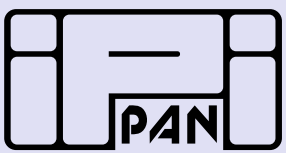
Identyfikacja terminów wielowyrazowych w tekście TermoPL

Małgorzata Marciniak, Agnieszka Mykowiecka, Piotr Rychlik



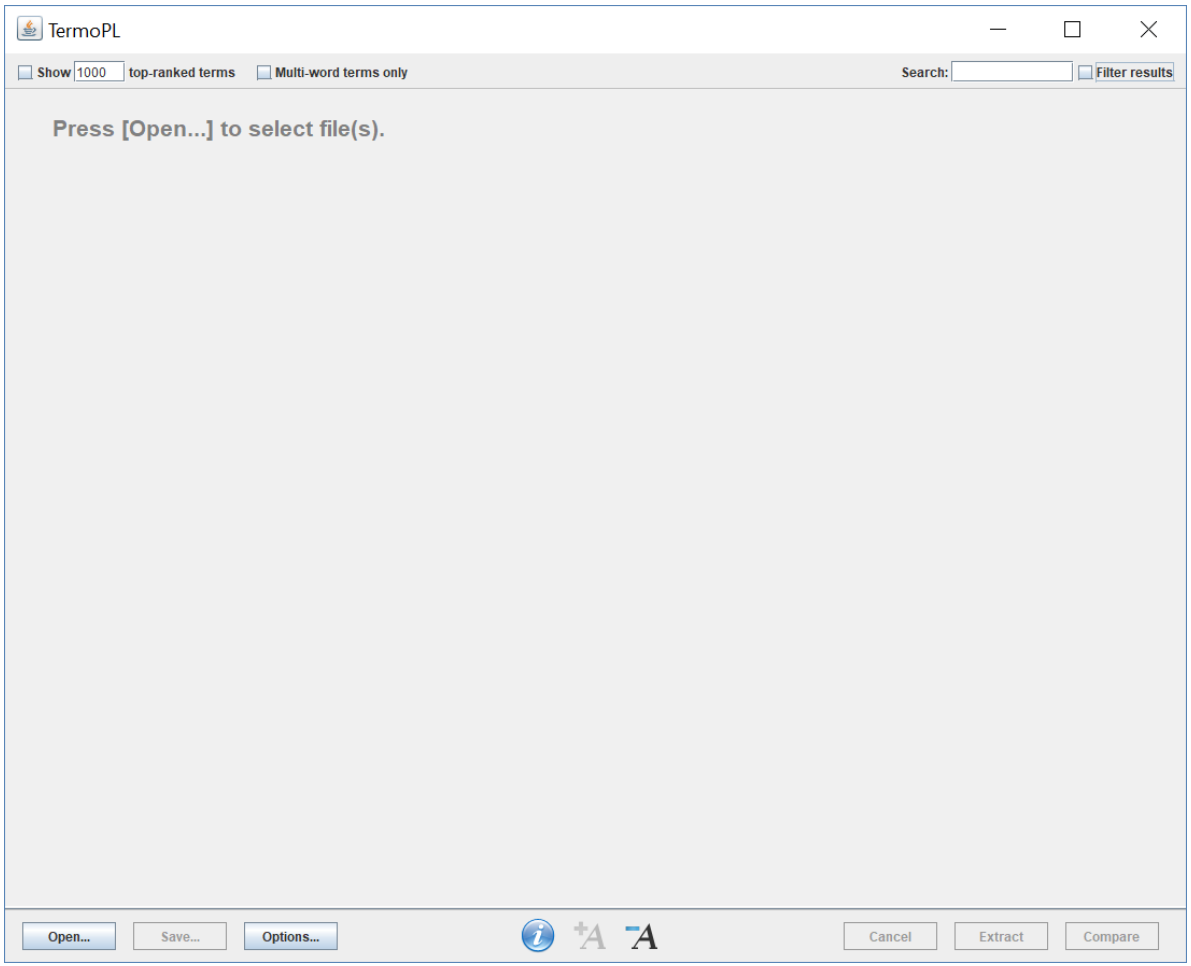
INSTITUTE OF COMPUTER SCIENCE
POLISH ACADEMY OF SCIENCES
ul. J. K. Ordona 21, 01-237 Warszawa

Warsztaty CLARIN-PL, Toruń 16-17 XI 2018

Program TermoPl

instalacja: http://zil.ipipan.waw.pl/TermoPL

uruchamianie: termopl.bat zamiast TermoPL.jar (.bat rozszerza dostępną pamięć)



Różne podejścia do wyróżniania grup słów

- kolokacje (kryteria leksykalne/korpusowe)
  klatka piersiowa konwalia majowa
- nazwy własne (kryteria dziedzinowe)
  New York Nowa Zelandia, Rio de Janeiro
- **terminy dziedzinowe** (kryteria dziedzinowe)
  skrzynia biegów, podatek od osób fizycznych, komedia romantyczna
- grupy składniowe ("słabe" kryteria lingwistyczne)
  [The debate]$_{NP}$ and [vote]$_{NP}$ [in Cyprus' parliament]$_{NP}$ has now been [postponed]$_{V}$ until [18:00 local time (16:00 GMT)]$_{NP}$ on [Tuesday]$_{NP}$
- frazy (kryteria lingwistyczne)
  [[The debate]$_{NP}$ and [vote]$_{NP}$]$_{NP\_SUBJ}$ [in [Cyprus']$_{ADJ\_MOD}$ parliament$_{NP}$]$_{Prep\_MOD}$ [has [now]$_{ADV\_MOD}$ been postponed]$_{VP}$ ...

Kolokator (http://www.nkjp.uni.lodz.pl/collocations.jsp)

Kryteria ośrodka kolokacji:

student

Maksymalny odstęp: 0 Zachowaj szyk: ☑

Kryteria kolokatu:

Część mowy: Przym./lmiesł. Kontekst z lewej: 1 Kontekst z prawej: 1

Wielkość próbki: 10000 Min. współwystąpień: 5

Ośrodek kolokacji występuje w korpusie 5153 razy. Podajemy wyniki na podstawie wszystkich współwystąpień. Znaleziono **56** kolokacji spełniających zadane kryteria.

