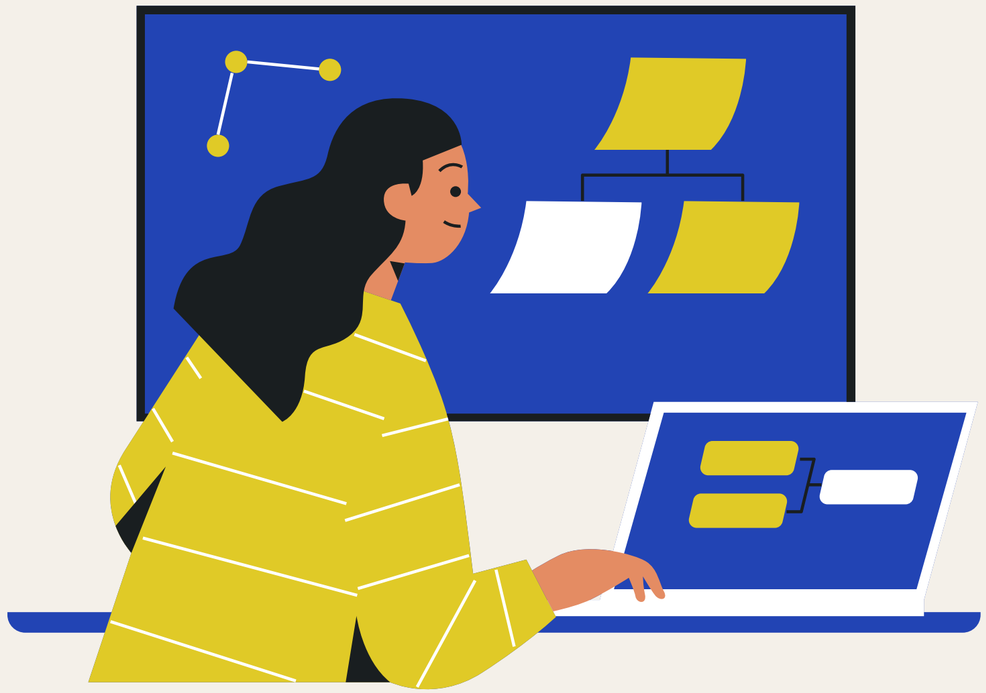
DS-DE CEDT Project

Teeihaiturtangjai group

Kawinwat	Phithukwonglerd	6633010521
Tinna	Jiarawapee	6633082721
Photchara	Kallayanasiri	6633158821
Penpitcha	Yoohoon	6633178321

Objective



Our group want to use data about researches from Scopus and Sprinker Nature link to predict open access of each research

Pipeline

Diagram

Web Scraping

-Selenium

Data Preparation

-Pandas

DE and AI / ML

-Spark

Data Visualization

-Power BI

01 - Web Scraping

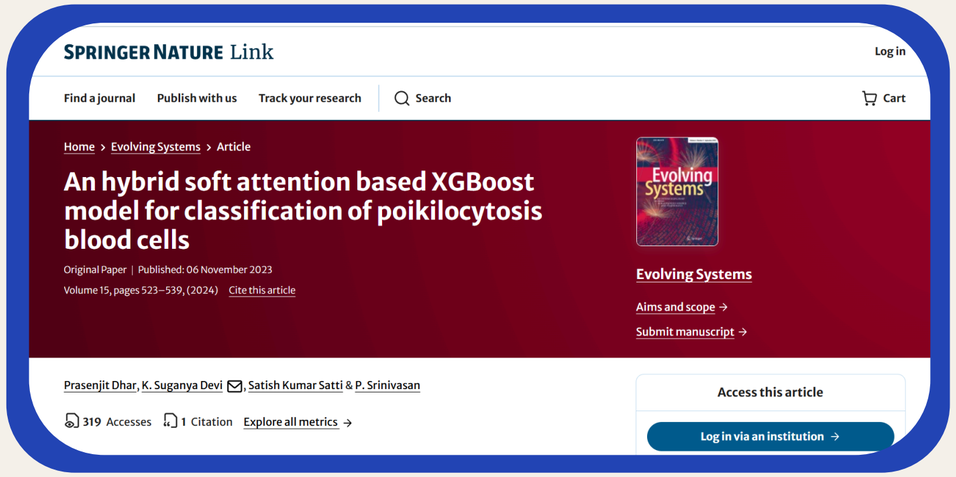
tool : selenium

scraping data :

title, publication_year, age_of_paper, aggregation_type, reference_count, publisher, has_funding_info, citation_count, open_access

source : https://link.springer.com

example of research :



