mapping_analysis

November 18, 2022

Table of Contents

1 Introduction

2 Library imports and configs

3 Setup

4 Load bibliographical data

4.1 Prepare data for further analysis

4.2 Remov unuseful columns

4.3 Add columns for mapping study facets

5 Final DF

5.1 Table describing the columns

6 Explore notes

6.1 Extract search terms from notes

7 Explore facets

7.1 Research Facet

7.2 Cont Facet

7.3 Domain Facet

8 Explore interaction of research and contribution facets

8.1 Correlations

8.2 Crosstab between research and contribution

8.2.1 Absolute numbers

8.2.1.1 Visualizations

8.2.2 Percent of Totals

8.2.2.1 Visualizations

9 Explore year

10 Explore Authors

11 Explore Keywords

12 Cluster similar words from keywords and identify groups

12.1 Load Model

12.2 Get keywords and compare to vocab in pre-trained model

12.3 PCA

12.4 Cluster

12.5 Plot

12.6 Conclusion

13 Cluster similar words from domain facet

13.1 Conclusion

14 Further work

1 Introduction

This is the corresponding notebook to the class "seminar in AI" at the Master's study program at Johannes Kepler University in Austria.

This notebook describes the mapping study conducted for my master thesis on MLOps.

The goal is to analyse various papers for the study field and provide a base for further research in presenting a reference architecture in MLOps.

2 Library imports and configs

Various used libraries are imported and certain imports configured.

[1]:
```
import pandas as pd
import bibtexparser
import numpy as np

from sklearn.decomposition import PCA
from sklearn import cluster
from sklearn.utils import check_random_state

import plotly.express as px
import plotly.graph_objects as go
import re

# for displaying plotly inside jupyter notebook
from plotly.offline import init_notebook_mode

init_notebook_mode(connected=True)

pd.set_option('display.max_columns', None)
```

```
pd.set_option('display.max_rows', None)

# %matplotlib notebook
```

3 Setup

Here we define the - project root - what columns from the bibliography are used - Random Seed for reproducability is set

[2]:
```
PROJECT_ROOT = '../'

# DATA_PATH = PROJECT_ROOT+'bibliography/zotero_collection_export.bib'
DATA_PATH = PROJECT_ROOT + 'bibliography/mapping_study/
 ↪zotero_collection_ieee_export.bib'
```

[3]:
```
search_words = ['research', 'cont', 'domain', 'summary']

cont_cols = [
    'approach', 'casestudy', 'experiment', 'literature', 'metric', 'model',
    'nonempirical', 'process', 'tool'
]

research_cols = [
    'evaluation', 'experience', 'opinion', 'philosophical', 'solution',
    'validation'
]
```

[4]:
```
SEED = 0
RNG = np.random.RandomState(SEED)

random_state = check_random_state(RNG)
print(RNG)
print(RNG.permutation(10))


if not set(RNG.permutation(10)).issubset(set([2, 8, 4, 9, 1, 6, 7, 3, 0, 5])):
    raise ValueError('RandomState not working')
```

```
RandomState(MT19937)
[2 8 4 9 1 6 7 3 0 5]
```

4 Load bibliographical data

From the bibliography tool Zotero the entries are important and put into a dataframe for further usage.

[5]:
```
with open(DATA_PATH) as bibtex_file:
    bib_database = bibtexparser.load(bibtex_file)

df = pd.DataFrame(bib_database.entries)
```

[6]:
```
if not len(df) >= 46:
    raise ValueError('Not enough content loaded')
```

4.1 Prepare data for further analysis

[7]:
```
df
```

[7]:

	file \
0	/Users/danieldeutsch/Zotero/storage/EV5GXSAC/A…
1	/Users/danieldeutsch/Zotero/storage/34Y42EMM/B…
2	/Users/danieldeutsch/Zotero/storage/568ZRGE4/B…
3	/Users/danieldeutsch/Zotero/storage/DAFF2ZLA/B…
4	/Users/danieldeutsch/Zotero/storage/NTHZ4YRC/B…
5	/Users/danieldeutsch/Zotero/storage/AKWWA25B/C…
6	/Users/danieldeutsch/Zotero/storage/I9XKIGI6/D…
7	/Users/danieldeutsch/Zotero/storage/IGEJHKUJ/F…
8	/Users/danieldeutsch/Zotero/storage/B79WJ55I/G…
9	/Users/danieldeutsch/Zotero/storage/R865VPXI/G…
10	/Users/danieldeutsch/Zotero/storage/XZWYSPK3/G…
11	/Users/danieldeutsch/Zotero/storage/CHK33FPI/G…
12	/Users/danieldeutsch/Zotero/storage/7D9CW726/I…
13	/Users/danieldeutsch/Zotero/storage/GCI58T9F/J…
14	/Users/danieldeutsch/Zotero/storage/DZLPUFER/J…
15	/Users/danieldeutsch/Zotero/storage/BSLEIXN2/J…
16	/Users/danieldeutsch/Zotero/storage/72KU2MN7/J…
17	/Users/danieldeutsch/Zotero/storage/ZL67CP4J/K…
18	/Users/danieldeutsch/Zotero/storage/738H3N3Y/K…
19	/Users/danieldeutsch/Zotero/storage/97R4XLA3/K…
20	/Users/danieldeutsch/Zotero/storage/VM7YURRL/K…
21	/Users/danieldeutsch/Zotero/storage/J8LP7KYD/L…
22	/Users/danieldeutsch/Zotero/storage/6KLVRWJN/L…
23	/Users/danieldeutsch/Zotero/storage/BMY73PWF/L…
24	/Users/danieldeutsch/Zotero/storage/R9ZSQ6FW/L…
25	/Users/danieldeutsch/Zotero/storage/4QQY49FX/M…
26	/Users/danieldeutsch/Zotero/storage/WLMXQ297/M…
27	/Users/danieldeutsch/Zotero/storage/43PVX7HA/M…
28	/Users/danieldeutsch/Zotero/storage/AII6ITD5/M…
29	/Users/danieldeutsch/Zotero/storage/BS3QCBZF/M…
30	/Users/danieldeutsch/Zotero/storage/Z29IS8GS/M…
31	/Users/danieldeutsch/Zotero/storage/TLPQ93QP/P…
32	/Users/danieldeutsch/Zotero/storage/UP23SHP2/R…
33	/Users/danieldeutsch/Zotero/storage/XEHAF4SR/R…

```
34  /Users/danieldeutsch/Zotero/storage/FETIRHB4/R…
35  /Users/danieldeutsch/Zotero/storage/IHKA3HXI/R…
36  /Users/danieldeutsch/Zotero/storage/Z9P444PP/R…
37  /Users/danieldeutsch/Zotero/storage/3IH9AT83/S…
38  /Users/danieldeutsch/Zotero/storage/3MC7REQC/S…
39  /Users/danieldeutsch/Zotero/storage/INKDAIM2/S…
40  /Users/danieldeutsch/Zotero/storage/BV8NBBDP/S…
41  /Users/danieldeutsch/Zotero/storage/P7NNEL6Z/T…
42  /Users/danieldeutsch/Zotero/storage/3E6GQZH7/T…
43  /Users/danieldeutsch/Zotero/storage/GZ4N5GE7/V…
44  /Users/danieldeutsch/Zotero/storage/85IFEMME/Y…
45  /Users/danieldeutsch/Zotero/storage/QEKZ5HPE/Z…
46  /Users/danieldeutsch/Zotero/storage/Q64Y5THD/Z…
47  /Users/danieldeutsch/Zotero/storage/GVWJ8Z6E/Z…
48  /Users/danieldeutsch/Zotero/storage/G25TUNNU/Z…

                                                 note  \
0   2\n\par\nresearch: solution\n\par\ncont: cases…
1   2\n\par\nresearch: philosophical\n\par\ncont: …
2   2\n\par\nresearch: solution\n\par\ncont: cases…
3   2\n\par\nresearch: validation\n\par\ncont: cas…
4   2\n\par\nresearch: philosophical\n\par\ncont: …
5   2\n\par\nresearch: evaluation\n\par\ncont: cas…
6   2\n\par\nresearch: philosophical\n\par\ncont: …
7   2\n\par\nresearch: validation\n\par\ncont: cas…
8   2\n\par\nresearch: philosophical\n\par\ncont: …
9   2\n\par\nresearch: philosophical\n\par\ncont: …
10  2\n\par\nresearch: solution\n\par\ncont: cases…
11  2\n\par\nresearch: solution\n\par\ncont: cases…
12  2\n\par\nresearch: philosophical\n\par\ncont: …
13  1\n\par\nresearch: solution\n\par\ncont: appro…
14  2\n\par\nresearch: validation\n\par\ncont: cas…
15  2\n\par\nresearch: solution\n\par\ncont: cases…
16  2\n\par\nresearch: philosophical\n\par\ncont: …
17  2\n\par\nresearch: solution\n\par\ncont: cases…
18  2\n\par\nresearch: solution\n\par\ncont: appro…
19  2\n\par\nresearch: solution\n\par\ncont: cases…
20  2\n\par\nresearch: philosophical\n\par\ncont: …
21  2\n\par\nresearch: philosophical\n\par\ncont: …
22  2\n\par\nresearch: philosophical\n\par\ncont: …
23  2\n\par\nresearch: solution\n\par\ncont: cases…
24  2\n\par\nresearch: philosophical\n\par\ncont: …
25  2\n\par\nresearch: evaluation\n\par\ncont: app…
26  2\n\par\nresearch: philosophical\n\par\ncont: …
27  2\n\par\nresearch: philosophical\n\par\ncont: …
28  2\n\par\nresearch: evaluation\n\par\ncont: app…
29  2\n\par\nresearch: evaluation\n\par\ncont: cas…
```

30	2\n\par\nresearch: philosophical\n\par\ncont: …
31	2\n\par\nresearch: evaluation\n\par\ncont: cas…
32	2\n\par\nresearch: evaluation\n\par\ncont: app…
33	2\n\par\nresearch: solution\n\par\ncont: cases…
34	2\n\par\nresearch: philosophical\n\par\ncont: …
35	1\n\par\nresearch: solution\n\par\ncont: cases…
36	2\n\par\nresearch: evaluation\n\par\ncont: app…
37	1\n\par\nresearch: solution\n\par\ncont: appro…
38	2\n\par\nresearch: evaluation\n\par\ncont: app…
39	2\n\par\nresearch: solution\n\par\ncont: cases…
40	2\n\par\nresearch: philosophical\n\par\ncont: …
41	2\n\par\nresearch: philosophical\n\par\ncont: …
42	2\n\par\nresearch: philosophical\n\par\ncont: …
43	2\n\par\nresearch: solution\n\par\ncont: cases…
44	2\n\par\nresearch: solution\n\par\ncont: appro…
45	2\n\par\nresearch: evaluation\n\par\ncont: cas…
46	2\n\par\nresearch: solution\n\par\ncont: cases…
47	2\n\par\nresearch: solution\n\par\ncont: cases…
48	2\n\par\nresearch: evaluation\n\par\ncont: cas…

	keywords \
0	Adaptation models,Costs,Image edge detection,R…
1	Cloud computing,Computational modeling,Compute…
2	action research,AI quality,Context modeling,co…
3	Automation,Conferences,Grad-CAM,Heating system…
4	5G mobile communication,B5G networks,Cloud com…
5	Benchmark,Benchmark testing,Machine learning,M…
6	Buildings,Denial-of-service attack,explanation…
7	Carbon dioxide,Estimation error,Intelligent Ve…
8	CI/CD,Conferences,Containerization,Deployment,…
9	Artificial intelligence; AI,Data models,Distri…
10	Conferences,Data drift,Data models,deep learni…
11	diversity,Germanium,Grammar,Grammatical evolut…
12	Analytical models,Arguments,Goal-Oriented Requ…
13	brain tumor,Classification algorithms,Convolut…
14	Architecture,Artificial Intelligence,Cloud,Com…
15	Action Research,Architectural alternatives,Art…
16	Bibliographies,Companies,Embedded systems,Fram…
17	Dimensionality reduction,Ensemble blending,Fea…
18	Clinical MLOps,Clinical research support,Cloud…
19	Benchmark testing,Conferences,continuous integ…
20	Adaptation models,Bibliographies,Computational…
21	AI Governance,Artificial intelligence,Deep lea…
22	Conferences,Distributed processing,Edge AI,Int…
23	Agent,Autonomic,Benchmark testing,Cloud comput…
24	AI,DevOps,ethics,Ethics,important,Industries,m…
25	AI,Artificial intelligence,Computational model…

```
26  Data models,development,DevOps,machine learnin…
27  Conferences,development,DevOps,machine learnin…
28  Automation,Databases,DevOps,Industries,Machine…
29  AI life-cycle,analytic pipeline,Atomic layer d…
30  AIOps,Codes,Conferences,DevOps,Machine learnin…
31  Companies,COVID-19,Deep Learning,Forecasting,L…
32  Automation,Bot (Internet),bots,Conferences,dee…
33  5G Networks,AI,Atmospheric modeling,Automation…
34  agile,Agile software development,Complexity th…
35  Automation,Codes,Continuous Integration (CI),M…
36  Collaboration,CPS,Deployment,Digital systems,E…
37  5G mobile communication,Automation,Quality of …
38  best practices,important,machine learning engi…
39  Artificial intelligence,beyond-schema inferenc…
40  AutoML,Computational modeling,Conferences,Depl…
41  DataOps,Decision making,Machine learning,Machi…
42  Automation,Business,continuous delivery,contin…
43  Automation,Azure,Big Data,Buildings,CI/CD,DevO…
44  AI-Powered Systems,Architecture,Artificial Int…
45  5G,Cloud computing,Collaborative work,Computat…
46  automation,catalogues,Codes,Computer architect…
47  Adaptation models,Analytical models,Computatio…
48  Computational modeling,continuous training,Dat…

                                                abstract  \
0   Empowering the Internet of Things devices with…
1   Machine Learning Operations (MLOps) is an appr…
2   Due to the migration megatrend, efficient and …
3   Machine Learning (ML) is a fundamental part of…
4   Open Radio Access Network (O-RAN) alliance was…
5   Machine learning (ML) is becoming critical to …
6   We have developed a Distributed Denial of Serv…
7   The increased number of sensors in modern cars…
8   In recent years, model deployment in machine l…
9   The emerging age of connected, digital world m…
10  Despite the significant improvements made by d…
11  The advent of cloud-based super-computing plat…
12  Requirements engineering for machine learning …
13  The research of brain tumor classification by …
14  Machine learning and deep learning techniques …
15  Since the advent of mobile computing and IoT, …
16  The adoption of continuous software engineerin…
17  The industrial machine learning applications t…
18  Epilepsy is a major neurological disorder affe…
19  In this paper, we present a coverage-based reg…
20  Deploying machine learning (ML) models to prod…
21  In this study we explore the incorporation of …
```

```
22  Deploying machine learning applications on edg…
23  Machine Learning (ML) projects are currently h…
24  Although AI is transforming the world, there a…
25  Following continuous software engineering prac…
26  DevOps practices have increasingly been applie…
27  DevOps practices have increasingly been applie…
28  DevOps and Machine Learning (ML) on their own …
29  In the last years, MLOps (Machine Learning Ope…
30  DevOps practices are the de facto sandard when…
31  In power grids, short-term load forecasting (S…
32  Machine learning (ML) operations or MLOps advo…
33  Artificial Intelligence of Things (AIoT) is th…
34  Software development teams are often hampered …
35  Machine Learning is a widely popular field tha…
36  The traditional field of industrial manufactur…
37  In this paper we present QMP\textemdash an AI-…
38  Background. The increasing reliance on applica…
39  AI convergence platforms such as Google's Unif…
40  This paper is an concentrated overview of the …
41  Even simply through a GoogleTrends search it b…
42  Over the past few decades, the substantial gro…
43  Microsoft Azure DevOps is a robust ,cross plat…
44  The research on engineering software applicati…
45  Machine Learning (ML) on the edge is key to en…
46  Nowadays, there are a variety of problems asso…
47  Real-world machine learning applications need …
48  The development and deployment of machine lear…

                                      doi       issn           pages  \
0              10.1109/EAIS51927.2022.9787703  2473-4691            1--8
1                 10.1109/ACCESS.2022.3206366  2169-3536    99337--99352
2                     10.1145/3522664.3528592        NaN          22--32
3               10.1109/ICSTW52544.2021.00039        NaN        175--181
4                 10.1109/OJCOMS.2022.3146618  2644-125X        228--250
5               10.1109/ICMLA51294.2020.00104        NaN        626--633
6            10.1109/UEMCON53757.2021.9666619        NaN      0171--0177
7   10.1109/MetroAutomotive54295.2022.9855110        NaN          18--23
8                10.1109/AIKE52691.2021.00010        NaN          25--28
9                10.1109/WAIN52551.2021.00019        NaN          82--88
10              10.1109/ICDMW53433.2021.00049  2375-9259        341--349
11                10.1109/ACCESS.2022.3166115  2169-3536    38694--38708
12                10.1109/RE48521.2020.00046  2332-6441        346--351
13            10.1109/ICBDS53701.2022.9936020        NaN            1--6
14              10.1109/APSEC51365.2020.00048  2640-0715        395--404
15               10.1109/SEAA51224.2020.00015        NaN          21--28
16               10.1109/SEAA53835.2021.00050        NaN            1--8
17                          10.15439/2022F296        NaN        403--412
```

