

Porównanie wybranych miar semantycznego podobieństwa / powiązania słów

Filip Dratwiński 237999, Piotr Gramacki 238493

24.11.2020

Wykorzystywane miary podobieństwa semantycznego

Wektory semantyki dystrybucyjnej

- Odległość euklidesowa:

$$Sim(s1, s2) = |emb(s1) - emb(s2)|, \quad (1)$$

gdzie $emb(x)$ oznacza wektor osadzeń dla słowa x . Miara ta jest odwrotnie proporcjonalna do podobieństwa słów ($Sim(x, x) = 0$), więc przeskalowano je liniowo aby $Sim(x, x)$ przyjmowało wartość maksymalną. Wykorzystano przekształcenie:

$$Sim(s1, s2) = 1.5 - Sim(s1, s2). \quad (2)$$

- Miara cosinusowa:

$$Sim(s1, s2) = \frac{emb(s1) \cdot emb(s2)}{|emb(s1)| |emb(s2)|}, \quad (3)$$

Miara ta przyjmuje wartości z zakresu $[-1, 1]$. Aby porównać ją łatwo z miarami ze zbioru ‘Simlex999’ przekształcono ją do zakresu $[0, 10]$ korzystając z przekształcenia:

$$Sim(s1, s2) = (Sim(s1, s2) + 1) * 5. \quad (4)$$

Sieć leksykalna

- Miara Wu-Palmer:

$$Sim(s1, s2) = 2 * \frac{depth(lcs(s1, s2))}{depth(s1) + depth(s2)}, \quad (5)$$

gdzie $lcs(s1, s2)$ to *Least Common Subsumer*. Można to przetłumaczyć jako najbliższy wspólny przodek wierzchołków $s1$ i $s2$. $depth(s1)$ to głębokość drzewa taksonomicznego, na którym znajduje się wierzchołek $s1$. Miara ta przyjmuje wartości od 0 do 1, gdzie 0 oznacza, że Synset’y są całkowicie niepodobne do siebie, a 1 oznacza, że jest to dokładnie ten sam Synset.

- Miara Leacock-Chodorow:

$$Sim(s1, s2) = -\log \frac{length(s1, s2)}{2D}, \quad (6)$$

gdzie $length(s1, s2)$ oznacza najkrótszą ścieżkę pomiędzy wierzchołkami $s1$ i $s2$. D oznacza największą głębokość drzewa taksonomicznego. Zakres miary zależy od współczynnika D . Minimalną wartością jest zawsze 0, co oznacza, że Synset’y są całkowicie niepodobne do siebie (ścieżka pomiędzy nimi jest maksymalna). W przypadku sieci *PLWN* największą głębokością jest 32. Jeżeli wierzchołki znajdują się koło siebie to wartość miary wyniesie 1.8062. Gdy porównywane jest ze sobą ten sam Synset to przyjmowana jest stała wartość 2.0, jako maksymalna.

Wyniki oceny siły podobieństwa dla par słów na zbiorze danych SimLex999

Miara cosinusowa została przeskalowana do zakresu $[0, 10]$ więc możliwe było porównanie jej bezpośrednio z wartościami miar w zbiorze *SimLex999*. Obliczono średnią wartość bezwzględną z różnicy między miarą cosinusową a miarami z *SimLex999*. Wyniki pokazano poniżej.

	cosine_metric
similarity	4.61717
relatedness	1.74491

Korelacja pomiędzy zdefiniowanymi miarami, a wartościami similarity i relatedness ze zbioru SimLex999

	euclidean_metric	cosine_metric	wu_palmer	leacock_chodorow
similarity	0.370449	0.39001	0.358705	0.436089
relatedness	0.599525	0.627853	0.37754	0.355072

Przykładowe wygenerowane listy k podobnych słów według wszystkich miar podobieństwa

W poniższym badaniu przyjęto $k=10$.

