Europäisches
Patentamt

European
Patent Office

Office européen
des brevets

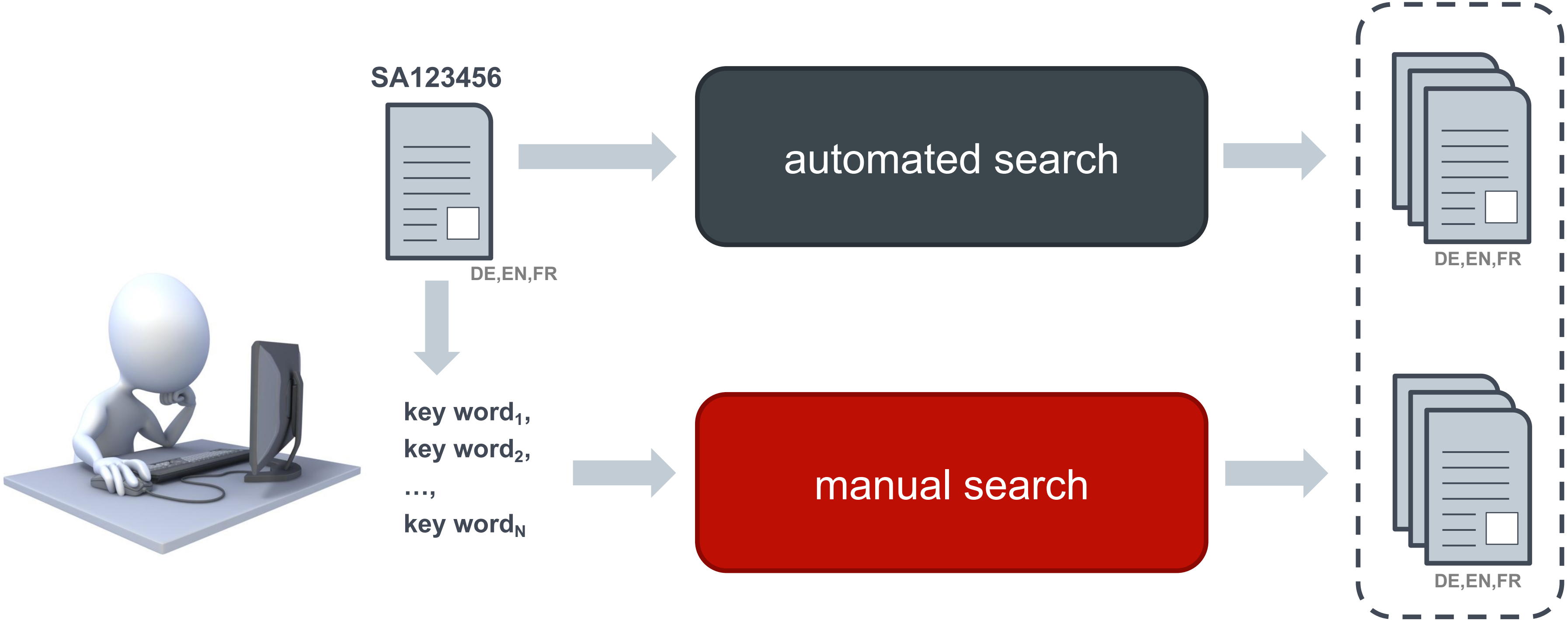
Query Terms Suggestion

M. Wirth & A. Klenner-Bajaja

Data Science

15.07.2022

Patent Search



SA123456

DE,EN,FR

key word$_1$,
key word$_2$,
…,
key word$_N$

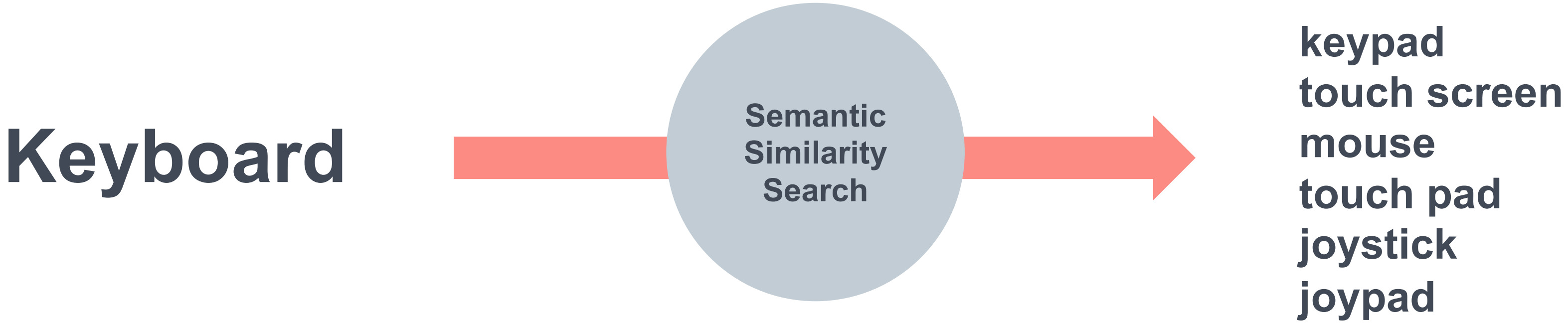
automated search

manual search

DE,EN,FR

DE,EN,FR

Query Terms Suggestion



Keyboard

Semantic Similarity Search

keypad
touch screen
mouse
touch pad
joystick
joypad

Word2Vec

word2vec* - trained auto-encoder

input

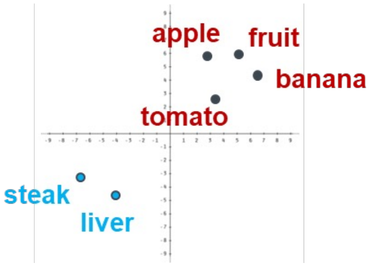
Word

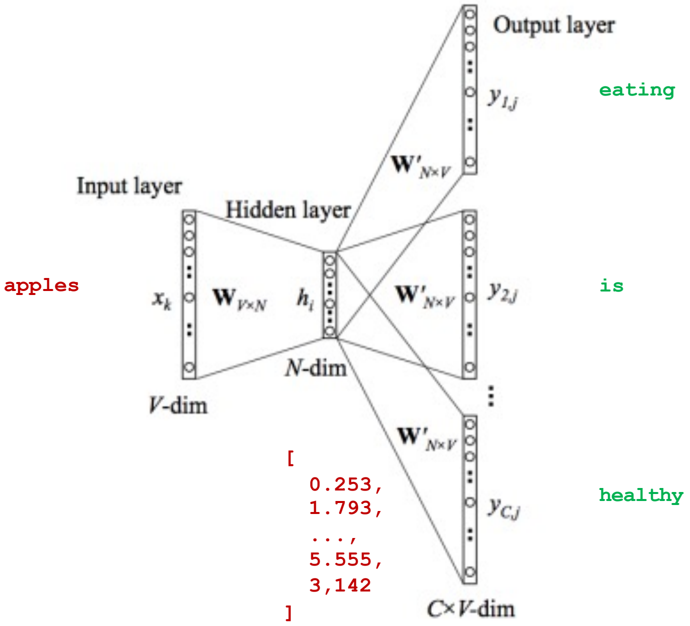
Context

```
eating apples is healthy

eating fruits is healthy
```

output

```
apples  = [ 0.253, 1.793, ..., 5.555, 3.142 ]

fruits = [ 0.254, 1.993, ..., 4.921, 2.999 ]
```





* Mikolov et al. (2013) https://arxiv.org/abs/1301.3781

PreProcessing - text quality control

- A range of filters were developed to detect problematic words

Examples:	OCR errors	e.g. *l1elp, rnyself, 1ever, deSense*	**Rule-based**
