OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Artificial Intelligence in LUDZIE NAUKI

Przemyslaw Korytkowski

Marek Michajłowicz

euroCRIS Strategic Membership Meeting 2024

Paris, 28.11.2024

Project goals

Promotion and opening up of data on Polish science

- Researchers and their achievements accessible to all
- Contact details for experts

Reliable source of data

- Scientific experts
- Trustworthy data
- Cooperation between science and business

E-services for researchers

- Grants application
- Degree obtaining procedures

Origins of Ludzie Nauki

NAUKA POLSKA

The database of:
- degrees awarded
- disciplines/disciplines
- employment history at Polish institutions
- doctoral thesis supervisors
- doctoral thesis reviews
- history of academic functions

PBN
Polska Bibliografia Naukowa

Collect and present metadata of scientific publications

polon 2.0

It collects data as in Nauka Polska and additionally:
- employment history of the scientist in Polish institutions
- patents (national, EPO, PCT)
- projects
- promotions of doctoral theses
- reviews of doctoral theses
- Information on research infrastructure
- ORCID of the researcher

LUDZIE NAUKI
PEOPLE OF SCIENCE

Portal with verified data on scientists and their achievements for:
- researchers
- governmental and local administration
- society
- journalists
- business

which enables the implementation of the State Informatisation Strategy

1999 — **2013** — **2014** — **2023**

www

reporting systems

