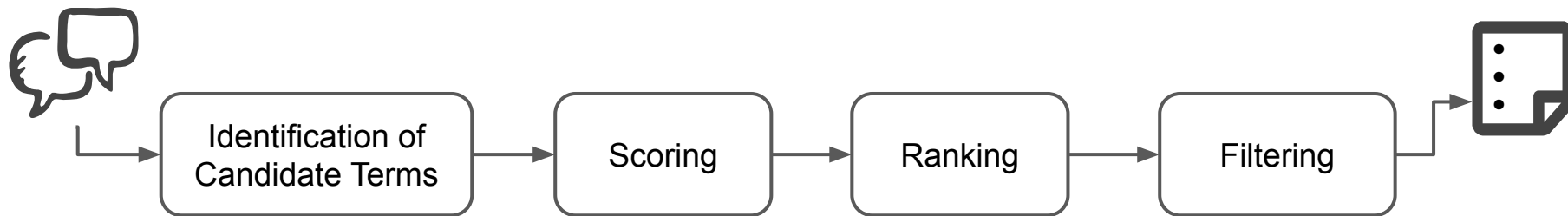


Domain-specific



Auditable by the  
organisation

# Automatic Term Recognition



# Automatic Term Recognition



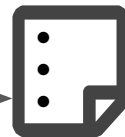
Identification of  
Candidate Terms

- Minimum frequency
- Part-of-speech patterns [9, 20]
- N-gram size
- Do not start or end with stopword [20]

Scoring

Ranking

Filtering



Hi, how can I deposit money from my old 401(k) plan into my retirement account?