18	10.1109/BHI50953.2021.9508555	2641-3604	1--5
19	10.1109/SANER50967.2021.00077	1534-5351	618--621
20	10.1145/3526073.3527584	NaN	1--8
21	10.1145/3522664.3528598	NaN	113--123
22	10.1109/IPDPSW55747.2022.00160	NaN	1003--1010
23	10.1109/CCGrid54584.2022.00047	NaN	376--385
24	10.1145/3522664.3528607	NaN	101--112
25	10.1109/WAIN52551.2021.00024	NaN	109--112
26	10.1145/3522664.3528611	NaN	33--34
27	10.1145/3526073.3527591	NaN	45--49
28	10.1109/ICECET55527.2022.9872968	NaN	1--6
29	10.23919/SpliTech55088.2022.9854211	NaN	1--6
30	10.1109/SANER53432.2022.00155	1534-5351	1293--1294
31	10.1109/IISA56318.2022.9904363	NaN	1--8
32	10.1109/COMPSAC54236.2022.00171	0730-3157	1093--1097
33	10.1109/IC2E52221.2021.00034	NaN	191--200
34	10.1109/SLAAI-ICAI54477.2021.9664736	NaN	1--6
35	10.1109/ICECAA55415.2022.9936252	NaN	1292--1297
36	10.1109/MECO55406.2022.9797080	2637-9511	1--6
37	10.1109/MeditCom55741.2022.9928678	NaN	86--89
38	10.1145/3382494.3410681	NaN	1--12
39	10.1109/SDS54800.2022.00020	NaN	69--70
40	10.1109/CCWC54503.2022.9720902	NaN	0453--0460
41	10.1109/SYNASC51798.2020.00015	NaN	17--23
42	10.1109/ACCESS.2022.3181730	2169-3536	63606--63618
43	10.1109/BigData50022.2020.9377755	NaN	2375--2384
44	10.1109/ASE51524.2021.9678647	2643-1572	1368--1372
45	10.1109/ACCESS.2022.3207200	2169-3536	100867--100877
46	10.1109/ICSA-C54293.2022.00047	2768-4288	206--209
47	10.1109/TVCG.2022.3209465	1941-0506	1--11
48	10.1109/ICAICE51518.2020.00102	NaN	494--500

	month	year	author
0	May	2022	Antonini, Mattia and Pincheira, Miguel and Vec…
1	NaN	2022	Barrak, Amine and Petrillo, Fabio and Jaafar, …
2	May	2022	Borg, Markus and Bengtsson, Johan and {\"O}ste…
3	April	2021	Borg, Markus and Jabangwe, Ronald and {\AA}ber…
4	NaN	2022	Brik, Bouziane and Boutiba, Karim and Ksentini…
5	December	2020	Cardoso Silva, Lucas and Rezende Zagatti, Fern…
6	December	2021	Das, Saikat and Shiva, Sajjan
7	July	2022	Flores, Thommas and Silva, Marianne and Andrad…
8	December	2021	Garg, Satvik and Pundir, Pradyumn and Rathee, …
9	May	2021	Granlund, Tuomas and Kopponen, Aleksi and Stir…
10	December	2021	Greco, Salvatore and Cerquitelli, Tania
11	NaN	2022	Gupt, Krishn Kumar and Raja, Muhammad Adil and…
12	August	2020	Ishikawa, Fuyuki and Matsuno, Yutaka
13	September	2022	Jain, Archit and Malviya, Adarsh and Bajaj, Di…

```
14   December  2020  John, Meenu Mary and Olsson, Helena Holmstr{\"…
15     August  2020  John, Meenu Mary and Holmstr{\"o}m Olsson, Hel…
16  September  2021  John, Meenu Mary and Olsson, Helena Holmstr{\"…
17  September  2022  Kannout, Eyad and Grodzki, Micha{\l} and Grzeg…
18       July  2021  Kar{\'a}csony, Tam{\'a}s and {Loesch-Biffar}, …
19      March  2021  Kauhanen, Eero and Nurminen, Jukka K. and Mikk…
20        May  2022             Kolltveit, Ask Berstad and Li, Jingyue
21        May  2022  Laato, Samuli and Birkstedt, Teemu and M{\"a}n…
22        May  2022  Leroux, Sam and Simoens, Pieter and Lootus, Me…
23        May  2022  Liu, Peini and {Bravo-Rocca}, Gusseppe and Gui…
24        May  2022  Lu, Qinghua and Zhu, Liming and Xu, Xiwei and …
25        May  2021  M{\"a}kinen, Sasu and Skogstr{\"o}m, Henrik an…
26        May  2022          Matsui, Beatriz M. A. and Goya, Denise H.
27        May  2022          Matsui, Beatriz M. A. and Goya, Denise H.
28       July  2022  Mboweni, Tsakani and Masombuka, Themba and Don…
29       July  2022  Mi{\~n}{\'o}n, Ra{\'u}l and {D{\'i}az-de-Arcay…
30      March  2022  Moreschini, Sergio and Lomio, Francesco and H{…
31       July  2022  Pelekis, Sotiris and Karakolis, Evangelos and …
32       June  2022  Rahman, Akond and Bhuiyan, Farzana Ahamed and …
33    October  2021  Raj, Emmanuel and Buffoni, David and Westerlun…
34   December  2021         Ranawana, Romesh and Karunananda, Asoka S.
35    October  2022                           R, Niranjan D and {Mohana}
36       June  2022  Ruf, Philipp and Reich, Christoph and {Ould-Ab…
37  September  2022  Samaras, Georgios and Theodorou, Vasileios and…
38    October  2020  Serban, Alex and {van der Blom}, Koen and Hoos…
39       June  2022                                      Spillner, Josef
40    January  2022  Symeonidis, Georgios and Nerantzis, Evangelos …
41  September  2020                                  Tamburri, Damian A.
42        NaN  2022  Testi, Matteo and Ballabio, Matteo and Fronton…
43   December  2020  Vuppalapati, Chandrasekar and Ilapakurti, Anit…
44   November  2021            Yasser, Ammar and {Abu-Elkhier}, Mervat
45        NaN  2022  Zaidi, Syed Ali Raza and Hayajneh, Ali M. and …
46      March  2022  Z{\'a}rate, Gorka and Mi{\~n}{\'o}n, Ra{\'u}l …
47        NaN  2022  Zhang, Xiaoyu and Ono, Jorge Piazentin and Son…
48    October  2020                 Zhou, Yue and Yu, Yue and Ding, Bo

                                          booktitle  \
0   2022 {{IEEE International Conference}} on {{Ev…
1                                                NaN
2   2022 {{IEEE}}/{{ACM}} 1st {{International Conf…
3   2021 {{IEEE International Conference}} on {{So…
4                                                NaN
5   2020 19th {{IEEE International Conference}} on…
6   2021 {{IEEE}} 12th {{Annual Ubiquitous Computi…
7   2022 {{IEEE International Workshop}} on {{Metr…
8   2021 {{IEEE Fourth International Conference}} …
9   2021 {{IEEE}}/{{ACM}} 1st {{Workshop}} on {{AI…
```

```
10  2021 {{International Conference}} on {{Data Mi…
11                                                NaN
12  2020 {{IEEE}} 28th {{International Requirement…
13  2022 {{IEEE International Conference}} on {{Bl…
14  2020 27th {{Asia-Pacific Software Engineering …
15  2020 46th {{Euromicro Conference}} on {{Softwa…
16  2021 47th {{Euromicro Conference}} on {{Softwa…
17  2022 17th {{Conference}} on {{Computer Science…
18  2021 {{IEEE EMBS International Conference}} on…
19  2021 {{IEEE International Conference}} on {{So…
20  2022 {{IEEE}}/{{ACM}} 1st {{International Work…
21  2022 {{IEEE}}/{{ACM}} 1st {{International Conf…
22  2022 {{IEEE International Parallel}} and {{Dis…
23  2022 22nd {{IEEE International Symposium}} on …
24  2022 {{IEEE}}/{{ACM}} 1st {{International Conf…
25  2021 {{IEEE}}/{{ACM}} 1st {{Workshop}} on {{AI…
26  2022 {{IEEE}}/{{ACM}} 1st {{International Conf…
27  2022 {{IEEE}}/{{ACM}} 1st {{International Work…
28  2022 {{International Conference}} on {{Electri…
29  2022 7th {{International Conference}} on {{Sma…
30  2022 {{IEEE International Conference}} on {{So…
31  2022 13th {{International Conference}} on {{In…
32  2022 {{IEEE}} 46th {{Annual Computers}}, {{Sof…
33  2021 {{IEEE International Conference}} on {{Cl…
34  2021 5th {{SLAAI International Conference}} on…
35  2022 {{International Conference}} on {{Edge Co…
36  2022 11th {{Mediterranean Conference}} on {{Em…
37  2022 {{IEEE International Mediterranean Confer…
38  Proceedings of the 14th {{ACM}} / {{IEEE Inter…
39  2022 9th {{Swiss Conference}} on {{Data Scienc…
40  2022 {{IEEE}} 12th {{Annual Computing}} and {{…
41  2020 22nd {{International Symposium}} on {{Sym…
42                                                NaN
43  2020 {{IEEE International Conference}} on {{Bi…
44  2021 36th {{IEEE}}/{{ACM International Confere…
45                                                NaN
46  2022 {{IEEE}} 19th {{International Conference}…
47                                                NaN
48  2020 {{International Conference}} on {{Artific…

                                          shorttitle  \
0                                     Tiny-{{MLOps}}
1                   Serverless on {{Machine Learning}}
2                                                 NaN
3                                                 NaN
4   Deep {{Learning}} for {{B5G Open Radio Access …
5                                                 NaN
```

```
6                                                    NaN
7                                                    NaN
8                                                    NaN
9   {{MLOps Challenges}} in {{Multi-Organization S…
10                                        Drift {{Lens}}
11                                                   NaN
12                                                   NaN
13                                                   NaN
14                      {{AI Deployment Architecture}}
15                              {{AI}} on the {{Edge}}
16                                   Towards {{MLOps}}
17                                                   NaN
18                                       {{DeepEpil}}
19                                                   NaN
20                                                   NaN
21  {{AI Governance}} in the {{System Development …
22                                       {{TinyMLOps}}
23                                    Scanflow-{{K8s}}
24                                                   NaN
25                                  Who {{Needs MLOps}}
26                                           {{MLOps}}
27                                           {{MLOps}}
28                                                   NaN
29                                       {{MLPacker}}
30                                                   NaN
31  In {{Search}} of {{Deep Learning Architectures…
32                Towards {{Automation}} for {{MLOps}}
33                                       Edge {{MLOps}}
34                                                   NaN
35                                Jenkins {{Pipelines}}
36                                                   NaN
37                                             {{QMP}}
38                                                   NaN
39                                                   NaN
40                                                   NaN
41                                Sustainable {{MLOps}}
42                                           {{MLOps}}
43                                                   NaN
44            Towards {{Fluid Software Architectures}}
45                                                   NaN
46                                             {{K2E}}
47                                       {{SliceTeller}}
48                                    Towards {{MLOps}}

                                                 title      ENTRYTYPE  \
0   Tiny-{{MLOps}}: A Framework for Orchestrating …  inproceedings
1   Serverless on {{Machine Learning}}: {{A System…        article
```

```
2   Quality {{Assurance}} of {{Generative Dialog M…  inproceedings
3   Test {{Automation}} with {{Grad-CAM Heatmaps}}…  inproceedings
4   Deep {{Learning}} for {{B5G Open Radio Access …        article
5   Benchmarking {{Machine Learning Solutions}} in…  inproceedings
6   Machine {{Learning}} Application Lifecycle Aug…  inproceedings
7   A {{TinyML Soft-Sensor}} for the {{Internet}} …  inproceedings
8   On {{Continuous Integration}} / {{Continuous D…  inproceedings
9   {{MLOps Challenges}} in {{Multi-Organization S…  inproceedings
10  Drift {{Lens}}: {{Real-time}} Unsupervised {{C…  inproceedings
11  {{GELAB}} \textendash{} {{The Cutting Edge}} o…        article
12  Evidence-Driven {{Requirements Engineering}} f…  inproceedings
13  Brain {{Tumor Detection}} Using {{MLops}} and …  inproceedings
14  {{AI Deployment Architecture}}: {{Multi-Case S…  inproceedings
15  {{AI}} on the {{Edge}}: {{Architectural Altern…  inproceedings
16  Towards {{MLOps}}: {{A Framework}} and {{Matur…  inproceedings
17  Considering Various Aspects of Models' Quality…  inproceedings
18  {{DeepEpil}}: {{Towards}} an {{Epileptologist-…  inproceedings
19  Regression {{Test Selection Tool}} for {{Pytho…  inproceedings
20  Operationalizing {{Machine Learning Models}} -…  inproceedings
21  {{AI Governance}} in the {{System Development …  inproceedings
22  {{TinyMLOps}}: {{Operational Challenges}} for …  inproceedings
23  Scanflow-{{K8s}}: {{Agent-based Framework}} fo…  inproceedings
24  Towards a {{Roadmap}} on {{Software Engineerin…  inproceedings
25  Who {{Needs MLOps}}: {{What Data Scientists Se…  inproceedings
26  {{MLOps}}: {{Five Steps}} to {{Guide}} Its {{E…  inproceedings
27  {{MLOps}}: {{A Guide}} to Its {{Adoption}} in …  inproceedings
28  A {{Systematic Review}} of {{Machine Learning …  inproceedings
29  {{MLPacker}}: {{A Unified Software Tool}} for …  inproceedings
30  {{MLOps}} for Evolvable {{AI}} Intensive Softw…  inproceedings
31  In {{Search}} of {{Deep Learning Architectures…  inproceedings
32  Towards {{Automation}} for {{MLOps}}: {{An Exp…  inproceedings
33  Edge {{MLOps}}: {{An Automation Framework}} fo…  inproceedings
34  An {{Agile Software Development Life Cycle Mod…  inproceedings
35  Jenkins {{Pipelines}}: {{A Novel Approach}} to…  inproceedings
36  Aspects of {{Module Placement}} in {{Machine L…  inproceedings
37  {{QMP}}: {{A Cloud-native MLOps Automation Pla…  inproceedings
38  Adoption and {{Effects}} of {{Software Enginee…  inproceedings
39  Tabular {{Data Insights}} and {{Synthesis}} wi…  inproceedings
40  {{MLOps}} - {{Definitions}}, {{Tools}} and {{C…  inproceedings
41  Sustainable {{MLOps}}: {{Trends}} and {{Challe…  inproceedings
42    {{MLOps}}: {{A Taxonomy}} and a {{Methodology}}        article
43  Automating {{Tiny ML Intelligent Sensors DevOP…  inproceedings
44  Towards {{Fluid Software Architectures}}: {{Bi…  inproceedings
45  Unlocking {{Edge Intelligence Through Tiny Mac…        article
46  {{K2E}}: {{Building MLOps Environments}} for {…  inproceedings
47  {{SliceTeller}} : {{A Data Slice-Driven Approa…        article
48  Towards {{MLOps}}: {{A Case Study}} of {{ML Pi…  inproceedings
```