FAIR data system, data can be found, understood and reused

Data wokrflow

BIBLIOGRAPHIC DATABASES

Crossref · nukat · BIBLIOTEKA NARODOWA · OpenAlex

NATIONAL SCINTIFIC PUBLICATIONS DATABASE

PBN
Polska Bibliografia Naukowa

INSTITUTIONAL CRISes

polon 2.0

LUDZIE NAUKI
PEOPLE OF SCIENCE

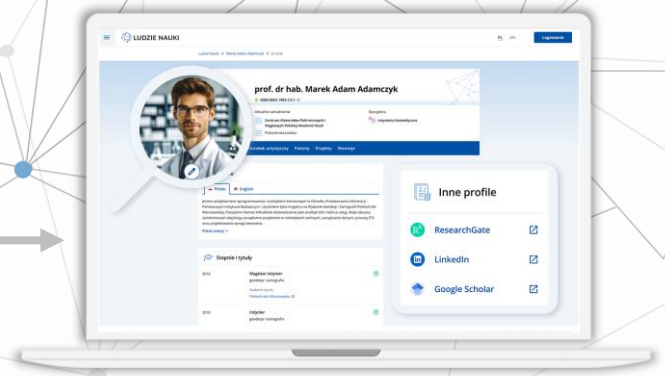
EDUCATION EMPLOYMENT PUBLICATIONS PATENTS AND PROJECTS

radon
raporty | analizy | dane

ORCID

API

e-services



ludzie.nauka.gov.pl

FOUNDING AGENCES

NARODOWE CENTRUM NAUKI

NAWA NARODOWA AGENCJA WYMIANY AKADEMICKIEJ

OSF OBSŁUGA STRUMIENI FINANSOWANIA

Researchers’ profiles

200 000+
profiles

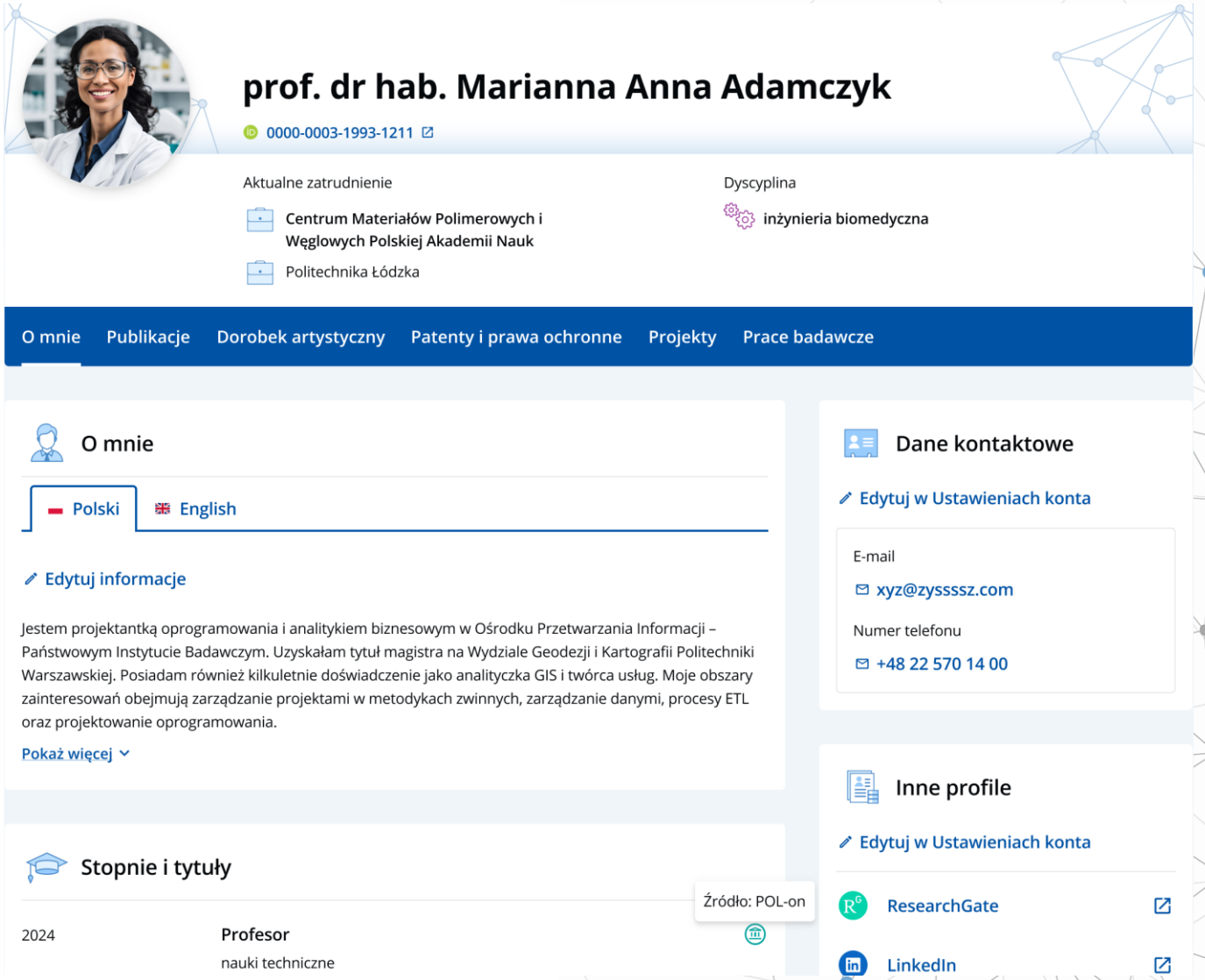
1 700 000+
publications

130 000
artistic achievements

12 000+
patents

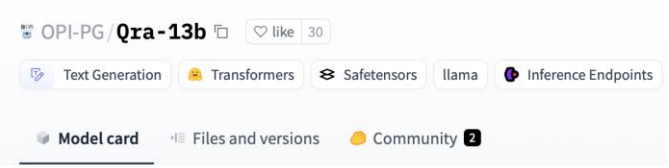
36 000+
projects

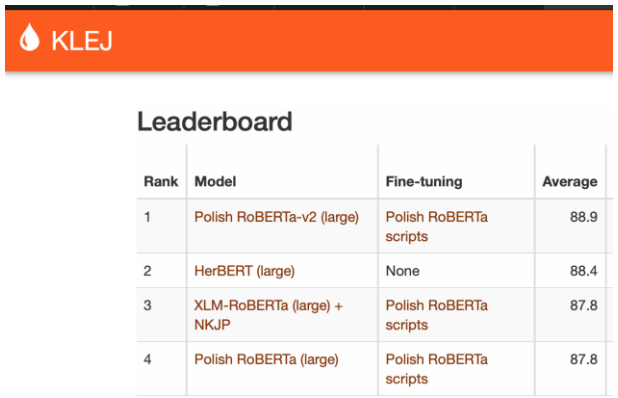
19 000+
PhD and DSc

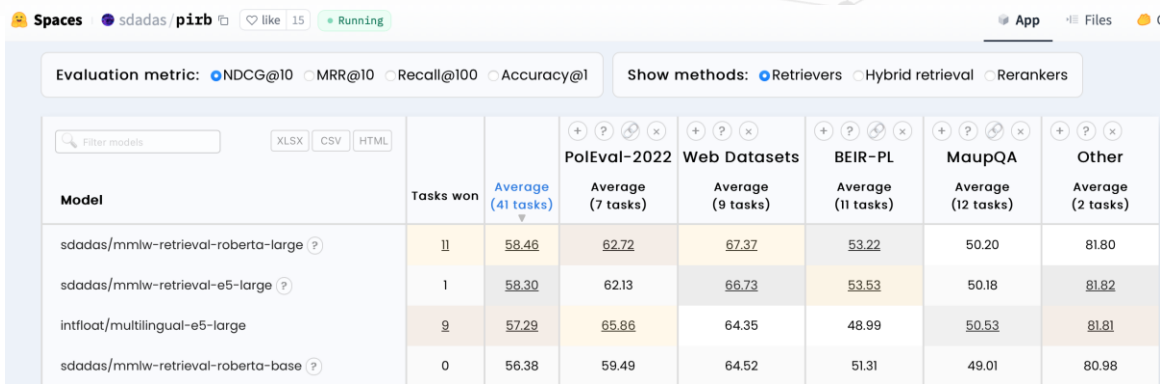


AI Lab at OPI

- We trained and fine-tuned many Polish Transformer models:
  - Collect and pre-process a large corpus of Polish documents (previously 300+ GB raw text, now we have 2+ TB)
  - Train different transformer-based language models from scratch (e.g. polish RoBERTa, BART, GPT2, Qra)
  - Train different neural text encoders (e.g. st-paraphrase, mmlw)
  - Fine-tune the models on a dozen of Polish linguistic tasks (e.g. classification, regression, QA, summarization tasks)
  - Different evaluations e.g. using Allegro KLEJ benchmark, or ours – PIRB
  - Release the pre-trained models to the public