01 - Web Scraping

Example of data after web scraping :

	title	publication_year	age_of_paper	aggregation_type
An hybrid soft attention based XGBoost model f...	2023	1	Original Paper
Parasitic egg recognition using convolution an...	2023	1	Article
Spatiotemporal changes of gross primary produc...	2024	0	Research Article
Cases of Castigation	2024	0	Chapter
Understanding Underdevelopment: A Study on Sel...	2024	0	Chapter

reference_count	open_access	has_funding_info	citation_count
38	2	NaN	1
62	1	Funding\nThis research project was funded by M...	5
83	2	NaN	4
18	2	NaN	0
34	2	NaN	0

02 - Data preparation

tool : pandas

selected column from given data :

filename, title, subject_name, subject_abbreviation, subject_code, keywords, publication_year, aggregation_type, reference_count, publisher, has_funding_info, citation_count, open_access

added column :

supergroup is from subject_name, subject_abbreviation, subject_code

keywords_list is from keywords

age_of_paper is from publication_year

citation_count_log is from citation_count

reference_count_log is from reference_count

02 - Data preparation

Example of data after data preparation :

	filename	title	subject_name	subject_abbreviation	subject_code	supergroup	keywords	publication_year
0	201800282	Recent developments in bifunctional air electr...	Renewable Energy, Sustainability and the Envir...	ENER, ENER, PHYS, ENER	2105, 2103, 3104, 2102	Physical Sciences	Bifunctional air electrode, Catalyst support, ...	2018
1	201801350	The benefit of punishment sensitivity on motor...	Social Psychology	PSYC	3207	Social Sciences	anxiety, defensive distance, performance, rein...	2018

age_of_paper	aggregation_type	reference_count	publisher	has_funding_info	citation_count	open_access	keywords_list
6	Journal	89	Elsevier Ltd	1	29	0	['Bifunctional air electrode', 'Catalyst suppo...
6	Journal	44	Blackwell Publishing Ltd	0	2	2	['anxiety', 'defensive distance', 'performance...

03 - DE and AI / ML

tool : Spark ML

training data : the given data and the scraping data

data transformation : use log transformation to reference_count and citation_count

algorithm : Random Forest Classifier

Validator : Params Grid, Cross Validator

features : age_of_paper, citation_count_log, reference_count_log

target : open_acess

04 - Data Visualization

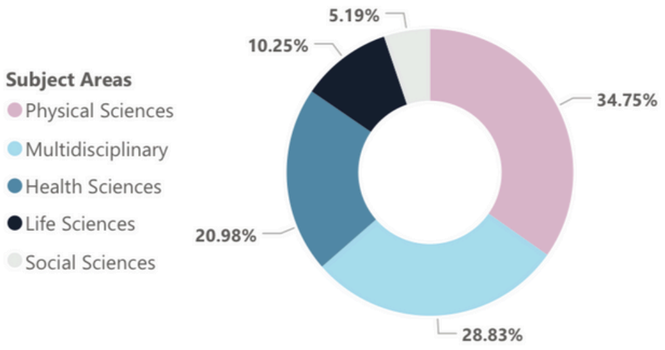
Top 10 Keywords by Year

year

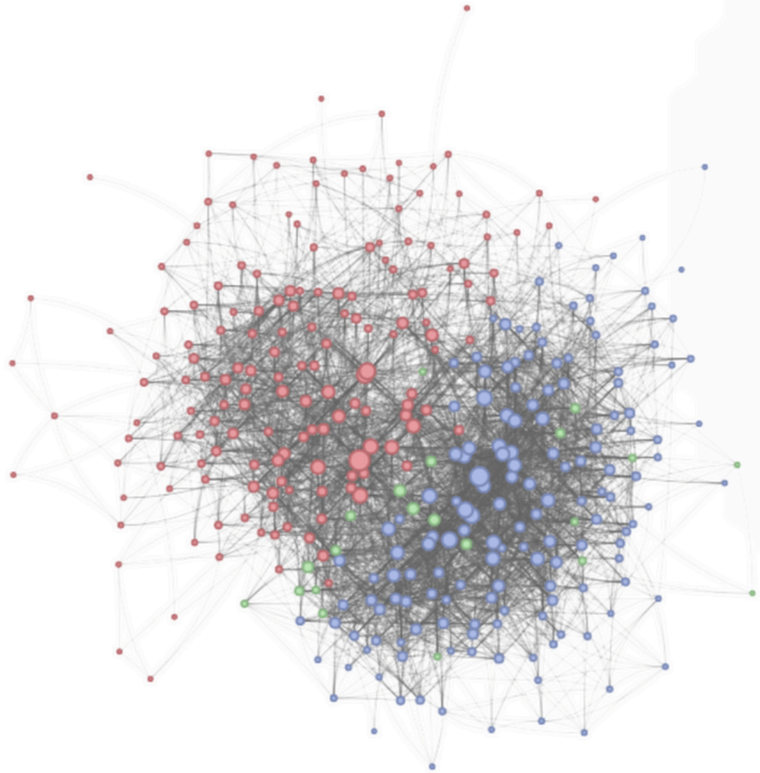
- 2018
- 2019
- 2020
- 2021
- 2022
- 2023

