Information Retrieval in Context

A Case Study with the COVID-19 Pandemic

Bachelor's thesis presentation

Marcel Braasch

Goethe Universität Frankfurt

11/06/2020

Contents

- Introduction
- Motivation
- Aim
- Recap: vector space models
- Context-based IR approach
- Context

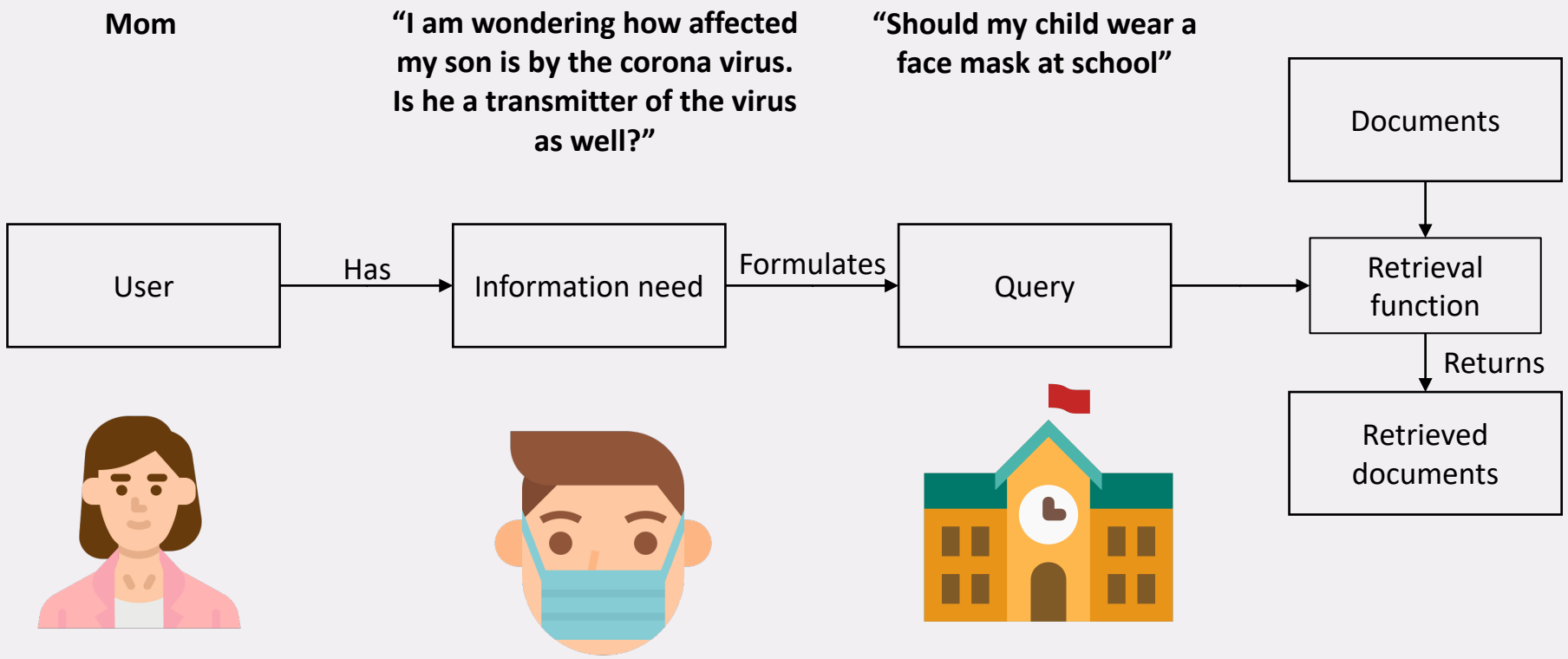
Introduction – Academics

- 2004 – 2015: Abitur in Butzbach
- 2011 – 2012: High school year in California, USA
- 2015 – 2018: Studies in business and physics (not completed)
- 2018 – 2020: Bachelor's in Computer Science with a minor in linguistics at Goethe University Frankfurt
- 2020 – now: Master's in Data Engineering and Analytics majoring in ML and NLP at Technical University Munich