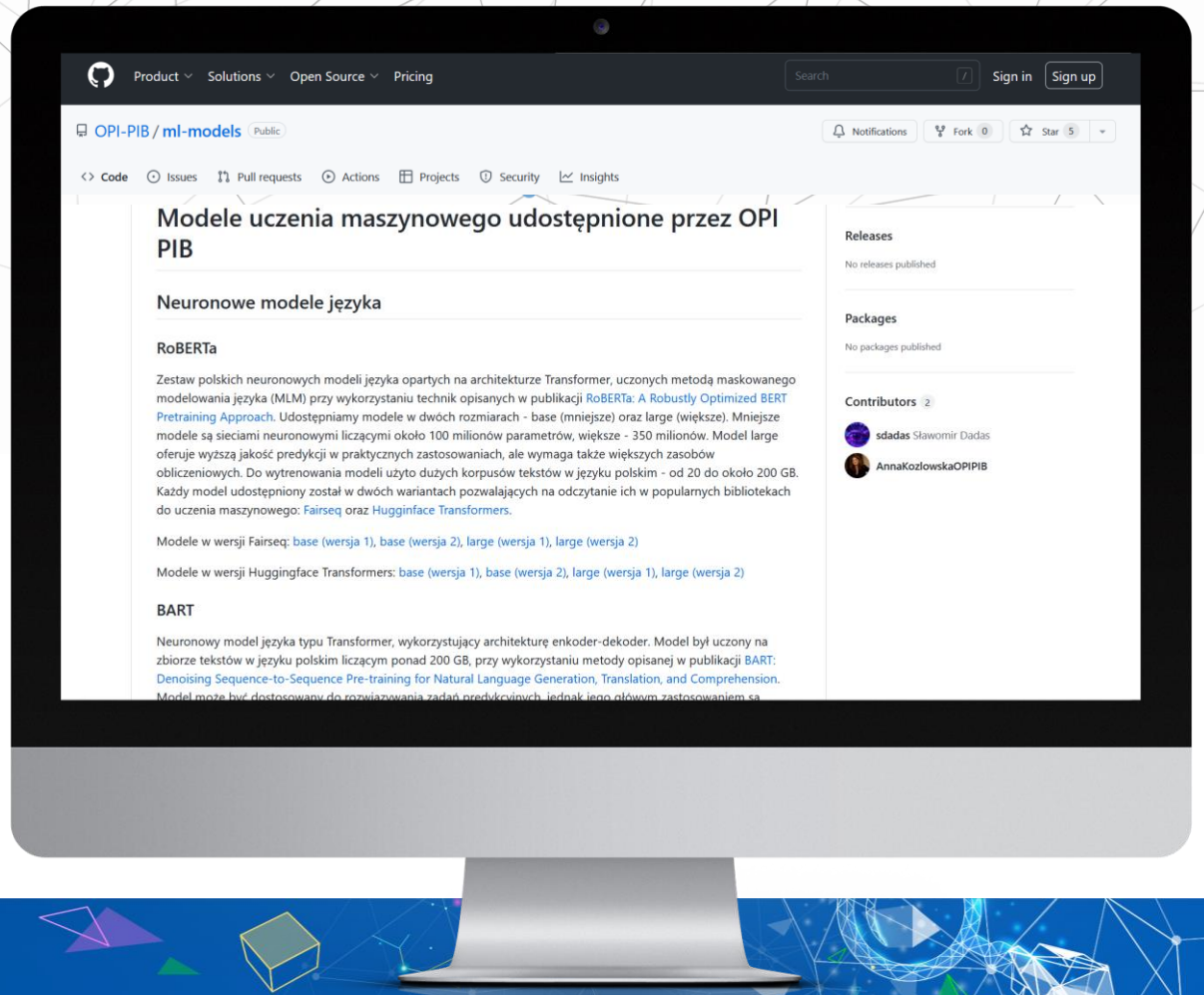




HOW TO DOWNLOAD OUR **Neural Language Models**

https://github.com/OPI-PIB/ml-models:

- Pre-trained models, such as RoBERTa, BART, GPT-2, ELMo, Qra
- Models in two versions:
  - Pytorch Fairseq
  - HuggingFace Transformers
- Evaluation codes are also available
- Detailed results of the experiments can be found in the corresponding papers



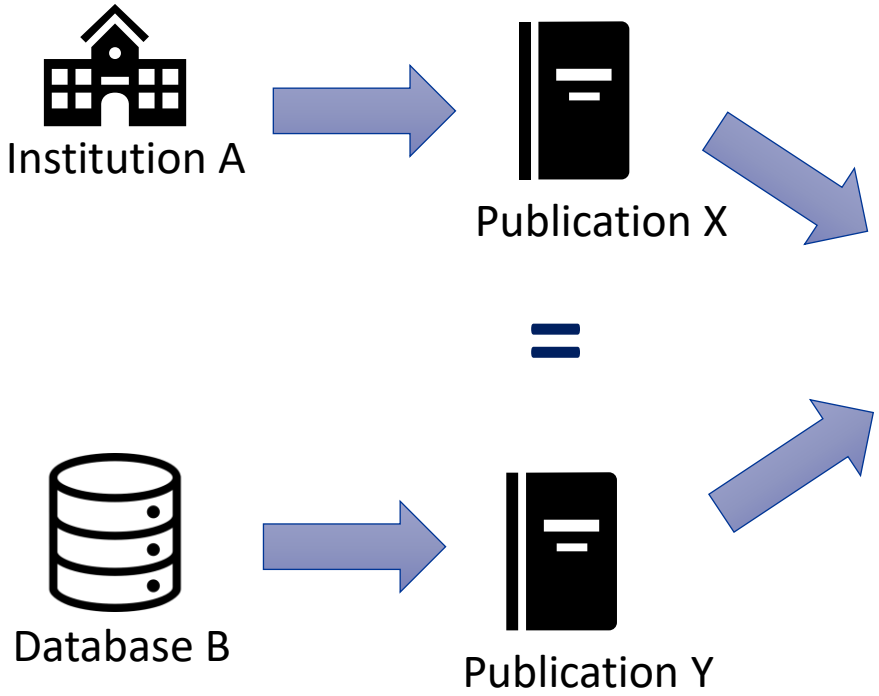
AI in Ludzie Nauki

1. Publication deduplication (working)
2. Semantic search engine (under development)
3. Scientific expert search (in the future)

Publication deduplication

The **Golden Publication Record** is the single, richest and most reliable representation of metadata about a scientific publication, at least one of whose authors is an employee of a Polish scientific institution.

