

MTAS: nowa wyszukiwarka korpusowa

Doświadczenia wdrożeniowe w kontekście polskich
korpusów językowych

Łukasz Kobyliński, Michał Wasiłuk
i Zbigniew Gawłowicz (IPI PAN)

Agenda

- Krótko o wyszukiwarkach korpusowych
- Wyszukiwanie pełnotekstowe - Apache Lucene i Solr
- MTAS - charakterystyka projektu
 - architektura
 - dostępne funkcje
 - rozwiązania problemów i ograniczeń w kontekście języka polskiego
- Polskie wdrożenia wyszukiwarki MTAS
 - język zapytań i możliwości wyszukiwania
 - wykorzystanie w kontekście analiz statystycznych
 - wyniki testów wydajnościowych
- Plany na przyszłość

Wyszukiwarki korpusowe

Jakie cechy powinna spełniać wyszukiwarka korpusowa?

- efektywnie przeszukiwać duże ilości tekstu,
- umożliwić przeszukiwanie i analizę dowolnej liczby warstw anotacji,
- udostępniać rozbudowany język zapytań (standard: CQL),
- wspierać metadane tekstów,
- pozwalać na analizę kontekstów wyszukiwanych fraz lub segmentów,
- oraz na zadawanie zapytań statystycznych.
- [...]

Oprogramowanie powinno być aktywnie utrzymywane i posiadać stabilną wersję.

Wyszukiwarki korpusowe

- [Corpus Workbench](#) (Christ, 1994), (Evert & Hardie, 2011),
- [Sketch Engine](#) / [Manatee](#) (Kilgarrieff, et al., 2004), (Rychlý, 2007),
- [Poliqarp](#) (Przepiórkowski, et al., 2004), (Janus & Przepiórkowski 2007),
- [Corpuscle](#) (Meurer, 2012),
- [KorAP](#) (Banski, et al., 2013),
- [GrETEL](#) (Vandeghinste, et al., 2014),
- [BlackLab](#) (Reynaert, et al., 2014),
- [PaQu](#) (Odijk, 2015),
- [Poliqarp 2](#) (Zaborowski & Zabłocki, 2016),
- [MTAS](#) (Brouwer, et al., 2017).

MTAS - porównanie

	Corpus Workbench	Sketch Engine	PaQu	GrETEL	BlackLab	Corpuscle	Neo4J	Solr	Solr + Mtas
Open source	✓	✗	✓	?	✓	?	✓	✓	✓
Highly scalable	✓	✓	✗	✗	✓	?	✓	✓	✓
Distributed search	?	?	✗	✗	✗	?	✓	✓	✓
Arbitrary (CMDI) metadata schemas	✗	✗	✗	✗	✗	?	✓	✓	✓
Annotated text	✓	✓	✓	✓	✓	✓	✗	✗	✓
- full support annotations	✗	✗	✓	✓	✗	✗	✗	✗	✓
- hierarchical structure	✗	✗	✓	✓	✗	✗	✗	✗	✓
- configurable mapping of input format on index	✗	✗	✗	✗	✗	✗	✗	✗	✓
- support FoLiA annotation format	✗	✗	✓	✗	✓	✗	✗	✗	✓
- corpus query language (CQL)	✓	✓	✗	✗	✓	✓	✗	✗	✓
- full statistics	✗	?	✗	✗	✗	✗	✗	✗	✓
- term vectors over any result document set	✓	✓	✗	✗	✗	✗	✗	✗	✓
- grouping	✓	✓	✗	✗	✓	✓	✗	✗	✓
- faceting	✗	✗	✗	✗	✗	✗	✗	✗	✓
- keyword in context (kwic)	✓	✓	✓	✓	✓	✓	✗	✗	✓

Apache Lucene

Skąd zainteresowanie projektem w kontekście wyszukiwarek korpusowych?

- implementuje indeksowanie i przeszukiwanie dowolnych dokumentów tekstowych,
- dojrzała biblioteka (rozwijana o 1997 roku, od 2004 projekt Apache),
- wykorzystywana produkcyjnie (bezpośrednio oraz poprzez inne projekty, np. Solr/Elasticsearch),
- rozbudowany ekosystem dodatków i rozszerzeń, wsparcie społeczności.

Unikalność podejścia: przyrostowe tworzenie indeksu, co zapewnia efektywność indeksowania i przeszukiwania dla zmieniającego się zbioru dokumentów.

Apache Lucene + Apache Solr = skalowalność

Apache Solr jest aplikacją serwerową, która rozszerza Lucene m.in. o możliwość tworzenia rozproszonych indeksów:

- kluczowe w kontekście dużych zbiorów danych,
- kluczowe w kontekście dużych indeksów (zawierających wiele indeksowanych pól).

Poza tym Solr dodaje:

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

MTAS

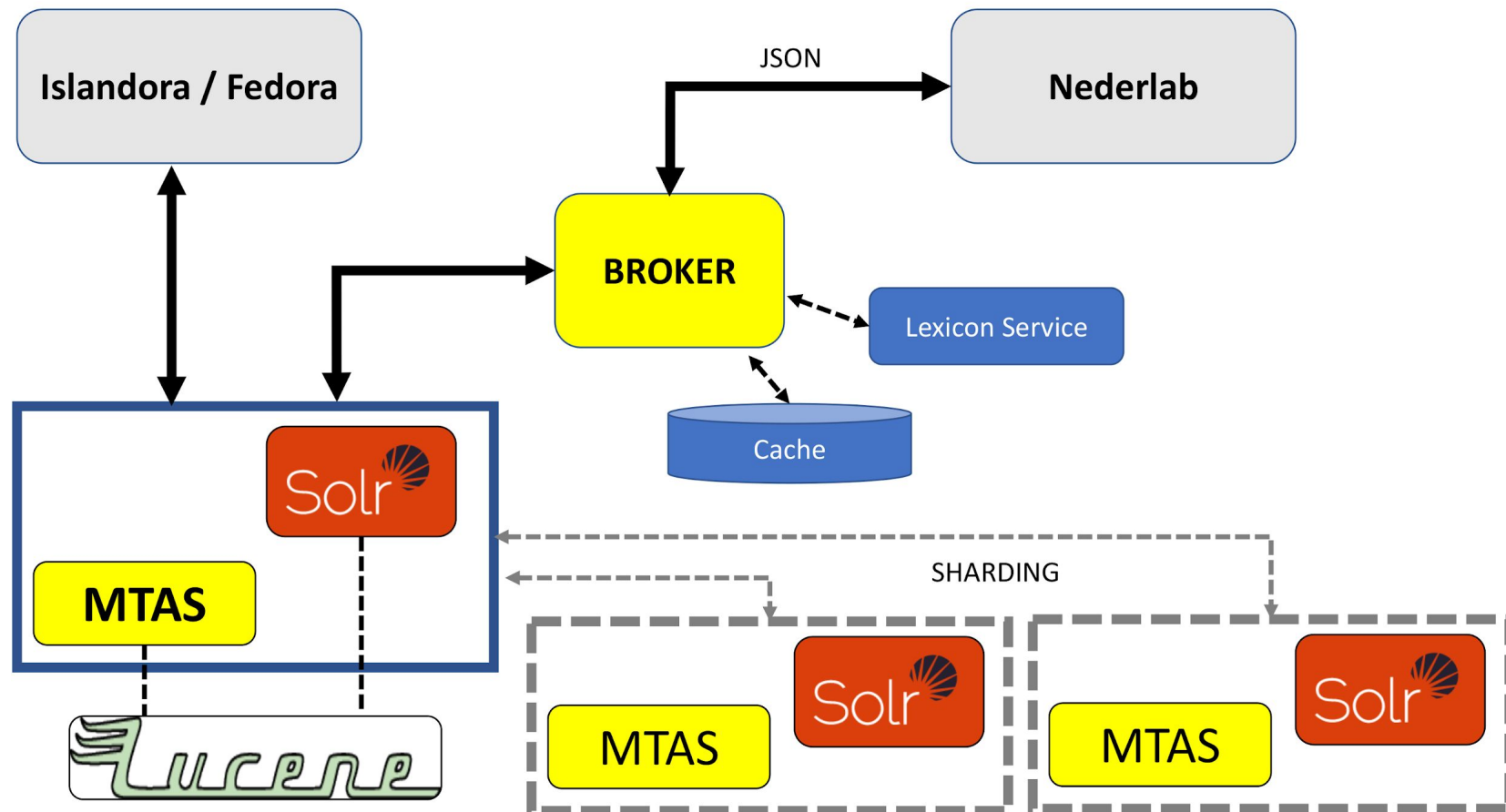
Czym jest MTAS?

- rozszerzony indeks Lucene - możliwość indeksowania anotowanego tekstu
- język zapytań CQL (Corpus Query Language) - częściowa implementacja
- rozbudowane parsowanie i indeksowanie dokumentów
- skalowalny, wspiera rozproszone wyszukiwanie Solr
- ekstensywna i elastyczna konfiguracja indeksowania dokumentów

Może być wykorzystany na dwa sposoby:

- biblioteka JAVA w połączeniu z Apache Lucene,
- plugin do Apache Solr.

MTAS - architektura (projekt Nederlab)





Index Structure

title: Max Havelaar

year: 1860

text: Ik ben makelaar in koffie,
en woon op de
Lauriergracht ...

Document	Field	Term	Position	Payload
1	title	Max Havelaar	-	-
1	year	1860	-	-
1	text	Ik	0	-
1	text	ben	1	-
1	text	makelaar	2	-
1	text	in	3	-
1	text	koffie	4	-
1	text	,	5	-
1	text	en	6	-
1	text	woon	7	-
...	...	...	...	...

Indeks jest **posortowany** według **poła** i **wartości (term)**, umożliwiając szybkie pobieranie dokumentów (i pozycji) dla danego terminu.

MTAS - rozszerzenie indeksu Lucene

- Rozszerza pole **term** w Lucene tworząc dwa osobne pola: **prefiks** i **postfiks**.
- Używa **payload** do przechowywania:
 - id poszczególnych tokenów,
 - hierarchicznych relacji,
 - dodatkowych pozycji (wsparcie elementów wielopozycyjnych).

Document	Field	Term	Position	Payload
1	text	0	s	-
1	text	1	t	lk
1	text	2	pos	VNW
1	text	3	t	ben
...	...	...	...	...

Document	Field	TokenId	Prefix	Postfix	Position	ParentId	Payload
1	text	0	s	-	0 - 10	-	-
1	text	1	t	lk	0	0	-
1	text	2	pos	VNW	0	0	-
1	text	3	t	ben	1	0	-
...	...	...	...	...	...	...	...