	ID	volume
0	antoniniTinyMLOpsFrameworkOrchestrating2022	NaN
1	barrakServerlessMachineLearning2022	10
2	borgQualityAssuranceGenerative2022	NaN
3	borgTestAutomationGradCAM2021	NaN
4	brikDeepLearningB5G2022	3
5	cardososilvaBenchmarkingMachineLearning2020	NaN
6	dasMachineLearningApplication2021	NaN
7	floresTinyMLSoftSensorInternet2022	NaN
8	gargContinuousIntegrationContinuous2021	NaN
9	granlundMLOpsChallengesMultiOrganization2021	NaN
10	grecoDriftLensRealtime2021	NaN
11	guptGELABCuttingEdge2022	10
12	ishikawaEvidencedrivenRequirementsEngineering2020	NaN
13	jainBrainTumorDetection2022	NaN
14	johnAIDeploymentArchitecture2020	NaN
15	johnAIEdgeArchitectural2020	NaN
16	johnMLOpsFrameworkMaturity2021	NaN
17	kannoutConsideringVariousAspects2022	NaN
18	karacsonyDeepEpilEpileptologistFriendlyAI2021	NaN
19	kauhanenRegressionTestSelection2021	NaN
20	kolltveitOperationalizingMachineLearning2022	NaN
21	laatoAIGovernanceSystem2022	NaN
22	lerouxTinyMLOpsOperationalChallenges2022	NaN
23	liuScanflowK8sAgentbasedFramework2022	NaN
24	luRoadmapSoftwareEngineering2022	NaN
25	makinenWhoNeedsMLOps2021	NaN
26	matsuiMLOpsFiveSteps2022	NaN
27	matsuiMLOpsGuideIts2022	NaN
28	mboweniSystematicReviewMachine2022	NaN
29	minonMLPackerUnifiedSoftware2022	NaN
30	moreschiniMLOpsEvolvableAI2022	NaN
31	pelekisSearchDeepLearning2022	NaN
32	rahmanAutomationMLOpsExploratory2022	NaN
33	rajEdgeMLOpsAutomation2021	NaN
34	ranawanaAgileSoftwareDevelopment2021	NaN
35	rJenkinsPipelinesNovel2022	NaN
36	rufAspectsModulePlacement2022	NaN
37	samarasQMPCloudnativeMLOps2022	NaN
38	serbanAdoptionEffectsSoftware2020	NaN
39	spillnerTabularDataInsights2022	NaN
40	symeonidisMLOpsDefinitionsTools2022	NaN
41	tamburriSustainableMLOpsTrends2020	NaN
42	testiMLOpsTaxonomyMethodology2022	10
43	vuppalapatiAutomatingTinyML2020	NaN
44	yasserFluidSoftwareArchitectures2021	NaN

```
45                 zaidiUnlockingEdgeIntelligence2022     10
46                         zarateK2EBuildingMLOps2022    NaN
47                 zhangSliceTellerDataSliceDriven2022    NaN
48                             zhouMLOpsCaseStudy2020    NaN

                                            journal               isbn  \
0                                               NaN                NaN
1                                       IEEE Access                NaN
2                                               NaN                NaN
3                                               NaN                NaN
4     IEEE Open Journal of the Communications Society                NaN
5                                               NaN                NaN
6                                               NaN                NaN
7                                               NaN                NaN
8                                               NaN                NaN
9                                               NaN                NaN
10                                              NaN                NaN
11                                      IEEE Access                NaN
12                                              NaN                NaN
13                                              NaN                NaN
14                                              NaN                NaN
15                                              NaN                NaN
16                                              NaN                NaN
17                                              NaN                NaN
18                                              NaN                NaN
19                                              NaN                NaN
20                                              NaN                NaN
21                                              NaN                NaN
22                                              NaN                NaN
23                                              NaN                NaN
24                                              NaN                NaN
25                                              NaN                NaN
26                                              NaN                NaN
27                                              NaN                NaN
28                                              NaN                NaN
29                                              NaN                NaN
30                                              NaN                NaN
31                                              NaN                NaN
32                                              NaN                NaN
33                                              NaN                NaN
34                                              NaN                NaN
35                                              NaN                NaN
36                                              NaN                NaN
37                                              NaN                NaN
38                                              NaN  978-1-4503-7580-1
39                                              NaN                NaN
40                                              NaN                NaN
```

41	NaN	NaN
42	IEEE Access	NaN
43	NaN	NaN
44	NaN	NaN
45	IEEE Access	NaN
46	NaN	NaN
47	IEEE Transactions on Visualization and Compute…	NaN
48	NaN	NaN

	address	publisher	series
0	NaN	NaN	NaN
1	NaN	NaN	NaN
2	NaN	NaN	NaN
3	NaN	NaN	NaN
4	NaN	NaN	NaN
5	NaN	NaN	NaN
6	NaN	NaN	NaN
7	NaN	NaN	NaN
8	NaN	NaN	NaN
9	NaN	NaN	NaN
10	NaN	NaN	NaN
11	NaN	NaN	NaN
12	NaN	NaN	NaN
13	NaN	NaN	NaN
14	NaN	NaN	NaN
15	NaN	NaN	NaN
16	NaN	NaN	NaN
17	NaN	NaN	NaN
18	NaN	NaN	NaN
19	NaN	NaN	NaN
20	NaN	NaN	NaN
21	NaN	NaN	NaN
22	NaN	NaN	NaN
23	NaN	NaN	NaN
24	NaN	NaN	NaN
25	NaN	NaN	NaN
26	NaN	NaN	NaN
27	NaN	NaN	NaN
28	NaN	NaN	NaN
29	NaN	NaN	NaN
30	NaN	NaN	NaN
31	NaN	NaN	NaN
32	NaN	NaN	NaN
33	NaN	NaN	NaN
34	NaN	NaN	NaN
35	NaN	NaN	NaN
36	NaN	NaN	NaN

```
37                  NaN                                        NaN           NaN
38  {New York, NY, USA}  {Association for Computing Machinery}  {{ESEM}} '20
39                  NaN                                        NaN           NaN
40                  NaN                                        NaN           NaN
41                  NaN                                        NaN           NaN
42                  NaN                                        NaN           NaN
43                  NaN                                        NaN           NaN
44                  NaN                                        NaN           NaN
45                  NaN                                        NaN           NaN
46                  NaN                                        NaN           NaN
47                  NaN                                        NaN           NaN
48                  NaN                                        NaN           NaN
```

4.2 Remov unuseful columns

[8]:
```
df.columns
```

[8]:
```
Index(['file', 'note', 'keywords', 'abstract', 'doi', 'issn', 'pages', 'month',
       'year', 'author', 'booktitle', 'shorttitle', 'title', 'ENTRYTYPE', 'ID',
       'volume', 'journal', 'isbn', 'address', 'publisher', 'series'],
      dtype='object')
```

[9]:
```
unuseful_cols = [
    'file', 'doi', 'issn', 'pages', 'booktitle', 'shorttitle', 'month',
    'volume', 'journal', 'isbn', 'address', 'publisher', 'series', 'ENTRYTYPE'
]
```

[10]:
```
useful_columns = [col for col in df.columns if col not in unuseful_cols]
useful_columns
```

[10]:
```
['note', 'keywords', 'abstract', 'year', 'author', 'title', 'ID']
```

[11]:
```
df = df[useful_columns]
```

[12]:
```
df.head()
```

[12]:
```
                                                note  \
0  2\n\par\nresearch: solution\n\par\ncont: cases…
1  2\n\par\nresearch: philosophical\n\par\ncont: …
2  2\n\par\nresearch: solution\n\par\ncont: cases…
3  2\n\par\nresearch: validation\n\par\ncont: cas…
4  2\n\par\nresearch: philosophical\n\par\ncont: …

                                            keywords  \
0  Adaptation models,Costs,Image edge detection,R…
1  Cloud computing,Computational modeling,Compute…
2  action research,AI quality,Context modeling,co…
```

```
3  Automation,Conferences,Grad-CAM,Heating system…
4  5G mobile communication,B5G networks,Cloud com…

                                              abstract  year  \
0  Empowering the Internet of Things devices with…  2022
1  Machine Learning Operations (MLOps) is an appr…  2022
2  Due to the migration megatrend, efficient and …  2022
3  Machine Learning (ML) is a fundamental part of…  2021
4  Open Radio Access Network (O-RAN) alliance was…  2022

                                               author  \
0  Antonini, Mattia and Pincheira, Miguel and Vec…
1  Barrak, Amine and Petrillo, Fabio and Jaafar, …
2  Borg, Markus and Bengtsson, Johan and {\"O}ste…
3  Borg, Markus and Jabangwe, Ronald and {\AA}ber…
4  Brik, Bouziane and Boutiba, Karim and Ksentini…

                                                title  \
0  Tiny-{{MLOps}}: A Framework for Orchestrating …
1  Serverless on {{Machine Learning}}: {{A System…
2  Quality {{Assurance}} of {{Generative Dialog M…
3  Test {{Automation}} with {{Grad-CAM Heatmaps}}…
4  Deep {{Learning}} for {{B5G Open Radio Access …

                                             ID
0  antoniniTinyMLOpsFrameworkOrchestrating2022
1          barrakServerlessMachineLearning2022
2           borgQualityAssuranceGenerative2022
3                borgTestAutomationGradCAM2021
4                      brikDeepLearningB5G2022
```

[13]:
```
df.shape
```

[13]:
```
(49, 7)
```

4.3 Add columns for mapping study facets

[14]:
```
# add new columns
only_notes_df = df[df.note.notna()].copy()

if len(only_notes_df) != len(df):
    raise ValueError('There are articles with missing notes')

only_notes_df = only_notes_df.reindex(only_notes_df.columns.tolist() +
                                      search_words,
                                      axis=1)
```

```
only_notes_df.shape  #.columns
```

[14]: (49, 11)

5 Final DF

[15]:
```
df = only_notes_df.copy()
df.head()
```

[15]:

	note	keywords	abstract	year	author	title	ID	research	cont	domain
0	2\n\par\nresearch: solution\n\par\ncont: cases…	Adaptation models,Costs,Image edge detection,R…	Empowering the Internet of Things devices with…	2022	Antonini, Mattia and Pincheira, Miguel and Vec…	Tiny-{{MLOps}}: A Framework for Orchestrating …	antoniniTinyMLOpsFrameworkOrchestrating2022	NaN	NaN	NaN
1	2\n\par\nresearch: philosophical\n\par\ncont: …	Cloud computing,Computational modeling,Compute…	Machine Learning Operations (MLOps) is an appr…	2022	Barrak, Amine and Petrillo, Fabio and Jaafar, …	Serverless on {{Machine Learning}}: {{A System…				
2	2\n\par\nresearch: solution\n\par\ncont: cases…	action research,AI quality,Context modeling,co…	Due to the migration megatrend, efficient and …	2022	Borg, Markus and Bengtsson, Johan and {\"O}ste…	Quality {{Assurance}} of {{Generative Dialog M…				
3	2\n\par\nresearch: validation\n\par\ncont: cas…	Automation,Conferences,Grad-CAM,Heating system…	Machine Learning (ML) is a fundamental part of…	2021	Borg, Markus and Jabangwe, Ronald and {\AA}ber…	Test {{Automation}} with {{Grad-CAM Heatmaps}}…				
4	2\n\par\nresearch: philosophical\n\par\ncont: …	5G mobile communication,B5G networks,Cloud com…	Open Radio Access Network (O-RAN) alliance was…	2022	Brik, Bouziane and Boutiba, Karim and Ksentini…	Deep {{Learning}} for {{B5G Open Radio Access …				

```
1          barrakServerlessMachineLearning2022       NaN   NaN     NaN
2           borgQualityAssuranceGenerative2022       NaN   NaN     NaN
3                borgTestAutomationGradCAM2021       NaN   NaN     NaN
4                      brikDeepLearningB5G2022       NaN   NaN     NaN

   summary
0      NaN
1      NaN
2      NaN
3      NaN
4      NaN
```

[16]:
```
df.shape
```

[16]:
```
(49, 11)
```

[17]:
```
# Export dataframe to excel
# df.to_excel("bibliography_dataframe.xlsx")
```

5.1 Table describing the columns

In the following cell a table with corresponding column descriptions is created. This is necessary for the mapping study.