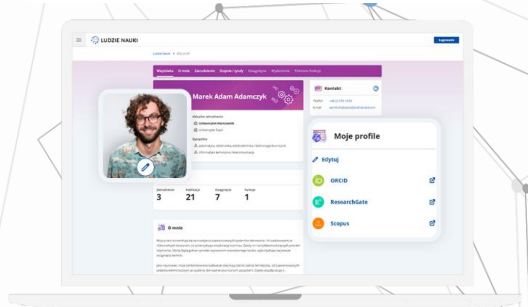
Publication deduplication



Golder Record

1. Publication X from Institution A
2. Publication Y from Database B

LUDZIE NAUKI
PEOPLE OF SCIENCE



Publication X from Institution A
=
Publication Y from Database B

Issues:

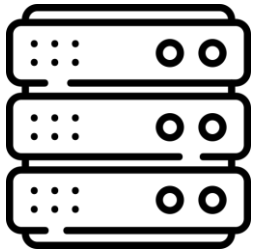
- Errors in metadata
- Missing metadata
- No identifiers for authors outside the institution

Publication deduplication

Standardisation of titles:

- Replacement of Roman numerals with Arabic numerals,
- Removal of non-alphanumeric characters (punctuation, inverted commas),
- Reduction to lower case and replacement of accents with basic characters
- Removal of HTML and XML tags
- Title should be longer than 16 characters

Publication deduplication – infrastructure



22 servers (each):