MTAS - ładowanie danych

Wbudowane wsparcie formatów:

- FoLiA
- ISO-TEI
- CHAT
- Sketch Engine
- CRM (Corpus Van Reenen-Mulder/Adelheid)

Możliwość łatwego dodawania obsługi innych formatów bazujących na XML,
np. TEI P5, XCES

Dostosowanie MTAS do polskiego TEI P5 (NKJP)

- nowy parser rozszerzający MtasXMLParser
- wstępne przetwarzanie składowych plików TEI do dwóch plików:
 - JSON - metadane dokumentu,
 - XML - rozszerzony *ann_morphosyntax.xml* - MTAS wymaga zebrania wszystkich indeksowanych informacji do jednego pliku
- konfiguracja indeksowania utworzonego pliku XML

MTAS - konfiguracja indeksowania

Plik konfiguracyjny pozwala na definicję:

- *segmenty (słowa)* - podstawowa warstwa tokenizacji, np. <seg/>
- *anotacje* pojawiające się w danym *segmentcie*
- *grupy* - zawierające jedno lub wiele *segmentów*, np. <p/> i <s/>
- *anotacje grupy*
- *referencje* - elementy odnoszące się do *segmentów* wg id
- *relacje* - zawierające jedno lub wiele *referencji*, np. <ne/>
- *anotacje relacji*

Konfiguracja indeksowania TEI

- specyfikacja poszczególnych warstw anotacji
 - segmenty: <seg/>
 - anotacje segmentów: *orth*, *base*, *nps*, *interpretation* itp.
 - grupy: paragrafy <p/>, zdania <s/>
 - relacje: <ne/> (jednostki nazewnicze)
- rozbicie znaczników morfosyntaktycznych (*interpretation*)
 - dwie reprezentacje:
 - base + tag
 - base + pos + feat
 - nazwy kategorii gramatycznych (feat) nie są przechowywane w indeksie

Wdrożenia MTAS dla korpusów w języku polskim

- KORBA – korba.edu.pl
 - korpus ręcznie anotowany (511 471 segmentów)
 - korpus anotowany automatycznie dwoma tagerami (13 453 367 segmentów)
- Korpus XIX w. – korpus19.nlp.ipipan.waw.pl
 - 635 934 segmentów
- Narodowy Korpus Języka Polskiego – nkjp.nlp.ipipan.waw.pl
 - 1 215 513 segmentów

Trwają prace nad:

- Korpusem Sejmowym
- Korpusomatem

KORBA: zapytanie i interfejs użytkownika

ELEKTRONICZNY KORPUS TEKSTÓW POLSKICH Z XVII I XVIII W. (DO 1772 R.)

Korpus

Korba automatyczna

Zapytanie

[base="Polska"]

ăăĂĂ

KONSTRUKTOR ZAPYTAŃ

ODRZUĆ OBCE SEGMENTY

Metadane

Ograniczenie

Skrót tekstu

zaczyna się od

Zapytanie o metadane

Ograniczenie wyszukiwania

Liczba wyników na stronę

Warstwa wyświetlania

Wszystkie teksty

10

uwspółcześniona

Wyszukaj

Znaleziono 500 wyników.

Lp	Lewy kontekst	Rezultat	Prawy kontekst	Skrót	Data
1	bowiem niepomału ściska, Bo na zajadłe wpadłem w	Polsce [Polska:subst:sg:loc:f]	zdrady. Uważaj rady. Już się tu postrzec w	WojDespBar_II	1700
2	czyje. Już bym i nazad wrócił tyło troje	Polsce [Polska:subst:sg:dat:f]	, byle mi darowała moje Zdrowie i więcej dałby	WojDespBar_II	1700
3	Nazad do ziemi wolne przeście dano, Wydałbym	Polsce [Polska:subst:sg:dat:f]	wszystko, com jej winien. Boś też	WojDespBar_II	1700
4	, Nieszczęsna Ryga tego narobiła. Bo polska była.	Polska [Polska:subst:sg:nom:f]	, ale ją mój dziad podbił sobie Pod swoje państwo	WojDespBar_II	1700

Litwie zostać? przy bliskości Fortunki mojej, czy do	Polski Concraft: [Polska:subst:sg:gen:f] Skrót: PotMuzaKarp Strona: 21 Autor: Wacław Potocki Tytuł: Muza polska	jachać in Spem JEj że pomnożenia. Ale i kraj	RzewKor
in postestatem redegerat. Udajmy się już do swej mitej		/ a weźmy przed się co o niej Kronikarze dają	PaxUlis
Bożej bywała/ i teraz jest. Czego znali oczywiste		w tak przestronym rozszerzeniu/ z wielu innych w językach	PaxUlis
jeżeli ją zdjąć ma poganin z głowy. Jeśli pół		, jako chce, zazonie, nie trzeba króla,	PotMuzaKarp

tu postrzec w tej mierze nie mogę. By nazad umknąć czy li zmyślić drogę, Żeby przynajmniej został się przy zdrowiu. Śmierć pogotowiu.

Trudno przy zdrowiu, wśród nieszczęścia wpadźszy, Mało polskiego chleba nie wyjadźszy, Nie wiem, co mnie tak srogo pobije. Konsylie czyje.

Już bym i nazad wrócił tyło troje **Polsce**, byle mi darowała moje Zdrowie i więcej dałbym co do tego. Spraw się wprzód z tego. Jakoż się sprawić, kiedy to Polacy Drwią głową zrazu i do końca tacy. Z kim co mówić i komu się zwierzyć? Chcą w cię uderzyć. Młode me lata wszystko to zrobiły, Że mnie tu

Skrót: WojDespBar_II
 Strona: 783
 Tytuł: Na desperacją szwedzką w Polsce echo odpowiada Szwedowi
 Autor: Anonim
 Miejsce wydania: nieznane
 Region: nieznane
 Typ tekstu: wiersz
 Rodzaj: liryka
 Poetyka żartu: nie
 Data wydania: 1700

Funkcje: język zapytań

Proste zapytania o różne warstwy - język zapytań CQL - podobnie jak w Poliqarpie:

1. `[orth="Polska"]`
2. wyrażenia regularne: `[translit="Pols.*"]`
3. łączenie znaczników: `[base="(ucho|oko)" & number="du"]`
4. kolejne segmenty: `[pos="subst"][pos="adv"]{0,1}[pos="adj"]`
5. negacja: `[!foreign=".*"]`
6. `[pos="subst" & !pos_c="subst"]`

KORBA: ograniczenie wg metadanych

Zapytanie

[base="Polska"]

á

é

Á

É

KONSTRUKTOR ZAPYTAŃ

ODRZUĆ OBCE SEGMENTY

Metadane

Data

Ograniczenie

z przedziału

Zapytanie o metadane

1640-1670

Ograniczenie wyszukiwania

Wszystkie teksty

Liczba wyników na stronę

10

Warstwa wyświetlania

uwspółcześniona

Wyszukaj

Znaleziono 19 wyników.