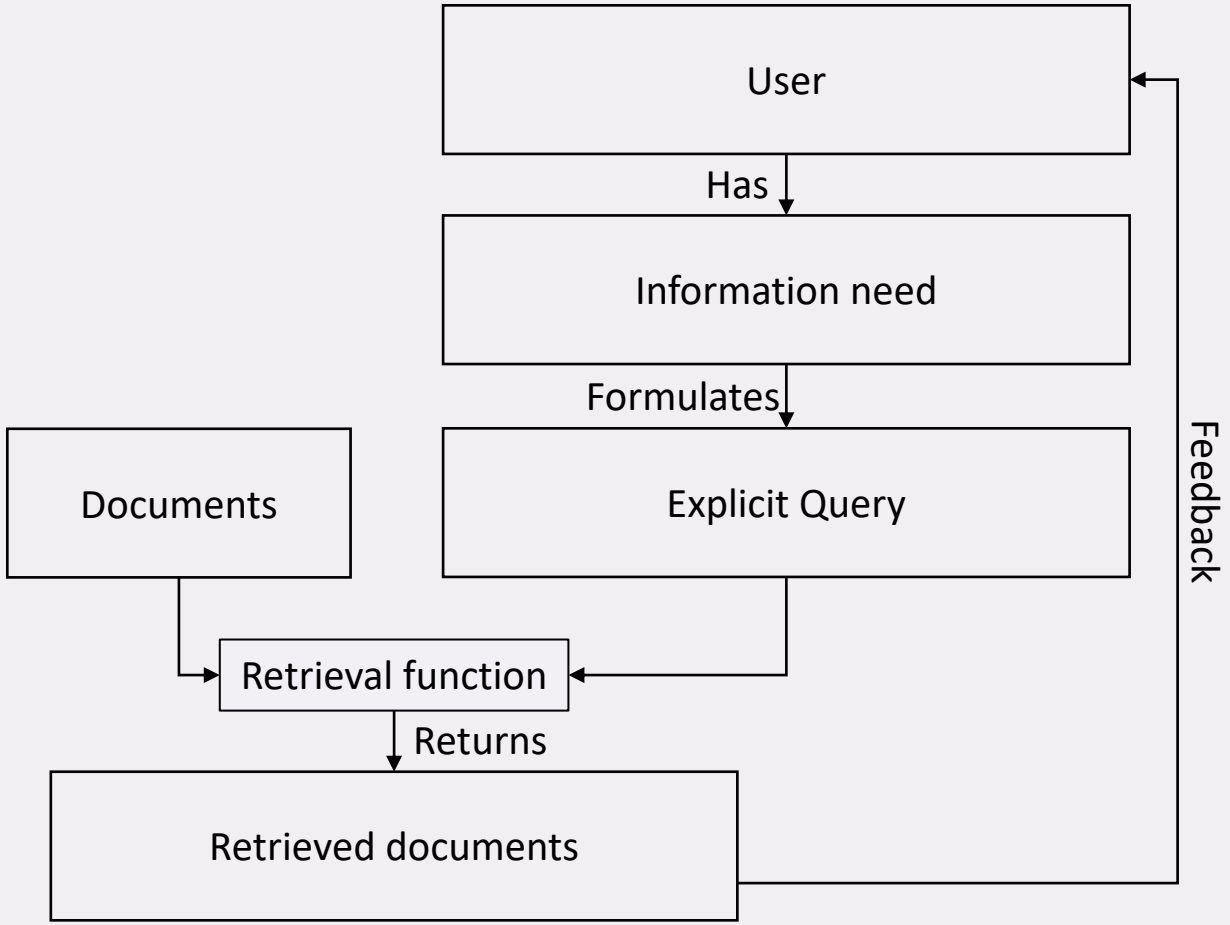
Motivation

- COVID-19 pandemic is a **new scenario** for most of us
- Uncertain how to **behave correctly**
- **Large volume** of information
- **Dynamic** information landscape
- People use **search tools** to complement their information need
- Search engines **possibly too general** to find fitting information
- Explicit search query **may not express intent** of the user
- Goal: find the context around the user

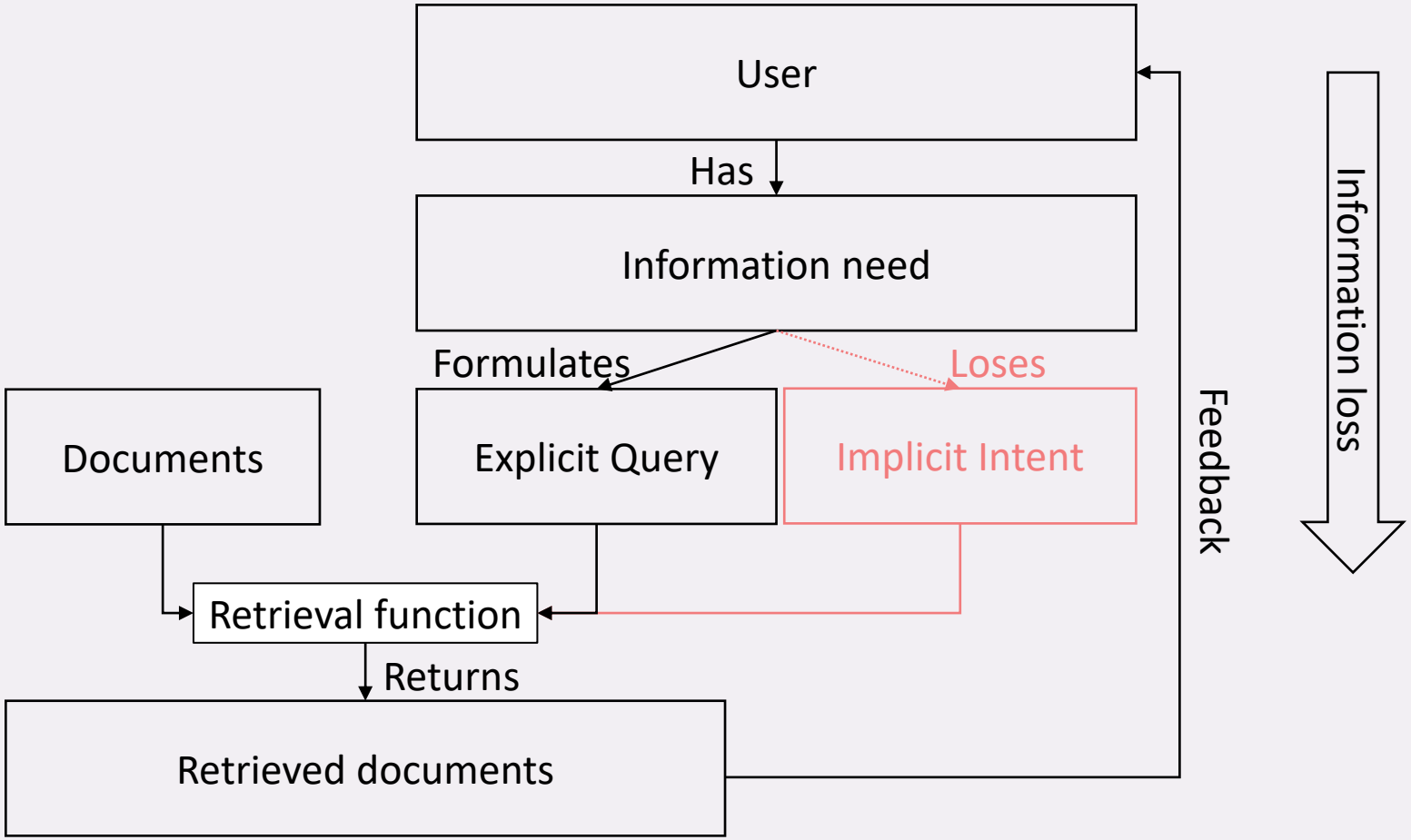
Aim


Mom
“I am wondering how affected my son is by the corona virus. Is he a transmitter of the virus as well?”
“Should my child wear a face mask at school”
User
Has
Information need
Formulates
Query
Documents
Retrieval function
Returns
Retrieved documents