retirement account

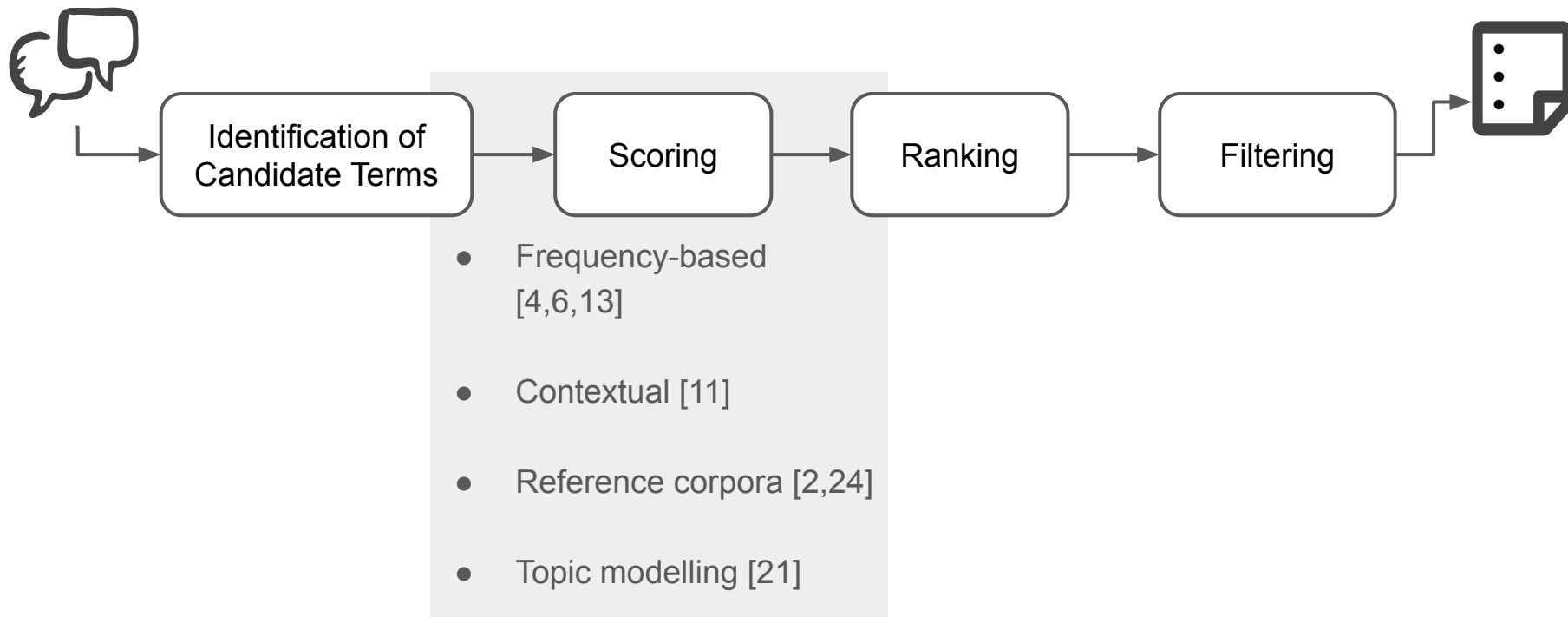


can I

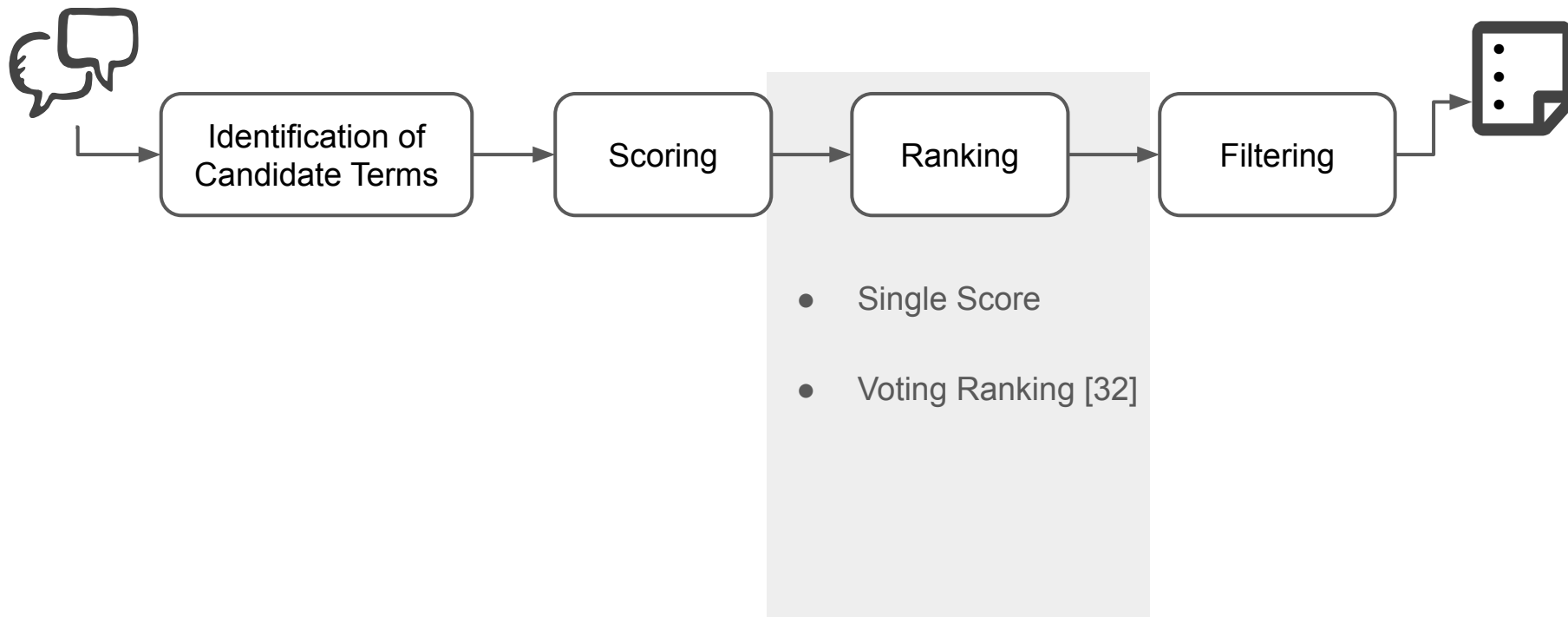


money from my old 401(k) plan into my retirement account

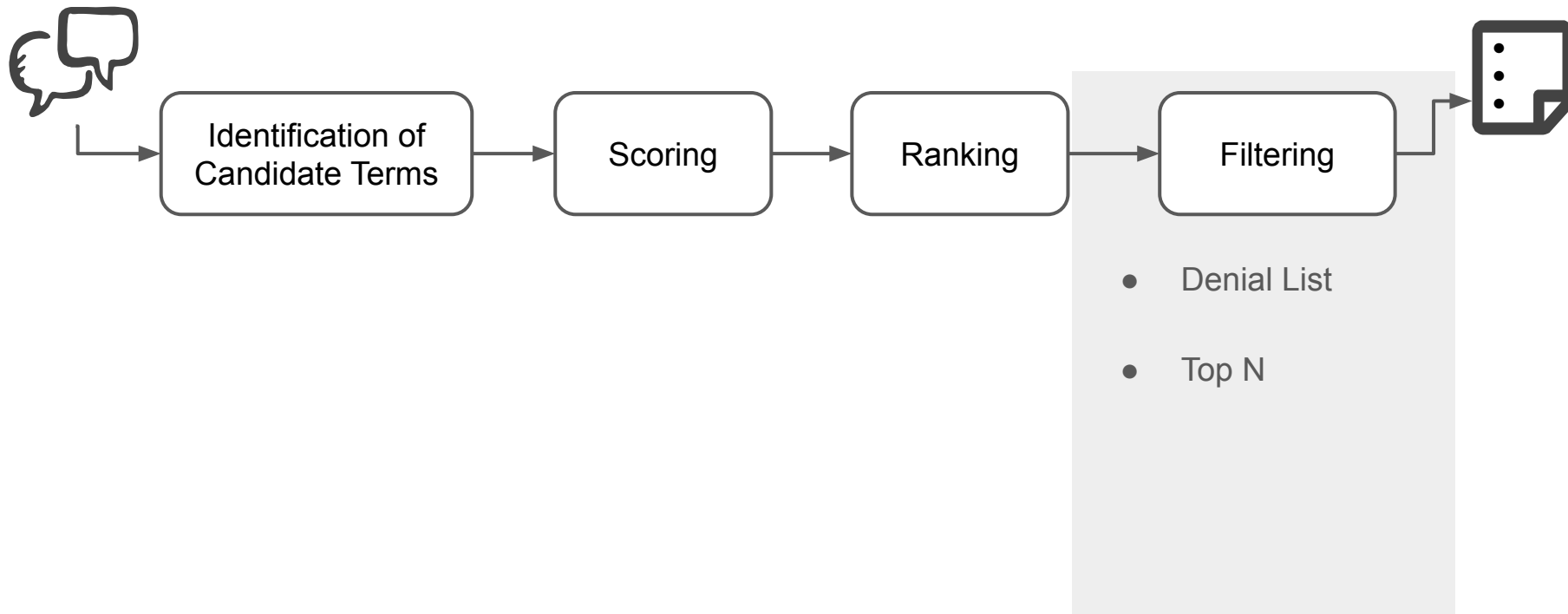
# Automatic Term Recognition



# Automatic Term Recognition



# Automatic Term Recognition



# Experiments

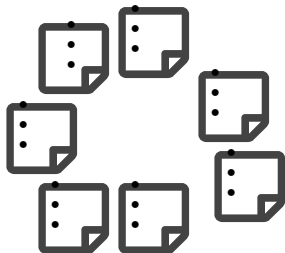
Experiment	n-gram		Scoring Function				Ranking
	Min	Max	Frequency	Context	Reference Corpora	Topic Modelling	
TE baseline	1	4	tf-idf	-	-	-	Single
TE baseline long	2	4	tf-idf	-	-	-	Single
TE <sub>1</sub>	1	4	tf-idf Residual Idf C Value ComboBasic	-	Weirdness	-	Voting
TE <sub>1</sub> long	2	4	tf-idf Residual Idf C Value ComboBasic	-	Weirdness	-	Voting
TE <sub>2</sub>	1	4	ComboBasic	-	-	-	Single
TE <sub>2</sub> long	2	4	ComboBasic	-	-	-	Single
TE <sub>3</sub> long	2	4	ComboBasic	-	Relevance	-	Voting
TE <sub>4</sub>	1	4	-	PostRankDC	-	-	Single
TE <sub>4</sub> long	2	4	-	PostRankDC	-	-	Single
TE <sub>5</sub>	1	4	-	-	-	NTM	Single
TE <sub>5</sub> long	2	4	-	-	-	NTM	Single