Lewy kontekst	Rezultat	Prawy kontekst	Skrót	Data
Ach, głupie śledzie! Co rok to w	Polsce [Polska:subst:sg:loc:f]	cierpicie, Dziwna rzecz, że się przecię do Polski	PrzetObBar_I	1653
Polsce cierpicie, Dziwna rzecz, że się przecię do	Polski [Polska:subst:sg:gen:f]	kwapicie. Sejmik raz miały ryby, w tym ich	PrzetObBar_I	1653
zgorszenie, Noteci go przyznawszy za pana, Na wszystkie	Polskę [Polska:subst:sg:acc:f]	wolny pas odkrywszy, Wpuścicieś wilka głodnego w koszary	NiebMSatBar_I	1658

Funkcje: filtrowanie wg metadanych

Metadane - dostępne z interfejsu:

1. Pola tekstowe: Skrót zaczyna się od Chmiel
2. Pola liczbowe: Data < 1700
3. Pola wielokrotnego wyboru: Region: Małopolska, Mazowsze

Korpus XIX w.: funkcje i interfejs użytkownika

Dodatkowa funkcjonalność: możliwość wygenerowania statystyk

Dwa typy statystyk: ze względu na metadane i ze względu na znaczniki segmentu

Statystyki:

1. Ze względu na warstwy - przykład: lista frekwencyjna rzeczowników
2. Ze względu na metadane
 - a. Data (pole liczbowe) - różne długości przedziału
 - b. Styl (string)

Zapytanie

[base="Polska"]

á

é

Á

É

ZAAWANSOWANE ▾

KONSTRUKTOR ZAPYTAŃ

Metadane

Etykieta ▾

Ograniczenie

zaczyna się od ▾

Zapytanie o metadane

☒ Statystyki

Grupowanie wg

w. współcześnieńona (ort

Liczba wyników

10 ▾

Liczba wyników na stronę

10 ▾

Warstwa wyświetlania

u współcześnieńona ▾

Wyszukaj

WYSTĄPIENIA ZE WZGLĘDU NA: WARSTWA UWSPÓŁCZEŚNIONA

1. Polski: 99
2. Polsce: 55
3. Polska: 41
4. Polskę: 25
5. Polską: 8
6. Polsko: 2
7. Polsci: 1

ZAAWANSOWANE ▾

KONSTRUKTOR ZAPYTAŃ

Metadane

Etykieta ▾

Ograniczenie

zaczyna się od ▾

Zapytanie o metadane

☒ Statystyki

Grupowanie wg

Data ▾

Długość przedziału

10 lat ▾

Liczba wyników na stronę

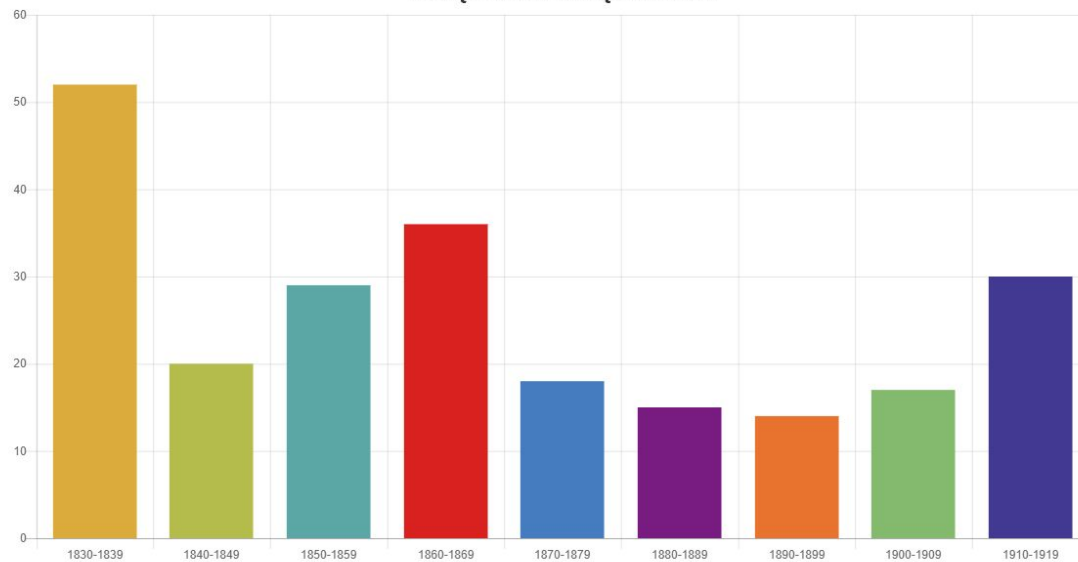
10 ▾

Warstwa wyświetlania

u współczesniona ▾

Wyszukaj

WYSTĄPIENIA ZE WZGLĘDU NA: DATA



Zapytanie
[base="Polska"]