#	Kolokacja	**Pasujące współwystąpienia** (kliknij na frekwencję, aby wyświetlić przykłady)	Ogółem	Chi^2
1.	**wrocławski**	wrocławski___student *(7)*, student___wrocławskiej *(4)*, student___wrocławskiego *(3)*, student___wrocławski *(1)*,	15	2,707,982.49
2.	**amerykański**	amerykański___student *(8)*, student___amerykański *(1)*, student___amerykańskiego *(1)*,	10	1,805,321.65
3.	**dziesiąty**	dziesiąty___student *(5)*,	5	902,660.82
4.	**pomysłowy**	pomysłowy___student *(4)*, student___pomysłowy *(1)*,	5	902,660.82
5.	**niepełnosprawny**	niepełnosprawny___student *(5)*, niepełnosprawnych___student *(1)*,	6	649,910.99
6.	**poznański**	student___poznańskiej *(4)*, poznańskiego___student *(1)*,	5	451,325.41
7.	**gdański**	student___gdańskiej *(4)*, gdańska___student *(1)*, gdańsku___student *(1)*, gdański___student *(1)*,	7	442,294.71
8.	**upiec**	upieczony___student *(10)*,	10	21,969.57
9.	**zaoczny**	student___zaocznego *(3)*, student___zaocznej *(2)*, student___zaoczny *(2)*, zaoczny___student *(1)*,	8	20,727.62
10.	**dwudziestoletni**	dwudziestoletni___student *(6)*,	6	13,360.91
11.	**prawy**	student___prawa *(95)*, prawa___student *(1)*,	96	4,909.26
12.	**wieczny**	wieczny___student *(11)*,	11	3,006.99
13.	**letni**	letni___student *(11)*,	11	2,966.79
14.	**biedny**	biedny___student *(12)*,	12	2,846.13
15.	**czwarty**	student___czwartego *(12)*, czwarty___student *(6)*,	18	2,223.5

Zadanie realizowane przez TermoPL

Cel

wydobycie specyficznej terminologii z tekstów dotyczących wybranej dziedziny.

Założenia:

- korzystamy z danych otagowanych morfologicznie,
- **identyfikacja terminów**: gramatyka opisującą składnię interesujących nas fraz,
- **szeregowanie terminów**: informacje o częstości występowania w tekście, C-value,
- uwzględniamy frazy zagnieżdżone,
- wykorzystujemy lematyzację do grupowania form fleksyjnych.

Co rozumiemy pod pojęciem terminu?

Definicja słownikowa

Wyraz albo połączenie wyrazowe o specjalnym, konwencjonalnie ustalonym znaczeniu naukowym lub technicznym; (Doroszewski)

Definicja robocza

Fraza rzeczownikowa, która w tekstach dziedzinowych występuje dostatecznie często by przypuszczać, że opisuje pojęcie istotne dla dziedziny. Częstość tej frazy w tekstach spoza dziedziny jest niższa.

Przykładowe wyniki TermoPl

"student", NKJP 1.2mln

student (122 wystąpienia)
student medycyny
student wszystkich uczelni Szczecina
student Wyższej Szkoły Zarządzania
student Akademii Ekonomicznej
student uczelni artystycznych
młody student ekonomii
student filologii rosyjskiej
student nauk społecznych
student siłaczy (student siłacz)
zaprzyjaźniony student
student architektury
student ASP
student AWF
afrykański student
student politologii
dzisiejszy student
student warszawski

Przykładowe wyniki TermoPL

"student", NKJP 1.2mln+Wikipedia

student (1089)
liczba studentów
pożyczka studencka
zrzeszenie studentów
wzrost liczby studentów
zrzeszenie studentów polskich
zwiększyć liczbę studentów
kształcenie studentów
student studium zaocznego
student studium
parlament studenta
student pierwszego roku
przejęcie majątku zrzeszenia studentów polskich
kredyt studencki

Realizacja zadania

- Zgromadzenie tekstów dziedzinowych.
- Wstępna analiza lingwistyczna — tagowanie (przypisanie formy podstawowej, części mowy oraz charakterystyki morfologicznej), można w tym celu użyć Korpusomatu.
- Identyfikacja fraz — kandydatów na terminy.
- Szeregowanie fraz.
- Selekcja fraz.

Wyniki, format wydruku, NKJP 1.2 mln.

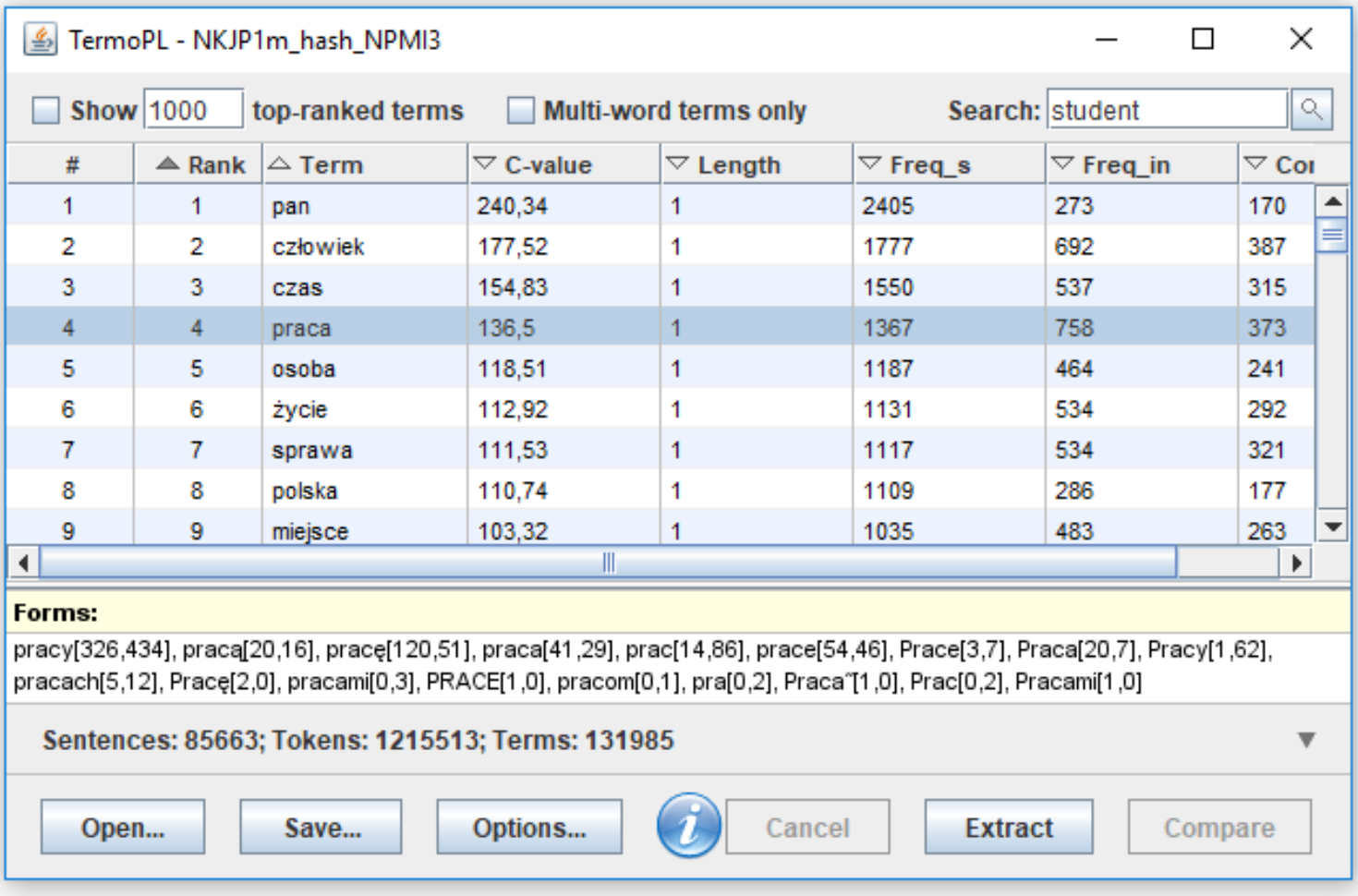
poz	termin	[forma podstawowa]	wsp.C	dl	#	#wew	kont.
1	pan	[pan]	240.3	1	2405	273	170
2	człowiek	[człowiek]	177.5	1	1777	692	387
3	czas	[czas]	154.8	1	1550	537	315
4	praca	[praca]	136.5	1	1367	758	373
5	osoba	[osoba]	118.5	1	1187	464	241
6	życie	[życie]	112.9	1	1131	534	292
7	sprawa	[sprawa]	111.5	1	1117	534	321
8	Polska	[polska]	110.7	1	1109	286	177
9	miejsce	[miejsce]	103.3	1	1035	483	263
10	dziecko	[dziecko]	102.3	1	1025	329	209
11	unia europejski	[unia europejska]	97.3	2	99	46	27