- 16 cores
- 90 GB RAM
- 2 TB disks

352 cores, 1.98TB RAM, 44TB disks

hadoop HDFS

DELTA LAKE

Apache Airflow

PySpark

Spark MLlib

Publication deduplication

PBN
Polska Bibliografia Naukowa

2 103 905

Crossref

343 773

Deduplication

Enrichment

Golder Record

1 774 311

Next sources

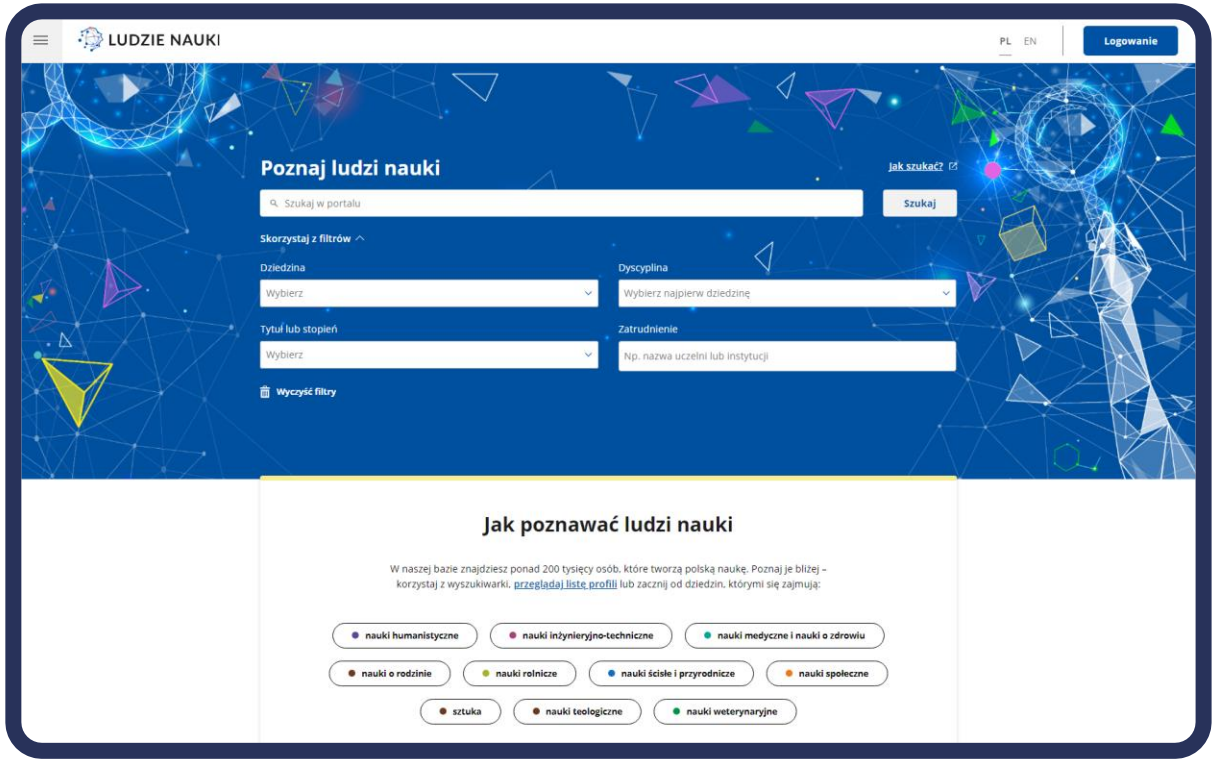
nukat

OpenAIRE

BIBLIOTEKA NARODOWA

OpenAlex

Semantic search engine (under development)



The goal is to achieve more accurate search results by better understanding user intensions and semantical relations between words in the query.

We are planning to employ Large Language Model based on Transformer architecture with attention mechanism.

Semantic search engine (under development)