Keyword	Sum of Count
thailand	624
covid-19	280
inflammation	201
hadron-hadron scattering (experiments)	152
machine learning	137
depression	116
oxidative stress	116
sars-cov-2	114
hiv	111
deep learning	106
Total	**1957**

Subject Areas



Subject Areas Network Analysis



Nodes	Edges	Density
337	5926	0.105

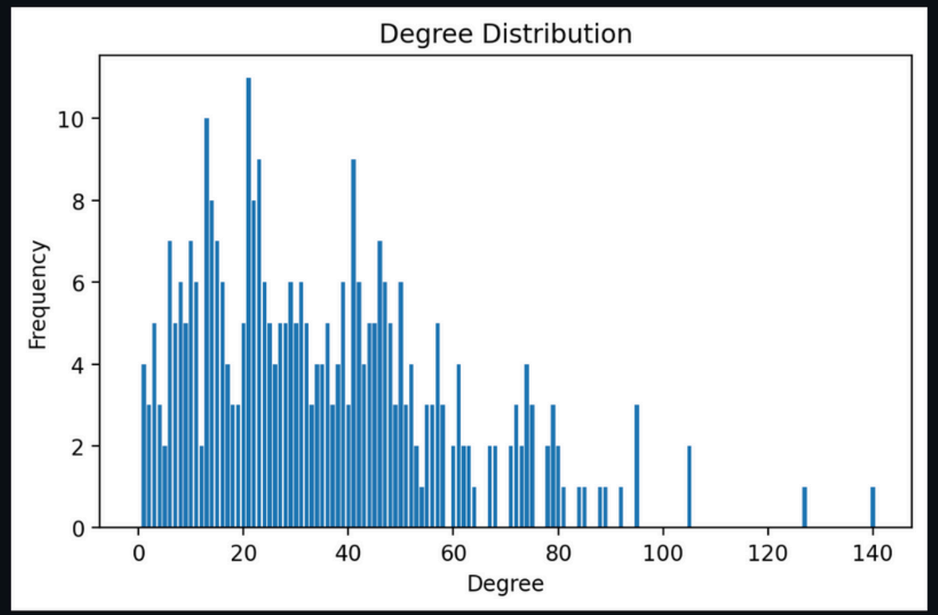
04 - Data Visualization

Network Analysis

Basic Statistics

Nodes	Edges	Density	Diameter
337	5926	0.105	5

Degree Distribution

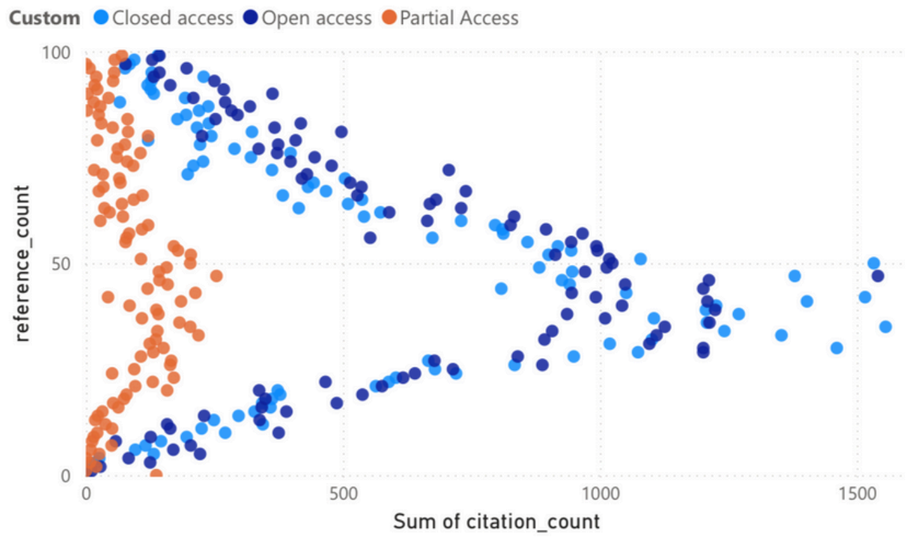


Centrality Analysis

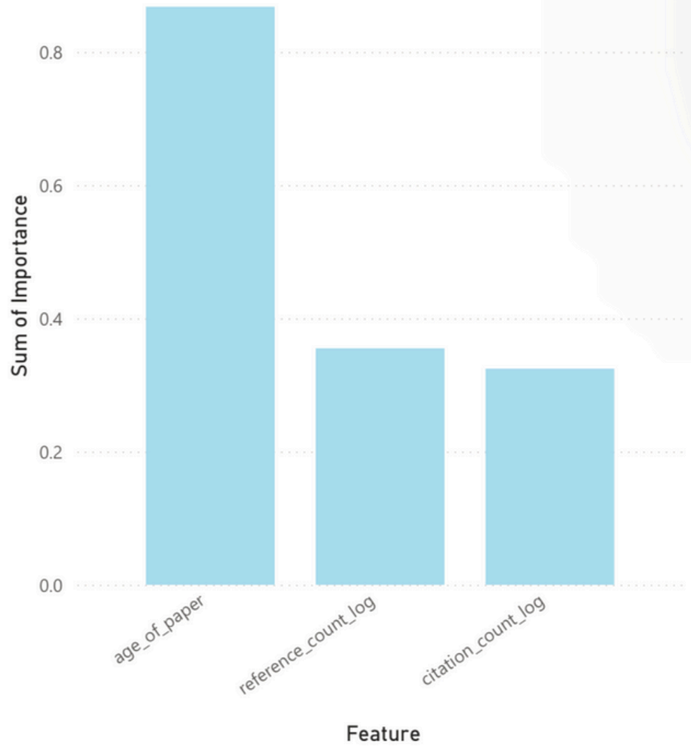
	Node	↓ Degree	Betweenness	Closeness	PageRank
5	Biochemistry	0.4167	0.058	0.6176	0.0104
57	Computer Science Applications	0.378	0.0372	0.5947	0.0091