	Space deletion	e.g. *dieArbeit, thenumber*	**CRF, Random Forest**
	Space insertion	e.g. *Austral_ia, bio_logy*	**Rule-based**
	DNA/Protein sequences	e.g. *AGGATTTCTAAAC, MVFPMWTLKR*	**Rule-based**

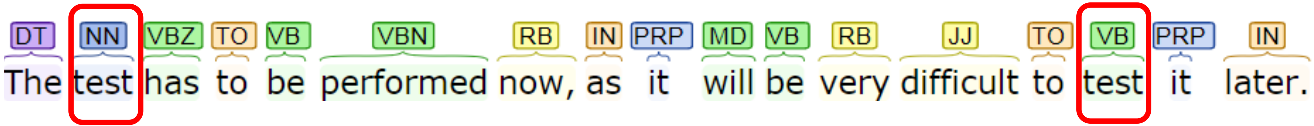
What to do with these cases ?

- Erroneous words are tagged with a PROB flag
- Number of erroneous words per sentence are counted
- **≥ 15% ?** Too much noise! The sentence is ignored for training of the models
- **< 15% ?** Sentence is used for training, but the resulting model is cleaned up from PROB flagged words

PreProcessing - Part-of-speech & N-gram detection

POS tagging

- Assignment of POS tags for clear distinguishment of word forms


DT
NN
VBZ
TO
VB
VBN
RB
IN
PRP
MD
VB
RB
JJ
TO
VB
PRP
IN
The test has to be performed now, as it will be very difficult to test it later.

evaluation, analysis

examine, analyze, check, assess

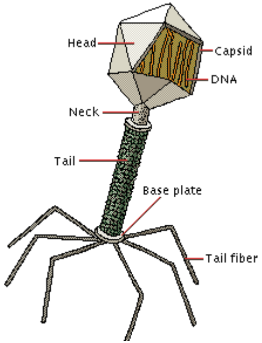
Bigrams

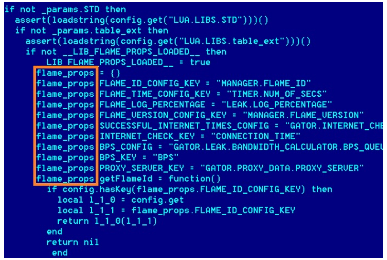
- Example: “windshield wiper” or “car wash”
- Detected by applying a *chi square* algorithm: How often occur words alone *vs.* together?