[18]:
```
descriptions = [
    'Those are the notes that are taken with Zotero. This column is used for␣
 ↪extracting further information for the facets',
    'Automatically extracted keywords via Zotero tool. Keywords are␣
 ↪categorizing the article to a degree.',
    'Abstract (summary) of the article. Extracted via Zotero tool.',
    'Year of publication. Extracted via Zotero tool.',
    'Author of publication. Extracted via Zotero tool.',
    'Title of publication. Extracted via Zotero tool.',
    'ID of publication in this dataframe. Extracted via Zotero tool.',
    'Research facet according to mapping study.',
    'Contribution facet according to mapping study',
    'Domain facet according to domain study',
    'Short summary notes'

]

explain_cols = pd.DataFrame(df.columns, columns=['Data Item'])
explain_cols['Description']= descriptions
explain_cols['Relevant RQ']= None

explain_cols.to_excel("dataframe_explanations.xlsx")
```

```
explain_cols
```

[18]:

	Data Item	Description	Relevant RQ
0	note	Those are the notes that are taken with Zotero…	None
1	keywords	Automatically extracted keywords via Zotero to…	None
2	abstract	Abstract (summary) of the article. Extracted v…	None
3	year	Year of publication. Extracted via Zotero tool.	None
4	author	Author of publication. Extracted via Zotero tool.	None
5	title	Title of publication. Extracted via Zotero tool.	None
6	ID	ID of publication in this dataframe. Extracted…	None
7	research	Research facet according to mapping study.	None
8	cont	Contribution facet according to mapping study	None
9	domain	Domain facet according to domain study	None
10	summary	Short summary notes	None

6 Explore notes

6.1 Extract search terms from notes

[19]:
```
def extract_search_terms(search_term, note, row):
    cols = {}
    res = None

    cleaned_note = note.replace("\\par", "")
    cleaned_note = cleaned_note.split('\n')

    for word in search_words:
        for content in cleaned_note:
            if word in content:
                try:
                    key, val = re.split(':', content)
                    cols[key] = val.replace(" ", "")

                except ValueError as e:
                    print(f'{word} not in {content} -> skip: ', row.title, e)
    #                   print(cleaned_note)

    try:
        res = cols[search_term]
    except KeyError as e:
        print(f'no {search_term} skip: ', row.title)

    return res


# print(extract_search_terms(df[df.note.notna()].note.iloc[0], df[df.note.
 ↪notna()].iloc[0])['research'])
```

[20]:
```
for term in search_words:
    df[term] = df.apply(
        lambda row: extract_search_terms(term, row['note'], row), axis=1)

df.head()
```

```
cont not in summary: data version control, model drift, SLR and GLR, nice
validation case study through companies, RQ: what is the state-of-the-art
regarding the adoption of MLOps in practice and the different stages that
companies go through in evolving their MLOps practices? -> skip:  Towards
{{MLOps}}: {{A Framework}} and {{Maturity Model}} too many values to unpack
(expected 2)
summary not in summary: data version control, model drift, SLR and GLR, nice
validation case study through companies, RQ: what is the state-of-the-art
regarding the adoption of MLOps in practice and the different stages that
companies go through in evolving their MLOps practices? -> skip:  Towards
{{MLOps}}: {{A Framework}} and {{Maturity Model}} too many values to unpack
(expected 2)
```

```
cont not in summary: data version control, model drift, SLR and GLR, nice
validation case study through companies, RQ: what is the state-of-the-art
regarding the adoption of MLOps in practice and the different stages that
companies go through in evolving their MLOps practices? -> skip:  Towards
{{MLOps}}: {{A Framework}} and {{Maturity Model}} too many values to unpack
(expected 2)
summary not in summary: data version control, model drift, SLR and GLR, nice
validation case study through companies, RQ: what is the state-of-the-art
regarding the adoption of MLOps in practice and the different stages that
companies go through in evolving their MLOps practices? -> skip:  Towards
{{MLOps}}: {{A Framework}} and {{Maturity Model}} too many values to unpack
(expected 2)
cont not in summary: data version control, model drift, SLR and GLR, nice
validation case study through companies, RQ: what is the state-of-the-art
regarding the adoption of MLOps in practice and the different stages that
companies go through in evolving their MLOps practices? -> skip:  Towards
{{MLOps}}: {{A Framework}} and {{Maturity Model}} too many values to unpack
(expected 2)
summary not in summary: data version control, model drift, SLR and GLR, nice
validation case study through companies, RQ: what is the state-of-the-art
regarding the adoption of MLOps in practice and the different stages that
companies go through in evolving their MLOps practices? -> skip:  Towards
{{MLOps}}: {{A Framework}} and {{Maturity Model}} too many values to unpack
(expected 2)
cont not in summary: data version control, model drift, SLR and GLR, nice
validation case study through companies, RQ: what is the state-of-the-art
regarding the adoption of MLOps in practice and the different stages that
companies go through in evolving their MLOps practices? -> skip:  Towards
{{MLOps}}: {{A Framework}} and {{Maturity Model}} too many values to unpack
(expected 2)
summary not in summary: data version control, model drift, SLR and GLR, nice
validation case study through companies, RQ: what is the state-of-the-art
regarding the adoption of MLOps in practice and the different stages that
companies go through in evolving their MLOps practices? -> skip:  Towards
{{MLOps}}: {{A Framework}} and {{Maturity Model}} too many values to unpack
(expected 2)
no summary skip:  Towards {{MLOps}}: {{A Framework}} and {{Maturity Model}}
```

[20]:

```
                                               note  \
0  2\n\par\nresearch: solution\n\par\ncont: cases…
1  2\n\par\nresearch: philosophical\n\par\ncont: …
2  2\n\par\nresearch: solution\n\par\ncont: cases…
3  2\n\par\nresearch: validation\n\par\ncont: cas…
4  2\n\par\nresearch: philosophical\n\par\ncont: …

                                           keywords  \
0  Adaptation models,Costs,Image edge detection,R…
```

```
1  Cloud computing,Computational modeling,Compute…
2  action research,AI quality,Context modeling,co…
3  Automation,Conferences,Grad-CAM,Heating system…
4  5G mobile communication,B5G networks,Cloud com…

                                             abstract  year  \
0  Empowering the Internet of Things devices with…  2022
1  Machine Learning Operations (MLOps) is an appr…  2022
2  Due to the migration megatrend, efficient and …  2022
3  Machine Learning (ML) is a fundamental part of…  2021
4  Open Radio Access Network (O-RAN) alliance was…  2022

                                               author  \
0  Antonini, Mattia and Pincheira, Miguel and Vec…
1  Barrak, Amine and Petrillo, Fabio and Jaafar, …
2  Borg, Markus and Bengtsson, Johan and {\"O}ste…
3  Borg, Markus and Jabangwe, Ronald and {\AA}ber…
4  Brik, Bouziane and Boutiba, Karim and Ksentini…

                                                title  \
0  Tiny-{{MLOps}}: A Framework for Orchestrating …
1  Serverless on {{Machine Learning}}: {{A System…
2  Quality {{Assurance}} of {{Generative Dialog M…
3  Test {{Automation}} with {{Grad-CAM Heatmaps}}…
4  Deep {{Learning}} for {{B5G Open Radio Access …

                                             ID       research  \
0  antoniniTinyMLOpsFrameworkOrchestrating2022       solution
1          barrakServerlessMachineLearning2022  philosophical
2           borgQualityAssuranceGenerative2022       solution
3                borgTestAutomationGradCAM2021     validation
4                      brikDeepLearningB5G2022  philosophical

                                                cont  \
0   casestudy,approach,model,tool,process,experiment
1            approach,model,metric,process,literature
2  casestudy,approach,model,metric,process,experi…
3              casestudy,approach,process,experiment
4  casestudy,approach,model,metric,tool,process,e…

                                               domain  \
0  iot,tinyml,mlops,deployment,anomalydetection,i…
1  mlops,pipeline,workflow,modelling,serverless,c…
2                                          qa,model
3                  deeplearning,modelling,automation
4                        deeplearning,infrastructure
```

```
                                summary
0           implementingtinymlformlops
1                     studymappingonml
2                   verylittleonmlops
3  exampleondeeplearningexplainability
4                                 None
```

[21]:
```
df.shape
```

[21]:
```
(49, 11)
```

7 Explore facets

7.1 Research Facet

[22]:
```
df.research.value_counts()
```

[22]:
```
solution         18
philosophical    18
evaluation       10
validation        3
Name: research, dtype: int64
```

[23]:
```
val_counts = df.research.str.split(',').explode().value_counts()

if len(val_counts) > 6:
    raise ValueError('Error with splitting. Too many values to for Research␣
 ↪facet', len(val_counts))

print(f'''Sum of research facet units

{val_counts}

in total: {val_counts.sum()}
''')
```

```
Sum of research facet units

solution         18
philosophical    18
evaluation       10
validation        3
Name: research, dtype: int64

in total: 49
```

[24]:
```
fig = px.bar(
    df.research.str.split(',').explode().value_counts())  #, barmode='group')
fig.show()
```

[25]:
```
df[df.research == 'solution'].title.head()
```

[25]:
```
0     Tiny-{{MLOps}}: A Framework for Orchestrating …
2     Quality {{Assurance}} of {{Generative Dialog M…
10    Drift {{Lens}}: {{Real-time}} Unsupervised {{C…
11    {{GELAB}} \textendash{} {{The Cutting Edge}} o…
13    Brain {{Tumor Detection}} Using {{MLops}} and …
Name: title, dtype: object
```

7.2 Cont Facet

[26]:
```
val_counts = df.cont.str.split(',').explode().value_counts()

if len(val_counts) > 9:
    raise ValueError('Error with splitting. Too many values to for Contribution␣
 ↪facet')

print(f'''Sum of contribution facet units

{val_counts}

in total: {val_counts.sum()}
''')
```

```
Sum of contribution facet units

approach        49
model           46
process         46
casestudy       30
metric          23
tool            18
experiment      15
literature      12
nonempirical     7
Name: cont, dtype: int64

in total: 246
```

[27]:
```
fig = px.bar(
    df.cont.str.split(',').explode().value_counts())  #, barmode='group')
fig.show()
```

[28]:
```
temp = df[df.cont.str.contains('model','literature')]##.title#.head()
temp[temp.research == 'philosophical'].title
```

[28]:
```
1     Serverless on {{Machine Learning}}: {{A System…
4     Deep {{Learning}} for {{B5G Open Radio Access …
6     Machine {{Learning}} Application Lifecycle Aug…
8     On {{Continuous Integration}} / {{Continuous D…
9     {{MLOps Challenges}} in {{Multi-Organization S…
12    Evidence-Driven {{Requirements Engineering}} f…
16    Towards {{MLOps}}: {{A Framework}} and {{Matur…
20    Operationalizing {{Machine Learning Models}} -…
21    {{AI Governance}} in the {{System Development …
22    {{TinyMLOps}}: {{Operational Challenges}} for …
24    Towards a {{Roadmap}} on {{Software Engineerin…
26    {{MLOps}}: {{Five Steps}} to {{Guide}} Its {{E…
27    {{MLOps}}: {{A Guide}} to Its {{Adoption}} in …
30    {{MLOps}} for Evolvable {{AI}} Intensive Softw…
34    An {{Agile Software Development Life Cycle Mod…
40    {{MLOps}} - {{Definitions}}, {{Tools}} and {{C…
41    Sustainable {{MLOps}}: {{Trends}} and {{Challe…
42      {{MLOps}}: {{A Taxonomy}} and a {{Methodology}}
Name: title, dtype: object
```

[29]:
```
# DF of counted units per facet


# res_cont = df.cont.str.split(',').explode().value_counts().to_frame(
#     'count').rename_axis('cont').reset_index()

# res_research = df.research.str.split(',').explode().value_counts().to_frame(
#     'count').rename_axis('research').reset_index()

# res = pd.concat([res_research, res_cont])
# res
```

7.3 Domain Facet

[30]:
```
val_counts = df.domain.str.strip().str.split(',').explode().value_counts()
val_counts


for val in val_counts.keys():
    if val == '':
        raise ValueError('Error with splitting. There are empty parts', val)

print(f'''Sum of domain facet units
```

```
{val_counts}

in total: {val_counts.sum()}
''')
```

Sum of domain facet units

mlops	33
pipeline	24
deeplearning	13
deployment	12
cd	11
ci	11
architecture	11
monitoring	9
iot	7
cloud	7
automation	7
infrastructure	7
docker	6
devops	6
data	6
modelling	6
tinyml	5
datapreparation	5
workflow	5
development	4
training	4
kubernetes	4
business	4
ethics	4
dataengineering	3
datacollection	3
validation	3
container	3
storage	3
pipelines	3
xai	3
privacy	3
governance	3
model	3
federatedlearning	3
fairness	3
edgecomputing	3
explainability	2
lstm	2
argo	2

mlflow	2
packaging	2
prometheus	2
grafana	2
ai	2
inference	2
kubeflow	2
ml	2
git	2
tools	2
security	2
architecure	2
tensorflow	2
datalake	1
modelpackaging	1
agile	1
audit	1
modeldrift	1
dataops	1
experimentation	1
dataflow	1
orchestration	1
versioncontrol	1
metainformation	1
opensource	1
bots	1
algorithm	1
mysql	1
forecasting	1
visualanalytics	1
optimization	1
gpu	1
aiops	1
tfx	1
bentoml	1
modelops	1
gitea	1
models	1
jenkins	1
transferlearning	1
interpretability	1
eda	1
datadrift	1
architecturea	1
artefacts	1
mlaas	1
datascience	1
iaac	1

hyperparametertuning	1
analysis	1
experimenttracking	1
versioning	1
datalabeling	1
registry	1
featreengineering	1
sagemaker	1
cm	1
featureengineering	1
sustainability	1
team	1
coding	1
svr	1
kafka	1
microservice	1
modularization	1
stages	1
pca	1
azure	1
python	1
responsibleai	1
worfklow	1
classification	1
labeling	1
bert	1
conceptdrift	1
challenges	1
delivery	1
aws	1
github	1
gitops	1
deplyoment	1
ann	1
~stages	1
shap	1
theory	1
mllifecycle	1
threads	1
algorithms	1
sklearn	1
keras	1
qa	1
mlalgorithms	1
batching	1
coldstart	1
scaling	1
serverless	1

```
anomalydetection            1
tool                        1
geneticalgorithm            1
IDE                         1
matlab                      1
humancontrol                1
pricinples                  1
seldon                      1
cluster                     1
distributed                 1
serving                     1
integration                 1
modelstorage                1
featurestore                1
testing                     1
modeltraining               1
featureselection            1
dataanalysis                1
modelregistry               1
release                     1
ops                         1
optimumsearch               1
costs                       1
dataprivacy                 1
modelperfromance            1
iam                         1
cnn                         1
imageprocessing             1
DAG                         1
requirementsengineering     1
drone                       1
Name: domain, dtype: int64

in total: 396
```

[31]:
```
fig = px.bar(
    df.domain.str.split(',').explode().value_counts())  #, barmode='group')
fig.show()
```

8 Explore interaction of research and contribution facets

[32]:
```
research_cols
```

[32]:
```
['evaluation',
 'experience',
```

```
 'opinion',
 'philosophical',
 'solution',
 'validation']
```

[33]:
```
applied_research_cols = df['research'].value_counts().keys()

if len(research_cols) > len(applied_research_cols):
    research_cols = applied_research_cols.to_list()
    print(f'''
Only found the following research facet cols:
{research_cols}
''')
```

```
Only found the following research facet cols:
['solution', 'philosophical', 'evaluation', 'validation']
```

8.1 Correlations

Sanity check of facets by correlation with crosstab with exploded content (1-hot encoded)

[34]:
```
exploded_cont = df['cont'].str.get_dummies(sep=',')
exploded_research = df['research'].str.get_dummies(sep=',')
exploded_cont

exploded = pd.concat([df, exploded_cont, exploded_research], axis=1)



exploded_facets = exploded[cont_cols + research_cols]
print(f'''
Overall sums:
{exploded_facets.sum()}

and total nr of facet units: {exploded_facets.sum().sum()}
''')
```

```
Overall sums:
approach         49
casestudy        30
experiment       15
literature       12
metric           23
model            46
nonempirical      7
```

```
process          46
tool             18
solution         18
philosophical    18
evaluation       10
validation        3
dtype: int64

and total nr of facet units: 295
```

[35]:
```
corr = exploded_facets.corr()
corr

corr = corr[abs(corr) >= 0.4]

fig = px.imshow(corr, text_auto=True, aspect='auto')
# fig.update_layout(
#     autosize=True,
# #     width=800,
# #     height=800,
# )
fig.show()
```

8.2 Crosstab between research and contribution

By copying As the research facet is only assigned once per article, we would need to copy the multilabel column "cont" to create a crosstab.