40	Environmental and Occupational Health	0.3125	0.0363	0.5773	0.0086
39	Public Health	0.3125	0.0363	0.5773	0.0086
60	Electrical and Electronic Engineering	0.2827	0.0128	0.549	0.0067
1	Sustainability and the Environment	0.2827	0.0085	0.5446	0.0066
0	Renewable Energy	0.2827	0.0085	0.5446	0.0066
86	Education	0.2738	0.0465	0.56	0.0079
88	Biomedical Engineering	0.2649	0.0325	0.5554	0.0069
119	Biotechnology	0.2619	0.014	0.5517	0.0064

04 - Data Visualization

Sum of Citation Count by Open Access and Reference Count



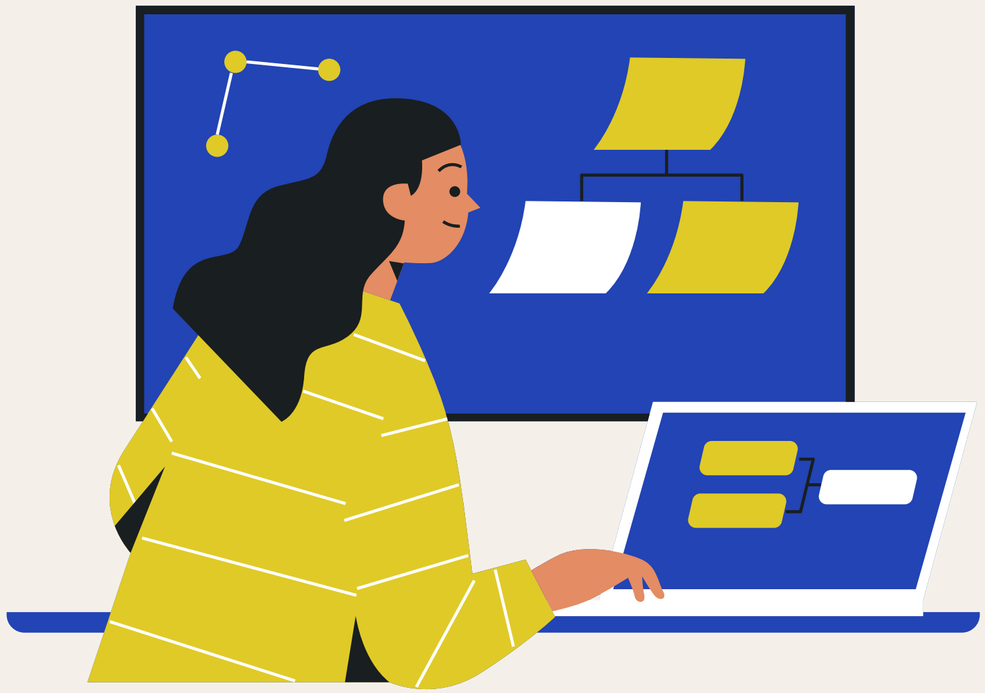
Importance of features used in predicting open access



Demonstration

Presentation

https://youtu.be/kVF2nZcYXXQ?si=4sTsJNfRgpJL4tVV



Thank you

Data visualization simplifies the communication of analysis findings