FR – n-grams

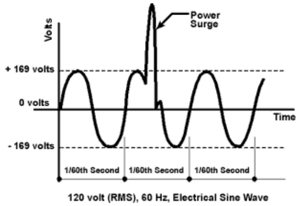
- **3-grams** bridged by *de*, *d’*, *à*, *du*, *en*, *des*, *au*, *aux*, *a*
- **4-grams** bridged by *de la*, *d’ un*, *à la*, *d' une*

How to deal with word ambiguity?

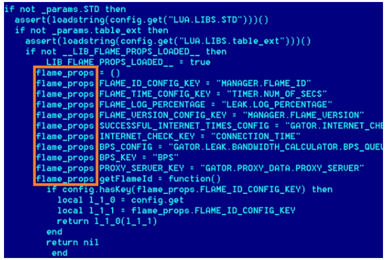




```
if not _params.STD then
  assert(loadstring(config.get("LUA.LIBS.STD")))()
  if not _params.table_ext then
    assert(loadstring(config.get("LUA.LIBS.table_ext")))()
    if not __LIB_FLAME_PROPS_LOADED__ then
        LIB_FLAME_PROPS_LOADED__ = true
      flame_props = {}
      flame_props FLAME_ID_CONFIG_KEY = "MANAGER.FLAME_ID"
      flame_props FLAME_TIME_CONFIG_KEY = "TIMER.NUM_OF_SECS"
      flame_props FLAME_LOG_PERCENTAGE = "LEAK.LOG_PERCENTAGE"
      flame_props FLAME_VERSION_CONFIG_KEY = "MANAGER.FLAME_VERSION"
      flame_props SUCCESSFUL_INTERNET_TIMES_CONFIG = "GATOR.INTERNET_CHE
      flame_props INTERNET_CHECK_KEY = "CONNECTION_TIME"
      flame_props BPS_CONFIG = "GATOR.LEAK.BANDWIDTH_CALCULATOR.BPS_QUEU
      flame_props BPS_KEY = "BPS"
      flame_props PROXY_SERVER_KEY = "GATOR.PROXY_DATA.PROXY_SERVER"
      flame_props getFlameId = function()
        if config.hasKey(flame_props.FLAME_ID_CONFIG_KEY) then
          local l_1_0 = config.get
          local l_1_1 = flame_props.FLAME_ID_CONFIG_KEY
          return l_1_0(l_1_1)
        end
        return nil
         end
```



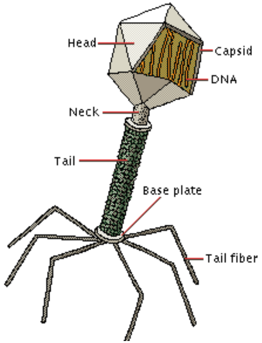
CPC to the rescue!



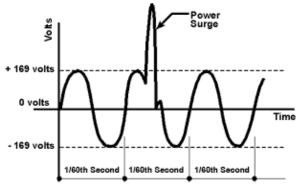
```
if not _params.STD then
  assert(loadstring(config.get("LUA.LIBS.STD")))()
  if not _params.table_ext then
    assert(loadstring(config.get("LUA.LIBS.table_ext")))()
    if not __LIB_FLAME_PROPS_LOADED__ then
        LIB FLAME PROPS_LOADED__ = true
      flame_props = {}
      flame_props FLAME_ID_CONFIG_KEY = "MANAGER.FLAME_ID"
      flame_props FLAME_TIME_CONFIG_KEY = "TIMER.NUM_OF_SECS"
      flame_props FLAME_LOG_PERCENTAGE = "LEAK.LOG_PERCENTAGE"
      flame_props FLAME_VERSION_CONFIG_KEY = "MANAGER.FLAME_VERSION"
      flame_props SUCCESSFUL_INTERNET_TIMES_CONFIG = "GATOR.INTERNET_CHE
      flame_props INTERNET_CHECK_KEY = "CONNECTION_TIME"
      flame_props BPS_CONFIG = "GATOR.LEAK.BANDWIDTH_CALCULATOR.BPS_QUEU
      flame_props BPS_KEY = "BPS"
      flame_props PROXY_SERVER_KEY = "GATOR.PROXY_DATA.PROXY_SERVER"
      flame_props getFlameId = function()
        if config.hasKey(flame_props.FLAME_ID_CONFIG_KEY) then
          local l_1_0 = config.get
          local l_1_1 = flame_props.FLAME_ID_CONFIG_KEY
          return l_1_0(l_1_1)
        end
        return nil
         end
```

G06F - ELECTRIC DIGITAL DATA PROCESSING …

A43B – CHARACTERISTICS OF FOOTWEAR …



C12N - MICROORGANISMS OR ENZYMES…



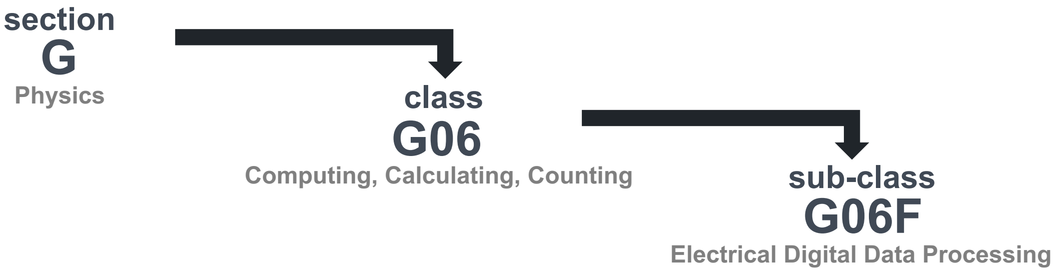
120 volt (RMS), 60 Hz, Electrical Sine Wave