Aim

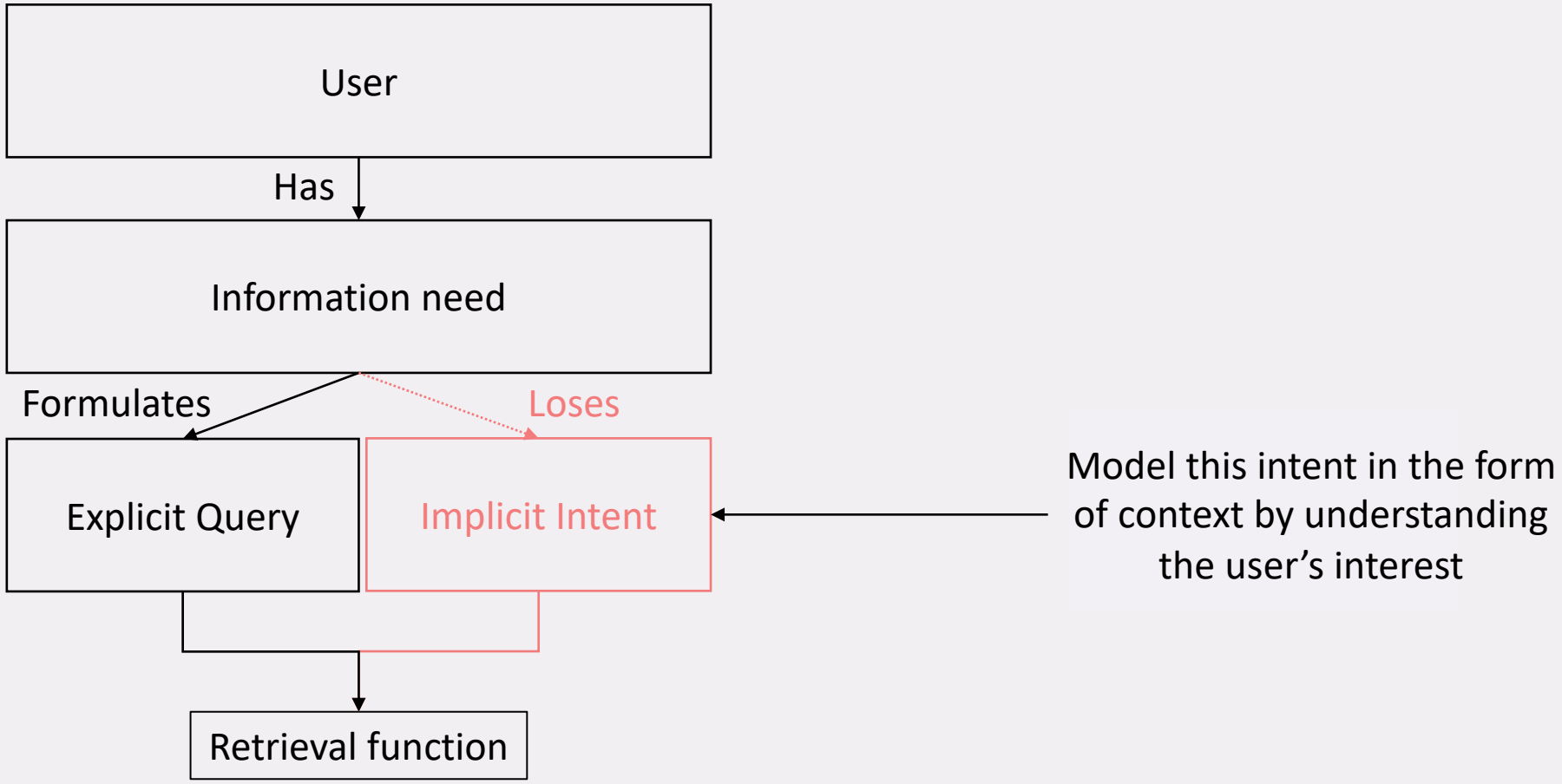

User
Has
Information need
Formulates
Documents
Explicit Query
Feedback
Retrieval function
Returns
Retrieved documents

Aim

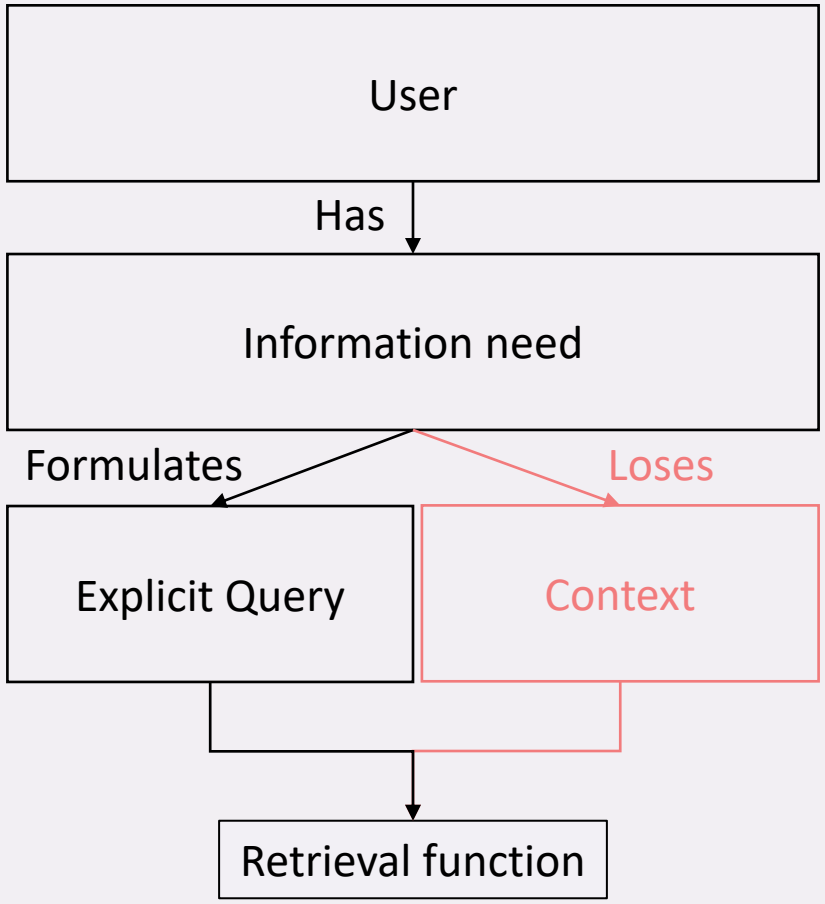


1. Ambiguity of words
2. Choosing words not present in target documents
3. Choosing incorrect words / missing background knowledge