This also implies that the **absolute numbers** for the research facet are not the ground truth.

[36]:
```
test = df.assign(cont=df.cont.str.split(',')).explode('cont')
test

cta = pd.crosstab(test.cont, test.research, margins=True)
cta
```

[36]:

research	evaluation	philosophical	solution	validation	All
cont					
approach	10	18	18	3	49
casestudy	5	8	14	3	30
experiment	3	1	9	2	15
literature	4	7	1	0	12
metric	6	5	11	1	23
model	9	18	17	2	46
nonempirical	1	6	0	0	7
process	10	17	16	3	46
tool	4	4	9	1	18

```
All                  52             84         95           15  246
```

[37]:
```
# cta = pd.crosstab(
#     index=[
#         df.assign(research=df.research.str.split(',')).explode(
#             'research').reset_index().research
#     ],
#     columns=[
#         df.assign(
#             cont=df.cont.str.split(',')).explode('cont').reset_index().cont
#     ],
#     margins=True)

# cta
```

8.2.1 Absolute numbers

[38]:
```
ct = cta[cta.columns[:-1]].iloc[:-1]
ct
```

[38]:

research	evaluation	philosophical	solution	validation
cont				
approach	10	18	18	3
casestudy	5	8	14	3
experiment	3	1	9	2
literature	4	7	1	0
metric	6	5	11	1
model	9	18	17	2
nonempirical	1	6	0	0
process	10	17	16	3
tool	4	4	9	1

Visualizations

[39]:
```
fig = px.imshow(ct, text_auto=True, aspect='auto', title='Contribution without␣
↪totals')
fig.show()
```

[40]:
```
#replaced_all_cell_for_heatmap = cta.replace(to_replace = cta.iloc[-1,-1],␣
↪value = 'All', inplace=False)

fig = px.imshow(cta.iloc[:-1], text_auto=True, aspect='auto',␣
↪title='Contribution with totals')
# fig = px.imshow(cta, text_auto=True)
fig.show()
```

[41]:
```
fig = px.bar(ct)  #, barmode='group')
fig.show()
```

[42]:
```
data = []
#use for loop on every zoo name to create bar data
for x in ct.columns:
    data.append(go.Bar(name=str(x), x=ct.index, y=ct[x]))

figure = go.Figure(data)
# figure.update_layout(barmode = 'stack')

#For you to take a look at the result use
figure.show()
```

8.2.2 Percent of Totals

[43]:
```
cta.pipe(
    lambda x: x.div(x['All'], axis='index')
).applymap('{:.0%}'.format).iloc[:-1]
```

[43]:

research	evaluation	philosophical	solution	validation	All
cont					
approach	20%	37%	37%	6%	100%
casestudy	17%	27%	47%	10%	100%
experiment	20%	7%	60%	13%	100%
literature	33%	58%	8%	0%	100%
metric	26%	22%	48%	4%	100%
model	20%	39%	37%	4%	100%
nonempirical	14%	86%	0%	0%	100%
process	22%	37%	35%	7%	100%
tool	22%	22%	50%	6%	100%

[44]:
```
cta.T.pipe(lambda x: x.div(x['All'], axis='index')).applymap('{:.0%}'.format).
 ↪iloc[:-1]
```

[44]:

cont	approach	casestudy	experiment	literature	metric	model	\
research							
evaluation	19%	10%	6%	8%	12%	17%	
philosophical	21%	10%	1%	8%	6%	21%	
solution	19%	15%	9%	1%	12%	18%	
validation	20%	20%	13%	0%	7%	13%	

cont	nonempirical	process	tool	All
research				
evaluation	2%	19%	8%	100%
philosophical	7%	20%	5%	100%
solution	0%	17%	9%	100%

```
validation              0%     20%   7%  100%
```

[45]:
```
cta.describe()
cta.T.describe()
```

[45]:
```
cont    approach  casestudy  experiment  literature     metric      model  \
count   5.000000   5.000000     5.00000    5.000000   5.000000   5.000000
mean   19.600000  12.000000     6.00000    4.800000   9.200000  18.400000
std    17.586927  10.885771     5.91608    4.868265   8.497058  16.742162
min     3.000000   3.000000     1.00000    0.000000   1.000000   2.000000
25%    10.000000   5.000000     2.00000    1.000000   5.000000   9.000000
50%    18.000000   8.000000     3.00000    4.000000   6.000000  17.000000
75%    18.000000  14.000000     9.00000    7.000000  11.000000  18.000000
max    49.000000  30.000000    15.00000   12.000000  23.000000  46.000000

cont   nonempirical    process       tool         All
count       5.000000   5.000000   5.000000    5.000000
mean        2.800000  18.400000   7.200000   98.400000
std         3.420526  16.410363   6.685806   88.194671
min         0.000000   3.000000   1.000000   15.000000
25%         0.000000  10.000000   4.000000   52.000000
50%         1.000000  16.000000   4.000000   84.000000
75%         6.000000  17.000000   9.000000   95.000000
max         7.000000  46.000000  18.000000  246.000000
```

Visualizations

[46]:
```
temp = cta.pipe(
    lambda x: x.div(x['All'], axis='index')
).applymap('{:.0%}'.format).iloc[:-1]#, :-1]
test = temp.copy()

for col in test.columns:
    test[col] = test[col].str.rstrip('%').astype('float') / 100.0

fig = px.imshow(test, text_auto=True, aspect='auto', title='Contribution facet␣
 ↪in percentages')
fig.show()
```

[47]:
```
temp = cta.T.pipe(lambda x: x.div(x['All'], axis='index')).applymap(
    '{:.0%}'.format).iloc[:-1]  #, :-1]
test = temp.copy()

for col in test.columns:
    test[col] = test[col].str.rstrip('%').astype('float') / 100.0

fig = px.imshow(test,
```

```
                text_auto=True,
                aspect='auto',
                title='Research facet in percentages')
fig.show()
```

9 Explore year

[48]:
```
fig = go.Figure()
fig.add_trace(go.Histogram(histfunc="count", x=df.year))
```

10 Explore Authors

[49]:
```
fig = px.bar(
    df.author.str.split(',').explode().value_counts())  #, barmode='group')
fig.show()
```

11 Explore Keywords

[50]:
```
df.keywords.str.split(',').explode().value_counts()
```

[50]:
```
MLOps                        27
Machine learning             24
Pipelines                    14
Training                     14
Software                     13
important                    13
DevOps                       11
Conferences                  11
machine learning              9
Data models                   9
Computational modeling        8
Automation                    8
Computer architecture         7
Cloud computing               7
Machine Learning              6
Tools                         6
Task analysis                 6
Production                    6
Deployment                    5
Predictive models             5
deep learning                 5
Industries                    5
Deep learning                 5
Software engineering          4
```

Companies	4
Artificial intelligence	4
Codes	4
Benchmark testing	3
Adaptation models	3
Monitoring	3
Buildings	3
Stakeholders	3
Artificial Intelligence	3
Deep Learning	3
Collaboration	3
Systematics	3
Organizations	3
AI	3
Interviews	3
Analytical models	2
software engineering	2
continuous training	2
Software systems	2
DataOps	2
CI/CD	2
sustainability	2
Forecasting	2
Market research	2
Kubernetes	2
SLR	2
Real-time systems	2
Bibliographies	2
Packaging	2
Business	2
Collaborative work	2
Embedded systems	2
Edge	2
Cloud	2
Optimization	2
Sociology	2
Architecture	2
Image segmentation	2
Transfer learning	2
transfer learning	2
IoT	2
continuous integration	2
TinyML	2
responsible AI	2
test automation	2
Time series analysis	2
development	2

Software architecture	2
model	2
software development life cycle	2
5G mobile communication	2
Visualization	2
ML	2
Robustness	2
Neural networks	2
Model Evaluation	1
Libraries	1
Focusing	1
Agile software development	1
mlops	1
agile	1
Regulation	1
delivery pipeline	1
continuous software engineering	1
Edge Computing	1
Model Validation	1
Digital Transformation	1
5G Networks	1
Atmospheric modeling	1
software architecture	1
Software algorithms	1
empirical study	1
Human-in-the-loop	1
Data-Centric AI	1
requirement engineering	1
Data Validation	1
Regulators	1
Text analysis	1
bots	1
devops	1
Load forecasting	1
AI life-cycle	1
analytic pipeline	1
Atomic layer deposition	1
end-to-end platform	1
deploying	1
Search problems	1
packaging	1
ML-DevOps	1
AIOps	1
Software Engineering	1
Databases	1
COVID-19	1
LSTM	1

deployment	1
NBEATS	1
Out-of-Distribution Generalization	1
Data collection	1
Writing	1
Pandemics	1
Standards organizations	1
Short-Term Load Forecasting	1
Smart Grid	1
Sustainability	1
Temporal Convolution	1
Bot (Internet)	1
testtag	1
Complexity theory	1
implementation	1
data-centric	1
data	1
dataset	1
Fluids	1
AI-Powered Systems	1
Planets	1
Intelligent sensors	1
Infrastructure as code	1
Big Data	1
Azure	1
management	1
XAI	1
Surgery	1
continuous monitoring	1
continuous delivery	1
Sustainable development	1
Software Sustainability	1
datalake	1
Human-AI Interaction	1
Middleware	1
catalogues	1
gesture recognition	1
federated learning	1
Internet of Things	1
Energy efficiency	1
Logic gates	1
LoRa	1
Memory management	1
Performance evaluation	1
energy efficiency	1
Edge computing	1
edge computing	1

5G	1
Software Architecture	1
Tiny machine learning	1
automation	1
Scientific computing	1
Machine-Learning Operations	1
experimentation	1
survey	1
best practices	1
Quality of service	1
Modularization	1
Manufacturing industries	1
Embedded computing	1
Digital systems	1
CPS	1
Source Code Management (SCM)	1
Software Development Lifecycle (SDLC)	1
Manuals	1
Machine Learning Operations (MLOps)	1
Machine Learning (ML)	1
Continuous Integration (CI)	1
Data Slicing	1
SDLC	1
machine learning engineering	1
beyond-schema inference	1
Decision making	1
Convergence	1
metadata	1
MlOps	1
models	1
robustness	1
re-training	1
monitoring	1
fairness	1
versioning	1
explainability	1
AutoML	1
Training data	1
pattern recognition	1
Ethics	1
Data science	1
Data intelligence	1
Internet	1
Video based diagnosis support	1
ethics	1
text topic classification	1
Soft-Sensor	1

Regression	1
Quantization (signal)	1
Quantization	1
OBD-II	1
Neurons	1
Intelligent Vehicles	1
Estimation error	1
Carbon dioxide	1
security augmented ML life cycle	1
Mobile communication	1
MLOpS	1
Machine learning life cycle	1
Intrusion detection	1
Fasteners	1
Temperature measurement	1
Torque	1
Containerization	1
information systems	1
Data drift	1
software engineering for AI/ML	1
multi-organisation	1
machine learning; ML	1
Learning (artificial intelligence)	1
integration	1
Distributed databases	1
Docker	1
Artificial intelligence; AI	1
Software development management	1
Orchestration	1
Kubeflow	1
Knowledge engineering	1
GitOps	1
explanation augmented ML life cycle	1
Denial-of-service attack	1
Systems operation	1
Serverless computing	1
Context modeling	1
AI quality	1
action research	1
systematic mapping	1
systematic literature review	1
SM	1
Serverless	1
generative dialog model	1
function as a service	1
FaaS	1
Transforms	1

System performance	1
Sensors	1
Image edge detection	1
conversational agent	1
Natural language processing	1
Benchmark	1
neural networks	1
Resource management	1
RAN intelligent controller	1
RAN	1
Radio access networks	1
open RAN architecture	1
B5G networks	1
machine learning testing	1
Quality assurance	1
image recognition	1
Heating systems	1
Grad-CAM	1
Testing	1
software testing	1
requirements engineering	1
Text categorization	1
transformer-based models	1
Self-Management	1
Transformers	1
Timing	1
Servers	1
Regression test selection	1
mutation testing	1
Measurement	1
Error analysis	1
Costs	1
Three-dimensional displays	1
Semiotics	1
Seizure semiology	1
Epilepsy	1
Clinical research support	1
Clinical MLOps	1
XGBoost	1
Stability analysis	1
Operationalization	1
Publishing	1
Systematic literature review	1
Reliability	1
Scanflow	1
Machine Learning Workflow	1
Image classification	1

Autonomic	1
Agent	1
TinyMLOps	1
Privacy	1
AI Governance	1
Intellectual property	1
Edge AI	1
Distributed processing	1
system development life cycle	1
software development	1
Encoding	1
MOO	1
MCDA	1
Logistics	1
Arguments	1
Classification algorithms	1
brain tumor	1
Uncertainty	1
Requirements Engineering	1
Requirements engineering	1
Goal-Oriented Requirements Analysis	1
Statistics	1
hybrid multi cloud	1
Matlab	1
hybrid optimization	1
Grammatical evolution	1
Grammar	1
Germanium	1
diversity	1
Convolutional neural networks	1
machine learning operations (MLOps)	1
Feature selection	1
Framework	1
Feature extraction	1
Ensemble blending	1
Dimensionality reduction	1
Validation Study	1
Maturity Model	1
GLR	1
Transfer Learning	1
Magnetic resonance imaging	1
Architectural alternatives	1
Action Research	1
Prototypes	1
smart healthcare	1
security	1
Object detection	1

```
Graphics processing units                1
Name: keywords, dtype: int64
```

[51]:
```
fig = px.bar(
    df.keywords.str.split(',').explode().value_counts())  #, barmode='group')
fig.show()
```

[52]:
```
df.keywords.str.split(',').explode().value_counts()
```

[52]:
```
MLOps                                   27
Machine learning                        24
Pipelines                               14
Training                                14
Software                                13
important                               13
DevOps                                  11
Conferences                             11
machine learning                         9
Data models                              9
Computational modeling                   8
Automation                               8
Computer architecture                    7
Cloud computing                          7
Machine Learning                         6
Tools                                    6
Task analysis                            6
Production                               6
Deployment                               5
Predictive models                        5
deep learning                            5
Industries                               5
Deep learning                            5
Software engineering                     4
Companies                                4
Artificial intelligence                  4
Codes                                    4
Benchmark testing                        3
Adaptation models                        3
Monitoring                               3
Buildings                                3
Stakeholders                             3
Artificial Intelligence                  3
Deep Learning                            3
Collaboration                            3
Systematics                              3
Organizations                            3
AI                                       3
```