Wczytanie danych

klikamy **Open**, a następnie wybieramy plik (NKJP1m_hash.txt)
klikamy **Extract**

Wczytanie danych

klikamy **Open**, a następnie wybieramy plik (NKJP1m_hash.txt)
klikamy **Extract**

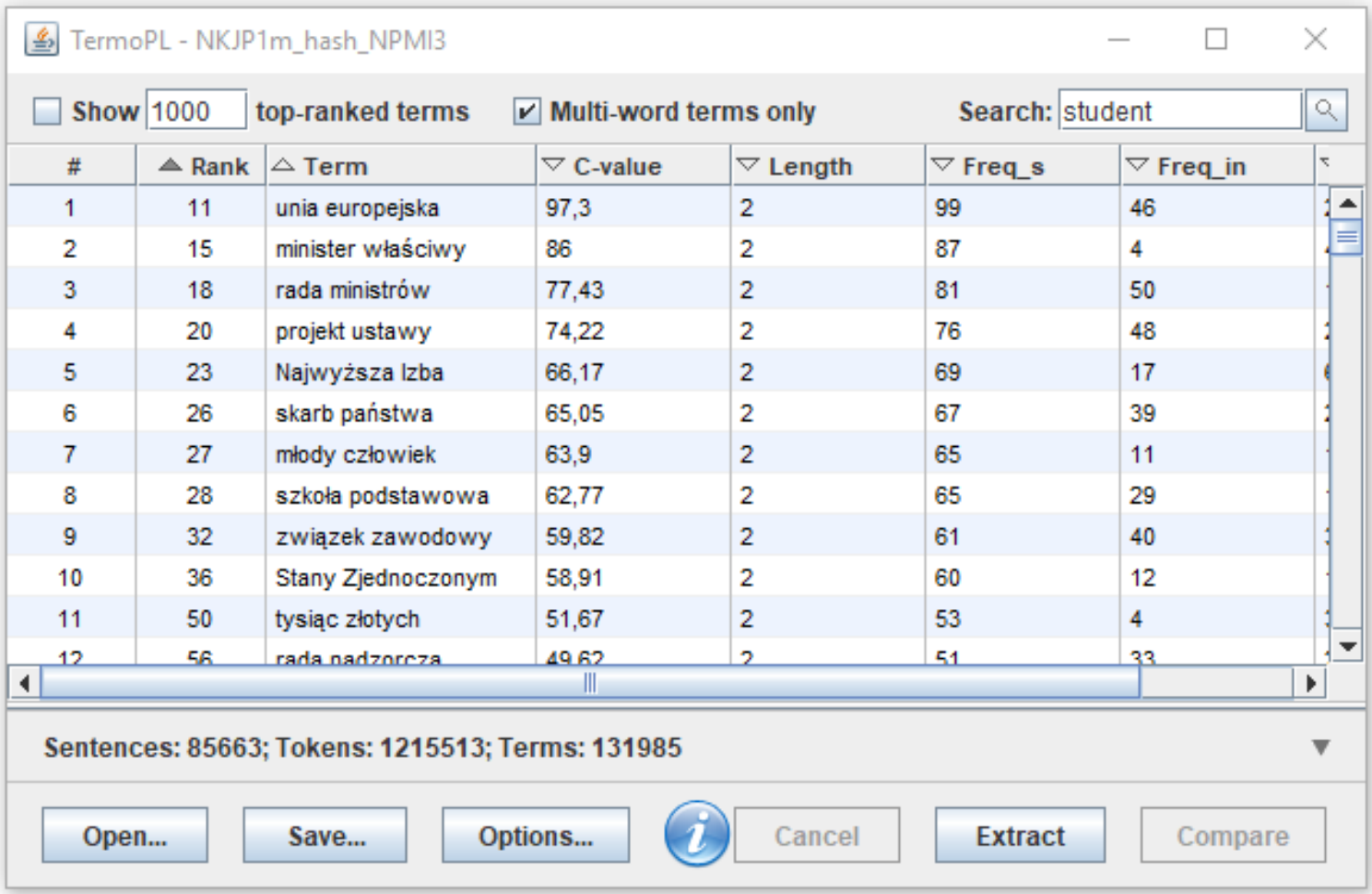


Selekcja terminów wielowyrazowych

klikamy 'Multi-word terms only'

Selekcja terminów wielowyrazowych

klikamy 'Multi-word terms only'



Wyszukiwanie

wpisujemy tekst w polu 'Search' i naciskamy 'Enter'

Wyszukiwanie

wpisujemy tekst w polu 'Search' i naciskamy 'Enter'

TermoPL - NKJP1m_hash_NPMI3

Show 1000 top-ranked terms ☑ Multi-word terms only 28 matches Search: student ☑ Filter results

#	Rank	Term	C-value	Length	Freq_s	Freq_in	Context #
1	2052	student medycyny	2	2	2	0	0
2	2052	student wszystkich uczelni Szczecina	2	4	1	0	0
3	2052	student Wyższej Szkoły Zarządzania	2	4	1	0	0
4	2179	student Akademii Ekonomicznej	1,58	3	1	0	0
5	2179	student uczelni artystycznych	1,58	3	1	0	0
6	2179	młody student ekonomii	1,58	3	1	0	0
7	2179	student filologii rosyjskiej	1,58	3	1	0	0
8	2301	student siłaczy	1	2	1	0	0
9	2301	zaprzyjaźniony student	1	2	1	0	0
10	2301	student architektury	1	2	1	0	0
11	2301	student ASP	1	2	1	0	0
12	2301	student AWF	1	2	1	0	0
13	2301	afrykański student	1	2	1	0	0
14	2301	student politologii	1	2	1	0	0
15	2301	wszystek student	1	2	1	0	0
16	2301	dzisiejszy student	1	2	1	0	0
17	2301	student warszawski	1	2	1	0	0
18	2301	student SZSP	1	2	1	0	0
19	2301	student izraelski	1	2	1	0	0
20	2301	student historii	1	2	1	0	0
21	2301	student fizyki	1	2	1	0	0
22	2301	student UMCS	1	2	1	0	0