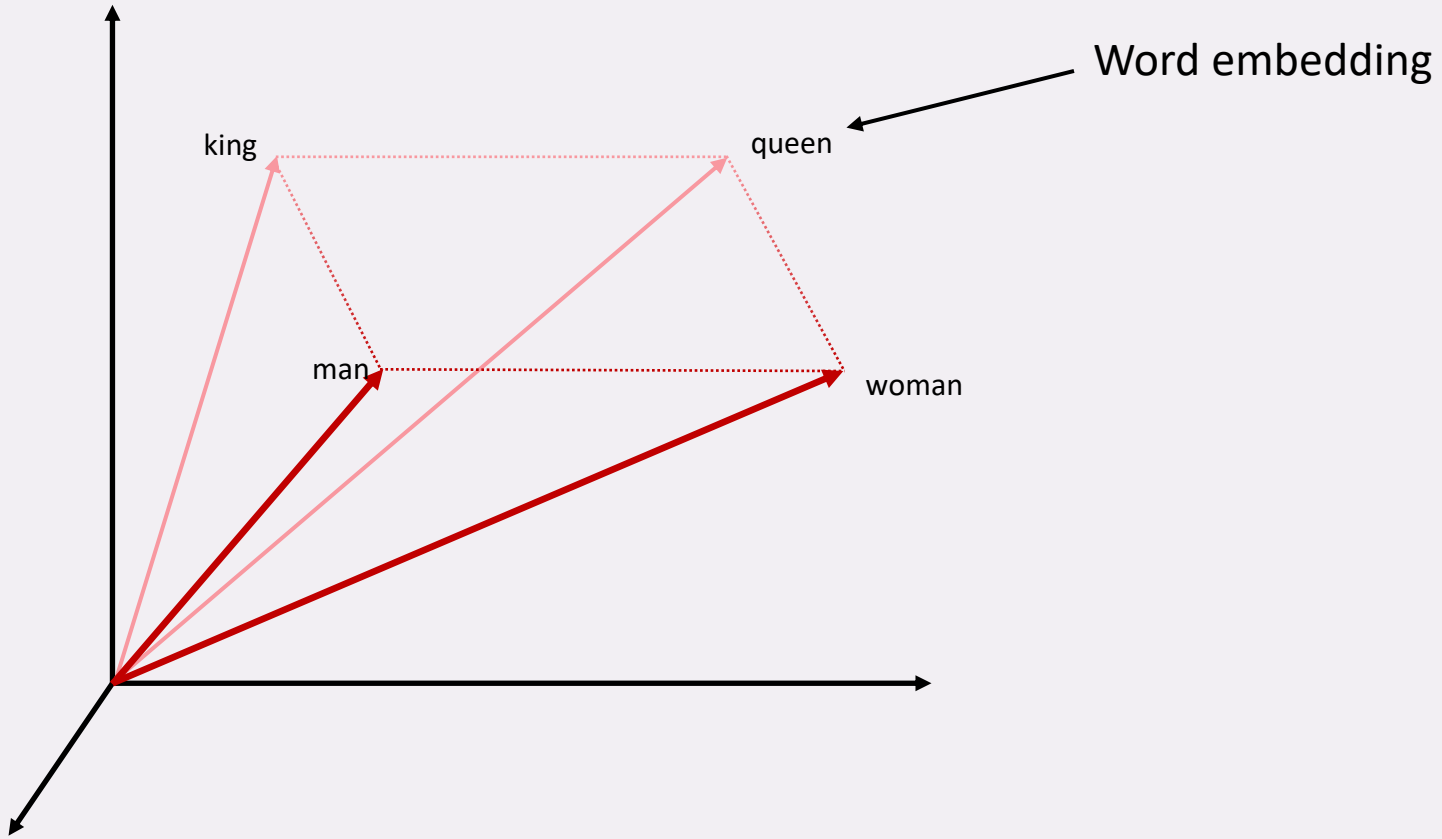
Aim


User
Has
Information need
Formulates
Loses
Explicit Query
Implicit Intent
Retrieval function
Model this intent in the form of context by understanding the user's interest

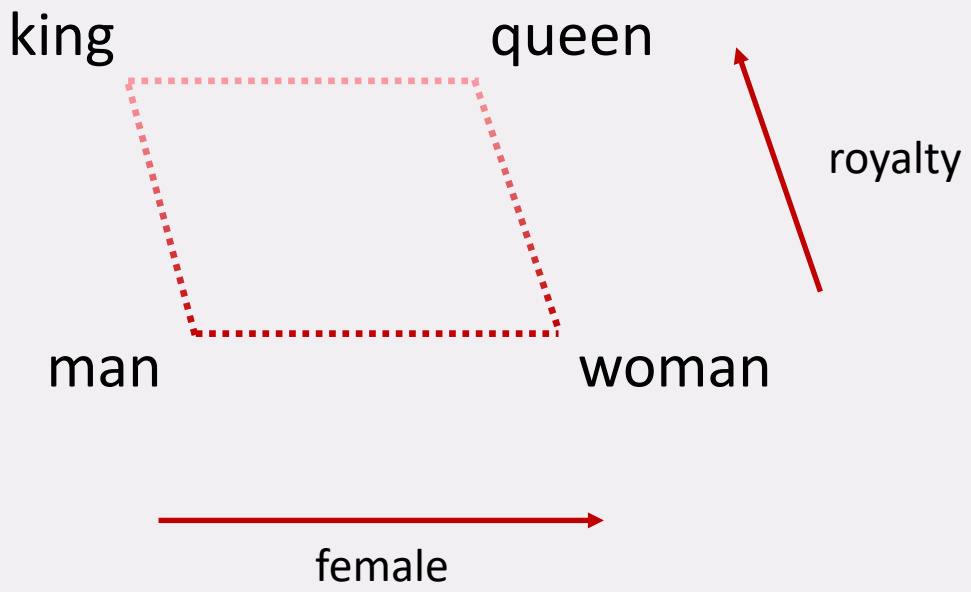
Aim


User
Has
Information need
Formulates
Loses
Explicit Query
Context
Retrieval function

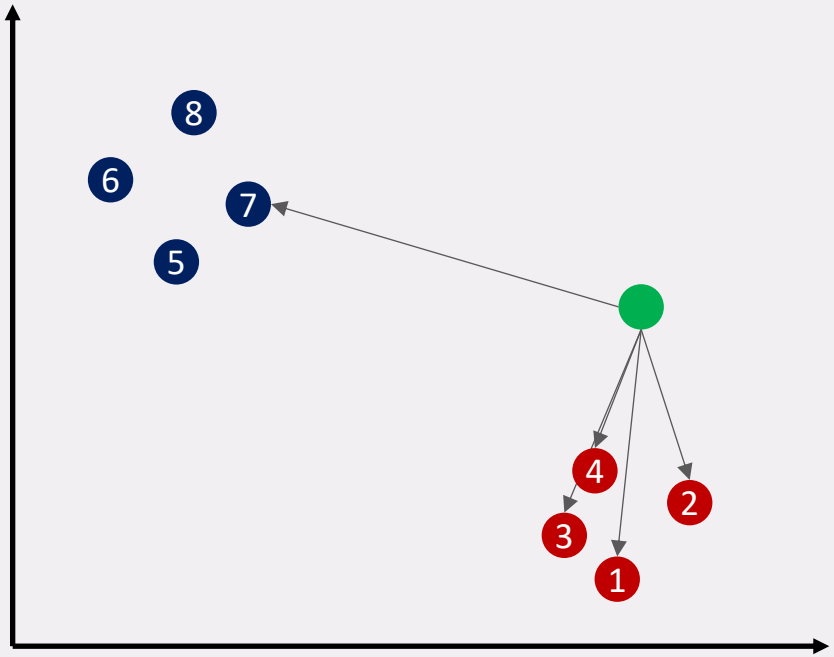
Vector space models



Vector space models



Vector space models: documents



Documents about dogs
Documents about cats

”Are cats good with kids?”