Interviews	3
Analytical models	2
software engineering	2
continuous training	2
Software systems	2
DataOps	2
CI/CD	2
sustainability	2
Forecasting	2
Market research	2
Kubernetes	2
SLR	2
Real-time systems	2
Bibliographies	2
Packaging	2
Business	2
Collaborative work	2
Embedded systems	2
Edge	2
Cloud	2
Optimization	2
Sociology	2
Architecture	2
Image segmentation	2
Transfer learning	2
transfer learning	2
IoT	2
continuous integration	2
TinyML	2
responsible AI	2
test automation	2
Time series analysis	2
development	2
Software architecture	2
model	2
software development life cycle	2
5G mobile communication	2
Visualization	2
ML	2
Robustness	2
Neural networks	2
Model Evaluation	1
Libraries	1
Focusing	1
Agile software development	1
mlops	1
agile	1

Regulation	1
delivery pipeline	1
continuous software engineering	1
Edge Computing	1
Model Validation	1
Digital Transformation	1
5G Networks	1
Atmospheric modeling	1
software architecture	1
Software algorithms	1
empirical study	1
Human-in-the-loop	1
Data-Centric AI	1
requirement engineering	1
Data Validation	1
Regulators	1
Text analysis	1
bots	1
devops	1
Load forecasting	1
AI life-cycle	1
analytic pipeline	1
Atomic layer deposition	1
end-to-end platform	1
deploying	1
Search problems	1
packaging	1
ML-DevOps	1
AIOps	1
Software Engineering	1
Databases	1
COVID-19	1
LSTM	1
deployment	1
NBEATS	1
Out-of-Distribution Generalization	1
Data collection	1
Writing	1
Pandemics	1
Standards organizations	1
Short-Term Load Forecasting	1
Smart Grid	1
Sustainability	1
Temporal Convolution	1
Bot (Internet)	1
testtag	1
Complexity theory	1

implementation	1
data-centric	1
data	1
dataset	1
Fluids	1
AI-Powered Systems	1
Planets	1
Intelligent sensors	1
Infrastructure as code	1
Big Data	1
Azure	1
management	1
XAI	1
Surgery	1
continuous monitoring	1
continuous delivery	1
Sustainable development	1
Software Sustainability	1
datalake	1
Human-AI Interaction	1
Middleware	1
catalogues	1
gesture recognition	1
federated learning	1
Internet of Things	1
Energy efficiency	1
Logic gates	1
LoRa	1
Memory management	1
Performance evaluation	1
energy efficiency	1
Edge computing	1
edge computing	1
5G	1
Software Architecture	1
Tiny machine learning	1
automation	1
Scientific computing	1
Machine-Learning Operations	1
experimentation	1
survey	1
best practices	1
Quality of service	1
Modularization	1
Manufacturing industries	1
Embedded computing	1
Digital systems	1

CPS	1
Source Code Management (SCM)	1
Software Development Lifecycle (SDLC)	1
Manuals	1
Machine Learning Operations (MLOps)	1
Machine Learning (ML)	1
Continuous Integration (CI)	1
Data Slicing	1
SDLC	1
machine learning engineering	1
beyond-schema inference	1
Decision making	1
Convergence	1
metadata	1
MlOps	1
models	1
robustness	1
re-training	1
monitoring	1
fairness	1
versioning	1
explainability	1
AutoML	1
Training data	1
pattern recognition	1
Ethics	1
Data science	1
Data intelligence	1
Internet	1
Video based diagnosis support	1
ethics	1
text topic classification	1
Soft-Sensor	1
Regression	1
Quantization (signal)	1
Quantization	1
OBD-II	1
Neurons	1
Intelligent Vehicles	1
Estimation error	1
Carbon dioxide	1
security augmented ML life cycle	1
Mobile communication	1
MLOpS	1
Machine learning life cycle	1
Intrusion detection	1
Fasteners	1

Temperature measurement	1
Torque	1
Containerization	1
information systems	1
Data drift	1
software engineering for AI/ML	1
multi-organisation	1
machine learning; ML	1
Learning (artificial intelligence)	1
integration	1
Distributed databases	1
Docker	1
Artificial intelligence; AI	1
Software development management	1
Orchestration	1
Kubeflow	1
Knowledge engineering	1
GitOps	1
explanation augmented ML life cycle	1
Denial-of-service attack	1
Systems operation	1
Serverless computing	1
Context modeling	1
AI quality	1
action research	1
systematic mapping	1
systematic literature review	1
SM	1
Serverless	1
generative dialog model	1
function as a service	1
FaaS	1
Transforms	1
System performance	1
Sensors	1
Image edge detection	1
conversational agent	1
Natural language processing	1
Benchmark	1
neural networks	1
Resource management	1
RAN intelligent controller	1
RAN	1
Radio access networks	1
open RAN architecture	1
B5G networks	1
machine learning testing	1

Quality assurance	1
image recognition	1
Heating systems	1
Grad-CAM	1
Testing	1
software testing	1
requirements engineering	1
Text categorization	1
transformer-based models	1
Self-Management	1
Transformers	1
Timing	1
Servers	1
Regression test selection	1
mutation testing	1
Measurement	1
Error analysis	1
Costs	1
Three-dimensional displays	1
Semiotics	1
Seizure semiology	1
Epilepsy	1
Clinical research support	1
Clinical MLOps	1
XGBoost	1
Stability analysis	1
Operationalization	1
Publishing	1
Systematic literature review	1
Reliability	1
Scanflow	1
Machine Learning Workflow	1
Image classification	1
Autonomic	1
Agent	1
TinyMLOps	1
Privacy	1
AI Governance	1
Intellectual property	1
Edge AI	1
Distributed processing	1
system development life cycle	1
software development	1
Encoding	1
MOO	1
MCDA	1
Logistics	1

```
Arguments                                  1
Classification algorithms                  1
brain tumor                                1
Uncertainty                                1
Requirements Engineering                   1
Requirements engineering                   1
Goal-Oriented Requirements Analysis        1
Statistics                                 1
hybrid multi cloud                         1
Matlab                                     1
hybrid optimization                        1
Grammatical evolution                      1
Grammar                                    1
Germanium                                  1
diversity                                  1
Convolutional neural networks              1
machine learning operations (MLOps)        1
Feature selection                          1
Framework                                  1
Feature extraction                         1
Ensemble blending                          1
Dimensionality reduction                   1
Validation Study                           1
Maturity Model                             1
GLR                                        1
Transfer Learning                          1
Magnetic resonance imaging                 1
Architectural alternatives                 1
Action Research                            1
Prototypes                                 1
smart healthcare                           1
security                                   1
Object detection                           1
Graphics processing units                  1
Name: keywords, dtype: int64
```

12 Cluster similar words from keywords and identify groups

12.1 Load Model

[53]:
```
try:
    from gensim.models import KeyedVectors
    model = KeyedVectors.load('models/word2vec-google-news-300.model')

except:
    print('Couldnt find a saved model')
    import gensim.downloader as api
```

```
    model = api.load('word2vec-google-news-300')
    model.save('models/word2vec-google-news-300.model')
```

12.2 Get keywords and compare to vocab in pre-trained model

[54]:
```
keywords_expl = df.keywords.str.split(',').explode()
listed_words = keywords_expl

print(keywords_expl)
```

```
0                         Adaptation models
0                                     Costs
0                     Image edge detection
0                        Real-time systems
0                                  Sensors
0                       System performance
0                                Transforms
1                           Cloud computing
1                  Computational modeling
1                   Computer architecture
1                             Data models
1                                    FaaS
1                  function as a service
1                                important
1                      machine learning
1                      Machine learning
1                             Serverless
1                    Serverless computing
1                                     SLR
1                                      SM
1           systematic literature review
1                     systematic mapping
2                        action research
2                             AI quality
2                          Context modeling
2                  conversational agent
2              generative dialog model
2                                Interviews
2                          Machine learning
2           Natural language processing
2                     Quality assurance
2                requirements engineering
2                    Software engineering
2                        software testing
2                                 Testing
3                              Automation
```

3	Conferences
3	Grad-CAM
3	Heating systems
3	image recognition
3	Image segmentation
3	machine learning testing
3	neural networks
3	Neural networks
3	Pipelines
3	test automation
3	Visualization
4	5G mobile communication
4	B5G networks
4	Cloud computing
4	Computer architecture
4	deep learning
4	Deep learning
4	Industries
4	MLOps
4	open RAN architecture
4	Radio access networks
4	RAN
4	RAN intelligent controller
4	Resource management
5	Benchmark
5	Benchmark testing
5	Machine learning
5	Machine Learning
5	MLOps
5	Monitoring
5	Production
5	Systems operation
5	Task analysis
5	Tools
6	Buildings
6	Denial-of-service attack
6	explanation augmented ML life cycle
6	Fasteners
6	Intrusion detection
6	Machine learning
6	Machine learning life cycle
6	MLOpS
6	Mobile communication
6	Predictive models
6	security augmented ML life cycle
7	Carbon dioxide
7	Estimation error
7	Intelligent Vehicles

7	MLOps
7	Neurons
7	OBD-II
7	Quantization
7	Quantization (signal)
7	Regression
7	Soft-Sensor
7	Temperature measurement
7	TinyML
7	Torque
7	Training
8	CI/CD
8	Conferences
8	Containerization
8	Deployment
8	DevOps
8	Docker
8	GitOps
8	Knowledge engineering
8	Kubeflow
8	Kubernetes
8	Machine learning
8	MLOps
8	Orchestration
8	Organizations
8	Pipelines
8	Software
8	Software development management
9	Artificial intelligence; AI
9	Data models
9	Distributed databases
9	information systems
9	integration
9	Learning (artificial intelligence)
9	Machine learning
9	machine learning; ML
9	multi-organisation
9	Organizations
9	Pipelines
9	Software
9	software engineering for AI/ML
10	Conferences
10	Data drift
10	Data models
10	deep learning
10	Deep learning
10	MLOps
10	Predictive models

10	Real-time systems
10	Text categorization
10	text topic classification
10	transformer-based models
10	Transformers
11	diversity
11	Germanium
11	Grammar
11	Grammatical evolution
11	hybrid optimization
11	Matlab
11	Optimization
11	Sociology
11	Software
11	Statistics
12	Analytical models
12	Arguments
12	Goal-Oriented Requirements Analysis
12	Machine learning
12	Machine Learning
12	Monitoring
12	Requirements engineering
12	Requirements Engineering
12	Stakeholders
12	Task analysis
12	Uncertainty
13	brain tumor
13	Classification algorithms
13	Convolutional neural networks
13	deep learning
13	hybrid multi cloud
13	Image segmentation
13	machine learning operations (MLOps)
13	Magnetic resonance imaging
13	Object detection
13	security
13	smart healthcare
13	Training
13	transfer learning
13	Transfer learning
14	Architecture
14	Artificial Intelligence
14	Cloud
14	Companies
14	Computer architecture
14	Conferences
14	Deep Learning
14	Deployment

14	Edge
14	Embedded systems
14	Interviews
14	Machine Learning
14	Prototypes
14	Software engineering
15	Action Research
15	Architectural alternatives
15	Artificial intelligence
15	Artificial Intelligence
15	Cloud
15	Cloud computing
15	Collaboration
15	Companies
15	Computer architecture
15	Deep Learning
15	Edge
15	important
15	Interviews
15	Machine Learning
15	Packaging
15	Transfer Learning
16	Bibliographies
16	Companies
16	Embedded systems
16	Framework
16	GLR
16	important
16	Machine learning
16	Maturity Model
16	MLOps
16	SLR
16	Software
16	Software engineering
16	Systematics
16	Validation Study
17	Dimensionality reduction
17	Ensemble blending
17	Feature extraction
17	Feature selection
17	Forecasting
17	Industries
17	Logistics
17	MCDA
17	MOO
17	Pipelines
17	Predictive models
17	Stability analysis

17	Training
17	XGBoost
18	Clinical MLOps
18	Clinical research support
18	Cloud computing
18	Collaboration
18	Epilepsy
18	Seizure semiology
18	Semiotics
18	Sociology
18	Three-dimensional displays
18	Tools
18	Video based diagnosis support
18	Visualization
19	Benchmark testing
19	Conferences
19	continuous integration
19	Error analysis
19	Measurement
19	mutation testing
19	Regression test selection
19	Servers
19	software engineering
19	test automation
19	Timing
19	Tools
20	Adaptation models
20	Bibliographies
20	Computational modeling
20	Deployment
20	important
20	Machine learning
20	MLOps
20	Operationalization
20	Publishing
20	Systematic literature review
20	Systematics
20	Training
21	AI Governance
21	Artificial intelligence
21	Deep learning
21	DevOps
21	Encoding
21	machine learning
21	MLOps
21	Neural networks
21	Pipelines
21	Software

21	software development
21	software development life cycle
21	Stakeholders
21	system development life cycle
22	Conferences
22	Distributed processing
22	Edge AI
22	Intellectual property
22	Machine learning
22	MLOps
22	Privacy
22	Reliability
22	Task analysis
22	TinyML
22	TinyMLOps
23	Agent
23	Autonomic
23	Benchmark testing
23	Cloud computing
23	Computer architecture
23	Image classification
23	important
23	Kubernetes
23	Machine learning
23	Machine Learning Workflow
23	MLOps
23	Pipelines
23	Robustness
23	Scanflow
23	Self-Management
24	AI
24	DevOps
24	ethics
24	Ethics
24	important
24	Industries
24	machine learning
24	MLOps
24	Regulation
24	Regulators
24	requirement engineering
24	responsible AI
24	Software
24	Software algorithms
24	software architecture
24	software engineering
24	Stakeholders
25	AI

```
25                  Artificial intelligence
25                   Computational modeling
25            continuous software engineering
25                              Data models
25                     delivery pipeline
25                                  DevOps
25                                Focusing
25                      machine learning
25                                     ML
25                                  MLOps
25                               Production
25                     Time series analysis
25                                  Tools
25                                Training
26                               Data models
26                               development
26                                    DevOps
26                           machine learning
26                           Machine learning
26                                  MLOps
26                                  model
26                               Pipelines
26                                Software
26                  Standards organizations
26                                Training
26                                 Writing
27                               Conferences
27                               development
27                                    DevOps
27                           machine learning
27                           Machine learning
27                                  MLOps
27                                  model
27                            responsible AI
27                                  Software
27                     Software engineering
28                                Automation
28                                 Databases
28                                    DevOps
28                               Industries
28                           Machine learning
28                           Machine Learning
28                                  ML-DevOps
28                                      MLOps
28                            Search problems
28                                Systematics
28                                    testtag
28                              Text analysis
```