ą é Ą ě

ZAAWANSOWANE ▾ KONSTRUKTOR ZAPYTAŃ

Metadane
Etykieta ▾

Ograniczenie
zaczyna się od ▾ Zapytanie o metadane

☒ Statystyki

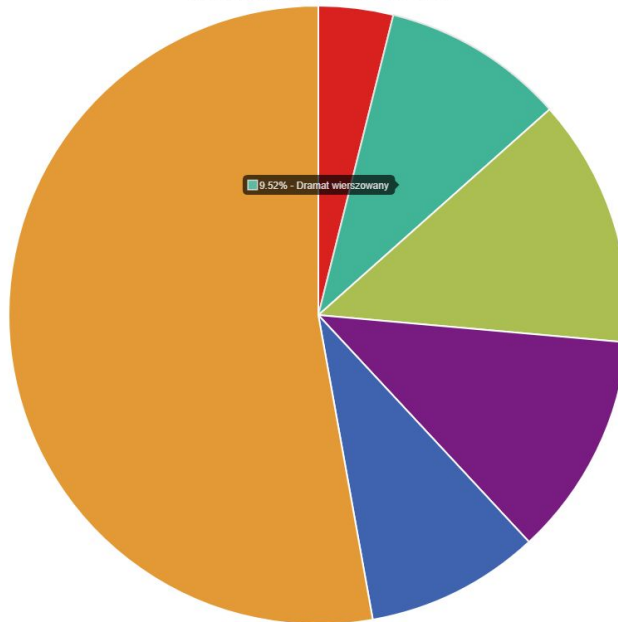
Grupowanie wg
Styl ▾

Liczba wyników na stronę
10 ▾

Warstwa wyświetlania
uwspółcześniona ▾

Wyszukaj

WYSTĄPIENIA ZE WZGLĘDU NA: STYL



Narodowy Korpus Języka Polskiego

Istnieje już interfejs webowy dla NKJP z wykorzystaniem Poliqarpa, ale:

- nie jest możliwe przeszukiwanie innych warstw poza warstwą morfosyntaktyczną,
- interfejs nie jest stabilny, ani szybki.

Udostępniamy nowy interfejs, który pozwala na przeszukiwanie i przeglądanie:

- warstw: morfosyntaktycznej, nazw własnych, sensów słów, grup składniowych (w trakcie prac)
- możliwa jest łatwa implementacja wcześniej omawianych funkcji (statystyki, wizualizacje)

Przykład: grupy (nazwy własne)

Zapytanie

<ne="persName"/>

ZAAWANSOWANE ▾

KONSTRUKTOR ZAPYTAŃ

Wyszukaj

Znaleziono 19280 wyników.

Lp	Lewy kontekst	Rezultat	Prawy kontekst	Etykieta
1	spoza "malborskiego podwórka" jest bardzo przydatne – twierdzi	Andrzej [Andrzej:subst:sg:nom:m1] Rychłowski [Rychłowski:subst:sg:nom:m1]	, burmistrz Malborka. - Pan Machczyński ma bardzo duże	IJPPAN_GazetaMalborska
2	– twierdzi Andrzej Rychłowski, burmistrz Malborka. - Pan	Machczyński [Machczyński:subst:sg:nom:m1]	ma bardzo duże doświadczenie i tego zamierzamy skorzystać – przekonuje	IJPPAN_GazetaMalborska
3	spółki oraz procedurą rokowań w malborskim magistracie. MALBORK.	Aneta [Aneta:subst:sg:nom:f] Florczyk [Florczyk:subst:sg:nom:f]	– siłaczka z Malborka – przeszła do następnego etapu turnieju	IJPPAN_GazetaMalborska
4	„Gwiazdy tańczą na lodzie” z udziałem malborskiej siłaczki	Anety [Aneta:subst:sg:gen:f] Florczyk [Florczyk:subst:sg:gen:f]	okazał się dla niej nie mniej fartownym niż poprzedni.	IJPPAN_GazetaMalborska
5	okazał się dla niej nie mniej fartownym niż poprzedni.	Aneta [Aneta:subst:sg:nom:f]	znalazła się w grupie szczęśliwców, którzy bawić będą widzów	IJPPAN_GazetaMalborska
6	pokłócić, bo życie byłoby nudne – mówi pani	Jadwiga [Jadwiga:subst:sg:nom:f] Studzińska [Studzińska:subst:sg:nom:f]	. Poseł Cymański docenił... Jadwiga i Władysław	IJPPAN_GazetaMalborska

Funkcje: grupy

Grupy - obecnie dotyczy zdań, akapitów i nazw własnych w NKJP

Operatory:

1. `within` - zawiera się w
2. `containing` - zawiera
3. `intersecting` - przecina
4. `fullyalignedwith` - w pełni uzgodniony
5. `followedby` - następujący
6. `precededby` - poprzedzony przez
7. `+` zaprzeczenie, np. `!intersecting`

Przykład: imię, po którym następuje nazwisko

Korpus

NKJP 1mIn

Zapytanie

<ne="persName.forename"/> followedby <ne="persName.surname"/>

ZAAWANSOWANE

KONSTRUKTOR ZAPYTAŃ

Wyszukaj

Znaleziono 6764 wyników.