H02H – EMERGENCY PROTECTIVE CIRCUIT ARRANGEMENTS …

F16B - DEVICES FOR FASTENING …

A23L - FOODS, FOODSTUFFS …

Deciding for the CPC level depth



section
G
Physics

class
G06
Computing, Calculating, Counting

sub-class
G06F
Electrical Digital Data Processing

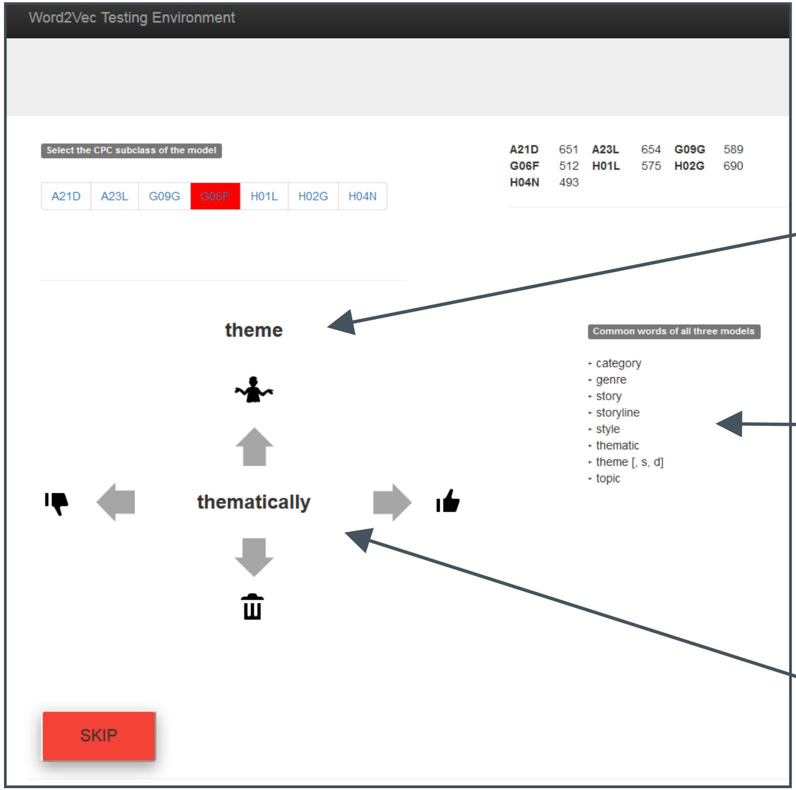
- Trade-off between training corpus size affecting quality and word disambiguation
- Will be input parameter at request time, or result in multiclass output
- What is the optimal level in the CPC tree (1-digit, 3-digit, 4-digit) ?
- Do we have to host 8, 124, or 663 vector sets for similarity search ?

Deciding for the CPC level depth

Query: ***“ theme ”***

against models

G
G06
G06F


Word2Vec Testing Environment
Select the CPC subclass of the model
A21D
A23L
G09G
G06F
H01L
H02G
H04N
Common words of all three models
SKIP

A21D	651	**A23L**	654	**G09G**	589
G06F	512	**H01L**	575	**H02G**	690
H04N	493				

theme

thematically

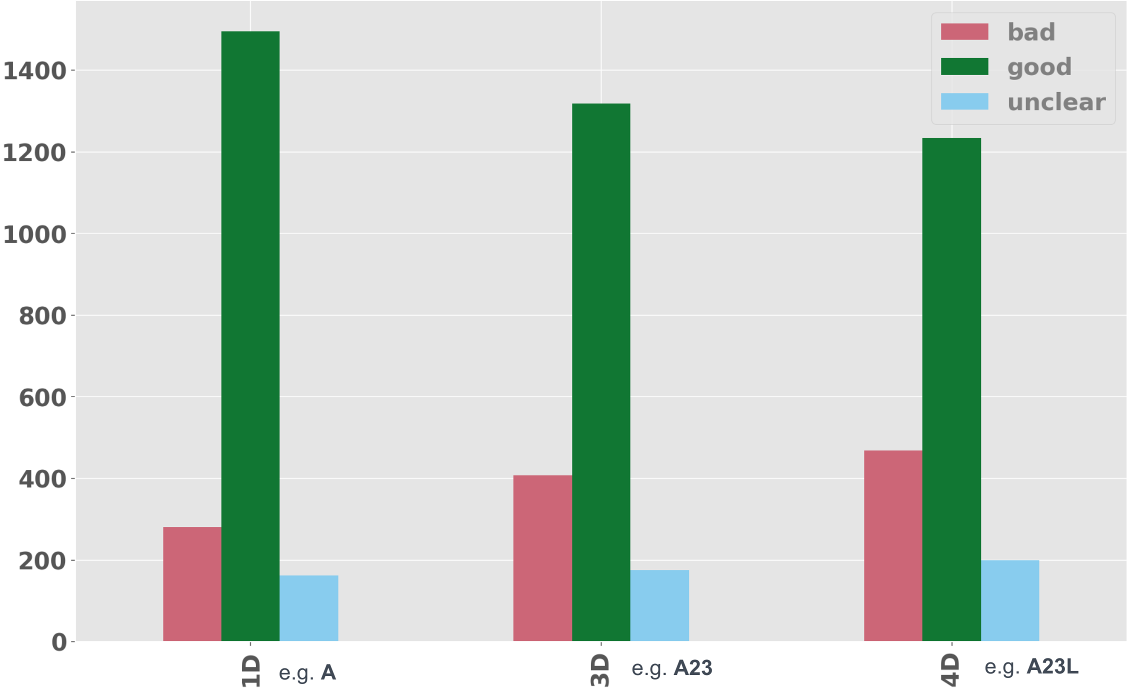
- category
- genre
- story
- storyline
- style
- thematic
- theme [, s, d]
- topic