Forms:

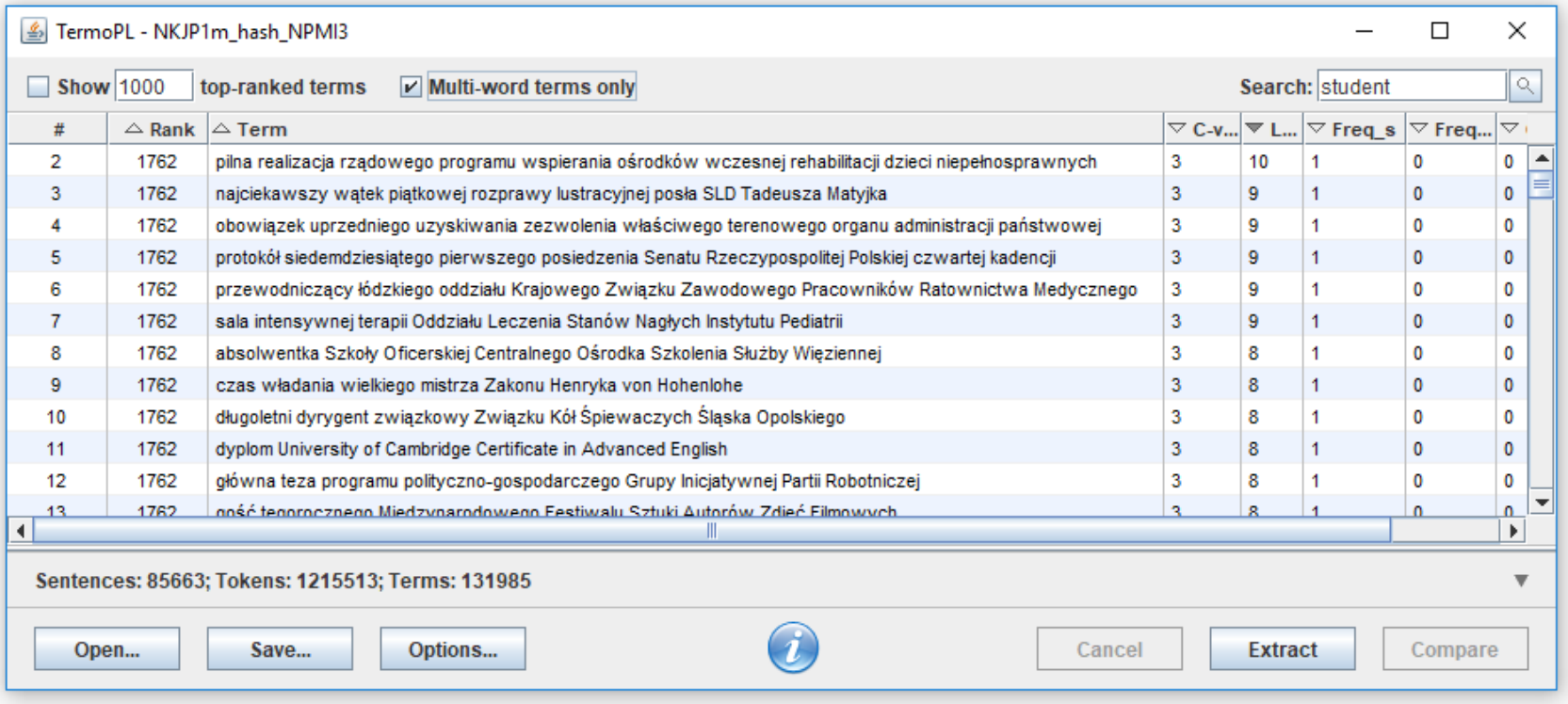
Studenci medycyny[1,0], studenta medycyny[1,0]

Sortowanie

klikamy 'trójkącik' przy nazwie kolumny, np. 'Length'