Top 5 documents: 4, 2, 3, 1, 7

Contextual problems in IR

- Same search queries may be expected to yield different results
- Search queries may be imprecise
- The weighting of terms in a search queries may be incorrect

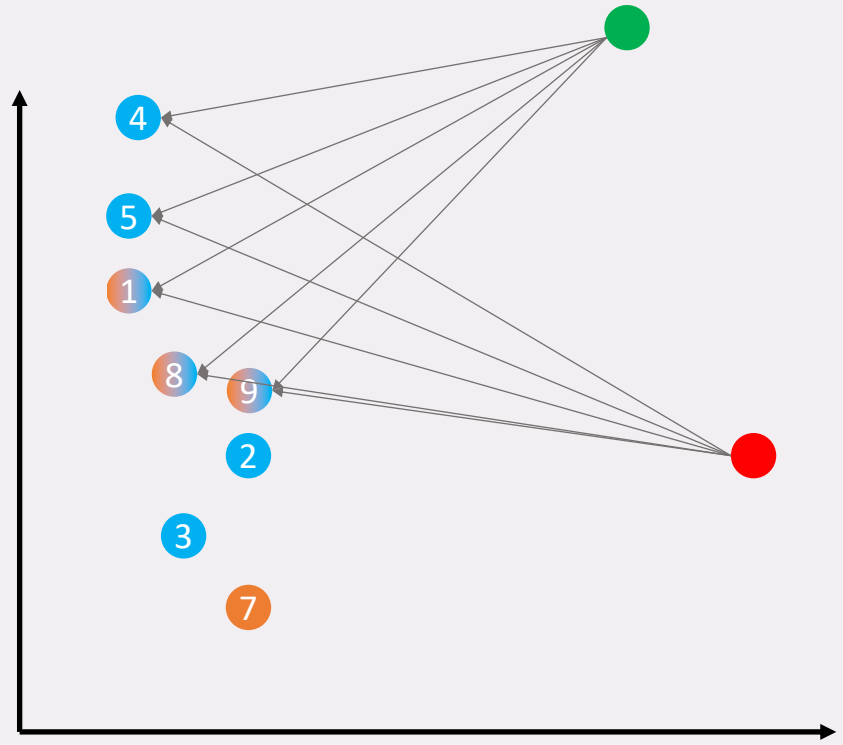
"Should my child wear a **face mask**"

Contextual problems in IR

- Same search queries may be expected to yield different results
- Search queries may be imprecise
- The weighting of terms in a search queries may be incorrect

  "Should my child wear a **face mask**"
- Seemingly unimportant sub-topics *may* not be captured well

Approach idea



- Documents with topic face mask
- Documents with topic children
- Documents with both mask / children

Step 1

- Query: ”Should my child wear a face mask?”
  4, 5, 1, 8, 9 are retrieved

Step 2

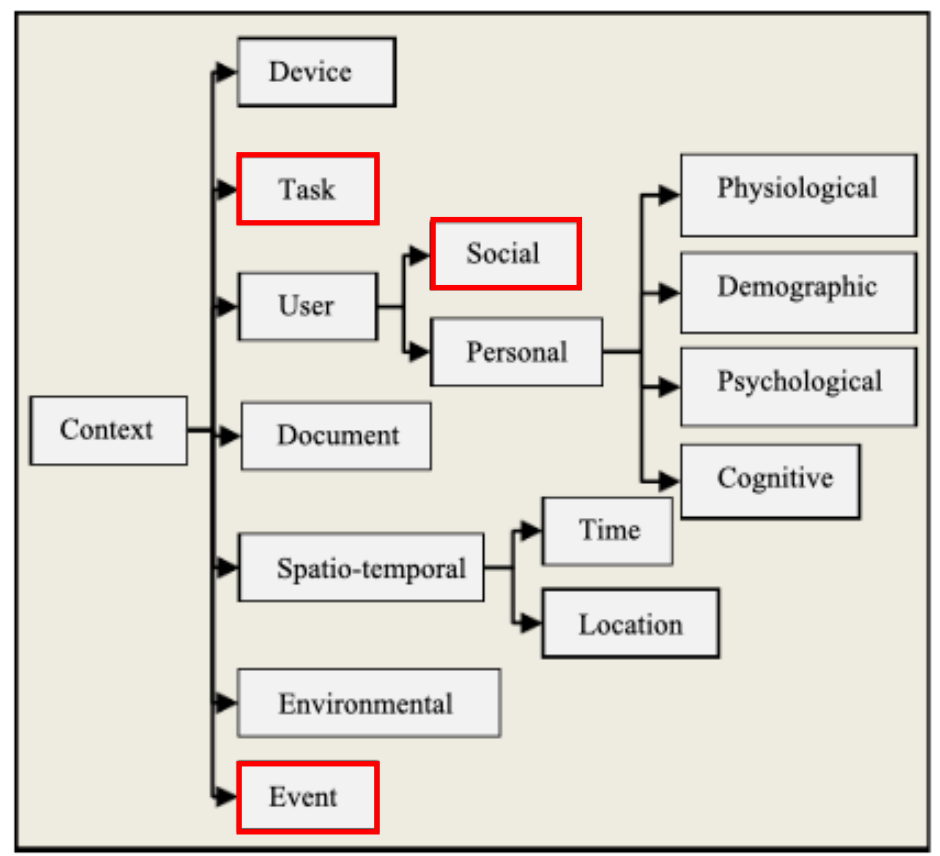
- Re-rank with context: ”School”
  9, 8, 1, 5, 4 is re-ranking result

Context

1. What contexts are relevant (in a pandemic setting)?
2. How to represent context (textually)?
3. How to find the correct context?

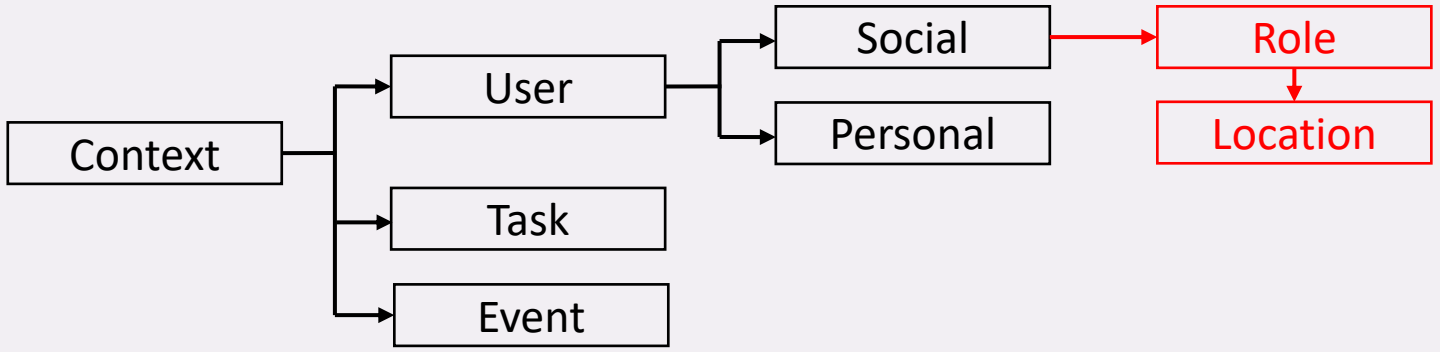
1. What contexts are relevant (in a pandemic setting)?