pre-selected query term

common terms of all three models

proposed term
(not found in top 15 of all models)

CPC level depth results



- Users prefer suggestions from section model
- Clear trend observable of decreasing ***good*** and increasing ***bad*** suggestions
- Least models to host and least complex input parameters

Section models (=8)

X

Office languages (=3)

=

24 independent models

Some Statistics EN

- Data ingested from 2015-01-01 (7+ years)
- 2,440,964 documents total
- 204 Gb of ingested data

processed sentences	
A	175,933,077
B	110,707,118
C	142,638,316
D	6,491,405
E	16,958,187
F	42,404,148
G	277,512,083
H	233,961,244
Total	**1,006,605,578**

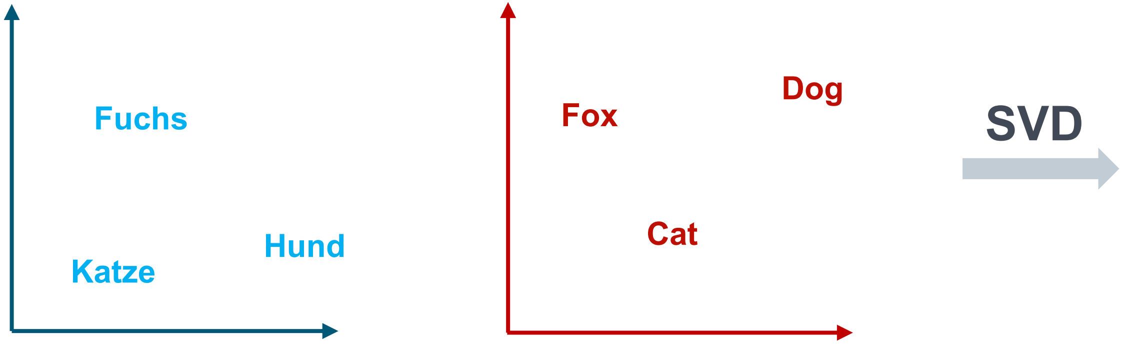
vector counts	
A	1,674,493
B	869,773
C	1,589,179
D	153,821
E	222,295
F	360,287
G	1,556,453
H	1,044,503

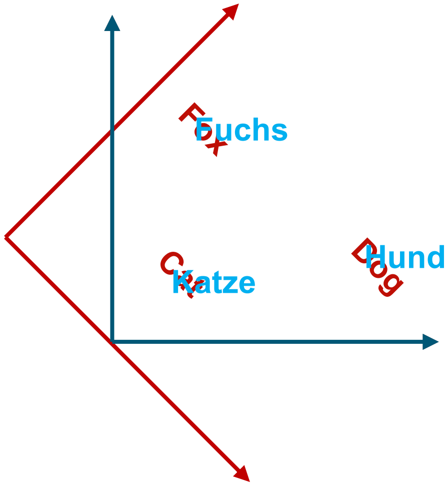
Finding similar terms in other languages