29	AI life-cycle
29	analytic pipeline
29	Atomic layer deposition
29	Codes
29	deploying
29	Industries
29	Machine learning
29	MLOps
29	packaging
29	Packaging
29	Pipelines
29	Training
30	AIOps
30	Codes
30	Conferences
30	DevOps
30	Machine learning
30	MLOps
30	Pipelines
30	Software
30	Software Engineering
30	Software systems
30	Task analysis
31	Companies
31	COVID-19
31	Deep Learning
31	Forecasting
31	Load forecasting
31	LSTM
31	MLOps
31	NBEATS
31	Out-of-Distribution Generalization
31	Pandemics
31	Predictive models
31	Short-Term Load Forecasting
31	Smart Grid
31	Sustainability
31	Temporal Convolution
31	Time series analysis
32	Automation
32	Bot (Internet)
32	bots
32	Conferences
32	deep learning
32	Deep learning
32	deployment
32	devops
32	empirical study

32	Libraries
32	machine learning
32	mlops
32	Software
32	Task analysis
33	5G Networks
33	AI
33	Atmospheric modeling
33	Automation
33	Cloud computing
33	Collaborative work
33	Computational modeling
33	Digital Transformation
33	Edge Computing
33	important
33	IoT
33	Machine Learning
33	MLOps
33	Pipelines
33	Training
34	agile
34	Agile software development
34	Complexity theory
34	Data collection
34	Data models
34	data-centric
34	DevOps
34	experimentation
34	Machine learning
34	MLOps
34	Production
34	SDLC
34	Software
34	software development life cycle
35	Automation
35	Codes
35	Continuous Integration (CI)
35	Machine learning
35	Machine Learning (ML)
35	Machine Learning Operations (MLOps)
35	Manuals
35	Pipelines
35	Software
35	Software Development Lifecycle (SDLC)
35	Source Code Management (SCM)
35	Training
36	Collaboration
36	CPS

36	Deployment
36	Digital systems
36	Embedded computing
36	Machine learning
36	Manufacturing industries
36	ML
36	MLOps
36	Modularization
36	Pipelines
36	Training
37	5G mobile communication
37	Automation
37	Quality of service
38	best practices
38	important
38	machine learning engineering
38	survey
39	Artificial intelligence
39	beyond-schema inference
39	Business
39	Convergence
39	Data intelligence
39	Data science
39	Internet
39	MLOps
39	pattern recognition
39	Training data
40	AutoML
40	Computational modeling
40	Conferences
40	Deployment
40	explainability
40	fairness
40	important
40	machine learning
40	Machine learning
40	Market research
40	MLOps
40	monitoring
40	Production
40	re-training
40	robustness
40	Robustness
40	sustainability
40	Training
41	DataOps
41	Decision making
41	Machine learning

41	Machine-Learning Operations
41	Market research
41	Middleware
41	MLOps
41	Scientific computing
41	Software Sustainability
41	Software systems
41	Sustainable development
42	Automation
42	Business
42	continuous delivery
42	continuous integration
42	continuous monitoring
42	continuous training
42	important
42	MLOps
42	Monitoring
42	Pipelines
42	Production
42	Surgery
42	sustainability
42	Training
42	XAI
43	Automation
43	Azure
43	Big Data
43	Buildings
43	CI/CD
43	DevOps
43	Infrastructure as code
43	Intelligent sensors
43	Machine learning
43	Planets
43	Tools
44	AI-Powered Systems
44	Architecture
44	Artificial Intelligence
44	Buildings
44	Computer architecture
44	Fluids
44	Human-AI Interaction
44	Machine learning
44	Production
44	Software architecture
44	Software Architecture
45	5G
45	Cloud computing
45	Collaborative work

45	Computational modeling
45	deep learning
45	Deep learning
45	edge computing
45	Edge computing
45	energy efficiency
45	Energy efficiency
45	federated learning
45	gesture recognition
45	implementation
45	Internet of Things
45	IoT
45	Logic gates
45	LoRa
45	Machine learning
45	Memory management
45	MLOps
45	Performance evaluation
45	Tiny machine learning
45	transfer learning
45	Transfer learning
46	automation
46	catalogues
46	Codes
46	Computer architecture
46	Conferences
46	data
46	Data models
46	datalake
46	DataOps
46	dataset
46	important
46	management
46	metadata
46	MlOps
46	models
46	Organizations
46	Software
46	Software architecture
46	versioning
47	Adaptation models
47	Analytical models
47	Computational modeling
47	Data models
47	Data Slicing
47	Data Validation
47	Data-Centric AI
47	Human-in-the-loop

```
47                              important
47                         Model Evaluation
47                         Model Validation
47                             Optimization
47                        Predictive models
47                                 Training
48                    Computational modeling
48                       continuous training
48                               Data models
48                                   DevOps
48                          end-to-end platform
48                  Graphics processing units
48                              important
48                         machine learning
48                                   MLOps
48                                Pipelines
48                           Task analysis
48                                   Tools
48                                Training
Name: keywords, dtype: object
```

[55]:
```
words = set(listed_words) & set(list(model.key_to_index.keys()))
vectors = list([model.get_vector(word) for word in words])

len(words), len(vectors)
```

[55]: (120, 120)

12.3 PCA

TSNE

https://stats.stackexchange.com/questions/263539/clustering-on-the-output-of-t-sne/264647#264647

only reproducable with high perplexity!

discouraged to be used with clustering

t-SNE is also a method to reduce the dimension. One of the most major differences between PCA and t-SNE is it preserves only local similarities whereas PA preserves large pairwise distance maximize variance.

- https://medium.com/analytics-vidhya/pca-vs-t-sne-17bcd882bf3d#:~:text=One%20of%20the%20most%20m

[56]:
```
pca = PCA(n_components=2, random_state=RNG)
pca_transformed = pca.fit_transform(vectors)
X_pca = pca_transformed
```

```
words = pd.DataFrame(words)
pca_df = pd.DataFrame(pca_transformed)
pca_df = pd.merge(words, pca_df, left_index=True, right_index=True)
pca_df.columns = ['words', 'x', 'y']

pca_df
```

[56]:

	words	x	y
0	experimentation	0.479172	-0.696349
1	Systematics	-0.109538	0.103520
2	Orchestration	0.710216	0.814788
3	integration	1.014274	-0.161324
4	robustness	1.059977	-0.680040
5	Companies	-0.606245	-0.060207
6	Business	-0.781408	0.063316
7	Stakeholders	-0.657384	-0.513601
8	Sensors	-0.632880	1.099867
9	ML	0.044387	-0.699915
10	Bibliographies	-0.008604	-0.359681
11	Encoding	0.692819	1.398205
12	Optimization	0.598027	1.242353
13	Reliability	-0.094660	0.489641
14	Production	-0.526405	0.374447
15	Middleware	1.006791	1.610532
16	Deployment	0.330417	0.882006
17	Torque	0.450761	0.196344
18	Industries	-1.330785	0.075619
19	Focusing	-0.183801	-0.554003
20	development	0.286127	-0.729154
21	data	0.553734	-0.422316
22	Privacy	-0.215381	-0.318253
23	Grammar	-0.447752	-0.240555
24	survey	-0.331875	-1.014053
25	RAN	0.678872	-0.893242
26	agile	1.098579	-0.314309
27	packaging	-0.029336	-0.442251
28	Sociology	-1.124959	-0.277150
29	Manuals	-0.199909	0.424420
30	Costs	-0.570416	0.239059
31	SLR	0.463587	-0.779939
32	Sustainability	-0.753130	-0.234583
33	MOO	0.373990	-0.997734
34	Codes	-0.230720	0.045811
35	Architecture	-0.121328	0.475496
36	versioning	2.279043	0.379008
37	Software	0.461481	1.055434
38	Servers	0.635757	1.008004

39	Fasteners	-1.075972	0.897930
40	Benchmark	-0.185844	0.175858
41	Pandemics	-0.793054	0.271990
42	Ethics	-0.995961	-0.727552
43	Serverless	0.314441	-0.297774
44	Fluids	-0.617162	0.911447
45	Robustness	0.730372	0.691305
46	GLR	0.179205	-0.811631
47	SDLC	1.199425	0.326177
48	Testing	-0.596668	0.414534
49	Internet	0.136960	-0.617710
50	Interviews	-1.113810	-0.486390
51	automation	1.044239	0.334569
52	Epilepsy	-1.171550	0.108959
53	Cloud	0.752292	0.115021
54	Regulation	-0.683954	-0.422788
55	Buildings	-0.951752	0.033059
56	Automation	-0.048252	1.256039
57	Logistics	-0.631514	0.521135
58	models	0.409738	-0.759249
59	Transformers	-0.086723	-0.631187
60	Conferences	-0.877325	0.087827
61	security	0.447678	-0.808339
62	fairness	-0.004936	-1.375538
63	Collaboration	0.066770	0.738917
64	DevOps	0.814766	-0.076258
65	Organizations	-0.105391	0.051824
66	Neurons	-0.113121	0.953543
67	Tools	-0.035128	1.039813
68	5G	0.257450	-0.573994
69	management	0.333942	-0.673455
70	ethics	-0.224658	-1.356046
71	Framework	0.649447	0.611592
72	SM	0.245650	-0.377783
73	Containerization	-0.112323	-0.151526
74	Prototypes	-0.277224	0.383764
75	Convergence	-0.068657	0.200362
76	deploying	0.995289	-0.348912
77	Arguments	-0.830795	-0.332765
78	Agent	-0.113036	-0.149078
79	monitoring	0.250282	-0.604213
80	MCDA	-0.184993	-0.621633
81	Autonomic	0.750456	0.508161
82	LSTM	-0.233332	-0.345717
83	Packaging	-1.078914	0.575858
84	Azure	0.998039	-0.237865
85	Surgery	-1.429866	0.282506

```
86                IoT  0.355856 -0.366343
87         deployment  1.119446 -0.411664
88                CPS -0.057774 -0.903955
89         Statistics -1.143261 -0.237288
90          important  0.141361 -0.597097
91               bots  0.869220 -0.390460
92          diversity  0.146108 -1.204207
93         Publishing -0.733686  0.252446
94        Measurement -0.596457  0.869325
95           Training -0.709120  0.251479
96            Writing -0.868043 -0.014604
97          Germanium -0.109735  0.273996
98               Edge  0.045002 -0.325179
99        Uncertainty -0.557754 -0.351688
100          metadata  2.048673  0.489678
101        Regression -0.086363  0.876600
102           Planets -0.047211  0.309067
103         Libraries -0.617319  0.312124
104    sustainability  0.056516 -1.087798
105         Pipelines -0.805116  0.747711
106     Visualization  0.445488  1.467344
107         Databases  0.029151  1.163607
108             model  0.439924 -1.017089
109            Docker -0.111136 -1.241025
110        Regulators -0.596595 -0.891750
111             Matlab  0.996214  0.514456
112        Transforms -0.107312  1.693265
113    implementation  0.730832 -0.485064
114        Monitoring -0.301218  0.272285
115            Timing -0.569479  0.171661
116           dataset  0.849675 -0.066859
117         Semiotics -0.203431 -0.099935
118       Forecasting -0.791867  0.289488
119                AI  0.543393 -0.556528
```

[57]:
```
pca_df[pca_df.words =='Tools']
```

[57]:
```
    words         x         y
67  Tools -0.035128  1.039813
```

[58]:
```
print(pca_df)
```

```
                words         x         y
0     experimentation  0.479172 -0.696349
1         Systematics -0.109538  0.103520
2       Orchestration  0.710216  0.814788
3         integration  1.014274 -0.161324
```

4	robustness	1.059977	-0.680040
5	Companies	-0.606245	-0.060207
6	Business	-0.781408	0.063316
7	Stakeholders	-0.657384	-0.513601
8	Sensors	-0.632880	1.099867
9	ML	0.044387	-0.699915
10	Bibliographies	-0.008604	-0.359681
11	Encoding	0.692819	1.398205
12	Optimization	0.598027	1.242353
13	Reliability	-0.094660	0.489641
14	Production	-0.526405	0.374447
15	Middleware	1.006791	1.610532
16	Deployment	0.330417	0.882006
17	Torque	0.450761	0.196344
18	Industries	-1.330785	0.075619
19	Focusing	-0.183801	-0.554003
20	development	0.286127	-0.729154
21	data	0.553734	-0.422316
22	Privacy	-0.215381	-0.318253
23	Grammar	-0.447752	-0.240555
24	survey	-0.331875	-1.014053
25	RAN	0.678872	-0.893242
26	agile	1.098579	-0.314309
27	packaging	-0.029336	-0.442251
28	Sociology	-1.124959	-0.277150
29	Manuals	-0.199909	0.424420
30	Costs	-0.570416	0.239059
31	SLR	0.463587	-0.779939
32	Sustainability	-0.753130	-0.234583
33	MOO	0.373990	-0.997734
34	Codes	-0.230720	0.045811
35	Architecture	-0.121328	0.475496
36	versioning	2.279043	0.379008
37	Software	0.461481	1.055434
38	Servers	0.635757	1.008004
39	Fasteners	-1.075972	0.897930
40	Benchmark	-0.185844	0.175858
41	Pandemics	-0.793054	0.271990
42	Ethics	-0.995961	-0.727552
43	Serverless	0.314441	-0.297774
44	Fluids	-0.617162	0.911447
45	Robustness	0.730372	0.691305
46	GLR	0.179205	-0.811631
47	SDLC	1.199425	0.326177
48	Testing	-0.596668	0.414534
49	Internet	0.136960	-0.617710
50	Interviews	-1.113810	-0.486390
51	automation	1.044239	0.334569

52	Epilepsy	-1.171550	0.108959
53	Cloud	0.752292	0.115021
54	Regulation	-0.683954	-0.422788
55	Buildings	-0.951752	0.033059
56	Automation	-0.048252	1.256039
57	Logistics	-0.631514	0.521135
58	models	0.409738	-0.759249
59	Transformers	-0.086723	-0.631187
60	Conferences	-0.877325	0.087827
61	security	0.447678	-0.808339
62	fairness	-0.004936	-1.375538
63	Collaboration	0.066770	0.738917
64	DevOps	0.814766	-0.076258
65	Organizations	-0.105391	0.051824
66	Neurons	-0.113121	0.953543
67	Tools	-0.035128	1.039813
68	5G	0.257450	-0.573994
69	management	0.333942	-0.673455
70	ethics	-0.224658	-1.356046
71	Framework	0.649447	0.611592
72	SM	0.245650	-0.377783
73	Containerization	-0.112323	-0.151526
74	Prototypes	-0.277224	0.383764
75	Convergence	-0.068657	0.200362
76	deploying	0.995289	-0.348912
77	Arguments	-0.830795	-0.332765
78	Agent	-0.113036	-0.149078
79	monitoring	0.250282	-0.604213
80	MCDA	-0.184993	-0.621633
81	Autonomic	0.750456	0.508161
82	LSTM	-0.233332	-0.345717
83	Packaging	-1.078914	0.575858
84	Azure	0.998039	-0.237865
85	Surgery	-1.429866	0.282506
86	IoT	0.355856	-0.366343
87	deployment	1.119446	-0.411664
88	CPS	-0.057774	-0.903955
89	Statistics	-1.143261	-0.237288
90	important	0.141361	-0.597097
91	bots	0.869220	-0.390460
92	diversity	0.146108	-1.204207
93	Publishing	-0.733686	0.252446
94	Measurement	-0.596457	0.869325
95	Training	-0.709120	0.251479
96	Writing	-0.868043	-0.014604
97	Germanium	-0.109735	0.273996
98	Edge	0.045002	-0.325179
99	Uncertainty	-0.557754	-0.351688

```
100          metadata  2.048673  0.489678
101        Regression -0.086363  0.876600
102           Planets -0.047211  0.309067
103         Libraries -0.617319  0.312124
104    sustainability  0.056516 -1.087798
105         Pipelines -0.805116  0.747711
106     Visualization  0.445488  1.467344
107         Databases  0.029151  1.163607
108             model  0.439924 -1.017089
109            Docker -0.111136 -1.241025
110        Regulators -0.596595 -0.891750
111            Matlab  0.996214  0.514456
112        Transforms -0.107312  1.693265
113    implementation  0.730832 -0.485064
114        Monitoring -0.301218  0.272285
115            Timing -0.569479  0.171661
116           dataset  0.849675 -0.066859
117         Semiotics -0.203431 -0.099935
118       Forecasting -0.791867  0.289488
119                AI  0.543393 -0.556528
```