Słowo	10 najbardziej podobnych słów - Euklides	10 najbardziej podobnych słów - Cosinus	10 najbardziej podobnych słów - Wu-Palmer	10 najbardziej podobnych słów - Leacock-Chodorow
drewno	tektura - 0.5727 cegła - 0.5506 drzewo - 0.5360 tkanina - 0.5193 węgiel - 0.5134 szkło - 0.4827 mech - 0.4590 mebel - 0.4562 żelazo - 0.4520 zboże - 0.4517	tektura - 7.8504 cegła - 7.7465 drzewo - 7.6769 tkanina - 7.5954 węgiel - 7.5667 szkło - 7.4126 mech - 7.2910 mebel - 7.2764 żelazo - 7.2541 zboże - 7.2525	drzewo - 1.0 kłoda - 0.8571 miód - 0.8333 słoma - 0.8333 materiał - 0.8 papier - 0.7273 perła - 0.7273 aluminium - 0.7273 bawełna - 0.7143 cegła - 0.6667	drzewo - 2.0 materiał - 1.5051 miód - 1.5051 kłoda - 1.5051 słoma - 1.5051 papier - 1.3291 cegła - 1.3291 korzeń - 1.3291 perła - 1.3291 aluminium - 1.3291
długopis	kubek - 0.6217 pudełko - 0.5718 torba - 0.5071 szuflada - 0.4813 torebka - 0.4722 tektura - 0.4566 słoik - 0.4556 cukierek - 0.4546 kanapka - 0.4493 butelka - 0.4445	kubek - 8.0713 pudełko - 7.8460 torba - 7.5353 szuflada - 7.4058 torebka - 7.3590 tektura - 7.2781 słoik - 7.2730 cukierek - 7.2678 kanapka - 7.2400 butelka - 7.2148	papier - 0.8571 tektura - 0.8 ser - 0.7143 cygaro - 0.7143 papieros - 0.7143 masło - 0.7143 fajka - 0.7143 pieniądz - 0.6667 drzwi - 0.6667 tkanina - 0.6667	papier - 1.5051 tektura - 1.3291 materiał - 1.3291 pieniądz - 1.2041 garnek - 1.2041 obiektyw - 1.2041 aluminium - 1.2041 ser - 1.2041 talerz - 1.2041 filizanka - 1.2041
ekran	kino - 0.4908 kamera - 0.4760 film - 0.4365 komputer - 0.4359 telewizja - 0.4285 obraz - 0.3816 aktor - 0.3769 kanapa - 0.3745 okno - 0.3685 mysz - 0.3654	kino - 7.4539 kamera - 7.3787 film - 7.1723 komputer - 7.1692 telewizja - 7.1297 obraz - 6.8727 aktor - 6.8467 kanapa - 6.8333 okno - 6.7995 mysz - 6.7817	komputer - 0.8571 maszyna - 0.8571 aparat - 0.8571 telefon - 0.8571 winda - 0.8 silnik - 0.8 klatka - 0.8 dzwon - 0.7692 dekoracja - 0.7692 pręt - 0.7692	komputer - 1.5051 maszyna - 1.5051 aparat - 1.5051 telefon - 1.5051 winda - 1.3291 dzwon - 1.3291 silnik - 1.3291 dekoracja - 1.3291 klatka - 1.3291 pręt - 1.3291
głupi	głupawy - 0.6743 mądry - 0.6419 głupol - 0.6357 tępy - 0.5048 zabawny - 0.4901 nudny - 0.4883 dziecinny - 0.4784 motłoch - 0.4586 facet - 0.4552 smutny - 0.4540	głupawy - 8.2955 mądry - 8.1593 głupol - 8.1325 tępy - 7.5240 zabawny - 7.4503 nudny - 7.4414 dziecinny - 7.3908 motłoch - 7.2888 facet - 7.2707 smutny - 7.2649	człowiek - 0.8889 pisarz - 0.8 mąż - 0.8 dziewczynka - 0.8 mężczyzna - 0.8 facet - 0.8 robotnik - 0.8 opiekun - 0.8 szef - 0.8 kandydat - 0.8	człowiek - 1.8062 pisarz - 1.5051 mąż - 1.5051 dziewczynka - 1.5051 mężczyzna - 1.5051 facet - 1.5051 zwierzę - 1.5051 robotnik - 1.5051 opiekun - 1.5051 szef - 1.5051