DE word2vec vector space

EN word2vec vector space

Aligned word2vec vector spaces





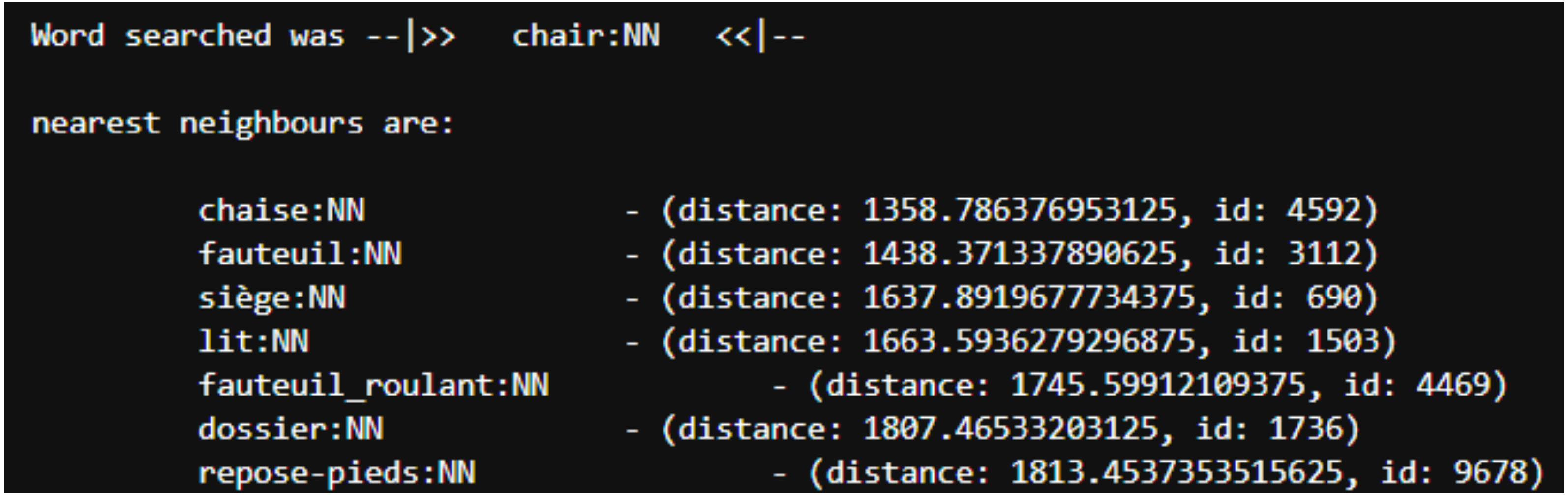
SVD

Given known correspondences, i.e. pre-calculated word translations

1. Translate vector space
2. Rotate vector space

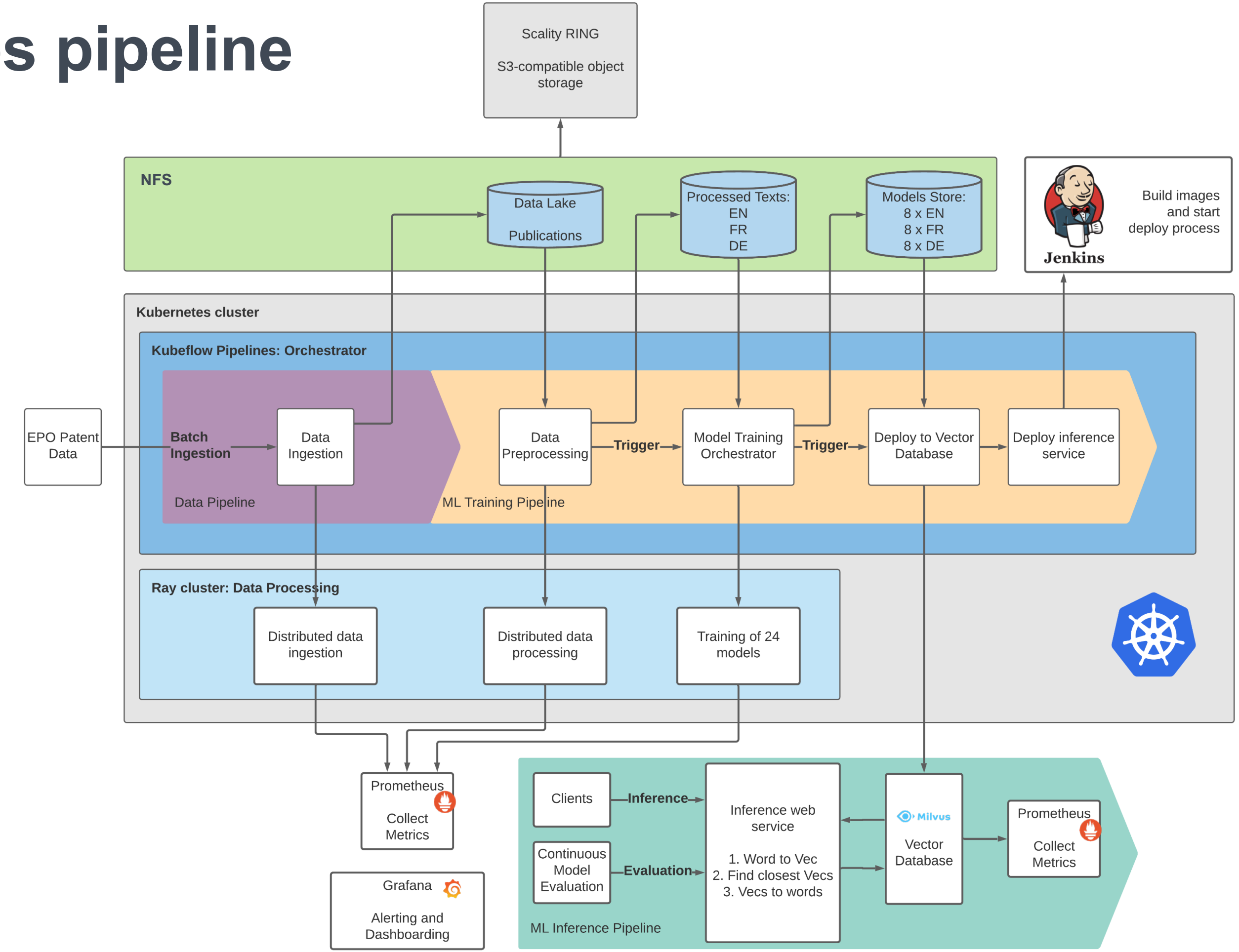
→ allows cross-lingual queries

Single term query in different language

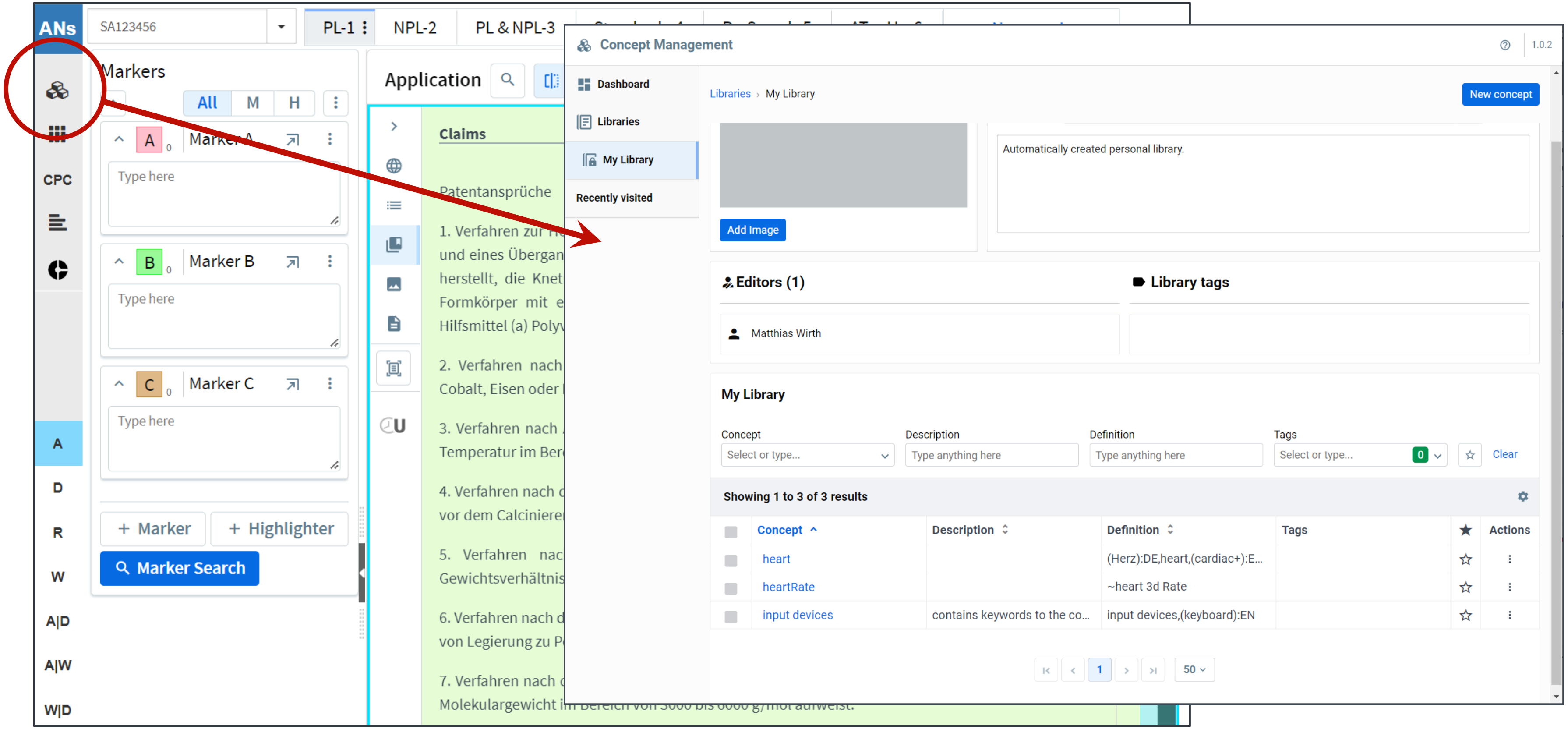


```
Word searched was --|>>   chair:NN   <<|--

nearest neighbours are:

         chaise:NN               - (distance: 1358.786376953125, id: 4592)
         fauteuil:NN             - (distance: 1438.371337890625, id: 3112)
         siège:NN                - (distance: 1637.8919677734375, id: 690)
         lit:NN                  - (distance: 1663.5936279296875, id: 1503)
         fauteuil_roulant:NN             - (distance: 1745.59912109375, id: 4469)
         dossier:NN              - (distance: 1807.46533203125, id: 1736)
         repose-pieds:NN                 - (distance: 1813.4537353515625, id: 9678)
```