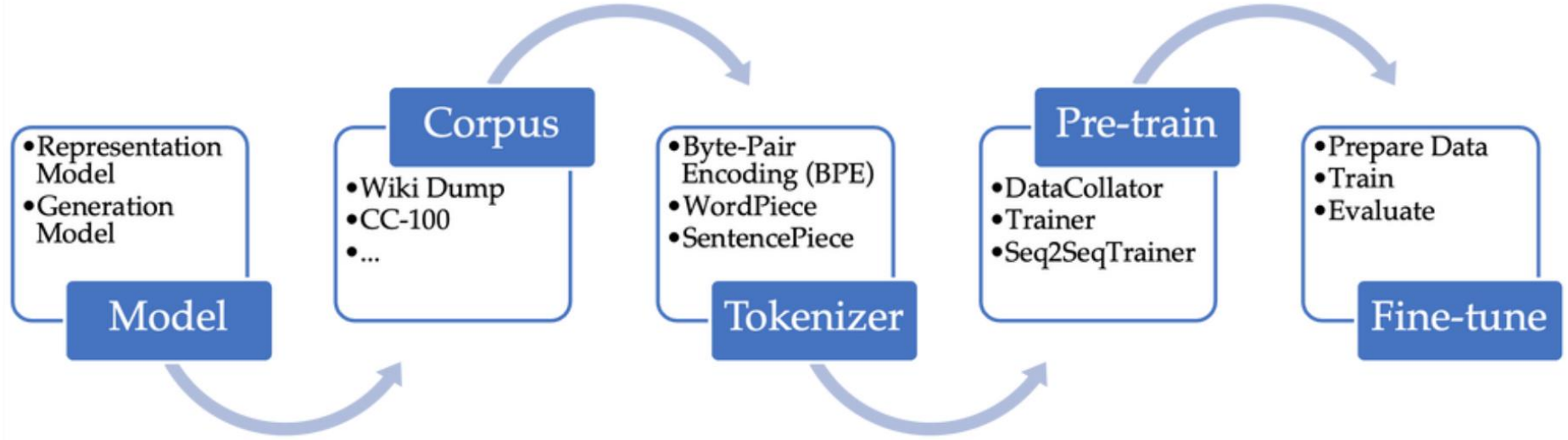


Figure 1. Overview of pre-training and fine-tuning with Hugging Face.

https://techcommunity.microsoft.com/t5/ai-customer-engineering-team/pre-train-and-fine-tune-language-model-with-hugging-face-and/ba-p/3590420

Scientific expert search (in the future)

Efficient selection of experts

- shorter selection times for experts
- increased availability of experts
- increasing the number of expert candidates

Increasing the transparency of the expert selection process

- Reducing subjectivity in the expert selection process

Efficient recruitment of domain experts

- Reducing the time and cost of recruiting and informing experts about the process

Thank You for Your attention!
Dziękuję za uwagę!