Contexts of relevance



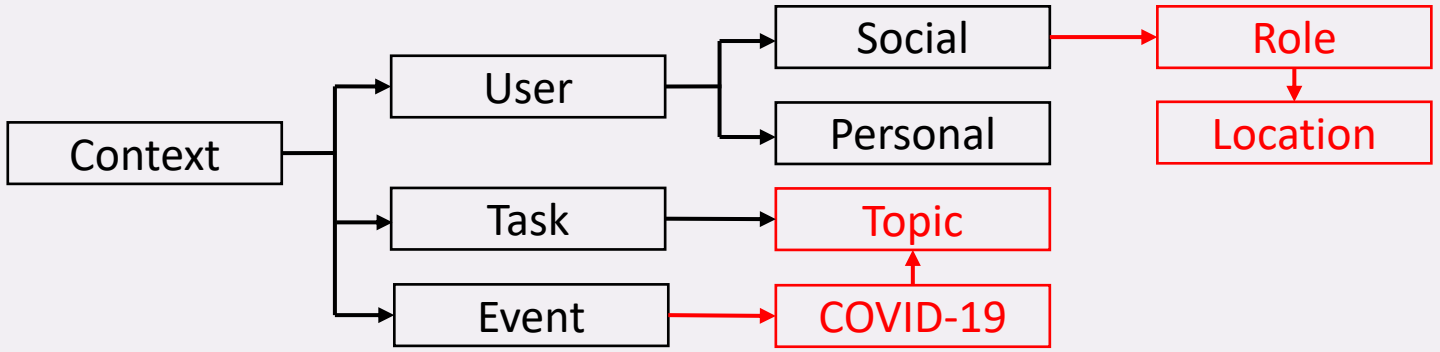
- Work our way through general taxonomies of contexts
- In the context of a pandemic, especially task, social, and event seem important

Contexts of relevance: social



- The role one takes may imply the locations one may be interested in
- Virus behaves differently in certain locations
- Possible locations to differentiate may be schools, hospitals, gyms, sports fields, airplanes, trains

Contexts of relevance: event & task



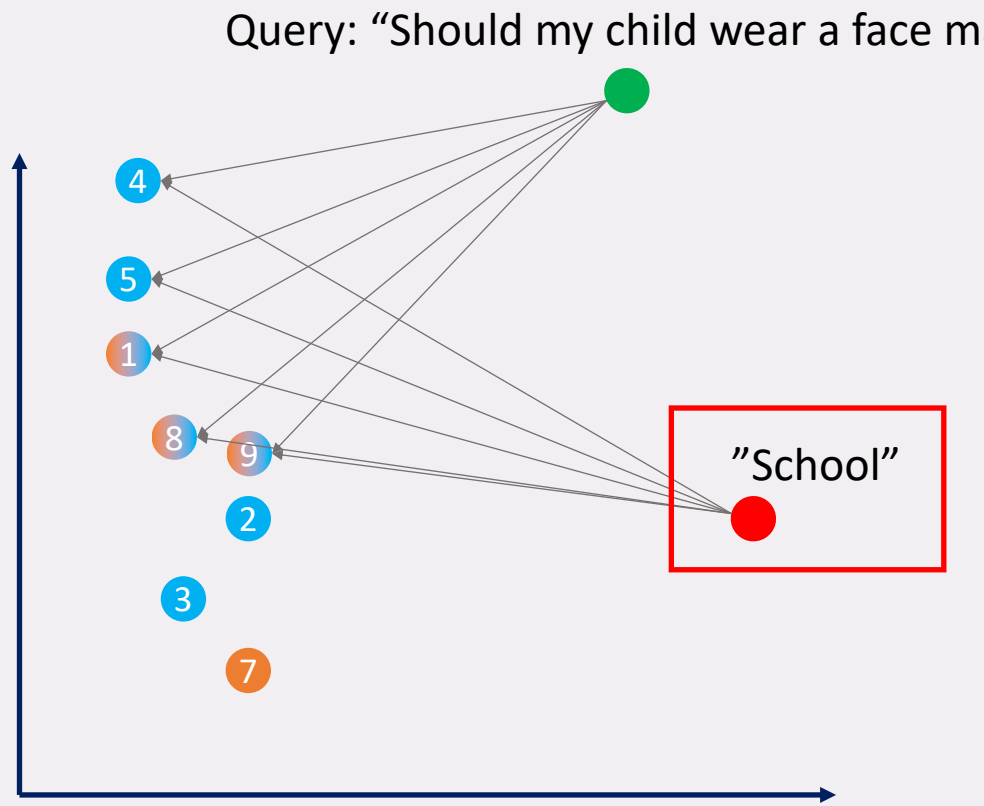
- The COVID-19 pandemic in principle brings two large classes of topics one may be interested in either
  - the **impact** it has or
  - the **mitigation measures** against it

Contexts of relevance: summary

- Though many contexts are presented only a few selected ones were investigated
- In this thesis I especially investigated a user's interest in
  - a location setting, often interest in *schools*
  - the mitigation measures, trying to find the correct realization out of *face masks, hand washing, social distancing, surface cleaning and air filtration*