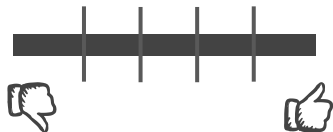
# Evaluation



300,000 anonymised chatlog conversations



Top 200 terms by Mean Reciprocal Rank



Manual validation by experts in the financial customer service domain

# Evaluation Results

Experiment	n-gram		Scoring Function				Ranking
	Min	Max	Frequency	Context	Reference Corpora	Topic Modelling	
TE baseline	1	4	tf-idf	-	-	-	Single
TE baseline long	2	4	tf-idf	-	-	-	Single
TE <sub>1</sub>	1	4	tf-idf Residual Idf C Value ComboBasic	-	Weirdness	-	Voting
TE <sub>1</sub> long	2	4	tf-idf Residual Idf C Value ComboBasic	-	Weirdness	-	Voting
TE <sub>2</sub>	1	4	ComboBasic	-	-	-	Single
TE <sub>2</sub> long	2	4	ComboBasic	-	-	-	Single
TE <sub>3</sub> long	2	4	ComboBasic	-	Relevance	-	Voting
TE <sub>4</sub>	1	4	-	PostRankDC	-	-	Single
TE <sub>4</sub> long	2	4	-	PostRankDC	-	-	Single
TE <sub>5</sub>	1	4	-	-	-	NTM	Single
TE <sub>5</sub> long	2	4	-	-	-	NTM	Single

Experiment	Coverage (%)	Precision (%)
TE baseline	36	48.6
TE baseline long	38.5	59.7
TE <sub>1</sub>	47	50.0
TE <sub>1</sub> long	41.5	59.0
TE <sub>2</sub>	30.5	54.1
TE <sub>2</sub> long	37.5	<b>65.3</b>
TE <sub>3</sub> long	37.5	<b>65.3</b>
TE <sub>4</sub>	45	56.7
TE <sub>4</sub> long	30.5	62.3
TE <sub>5</sub>	21.5	30.2
TE <sub>5</sub> long	21.5	30.2

# Automatic Term Recognition - Challenges

- Conversational Terms

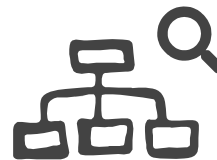
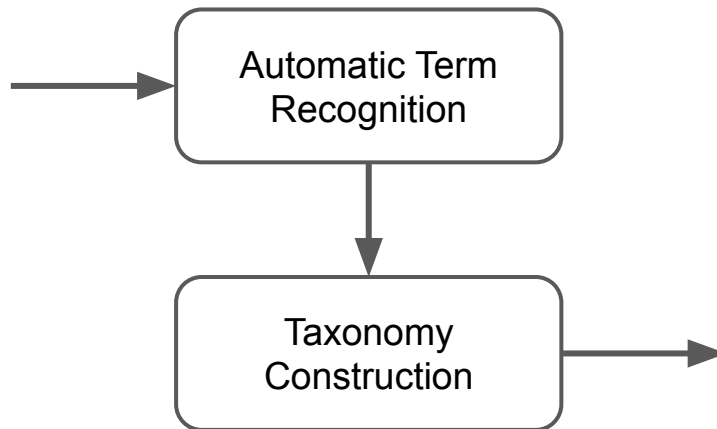
“good morning”, “thanks”, ...