MLOps pipeline



Tool Integration



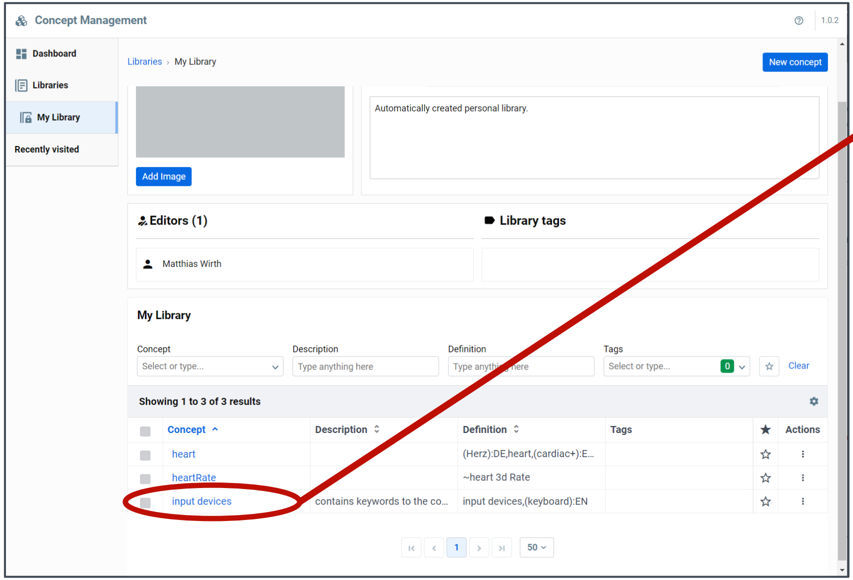
Automatically created personal library.

My Library

Showing 1 to 3 of 3 results

Concept	Description	Definition	Tags	★	Actions
heart		(Herz):DE,heart,(cardiac+):E...		☆	⋮
heartRate		~heart 3d Rate		☆	⋮
input devices	contains keywords to the co...	input devices,(keyboard):EN		☆	⋮

Tool Integration



Libraries › My Library

Automatically created personal library.

Editors (1)

Matthias Wirth

Library tags

My Library

Concept: Select or type...

Description: Type anything here

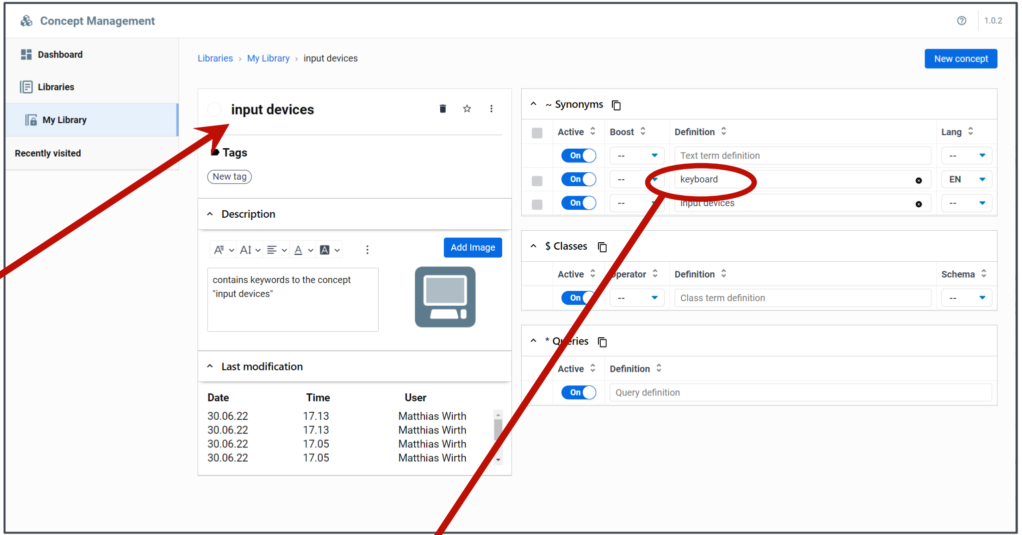
Definition: Type anything here

Tags: Select or type... 0

Showing 1 to 3 of 3 results

	Concept	Description	Definition	Tags	★	Actions
□	heart		(Herz):DE,heart,(cardiac+):E...		☆	⋮
□	heartRate		~heart 3d Rate		☆	⋮
□	input devices	contains keywords to the co...	input devices,(keyboard):EN		☆	⋮

1 · 50



Libraries › My Library › input devices

input devices

Tags

Description

contains keywords to the concept "input devices"

Last modification

Date	Time	User
30.06.22	17.13	Matthias Wirth
30.06.22	17.13	Matthias Wirth
30.06.22	17.05	Matthias Wirth
30.06.22	17.05	Matthias Wirth

~ Synonyms

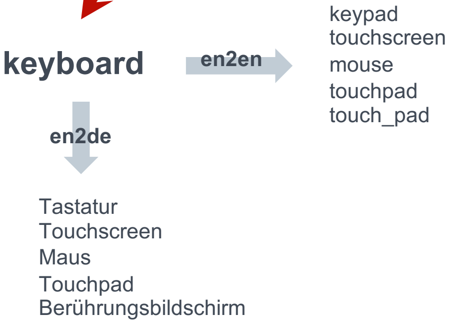
	Active	Boost	Definition	Lang
	On	--	Text term definition	--
□	On	--	keyboard	EN
□	On	--	input devices	--

$ Classes

Active	Operator	Definition	Schema
On	--	Class term definition	--

* Queries

Active	Definition
On	Query definition



Thank you very much for your attention!

Acknowledgements

Daniel Schneider

Hennadii Stas

Abdelkader Kouhli