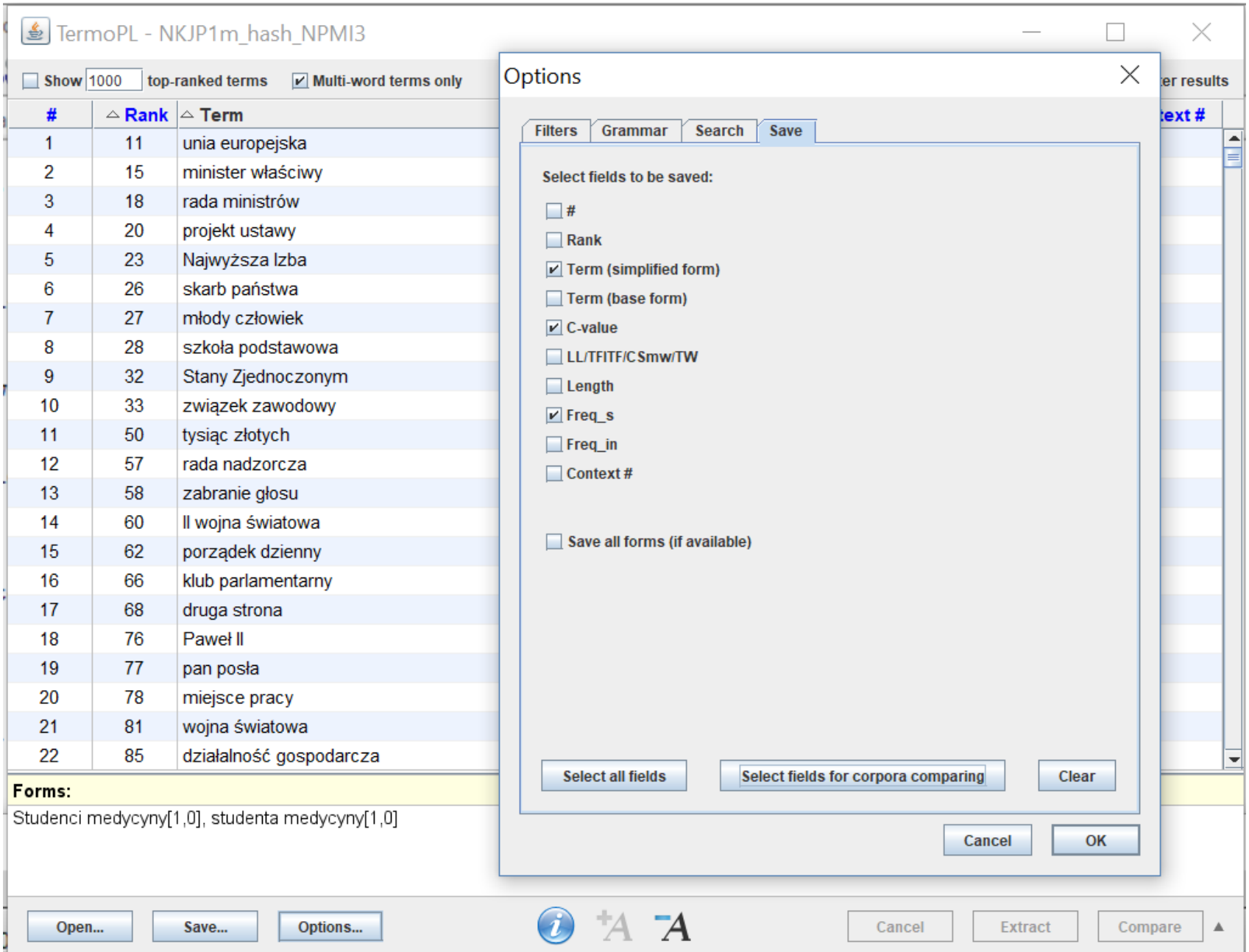
Sortowanie

klikamy 'trójkącik' przy nazwie kolumny, np. 'Length'



Zapisanie listy terminów

Zapisanie listy terminów



Struktura gramatyczna terminów w języku polskim

- rzeczownik, akronim lub skrót rzeczownika:
  - *podatek*, *angiografia*,
  - *PKB*, *USG*
  - *ust.*(awa),
- rzeczownik z przymiotnikiem (który wystąpił po lub rzadziej przed rzeczownikiem):
  - *stosunki gospodarcze*,
  - *granulocyty obojętnochłonne*;
- sekwencja rzeczownika z rzeczownikiem w dopełniaczu:
  - $udar_{n,nom}$ $mózgu_{n,gen}$;
  - $kodeks_{n,nom}$ $pracy_{n,gen}$;
- kombinacja powyższych dwóch struktur:
  - $europejski_{adj}$ $rynek_{n,nom}$ $usług_{n,gen}$ $finansowych_{adj}$,
  - *wodonercze niewielkiego stopnia dolnego układu podwójnego nerki prawej*;

Struktura gramatyczna terminów w języku polskim

- fraza rzeczownikowa modyfikowana frazą przyimkową:
  - *wierzytelność podatnika wobec skarbu państwa*,
  - *podatek dochodowy od osoby fizycznej*;
  - *poziom hormonów we krwi*;
- można uwzględnić koordynację:
  - *bezsporna i wymagalna wierzytelność podatnika wobec skarbu państwa*,
  - *zapalenie mózgu i rdzenia*,
  - *oddział alergologii, endokrynologii i pediatrii ogólnej*.

Opis stuktury terminów– gramatyka domyślna

$$
\begin{aligned}
NPP &: \$NAP \;\; NAP\_GEN^*; \\
NAP[agreement] &: AP^* \;\; N \;\; AP^*; \\
NAP\_GEN[case = gen] &: NAP; \\
AP &: ADJ \mid ADJA \;\; DASH \;\; ADJ \mid PPAS; \\
N[pos = subst, ger]; & \\
ADJ[pos = adj]; & \\
ADJA[pos = adja]; & \\
PPAS[pos = ppas]; & \\
DASH[form = \text{"-"}]; &
\end{aligned}
$$

Zagnieżdżone terminy wielowyrzowe

jama ustny [jama ustna] 241

śluzówka jama ustny [śluzówka jamy ustnej] 79
jama ustny czysty [jama ustna czysta] 4
suchość jama ustny [suchość jamy ustnej] 4
błona śluzowy jama ustny [błona śluzowa jamy ustnej] 3
grzybica śluzówka jama ustny [Grzybica śluzówek jamy ustnej] 2
pielęgnacja jama ustny [pielęgnacji jamy ustnej] 2
płukanie jama ustny [Płukanie jamy ustnej] 2
grzybica jama ustny [Grzybica jamy ustnej] 2
pędzlować jama ustny [pędzlowanie jamy ustnej] 2
sanacja jama ustny [sanacji jamy ustnej] 1
silny grzybica jama ustny [silna grzybica jamy ustnej] 1
jama ustny pleśniawka [jama ustnej pleśniawki] 1
zapalenie jama ustny [zapalenie jamy ustnej] 1
dno jama ustny [dno jamy ustnej] 1

Wykluczenie niektórych słów/fraz

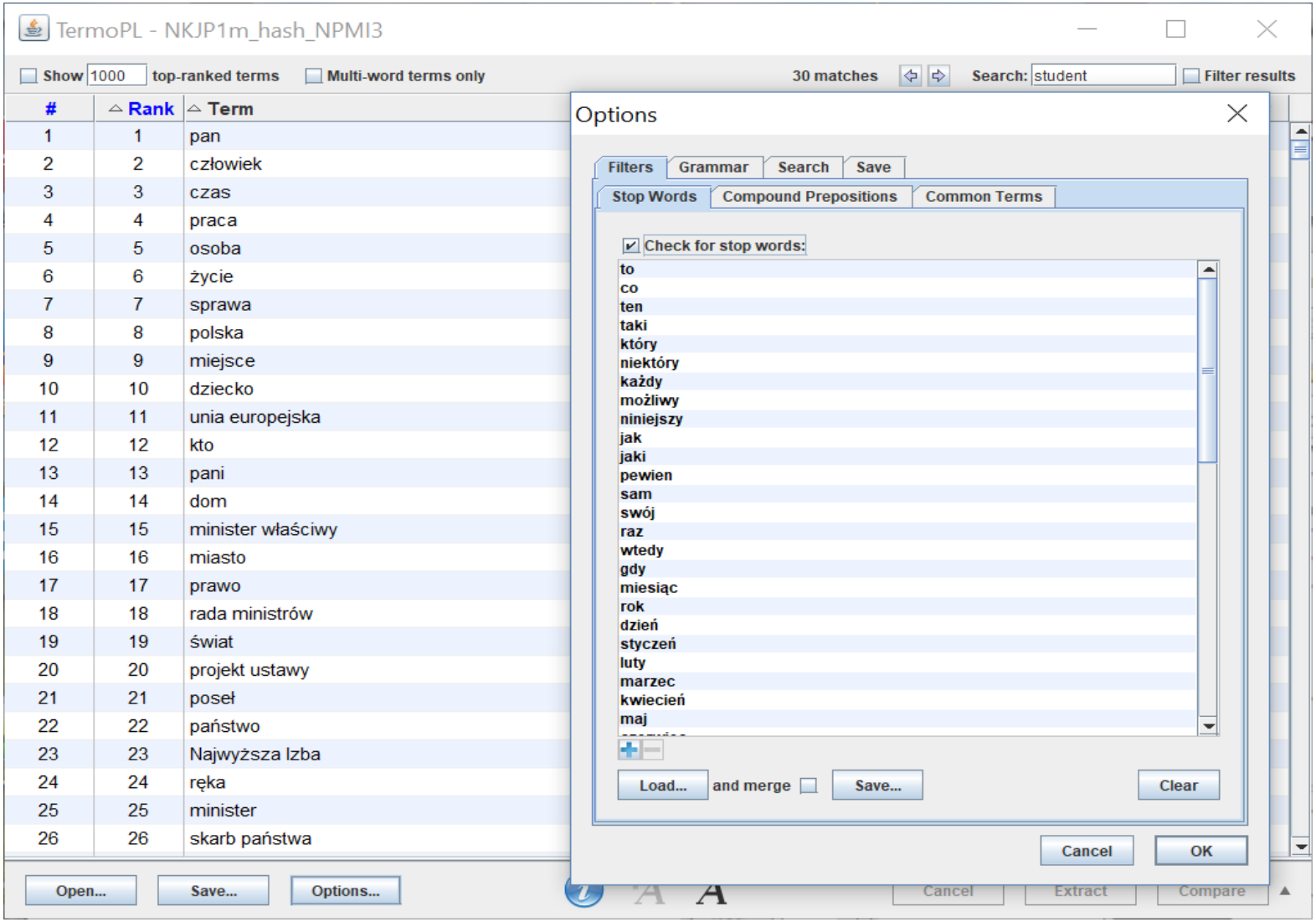
Zwykle nie traktujemy jako terminów dziedzinowych fraz składających się z:

- słów wskazujących na określenie czasu, jak np: *miesiąc, dzień*;
- nazwy dni i miesięcy, np: *styczeń, poniedziałek*;
- przymiotników wymagających kontekstu do interpretacji np: *inny, niektóry, jakiś, pewien*.

Chcemy też wykluczyć przyimki złożone:

- *[w kierunku] zapalenia nerek —> kierunek zapalenia nerek*;
- *[pod postacią] podatku VAT —> postać podatku VAT*;
- *[pod kątem] diagnostyki obrazowej —> kąt diagnostyki obrazowej*;
- *[pod kątem] prostym —> kąt prosty*.

Wykluczenie niektórych słów – opcje



Szeregowanie terminów — C-value

Bierzemy pod uwagę:

- częstość frazy;
- długość frazy;
- liczbę kontekstów.

referencja bibliograficzna

Frantzi, K., Ananiadou, S., Mima, H.: Automatic recognition of multi-word terms: the C-value/NC-value method. Int. Journal on Digital Libraries 3 (2000) 115–130

Identyfikacja fraz

	pojedyncza	mnoga
nom	*przewlekły nieżyt żołądka*	*przewlekłe nieżyty żołądka*
gen	*przewlekłego nieżytu żołądka*	*przewlekłych nieżytów żołądka*
dat	*przewlekłemu nieżytowi żołądka*	*przewlekłym nieżytom żołądka*
acc	*przewlekły nieżyt żołądka*	*przewlekłe nieżyty żołądka*
inst	*przewlekłym nieżytem żołądka*	*przewlekłymi nieżytami żołądka*
loc	*przewlekłym nieżycie żołądka*	*przewlekłych nieżytach żołądka*

Wykorzystujemy uproszczoną formę podstawową:

- *przewlekły nieżyt żołądka* –> *przewlekły nieżyt żołądek*;
- *ostra niewydolność nerek* –> *ostry niewydolność nerka*.

Problemy z uproszczoną formą podstawową

Taką sama uproszczoną formę podstawową mają:

- frazy w liczbie mnogiej i pojedynczej np. *zapalenie ucha* i *zapalenie uszu*, uproszczona: *zapalenie ucho*;
- przymiotniki w różnych stopniach (mały, mniejszy) np. *miednica mała* (częściej *mała miednica* — opisuje rozmiar) podczas gdy *miednica mniejsza* (określenie anatomiczne), uproszczona: *miednica mały*;
- pozytywne i zanegowane imiesłowy przymiotnikowe . *powiększony*/*niepowiększony* mają formę podstawową $powiększyć_{inf}$;
- gerundia i imiesłowy maja bezokoliczniki jako formy podstawowe:
  - $usunięcie_{ger}$ $kamienia_{subst:gen}$ — operacja,
  - $usunięty_{ppas}$ $kamień_{subst:nom}$ — opis kamienia,

  forma uproszczona: $usunąć_{inf}$ $kamień_{subst}$.

Konteksty

planowa	*operacja*	*przepukliny*	*pachwinowej*	*lewostronnej*
	operacja	*przepukliny*	*pachwinowej*	*lewostronnej*
planowa	*operacja*	*przepukliny*	*pachwinowej*	
	operacja	*przepukliny*	*pachwinowej*	
		przepuklina	*pachwinowa*	*lewostronna*
	lewostronna	*przepuklina*	*pachwinowa*	
		przepuklina	*pachwinowa*	*prawostronna*
		przepuklina	*pachwinowa*	*obustronna*
	prawostronna	*przepuklina*	*pachwinowa*	
	uwięźnięta	*przepuklina*	*pachwinowa*	*prawostronna*

Liczenie kontekstów

Metody liczenia kontekstów (ograniczamy do jednego słowa):

1. liczba różnych kontekstów liczona po obu stronach razem;
2. suma różnych kontekstów po obu stronach;
3. maksimum z kontekstów liczonych z lewej i prawj strony osobno.

Konteksty dla frazy: *przepuklina pachwinowa*:

1. ‘operacja’–‘lewostronny’, ‘operacja’–[pusty], [pusty]–‘lewostronny’, ‘lewostronny’–[pusty], [pusty]–‘prawostronny’, [pusty]–‘obustronny’, ‘prawostronny’–[pusty], ‘uwięźnięty’–‘prawostronny’;
2. ‘operacja’, ‘lewostronny’, ‘prawostronny’, ‘obustronny’, ‘uwięźnięty’;
3. ‘operacja’, ‘lewostronny’, ‘prawostronny’, ‘uwięźnięty’ (lewych o jeden więcej).

Problem uciętych fraz

Przykłady frazy o silnym powiązaniu słów:

- w medycynie: *pęcherzyk żółciowy*, *jama brzuszna*, *staw kolanowy*;
- w ekonomii: *papiery wartościowe*, *fundusz inwestycyjny*;
- w angielskim: *contact lens*.

Gramatycznie poprawne zagnieżdżone frazy:

- *[zapalenie pęcherzyka] żółciowego*;
- *[USG jamy] brzusznej*;
- *[operacja lewego stawu] kolanowego*;
- *[giełda papierów] wartościowych*;
- *[uczestnik funduszu] inwestycyjnego*;
- *[soft contact] lens*.

NPMI – Normalised Pointwise Mutual Information

NPMI wykorzystujemy do oceny siły powiązania pomiędzy słowami.

referencja bibliograficzna

Gerlof Bouma, 2009, *Normalized (pointwise) mutual information in collocation extraction.*, w: *Proceedings of the Biennial GSCL Conference 2009*, strony 31—40.