Lp	Lewy kontekst	Rezultat	Prawy kontekst	Etykieta
1	wedzą do Polski z takim impetem jak np. tytoniowy	Phillip [Phillip:subst:sg:nom:m2]	Morris. Wszystko rozegra się w zaciszu ministerialnych gabinetów.	130-3-900001
2	Nie ma takiego leku ani terapii - mówi psychoterapeuta Lin	Jing [Jing:subst:sg:nom:m1]	Hua. - Nie mogę jednak zawieść swoich pacjentów.	130-3-900001
3	założono go 22 czerwca br. na polecenie szefa UOP	Andrzeja [Andrzej:subst:sg:gen:m1]	Kapkowskiego. Podłuchiowano nie tylko Smółkę, ale też b	130-3-900001
4	Smółkę, ale też b. szefa białostockiej delegatury UOP	Krzysztofa [Krzysztof:subst:sg:acc:m1]	Bondaryka. Urządzenia zainstalowali funkcjonariusze UOP spoza Białegostoku (informator	130-3-900001
5	ten wysoki w krawacie - to moja ochrona".	Halina [Halina:subst:sg:nom:f]	Kuban wraca do domu z trzema pekińczykami: Czina,	130-3-900001
6	. - Pierwsze rozliczenie remontu budynku cofnęliśmy - mówi	Halina [Halina:subst:sg:nom:f]	Czerniowska, kierownik wydziału pomocy społecznej Urzędu Miasta. -	130-3-900001
7	Gazety" było "uśmiercenie" w 1993 r.	Janusza [Janusz:subst:sg:gen:m1]	Leksztonia, właściciela firmy Elgaz. 26 marca poinformowaliśmy	130-3-900001
6	. - Pierwsze rozliczenie remontu budynku cofnęliśmy - mówi	Halina [Halina:subst:sg:nom:f]	Czerniowska, kierownik wydziału pomocy społecznej Urzędu Miasta. -	130-3-900001
7	Gazety" było "uśmiercenie" w 1993 r.	Janusza [Janusz:subst:sg:gen:m1]	Leksztonia, właściciela firmy Elgaz. 26 marca poinformowaliśmy	130-3-900001

MTAS: ograniczenia

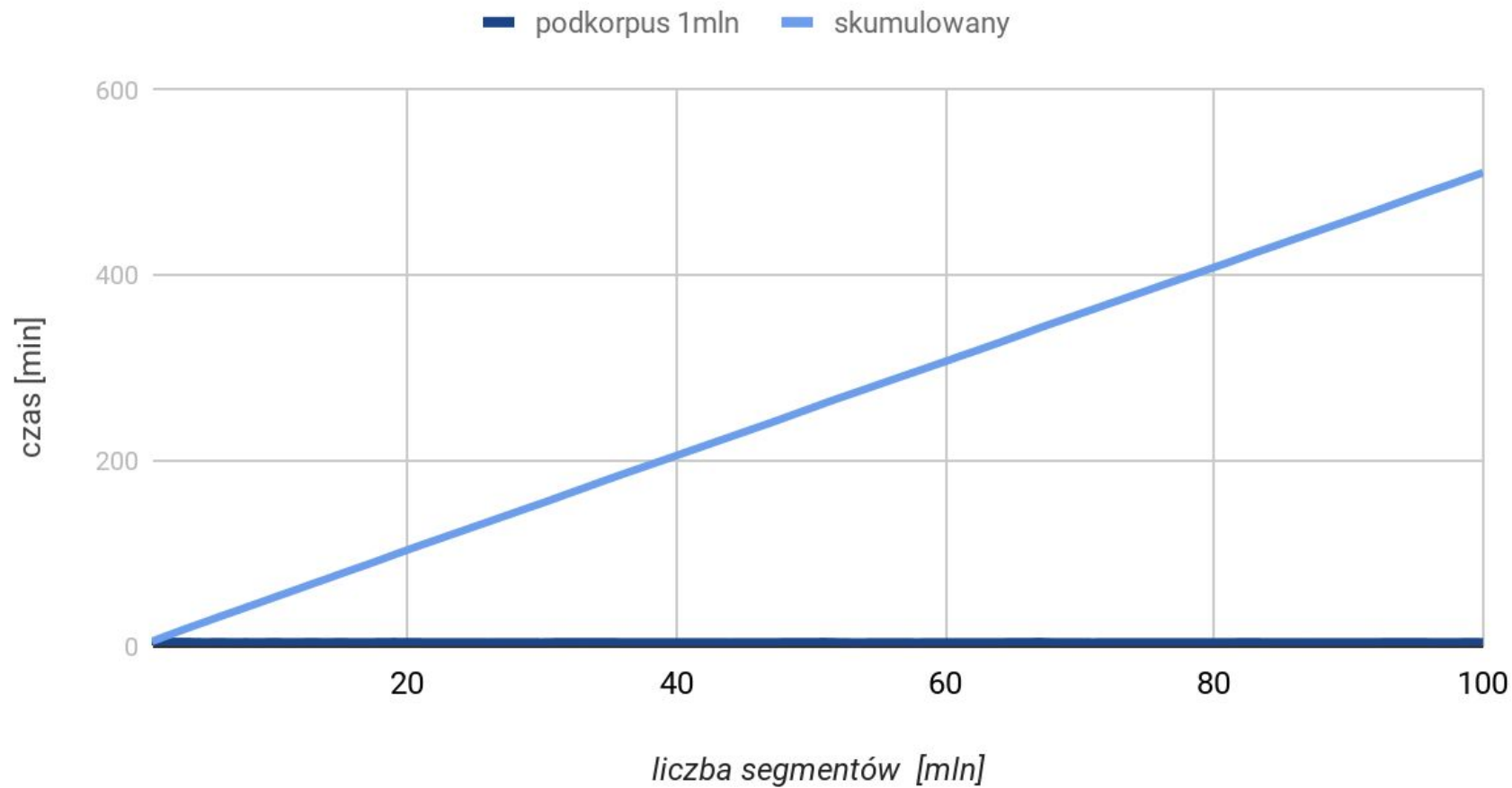
Czego nie ma:

1. **zmienne** - `[pos=subst & case=$1][pos=adj & case=$1]`
2. jednoczesne stronicowanie wyników oraz sortowanie wg klucza z metadanych
3. przedziały nieskończone - `[pos=subst]{5,}`
4. flagi do ignorowania wielkości liter `[orth=ktoś/i]`
zamiast tego dodatkowe pole: `[orth_1c="ktoś"]`
5. brak wsparcia dla niewybranych interpretacji - rozwiązanie: dodatkowa warstwa anotacji

Testy wydajności na 1 rdzeniu Solr

- Podkorporus NKJP:
 - 1 000 094 segmentów
 - 12 063 942 tokenów indeksu
 - 3 724 dokumentów
 - średnio 268 segmentów na dokument
- Podkorporus dodawany 100 krotnie do indeksu ze zmienionymi ID dokumentów
- Rozmiar (zoptymalizowanego) indeksu na dysku:
 - 1 mln segmentów - 306.5 MB
 - 100 mln segmentów - 27,58 GB
- Środowisko testowe:
 - Intel Core i7-6700K 4.5GHz OC, 32 GB RAM, dysk SSD
 - 24 GB RAM przydzielone dla Solr

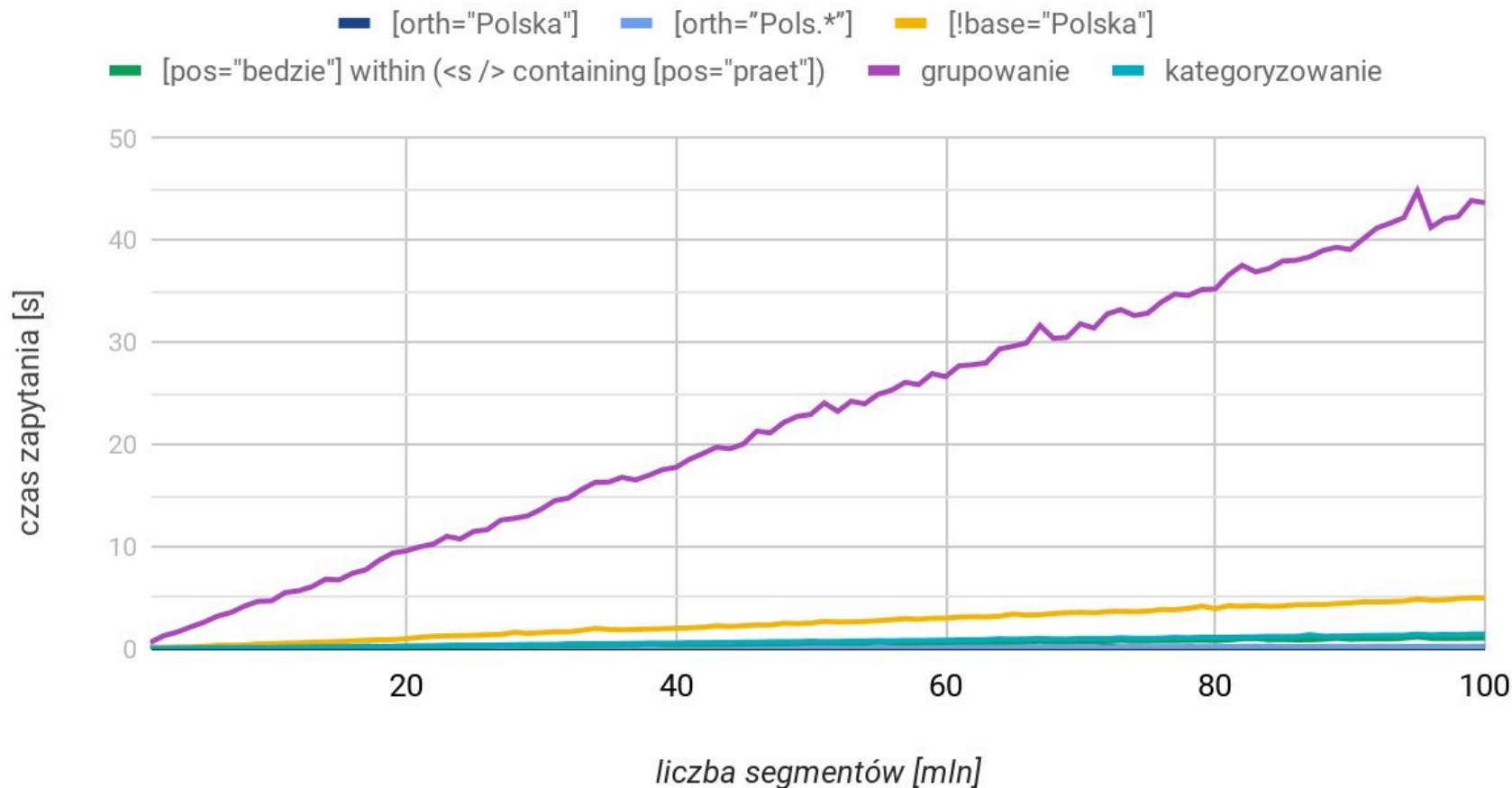
Czas indeksowania

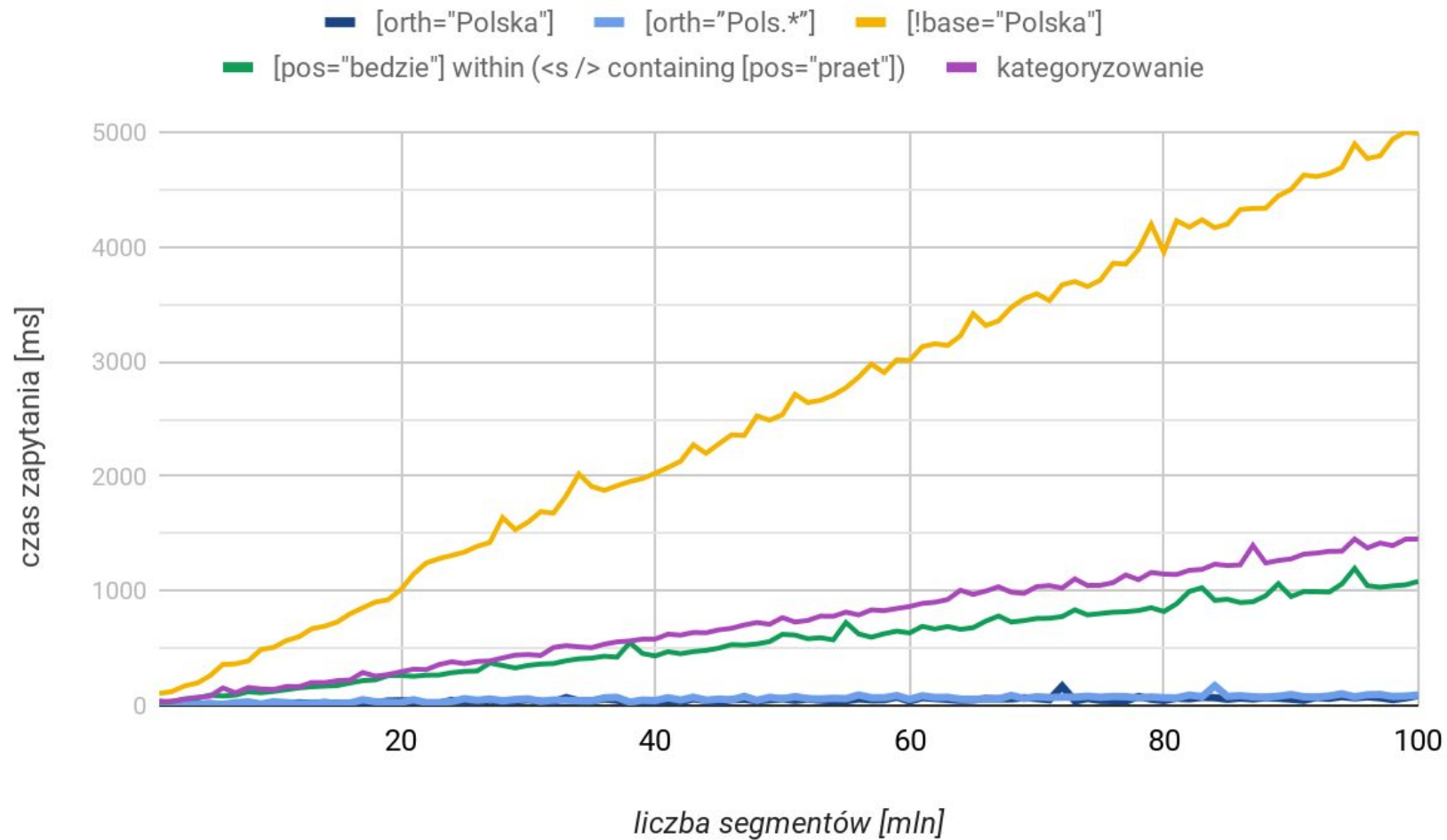


Testowane zapytania

- `[orth="Polska"]`
- `[orth="Pols.*"]`
- `[!base="Polska"]`
- `[pos="bedzie"]` within (`<s />` containing `[pos="praet"]`)
- grupowanie wg wartości występujących w warstwie anotacji pos
- grupowanie (kategoryzowanie) `[pos="subst"]` wg metadanych (daty)

Średni czas zapytania w zależności od wielkości korpusu





Projekt Nederlab