2. How to represent context (textually)?

Representation of Context



Representation of Context

Experimental Setting

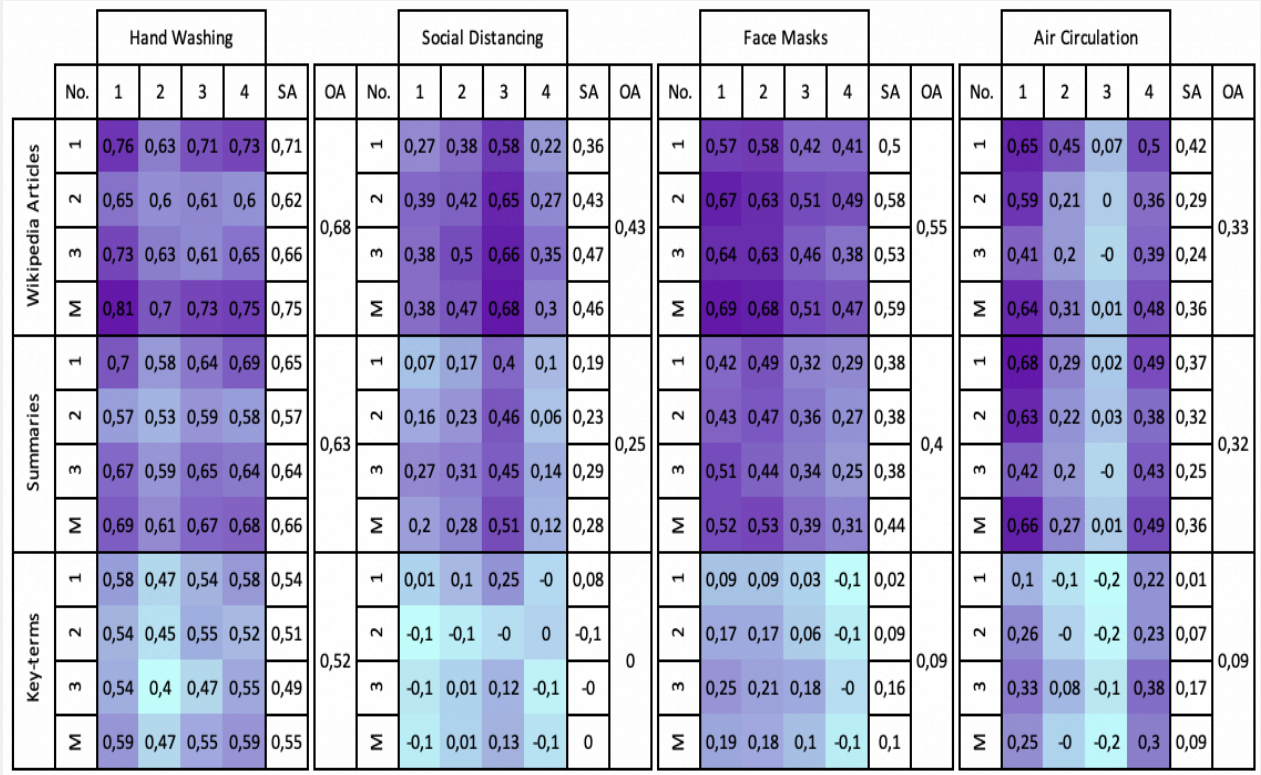
- Four concept topics hand washing, social distancing, face masks and air circulation were chosen
- For each topic four hand picked articles were chosen
- For each of the topics we try to find possible good representations
- Finally the representations are compared with the articles. The closest are the best

	Hand washing				Social Distancing				Face masks				Air circulation			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Terms																
Wikipedia																
Summaries																

Representation of Context

Interpretations

- Wikipedia articles are closest
- Summaries are not far off and might be fine-tuned to get better results
- Some articles are closer/further to all embeddings → captures all concepts better



		Hand Washing						Social Distancing						Face Masks						Air Circulation					
	No.	1	2	3	4	SA	OA	1	2	3	4	SA	OA	1	2	3	4	SA	OA	1	2	3	4	SA	OA
Wikipedia Articles	1	0,76	0,63	0,71	0,73	0,71	0,68	0,27	0,38	0,58	0,22	0,36	0,43	0,57	0,58	0,42	0,41	0,5	0,55	0,65	0,45	0,07	0,5	0,42	0,33
	2	0,65	0,6	0,61	0,6	0,62		0,39	0,42	0,65	0,27	0,43		0,67	0,63	0,51	0,49	0,58		0,59	0,21	0	0,36	0,29	
	3	0,73	0,63	0,61	0,65	0,66		0,38	0,5	0,66	0,35	0,47		0,64	0,63	0,46	0,38	0,53		0,41	0,2	-0	0,39	0,24	
	M	0,81	0,7	0,73	0,75	0,75		0,38	0,47	0,68	0,3	0,46		0,69	0,68	0,51	0,47	0,59		0,64	0,31	0,01	0,48	0,36	
Summaries	1	0,7	0,58	0,64	0,69	0,65	0,63	0,07	0,17	0,4	0,1	0,19	0,25	0,42	0,49	0,32	0,29	0,38	0,4	0,68	0,29	0,02	0,49	0,37	0,32
	2	0,57	0,53	0,59	0,58	0,57		0,16	0,23	0,46	0,06	0,23		0,43	0,47	0,36	0,27	0,38		0,63	0,22	0,03	0,38	0,32	
	3	0,67	0,59	0,65	0,64	0,64		0,27	0,31	0,45	0,14	0,29		0,51	0,44	0,34	0,25	0,38		0,42	0,2	-0	0,43	0,25	
	M	0,69	0,61	0,67	0,68	0,66		0,2	0,28	0,51	0,12	0,28		0,52	0,53	0,39	0,31	0,44		0,66	0,27	0,01	0,49	0,36	
Key-terms	1	0,58	0,47	0,54	0,58	0,54	0,52	0,01	0,1	0,25	-0	0,08	0	0,09	0,09	0,03	-0,1	0,02	0,09	0,1	-0,1	-0,2	0,22	0,01	0,09
	2	0,54	0,45	0,55	0,52	0,51		-0,1	-0,1	-0	0	-0,1		0,17	0,17	0,06	-0,1	0,09		0,26	-0	-0,2	0,23	0,07	
	3	0,54	0,4	0,47	0,55	0,49		-0,1	0,01	0,12	-0,1	-0		0,25	0,21	0,18	-0	0,16		0,33	0,08	-0,1	0,38	0,17	
	M	0,59	0,47	0,55	0,59	0,55		-0,1	0,01	0,13	-0,1	0		0,19	0,18	0,1	-0,1	0,1		0,25	-0	-0,2	0,3	0,09	

Representation of Context

Experimental Setting

- Make sure not every Wikipedia article is close to every newspaper document
- Compare every article with every document
- Normalized for visualization purposes

Interpretations

- Dark diagonal → most Wikipedia articles are close to their respective topic
- Intra-class Wikipedia article quality may differ significantly