Przykład9q

infekcja górnych dróg oddechowych

Noun_j Adj_i Noun_i Adj_i

infekcja | górnych dróg | oddechowych

infekcja górny droga oddechowy

bigram	NPMI
infekcja górny	0.66
górny droga	0.79
droga oddechowy	0.95

Porównanie dwóch metod

Poprawne gramatycznie podfrazy				Podfrazy z wykorzystaniem NPMI			
‘*infekcja*’	‘*górny*’	‘*droga*’	‘*oddechowy*’	‘*infekcja*’	‘*górny*’	‘*droga*’	‘*oddechowy*’
infekcja	*górnych*	*dróg*	*oddechowych*	*infekcja*	*górnych*	*dróg*	*oddechowych*
infekcja	*górnych*	*dróg*		—			
infekcja				*infekcja*			
	górne	*drogi*	*oddechowe*		*górne*	*drogi*	*oddechowe*
	górne	*drogi*		—			
		drogi	*oddechowe*			*drogi*	*oddechowe*
		drogi				*drogi*	

Preferowanie podziału na dwie frazy rzeczownikowe

prawidłowa$_{adj}$ *mikroflora*$_{noun}$ *górnych*$_{adj}$ *dróg*$_{noun}$ *oddechowych*$_{adj}$ —> *prawidłowa mikroflora* oraz *górne drogi oddechowe*

częste$_{adj}$ *infekcje*$_{noun}$ *górnych*$_{adj}$ *dróg*$_{noun}$ *oddechowych*$_{adj}$ —> *częste* modyfikuje całą frazę *infekcje górnych dróg oddechowych*

Modyfikacja:

- szukamy najsłabszej pozycji pozwalającej podzielić frazę na dwie podfrazy rzeczownikowe;
- jeśli różnica pomnięcy nasłabszym miejscem podziału a tym dzielącym na dwie frazy rzeczownikowe jest mniejsza od ustalonego progu to preferujemy podział na dwie frazy rzeczownikowe.

Analiza kontrastywna terminologii

Cel

Na podstawie porównania wyników ekstrakcji terminologii dla dwóch korpusów mają być wskazane frazy:

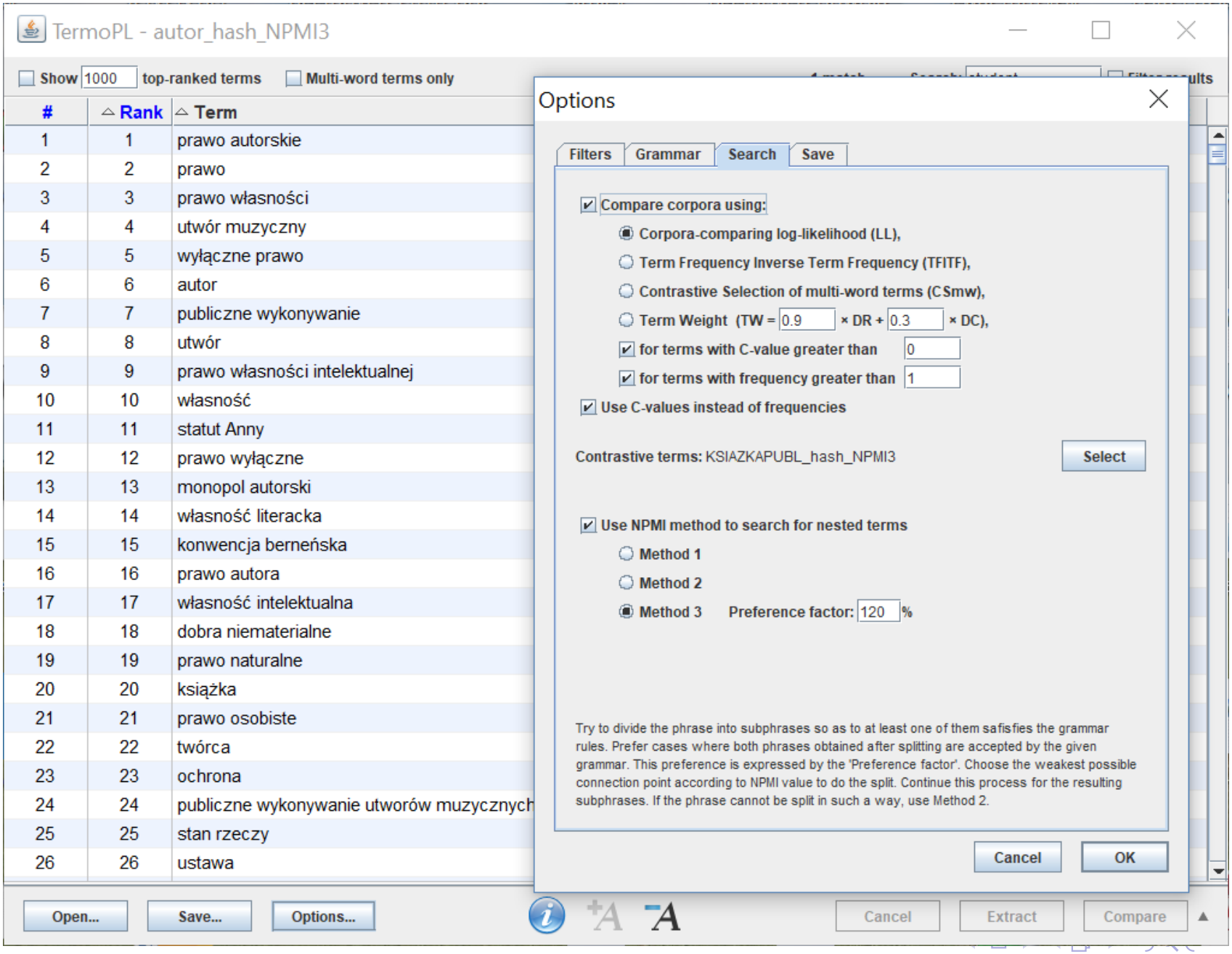
- bardziej specyficzne dla innej dziedziny (porównanie z terminologią wydobytą z innego korpusu dziedzinowego)
- terminy ogólne np.” *własny sposób*, *lewa strona*, *trudne zadanie* (porównanie z korpusem języka ogólnego).