- 15 mln dokumentów
 - ponad 2,2 mln dokumentów z anotacją morfosyntaktyczną
- 10 mld segmentów
- 35 mld anotacji (tokenów m_{tas})
- korpus rozproszony na 23 rdzenie Solr
- całkowity rozmiar indeksu ponad 1 TB
- serwer Xeon E5 z 128 GB RAM
- 67 GB RAM przydzielone dla Solr
- wyszukanie [pos="ADJ"][pos="N"] trwa około 4s

MTAS w Korpusomacie

Plan wdrożenia:

- wiele niezależnych rdzeni Solr:
 - prywatne korpusy użytkowników,
 - korpusy współdzielone;
- możliwość wyszukiwania w korpusie z użyciem metadanych zdefiniowanych przez użytkownika - dynamiczny schemat indeksowania Solr.

Zalety względem obecnego rozwiązania (Poliqarp):

- łatwe zarządzanie współbieżnością,
- łatwe rozpraszanie na wiele maszyn,
- indeksowanie przyrostowe.

Podsumowanie i dalsze kierunki rozwoju

Co zostało zrobione:

1. Wdrożenie MTAS w 3 polskich korpusach językowych: KORBA, XIX w., NKJP, wraz z interfejsem webowym.
2. Dostosowanie do specyfiki polskiego tagsetu, sposobu anotacji i formatu danych stosowanego w polskich korpusach (TEI).
3. Testy wydajności dla typowych zapytań.

Dalsze plany:

4. Wdrożenie dla “dużych” korpusów: NKJP 1,8 mld segmentów; wykorzystanie możliwości rozpraszania indeksu.
5. Implementacja zmiennych w CQL MTAS.
6. Rozwiązanie problemu z sortowaniem wyników.

Dziękujemy za uwagę!