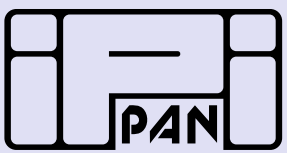
MTAS : nowa wyszukiwarka korpusowa

Łukasz Kobyliński Michał Wasiluk
Zbigniew Gawłowicz



INSTITUTE OF COMPUTER SCIENCE
POLISH ACADEMY OF SCIENCES
ul. Jana Kazimierza 5, 01-248 Warszawa

18 czerwca 2018

Czym jest MTAS?

- wyszukiwarka stworzona przez Meertens Institute (Royal Netherlands Academy of Arts and Sciences)
- oparta na oprogramowaniu tworzonym przez fundację Apache:
  - Lucene – biblioteka do przeszukiwania dokumentów,
  - Solr – aplikacja serwerowa oparta na Lucene.
- aktywnie rozwijana, wykorzystywana w dużych projektach (m.in. Nederlab – duża baza dokumentów dotyczących Holandii, ok. 10 mld. słów)

Apache Lucene i Solr

Apache Lucene – biblioteka:

- implementuje indeksowanie i przeszukiwanie dowolnych dokumentów tekstowych,
- dojrzała biblioteka (rozwijana od 1997 roku), wykorzystywana produkcyjnie.

Apache Solr – aplikacja serwerowa:

- wyszukiwanie i filtrowanie,
- stronicowanie wyników wyszukiwania,
- API XML/HTTP i JSON,
- panel administracyjny w interfejsie webowym.

Cechy wyszukiwarki

- możliwość indeksowania znakowanego tekstu oraz definiowania własnych warstw, w tym jednostek wielowyrazowych
- język zapytań CQL (Corpus Query Language)
- przyrostowy indeks – brak konieczności przetwarzania całego korpusu po dodaniu jednego tekstu
- skalowalność, także pomiędzy wieloma serwerami
- możliwość filtrowania wyników wg metadanych
- wbudowane proste zapytania statystyczne
- wymaga preprocessingu plików do specyficznego formatu XML (jeden plik zawierający wszystkie dane)

Wykorzystanie w polskich korpusach

- Korpus Barokowy — dwie warstwy tekstowe: transliteracyjna i transkrypcyjna, w Korbie automatycznej dwie warstwy fleksyjne (z dwóch różnych tagerów), `korba.edu.pl`,
- Korpus tekstów polskich z lat 1830-1918 — dwie warstwy tekstowe oraz warstwa ręcznego znakowania fleksyjnego, `korpus19.nlp.ipipan.waw.pl`,
- NKJP1M – warstwa tekstowa oraz trzy warstwy ręcznego znakowania: warstwa fleksyjna, warstwa nazw własnych, warstwa sensów słów, `nkjp.nlp.ipipan.waw.pl`.

Więcej szczegółów

- Matthijs Brouwer, Hennie Brugman, Marc Kemps-Snijders, *MTAS: A Solr/Lucene based Multi Tier Annotation Search solution* , Selected papers from the CLARIN Annual Conference 2016.
- Seminarium ZIL 14 maja 2018 r., nagranie dostępne na kanale IPI PAN w serwisie YouTube.