			Documents															
			Hand Washing				Social Distancing				Face Masks				Air Circulation			
			1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Wikipedia Articles	Hand Washing	1	0,96	0,92	0,97	1	0,48	0,44	0,31	0,32	0,57	0,36	0,39	0,57	0,61	0,42	0,4	0,41
		2	0,82	0,88	0,84	0,82	0,71	0,62	0,58	0,68	0,71	0,61	0,55	0,81	0,79	0,67	1	0,52
		3	0,96	0,95	0,93	0,89	0,96	0,79	0,68	0,61	0,81	0,6	0,68	0,97	0,83	0,79	0,71	0,71
		M	1	1	1	0,99	0,78	0,67	0,56	0,58	0,76	0,57	0,59	0,85	0,81	0,68	0,76	0,59
	Social Distancing	1	0,48	0,55	0,65	0,48	0,69	0,76	0,85	0,61	0,61	0,77	0,65	0,77	0,8	0,73	0,34	0,88
		2	0,45	0,56	0,55	0,43	1	0,84	0,96	0,77	0,6	0,67	0,53	0,69	0,73	0,88	0,68	0,75
		3	0,65	0,7	0,79	0,56	0,97	1	0,97	1	0,69	0,85	0,8	0,94	1	0,86	0,66	0,99
		M	0,57	0,65	0,71	0,53	0,96	0,94	1	0,86	0,68	0,83	0,71	0,86	0,91	0,89	0,6	0,94
	Face Masks	1	0,58	0,63	0,64	0,59	0,89	0,69	0,74	0,7	0,82	0,85	0,82	0,72	0,85	0,86	0,45	0,85
		2	0,65	0,7	0,62	0,72	0,67	0,66	0,54	0,58	0,97	0,93	1	0,85	0,85	1	0,79	0,73
		3	0,46	0,49	0,52	0,43	0,45	0,44	0,55	0,44	0,92	0,93	0,89	0,66	0,69	0,75	0,39	0,92
		M	0,62	0,67	0,66	0,64	0,74	0,66	0,68	0,63	1	1	1	0,82	0,88	0,96	0,6	0,92
	Air Circulation	1	0,59	0,7	0,65	0,58	0,75	0,68	0,53	0,65	0,68	0,6	0,55	1	0,88	0,78	0,25	1
		2	0,31	0,38	0,48	0,34	0,58	0,32	0,13	0,56	0,52	0,26	0,29	0,83	0,89	0,37	0	0,72
		3	0,32	0,42	0,54	0,4	0,08	0,25	0,23	0,08	0,48	0,33	0,24	0,53	0,61	0,37	-0,1	0,79
		M	0,47	0,58	0,64	0,51	0,54	0,48	0,34	0,5	0,65	0,45	0,42	0,91	0,92	0,58	0,05	0,97

3. How to find the correct context?

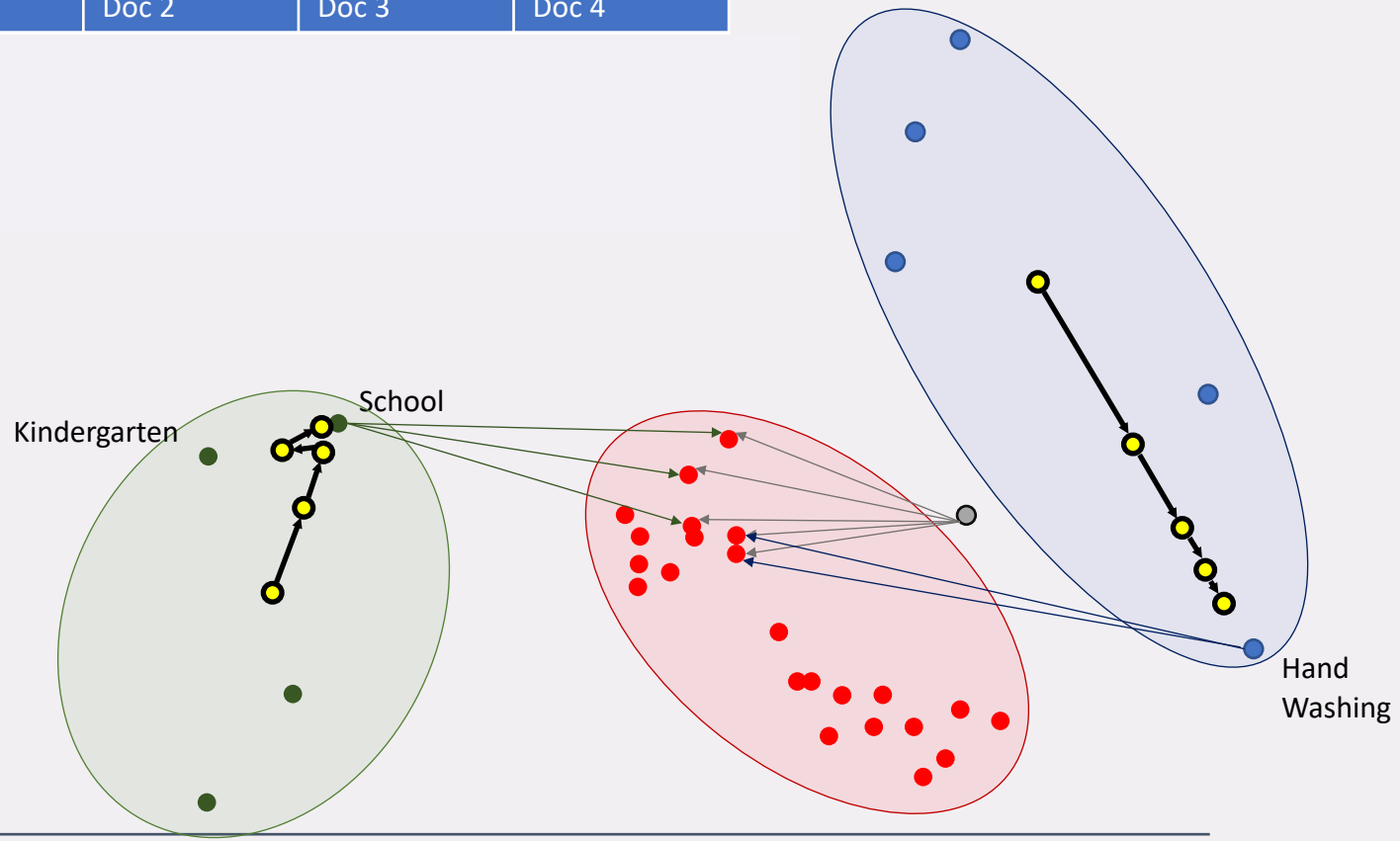
Optimization of Context

- Explicitly gaining user information has been in use for decades
  One could explicitly ask for more user context
- User does not always want to fine-tune the system due to
  - time constraints
  - privacy concerns
- One can implicitly infer the user's interest

Optimization of Context: example

- Assume a corpus of documents embedded in space
- Assume a context of mitigation measures embedded in space
- Assume a context of locations embedded in space
- Assume a previous sequence of clicked documents
- For each context space map each document to its closest corresponding topic
- Place a searcher amidst the contexts and move it in the direction of the mapped sequence
- For the next query, re-rank according to the topic which is closest to the searcher

Documents	Doc 1	Doc 2	Doc 3	Doc 4



Limitations

1. Last scenario is a theoretical construct and not thoroughly tested
2. Assumed clean sequence – noisy sequences will certainly occur
3. Assumed distinct topic borders

Thank you for your attention.

Questions?

References

Slide 12:

[1] Ethayarajh, K., Duvenaud, D., & Hirst, G. (2018). Towards understanding linear word analogies. *arXiv preprint arXiv:1810.04882*.

Slide 19:

[2] Boughareb, D., & Farah, N. (2014). Context in information retrieval.