- Detect topics with different n-grams length
- Terms with numbers (e.g. 401(k) plan)

# Objective



Domain-specific



Auditable by the  
organisation

# Taxonomy Construction



Only one root

Unique parent

Reflexivity

Antisymmetry

Transitivity

# Taxonomy Construction



Only one root

Unique parent

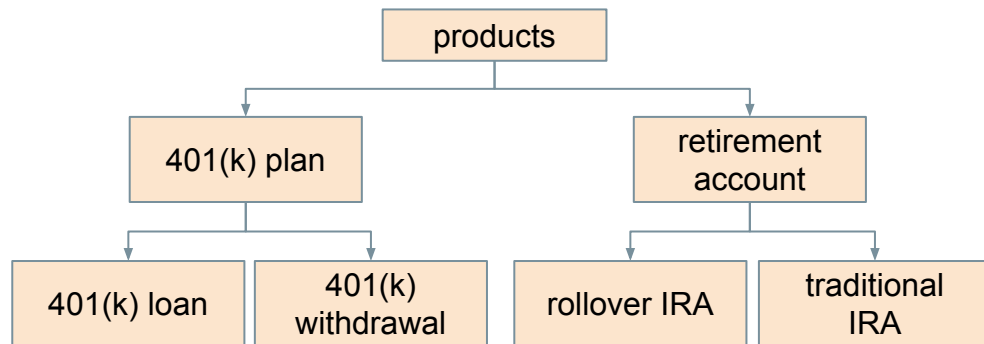
Reflexivity

Antisymmetry

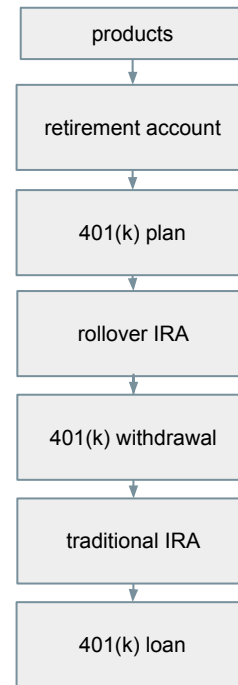
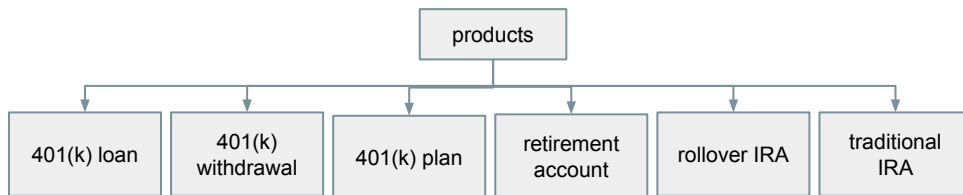
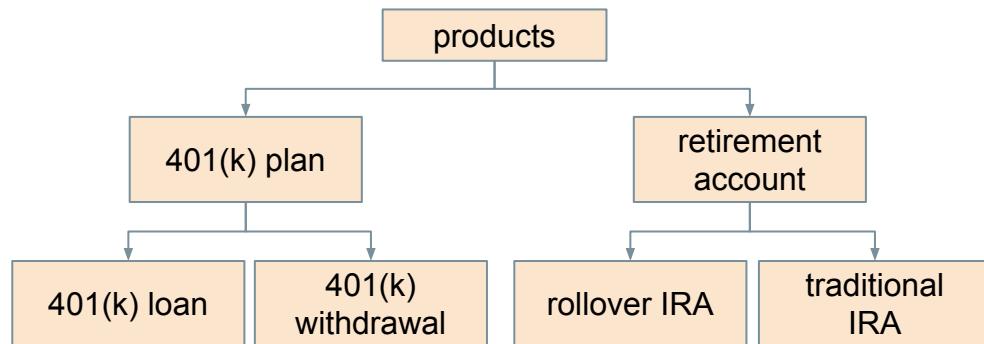
Transitivity

