

Artificial Intelligence in Ludzie Nauki

Przemysław Korytkowski, Marek Michajłowicz

National Information Processing Institute – National Research Institute, Warsaw, Poland

Ludzie Nauki (People of Science) portal (<https://ludzie.nauka.gov.pl>) is a knowledge base on Polish science. The portal makes available data dispersed in the state science and higher education system.

One of the objectives of the *Ludzie Nauki* portal is to collect the most complete and reliable information on publications of scientists employed in Polish scientific institutions. Unfortunately, there is no single data source on publications which would contain metadata of all scientific publications. For this reason, collecting and processing data from many different sources is necessary. Because information on scientific publications circulates around the world and has been entered into various databases, information on the same publication often appears many times.

Since high metadata reliability about scientific publications is one of the goals of the *Ludzie Nauki* portal, a decision was made to build the so-called Publication's Golden Record, a single, the 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.

The basis for building the Golden Publication Record is the Polish Scientific Bibliography (PBN). It is a database of metadata about publications created for the evaluation process of scientific activities of Polish scientific institutions, and each institution independently provides a set of metadata. Unfortunately, the data provided vary in quality (from very good, with no errors, to low quality, with many errors), and publications that have been prepared as part of inter-institutional collaboration have been reported several times.

As part of the work on the Golden Record, a process of automatic data cleaning and deduplication, i.e. detection of metadata relating to the same publications, takes place.

We use natural language processing methods to identify metadata that describe the same scientific publication. Many publications do not have a DOI, so using the title as a field to find duplicates is necessary. Unfortunately, titles are entered in different ways. Some metadata contains only abbreviated versions, or special symbols, e.g. chemical or mathematical, are coded differently. For this reason, it is not possible to compare titles as simple character strings. In *Ludzie Nauki*, we use the Hierarchical navigable small world (HNSW) algorithm and cosine distance to identify possible duplicates.