Metody porównań

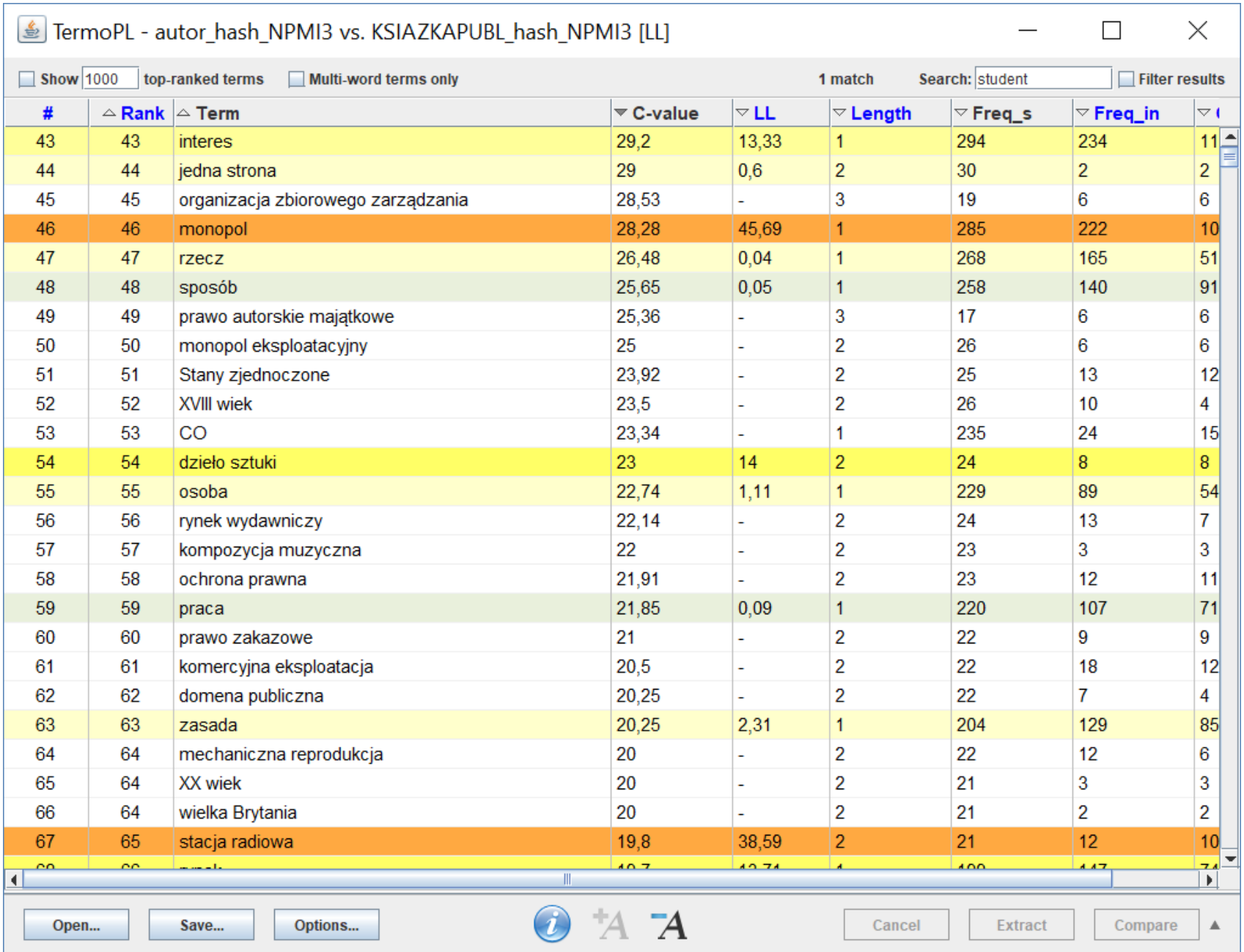
Zaimplementowane metody wykorzystują:

- Log-Likelihood (LL – logarytm wiarygodności): na ile różni się częstość konkretnego terminu w dwóch porównywanych korpusach;
- Term Frequency Inverse Term Frequency (TFITF): łączy częstość występowania w korpusie dziedzinowym z odwrotną częstością występowania w korpusie ogólnym (liczoną jako stosunek wielkości korpusu do częstości badanego terminu);
- Contrastive Selection of Multi-Word Terms(CSmw): dla terminów wielowyrazowych, uwzględnia zarówno częstość występowania pełnych terminów, ale też częstość występowania słów stanowiących element główny badanej frazy.

Porównanie dwóch list terminów



Porównanie dwóch list terminów



#	Rank	Term	C-value	LL	Length	Freq_s	Freq_in	
43	43	interes	29,2	13,33	1	294	234	11
44	44	jedna strona	29	0,6	2	30	2	2
45	45	organizacja zbiorowego zarządzania	28,53	-	3	19	6	6
46	46	monopol	28,28	45,69	1	285	222	10
47	47	rzecz	26,48	0,04	1	268	165	51
48	48	sposób	25,65	0,05	1	258	140	91
49	49	prawo autorskie majątkowe	25,36	-	3	17	6	6
50	50	monopol eksploatacyjny	25	-	2	26	6	6
51	51	Stany zjednoczone	23,92	-	2	25	13	12
52	52	XVIII wiek	23,5	-	2	26	10	4
53	53	CO	23,34	-	1	235	24	15
54	54	dzieło sztuki	23	14	2	24	8	8
55	55	osoba	22,74	1,11	1	229	89	54
56	56	rynek wydawniczy	22,14	-	2	24	13	7
57	57	kompozycja muzyczna	22	-	2	23	3	3
58	58	ochrona prawna	21,91	-	2	23	12	11
59	59	praca	21,85	0,09	1	220	107	71
60	60	prawo zakazowe	21	-	2	22	9	9
61	61	komercyjna eksploatacja	20,5	-	2	22	18	12
62	62	domena publiczna	20,25	-	2	22	7	4
63	63	zasada	20,25	2,31	1	204	129	85
64	64	mechaniczna reprodukcja	20	-	2	22	12	6
65	64	XX wiek	20	-	2	21	3	3
66	64	wielka Brytania	20	-	2	21	2	2
67	65	stacja radiowa	19,8	38,59	2	21	12	10

Podsumowanie

Cel ekstrakcji terminologii:

wydobycie specyficznej terminologii z tekstów dotyczących wybranej dziedziny.

Co otrzymujemy:

- listę terminów w formie podstawowej lub w uproszczonej formie podstawowej;
- listę możemy sortować według:
  - adekwatności frazy jako terminu;
  - częstości w tekstach;
  - długości frazy
- możliwość wyboru fraz wielowyrazowyvh;
- możliwość zebrania wystąpień fraz w korpusie wraz z częstościami;
- możliwość porównań dwóch zbiorów terminów.

Program Termo.PL

Opracowany w ramach projektu Clarin.PI

- Java Runtime Environment w wersji 7 lub nowszej;
- Wymaga Morfeusza 2 do wygenerowania formy podstawowej z uproszczonej formy;
- Wymaga otagowanego i ujednoznacznionego korpusu danych w jednym z formatów:
  - NKJP;
  - XCES;
  - XML-owe wyjście z Korpusomatu;
  - zapis uproszczony: token # lemat # tag.
- na wyjściu: lista uporządkowanych terminów (w uproszczonych formach lub zrekonstruowanych formach podstawowych wraz z formami znalezionych fraz).

Przydatne adresy internetowe

- `http://clarin-pl.eu/en/uslugi/`
- `http://zil.ipipan.waw.pl/TermoPL`