**Required but not sufficient**

# Taxonomy Construction



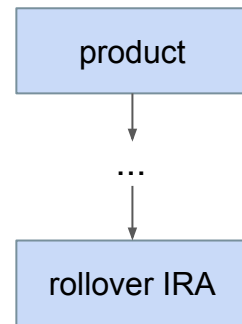
# Taxonomy Construction





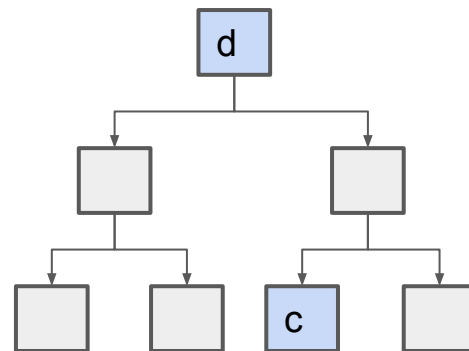
# Pairwise Probability

$p(\text{rollover IRA} \sqsubseteq \text{product})$



# Pairwise Probability

$$p(c \sqsubseteq d)$$



# Taxonomy Likelihood Score - Transitive

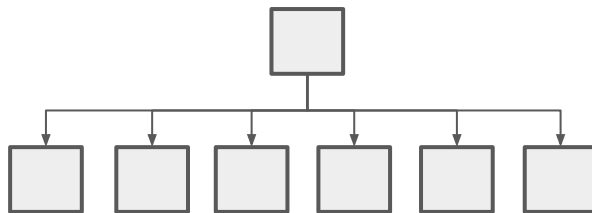
$$\max_T S(T)$$

$$S(T) = \prod_{c \sqsubseteq d} p(c \sqsubseteq d)$$

# Taxonomy Likelihood Score - Transitive

$$\max_T S(T)$$

$$S(T) = \prod_{c \sqsubseteq d} p(c \sqsubseteq d)$$

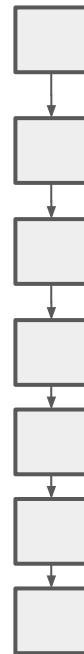


# Taxonomy Likelihood Score - Intransitive

$$S_{nonTransitive}(T) = \prod_{c \leq d} p(c \sqsubseteq d)$$

# Taxonomy Likelihood Score - Intransitive

$$S_{nonTransitive}(T) = \prod_{c \leq d} p(c \sqsubseteq d)$$

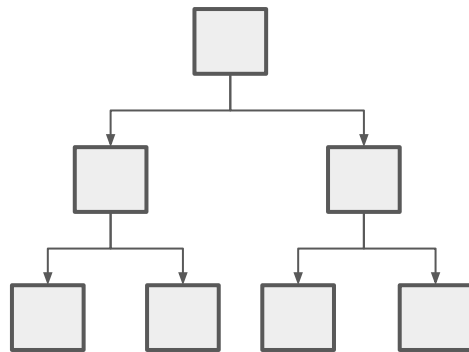


# Taxonomy Likelihood Score - Battacharyya-Poisson

$$S_{BP}(T) = \prod_{c \leq d} p(c \sqsubseteq d) + \alpha N \times BP(T)$$

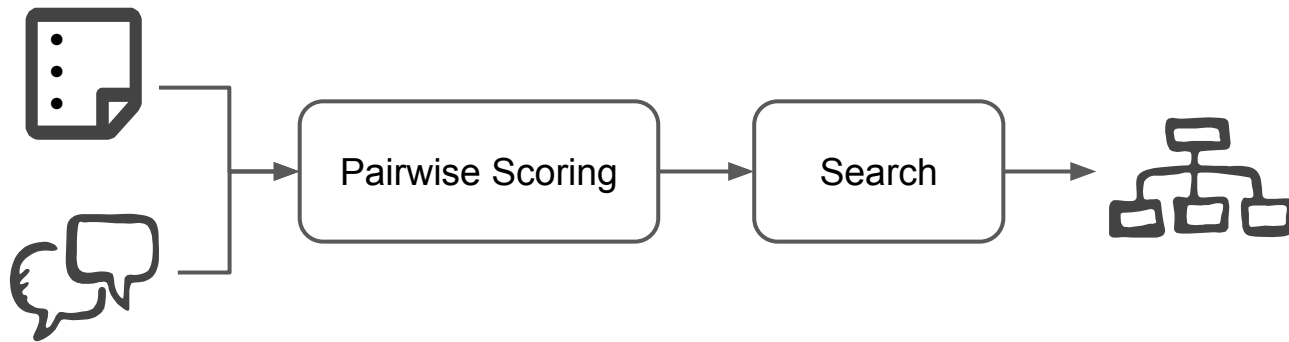
# Taxonomy Likelihood Score - Battacharyya-Poisson