12.4 Cluster

[59]:
```
NUM_CLUSTERS = 5

kmeans = cluster.KMeans(n_clusters=NUM_CLUSTERS,
                        random_state=RNG,
                        n_init=1000,
                        max_iter=1000)
kmeans.fit(X_pca)

labels = kmeans.labels_
centroids = kmeans.cluster_centers_

# print("Cluster id labels for inputted data")
# print(labels)
# print("Centroids data")
# print(centroids)

# print(
#     "Score (Opposite of the value of X on the K-means objective which is Sum␣
 ↪of distances of samples to their closest cluster center):"
# )
# print(kmeans.score(X))

# silhouette_score = metrics.silhouette_score(X, labels, metric='euclidean')
```

```
# print("Silhouette_score: ")
# print(silhouette_score)
```

[60]:
```
pca_df['cluster'] = labels
# to make clusters categorical for plotting
pca_df.cluster = pca_df.cluster.astype(str)

pca_df.sort_values(by=['cluster'])
```

[60]:

	words	x	y	cluster
28	Sociology	-1.124959	-0.277150	0
96	Writing	-0.868043	-0.014604	0
23	Grammar	-0.447752	-0.240555	0
55	Buildings	-0.951752	0.033059	0
32	Sustainability	-0.753130	-0.234583	0
89	Statistics	-1.143261	-0.237288	0
18	Industries	-1.330785	0.075619	0
54	Regulation	-0.683954	-0.422788	0
52	Epilepsy	-1.171550	0.108959	0
42	Ethics	-0.995961	-0.727552	0
85	Surgery	-1.429866	0.282506	0
110	Regulators	-0.596595	-0.891750	0
60	Conferences	-0.877325	0.087827	0
7	Stakeholders	-0.657384	-0.513601	0
6	Business	-0.781408	0.063316	0
5	Companies	-0.606245	-0.060207	0
77	Arguments	-0.830795	-0.332765	0
50	Interviews	-1.113810	-0.486390	0
99	Uncertainty	-0.557754	-0.351688	0
47	SDLC	1.199425	0.326177	1
76	deploying	0.995289	-0.348912	1
87	deployment	1.119446	-0.411664	1
36	versioning	2.279043	0.379008	1
91	bots	0.869220	-0.390460	1
81	Autonomic	0.750456	0.508161	1
51	automation	1.044239	0.334569	1
64	DevOps	0.814766	-0.076258	1
84	Azure	0.998039	-0.237865	1
26	agile	1.098579	-0.314309	1
100	metadata	2.048673	0.489678	1
116	dataset	0.849675	-0.066859	1
3	integration	1.014274	-0.161324	1
4	robustness	1.059977	-0.680040	1
113	implementation	0.730832	-0.485064	1
111	Matlab	0.996214	0.514456	1
17	Torque	0.450761	0.196344	1
53	Cloud	0.752292	0.115021	1

37	Software	0.461481	1.055434	2
101	Regression	-0.086363	0.876600	2
2	Orchestration	0.710216	0.814788	2
63	Collaboration	0.066770	0.738917	2
106	Visualization	0.445488	1.467344	2
107	Databases	0.029151	1.163607	2
112	Transforms	-0.107312	1.693265	2
66	Neurons	-0.113121	0.953543	2
45	Robustness	0.730372	0.691305	2
11	Encoding	0.692819	1.398205	2
12	Optimization	0.598027	1.242353	2
56	Automation	-0.048252	1.256039	2
15	Middleware	1.006791	1.610532	2
16	Deployment	0.330417	0.882006	2
38	Servers	0.635757	1.008004	2
71	Framework	0.649447	0.611592	2
67	Tools	-0.035128	1.039813	2
0	experimentation	0.479172	-0.696349	3
72	SM	0.245650	-0.377783	3
73	Containerization	-0.112323	-0.151526	3
78	Agent	-0.113036	-0.149078	3
90	important	0.141361	-0.597097	3
80	MCDA	-0.184993	-0.621633	3
82	LSTM	-0.233332	-0.345717	3
86	IoT	0.355856	-0.366343	3
88	CPS	-0.057774	-0.903955	3
92	diversity	0.146108	-1.204207	3
98	Edge	0.045002	-0.325179	3
104	sustainability	0.056516	-1.087798	3
108	model	0.439924	-1.017089	3
109	Docker	-0.111136	-1.241025	3
79	monitoring	0.250282	-0.604213	3
70	ethics	-0.224658	-1.356046	3
59	Transformers	-0.086723	-0.631187	3
68	5G	0.257450	-0.573994	3
9	ML	0.044387	-0.699915	3
10	Bibliographies	-0.008604	-0.359681	3
19	Focusing	-0.183801	-0.554003	3
20	development	0.286127	-0.729154	3
21	data	0.553734	-0.422316	3
22	Privacy	-0.215381	-0.318253	3
24	survey	-0.331875	-1.014053	3
25	RAN	0.678872	-0.893242	3
27	packaging	-0.029336	-0.442251	3
69	management	0.333942	-0.673455	3
31	SLR	0.463587	-0.779939	3
33	MOO	0.373990	-0.997734	3

```
43         Serverless  0.314441 -0.297774       3
119                AI  0.543393 -0.556528       3
62           fairness -0.004936 -1.375538       3
58             models  0.409738 -0.759249       3
61           security  0.447678 -0.808339       3
49           Internet  0.136960 -0.617710       3
46                GLR  0.179205 -0.811631       3
75        Convergence -0.068657  0.200362       4
74         Prototypes -0.277224  0.383764       4
14         Production -0.526405  0.374447       4
13        Reliability -0.094660  0.489641       4
41          Pandemics -0.793054  0.271990       4
65      Organizations -0.105391  0.051824       4
8             Sensors -0.632880  1.099867       4
114        Monitoring -0.301218  0.272285       4
115            Timing -0.569479  0.171661       4
1          Systematics -0.109538  0.103520       4
117         Semiotics -0.203431 -0.099935       4
105         Pipelines -0.805116  0.747711       4
118       Forecasting -0.791867  0.289488       4
102           Planets -0.047211  0.309067       4
44             Fluids -0.617162  0.911447       4
57           Logistics -0.631514  0.521135       4
29             Manuals -0.199909  0.424420       4
97          Germanium -0.109735  0.273996       4
30              Costs -0.570416  0.239059       4
48            Testing -0.596668  0.414534       4
94        Measurement -0.596457  0.869325       4
93         Publishing -0.733686  0.252446       4
83          Packaging -1.078914  0.575858       4
34              Codes -0.230720  0.045811       4
35       Architecture -0.121328  0.475496       4
39          Fasteners -1.075972  0.897930       4
40          Benchmark -0.185844  0.175858       4
103         Libraries -0.617319  0.312124       4
95           Training -0.709120  0.251479       4
```

[61]:
```
pca_df.cluster.value_counts()
```

[61]:
```
3    37
4    29
0    19
1    18
2    17
Name: cluster, dtype: int64
```

12.5 Plot

[62]:
```
fig = px.scatter(
    pca_df,
    x="x",
    y="y",
    color="cluster",
    #                   size='petal_length',
    hover_data=['words'],
    text=pca_df['words'])

# fig.update_layout(height=1600, width=1600, title_text='Vector Clusters')
fig.update_traces(textposition='bottom center', textfont_size=5)
fig.show()
```

12.6 Conclusion

To lay base for the further work on the domain research, I chose 5 clusters and wanted to see if those clusters align with my assumptions on my initially designed domain tables.

We can see the following, and already associate umbrella terms for the clusters:

```
Cluster 0 - BUSINESS:
['Companies' 'Industries' 'Business' 'Uncertainty' 'Conferences' 'Ethics'
 'Buildings' 'Sociology' 'Surgery' 'Interviews']
Cluster 1 - OPS:
['DevOps' 'robustness' 'Autonomic' 'Matlab' 'implementation' 'Azure'
 'bots' 'SDLC' 'agile' 'deploying']
Cluster 2 - DATA:
['Training' 'Logistics' 'Codes' 'Publishing' 'Timing' 'Pipelines'
 'Forecasting' 'Measurement' 'Fasteners' 'Organizations']
Cluster 3 - ML:
['MCDA' 'AI' 'Docker' 'management' 'monitoring' 'important' 'LSTM' 'SLR'
 'MOO' 'fairness']
Cluster 4 - DEV:
['Automation' 'Collaboration' 'Databases' 'Orchestration' 'Framework'
 'Optimization' 'Middleware' 'Servers' 'Deployment' 'Regression']
```

This indicates that the developed intuition of the first research iteration in regards to designing a reference architecture is promising.

[63]:
```
NR_SAMPLES = 10
print(f'''
Cluster 0 - BUSINESS:\n{pca_df[pca_df.cluster == '0'].sample(n=NR_SAMPLES,␣
 ↪random_state=RNG).words.values}
Cluster 1 - OPS:\n{pca_df[pca_df.cluster == '1'].sample(n=NR_SAMPLES,␣
 ↪random_state=RNG).words.values}
```

```
Cluster 2 - DATA:\n{pca_df[pca_df.cluster == '2'].sample(n=NR_SAMPLES,␣
 ↪random_state=RNG).words.values}
Cluster 3 - ML:\n{pca_df[pca_df.cluster == '3'].sample(n=NR_SAMPLES,␣
 ↪random_state=RNG).words.values}
Cluster 4 - DEV:\n{pca_df[pca_df.cluster == '4'].sample(n=NR_SAMPLES,␣
 ↪random_state=RNG).words.values}
''')
```

```
Cluster 0 - BUSINESS:
['Surgery' 'Conferences' 'Industries' 'Writing' 'Interviews' 'Sociology'
 'Arguments' 'Ethics' 'Regulation' 'Statistics']
Cluster 1 - OPS:
['DevOps' 'Cloud' 'Matlab' 'integration' 'deploying' 'robustness'
 'deployment' 'versioning' 'Autonomic' 'agile']
Cluster 2 - DATA:
['Robustness' 'Databases' 'Middleware' 'Collaboration' 'Neurons'
 'Orchestration' 'Deployment' 'Servers' 'Visualization' 'Optimization']
Cluster 3 - ML:
['RAN' 'Bibliographies' 'ethics' 'AI' 'GLR' 'ML' 'Docker' 'MOO'
 'packaging' 'Focusing']
Cluster 4 - DEV:
['Planets' 'Publishing' 'Costs' 'Semiotics' 'Fasteners' 'Libraries'
 'Production' 'Systematics' 'Pandemics' 'Reliability']
```

13 Cluster similar words from domain facet

[64]:
```
keywords_expl = df.domain.str.split(',').explode()
listed_words = keywords_expl

# print(keywords_expl)

words = set(listed_words) & set(list(model.key_to_index.keys()))
vectors = list([model.get_vector(word) for word in words])

print(len(words), len(vectors), 'words')

# pca = PCA(n_components=2, random_state=RNG)
pca_transformed = pca.fit_transform(vectors)
X_pca = pca_transformed

words = pd.DataFrame(words)
pca_df = pd.DataFrame(pca_transformed)
pca_df = pd.merge(words, pca_df, left_index=True, right_index=True)
pca_df.columns = ['words', 'x', 'y']
```

```
# pca_df

NUM_CLUSTERS = 5

kmeans = cluster.KMeans(n_clusters=NUM_CLUSTERS,
                        random_state=RNG,
                        n_init=1000,
                        max_iter=1000)
kmeans.fit(X_pca)

labels = kmeans.labels_
centroids = kmeans.cluster_centers_

pca_df['cluster'] = labels
# to make clusters categorical for plotting
pca_df.cluster = pca_df.cluster.astype(str)

pca_df.cluster.value_counts()
```

```
92 92 words
```

[64]:
```
3    37
2    21
0    17
4     9
1     8
Name: cluster, dtype: int64
```

[65]:
```
fig = px.scatter(
    pca_df,
    x="x",
    y="y",
    color="cluster",
    #                   size='petal_length',
    hover_data=['words'],
    text=pca_df['words'])

# fig.update_layout(height=1600, width=1600, title_text='Vector Clusters')
fig.update_traces(textposition='bottom center', textfont_size=5)
fig.show()
```

13.1 Conclusion

[66]:
```
NR_SAMPLES = 8
print(f'''
Cluster 0 :\n{pca_df[pca_df.cluster == '0'].sample(n=NR_SAMPLES,␣
 ↪random_state=RNG).words.values}
Cluster 1 :\n{pca_df[pca_df.cluster == '1'].sample(n=NR_SAMPLES,␣
 ↪random_state=RNG).words.values}
Cluster 2 :\n{pca_df[pca_df.cluster == '2'].sample(n=NR_SAMPLES,␣
 ↪random_state=RNG).words.values}
Cluster 3 :\n{pca_df[pca_df.cluster == '3'].sample(n=NR_SAMPLES,␣
 ↪random_state=RNG).words.values}
Cluster 4 :\n{pca_df[pca_df.cluster == '4'].sample(n=NR_SAMPLES,␣
 ↪random_state=RNG).words.values}
''')
```

```
Cluster 0 :
['data' 'scaling' 'architecture' 'registry' 'integration' 'github'
 'classification' 'threads']
Cluster 1 :
['bots' 'mysql' 'cnn' 'cd' 'opensource' 'git' 'gpu' 'matlab']
Cluster 2 :
['bert' 'azure' 'eda' 'shap' 'ci' 'aws' 'artefacts' 'drone']
Cluster 3 :
['forecasting' 'team' 'sustainability' 'pipelines' 'interpretability'
 'testing' 'development' 'training']
Cluster 4 :
['dataflow' 'algorithms' 'automation' 'workflow' 'coding' 'algorithm'
 'versioning' 'optimization']
```

14 Further work

There are many ways on how this base can be used for further work. Consider the following ideas:

- Cluster corpus of abstracts
- Plot interaction between domain and other facets
- Built domain model (will be done in master thesis)

[ ]:
```

```