$$S_{BP}(T) = \prod_{c \leq d} p(c \sqsubseteq d) + \alpha N \times BP(T)$$

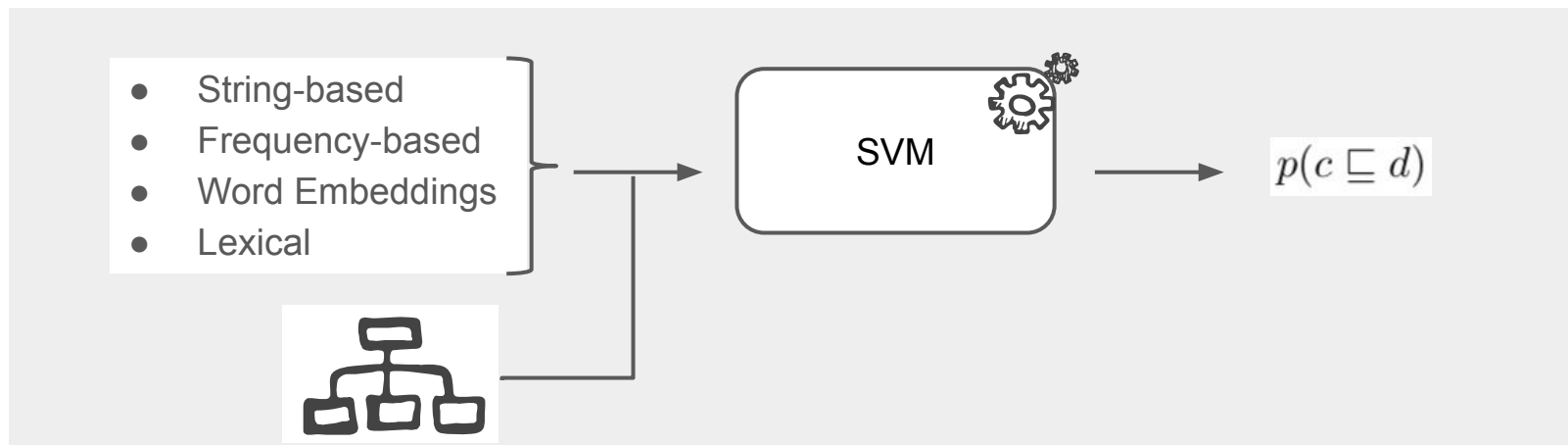
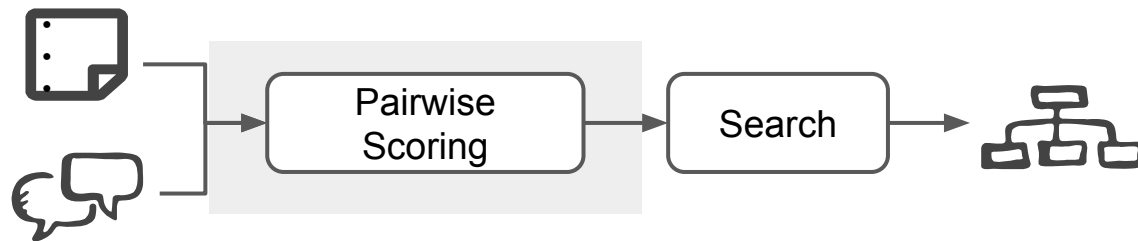




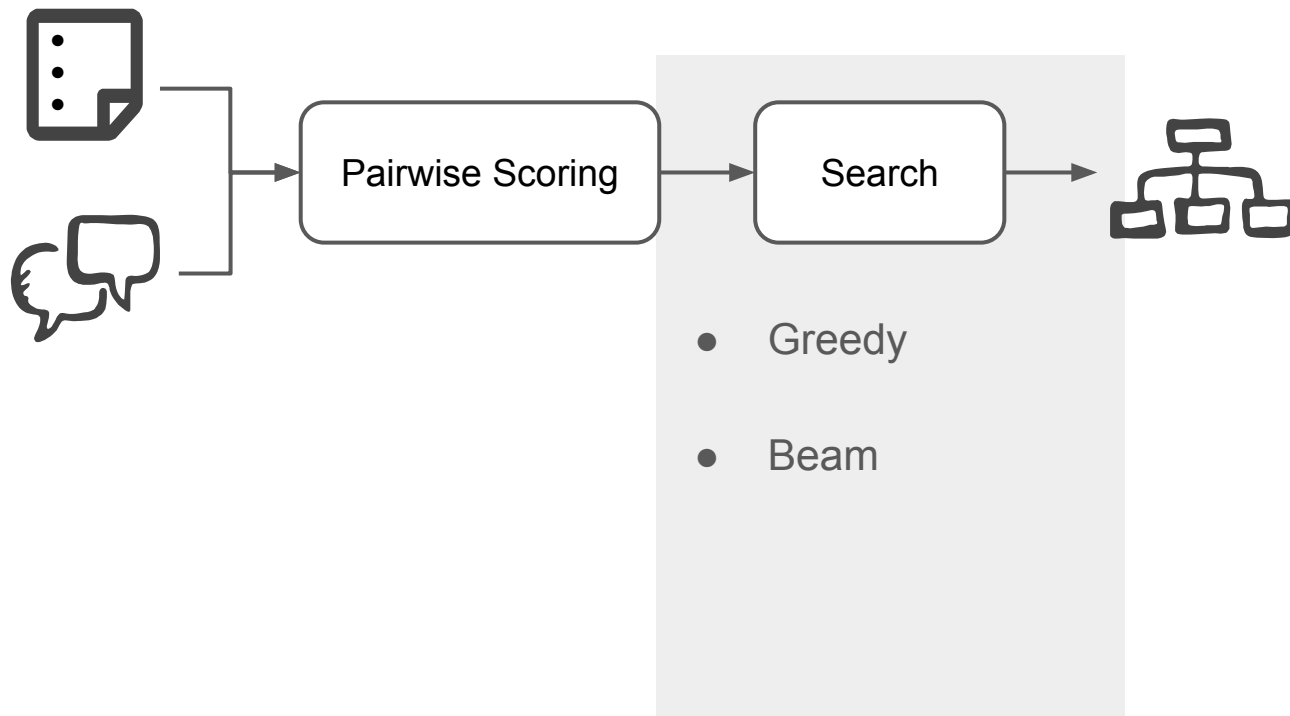
# Taxonomy Construction



# Taxonomy Construction



# Taxonomy Construction



# Evaluation Results

Search Method	Likelihood Score Function	Out-of-Domain Model			In-domain Model		
		Products	Sectors	Topics	Products	Sectors	Topics
Greedy	Transitive	16.86	11.25	11.41	7.28	8.75	7.38
	Non-Transitive	17.62	12.50	11.41	21.84	15.00	13.42
	BP	17.62	12.50	12.08	22.22↑	16.25↑	13.42↑
Beam	Transitive	4.21 ↓	1.25 ↓	1.34 ↓	3.83 ↓	1.25 ↓	0.67 ↓
	Non-Transitive	16.86	11.25	10.74	16.48	8.75	8.72
	BP	18.01	12.50	10.74	15.33	8.75	9.40

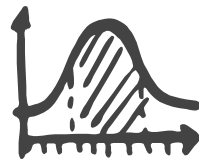
# Evaluation Results

Search Method	Likelihood Score Function	Out-of-Domain Model			In-domain Model		
		Products	Sectors	Topics	Products	Sectors	Topics
Greedy	Transitive	16.86	11.25	11.41	7.28	8.75	7.38
	Non-Transitive	17.62	12.50	11.41	21.84	15.00	13.42
	BP	17.62	12.50	12.08	22.22	16.25	13.42
Beam	Transitive	4.21	1.25	1.34	3.83	1.25	0.67
	Non-Transitive	16.86	11.25	10.74	16.48	8.75	8.72
	BP	18.01	12.50	10.74	15.33	8.75	9.40

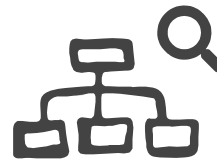
# Conclusion



Domain-specific



Identification of  
key concepts



Auditable by the  
organisation

Try it now!



<https://github.com/insight-centre/saffron>