Słowo	10 najbardziej podobnych słów - Euklides	10 najbardziej podobnych słów - Cosinus	10 najbardziej podobnych słów - Wu-Palmer	10 najbardziej podobnych słów - Leacock-Chodorow
kapusta	ziemniak - 0.7577 fasola - 0.6846 sałatka - 0.6564 zupa - 0.6180 masło - 0.5980 cebulka - 0.5853 ryż - 0.5777 jabłko - 0.5637 indyk - 0.5613 ciasto - 0.5500	ziemniak - 8.6226 fasola - 8.3377 sałatka - 8.2210 zupa - 8.0552 masło - 7.9662 cebulka - 7.9085 ryż - 7.8735 jabłko - 7.8083 indyk - 7.7972 ciasto - 7.7438	fiołek - 0.8235 len - 0.8235 ziemniak - 0.7778 zboże - 0.75 tytoń - 0.75 kwiat - 0.75 fasola - 0.75 trawa - 0.75 pszenica - 0.7059 ryż - 0.7059	fiołek - 1.3291 len - 1.3291 zboże - 1.2041 tytoń - 1.2041 kwiat - 1.2041 fasola - 1.2041 trawa - 1.2041 ziemniak - 1.2041 ser - 1.1072 posiłek - 1.1072
kot	zwierzak - 0.7099 pies - 0.6793 królik - 0.5902 szczur - 0.5815 zwierzę - 0.5492 lis - 0.5081 kura - 0.5005 koza - 0.4727 mysz - 0.4520 ptak - 0.4517	zwierzak - 8.4393 pies - 8.3160 królik - 7.9309 szczur - 7.8911 zwierzę - 7.7402 lis - 7.5401 kura - 7.5027 koza - 7.3614 mysz - 7.2540 ptak - 7.2525	lew - 0.9412 norka - 0.875 pies - 0.875 lis - 0.8235 jastrząb - 0.75 ogar - 0.7 zwierzę - 0.6667 ogier - 0.5714 królik - 0.5714 cielak - 0.5714	lew - 1.8062 norka - 1.5051 pies - 1.5051 lis - 1.3291 potwór - 1.2041 zboże - 1.2041 głupi - 1.2041 ogier - 1.2041 jastrząb - 1.2041 królik - 1.2041
książka	esej - 0.7001 opowiadanie - 0.6895 czasopismo - 0.5978 literatura - 0.5460 biografia - 0.5348 autor - 0.5239 opowieść - 0.5202 film - 0.5184 pisarz - 0.5121 artykuł - 0.4682	esej - 8.4004 opowiadanie - 8.3579 czasopismo - 7.9651 literatura - 7.7248 biografia - 7.6708 autor - 7.6183 opowieść - 7.6000 film - 7.5910 pisarz - 7.5601 artykuł - 7.3385	artykuł - 0.8333 esej - 0.8333 rozprawa - 0.8333 tekst - 0.8 piosenka - 0.7692 uwaga - 0.7273 hymn - 0.7143 film - 0.7143 duma - 0.7143 drzwi - 0.6667	artykuł - 1.5051 tekst - 1.5051 esej - 1.5051 rozprawa - 1.5051 uwaga - 1.3291 piosenka - 1.3291 drzwi - 1.2041 kalendarz - 1.2041 woda - 1.2041 piers - 1.2041
rower	samochód - 0.5374 pojazd - 0.4814 taksówka - 0.4497 autobus - 0.4469 wycieczka - 0.4365 powóz - 0.4309 wóz - 0.4215 torba - 0.4163 barierka - 0.4006 kurtka - 0.3964	samochód - 7.6833 pojazd - 7.4062 taksówka - 7.2423 autobus - 7.2274 wycieczka - 7.1723 powóz - 7.1427 wóz - 7.0920 torba - 7.0638 barierka - 6.9783 kurtka - 6.9551	powóz - 0.875 motor - 0.875 pojazd - 0.7692 wóz - 0.7059 samochód - 0.7059 autobus - 0.6667 rakieta - 0.5714 łódź - 0.5714 balon - 0.5 samolot - 0.4706	powóz - 1.5051 motor - 1.5051 pojazd - 1.3291 wóz - 1.2041 samochód - 1.2041 rakieta - 1.1072 maszyna - 1.1072 autobus - 1.1072 wieża - 1.028 rzadki - 1.028
tapeta	karnisz - 0.5376 sufit - 0.5314 dywan - 0.5082 tkanina - 0.4971 podłoga - 0.4878 mebel - 0.4752 tektura - 0.4451 paznokcie - 0.4193 ściana - 0.4193 ubranie - 0.4176	karnisz - 7.6844 sufit - 7.6547 dywan - 7.5406 tkanina - 7.4855 podłoga - 7.4388 mebel - 7.3743 tektura - 7.2178 paznokcie - 7.0804 ściana - 7.0802 ubranie - 7.0711	wierzch - 0.8 strona - 0.6667 ką - 0.4444 stopa - 0.4444 brzuch - 0.4444 piers - 0.4444 kolano - 0.4444 okoliczność - 0.4 serce - 0.4 jelito - 0.4	materiał - 1.5051 wierzch - 1.5051 papier - 1.3291 aluminium - 1.3291 strona - 1.3291 perła - 1.3291 drzewo - 1.2041 diament - 1.2041 tkanina - 1.2041 miód - 1.2041

Słowo	10 najbardziej podobnych słów - Euklides	10 najbardziej podobnych słów - Cosinus	10 najbardziej podobnych słów - Wu-Palmer	10 najbardziej podobnych słów - Leacock-Chodorow
wycieczka	drzwi - 0.5014 kanapa - 0.4533 podłoga - 0.4424 pudełko - 0.4355 torebka - 0.4293 klakson - 0.4285 miska - 0.4230 torba - 0.4225 szuflada - 0.4200 kabina - 0.4166	drzwi - 7.5072 kanapa - 7.2613 podłoga - 7.2036 pudełko - 7.1672 torebka - 7.1341 klakson - 7.1295 miska - 7.1004 torba - 7.0975 szuflada - 7.0837 kabina - 7.0658	podróż - 0.8571 stan - 0.8333 kąpiel - 0.8 lot - 0.8 emocja - 0.7692 spokój - 0.7692 ignorancja - 0.7692 postawa - 0.7692 obecność - 0.7692 nastrój - 0.7692	podróż - 1.5051 stan - 1.5051 emocja - 1.3291 spokój - 1.3291 okoliczność - 1.3291 ignorancja - 1.3291 kąpiel - 1.3291 postawa - 1.3291 lot - 1.3291 obecność - 1.3291
zamek	wieża - 0.4756 książę - 0.4606 kaplica - 0.4266 wzgórze - 0.4233 brama - 0.4125 kościół - 0.3784 most - 0.3773 zajazd - 0.3718 król - 0.3554 ogród - 0.3551	wieża - 7.3766 książę - 7.2989 kaplica - 7.1197 wzgórze - 7.1017 brama - 7.0431 kościół - 6.8552 most - 6.8489 zajazd - 6.8179 król - 6.7249 ogród - 6.7228	szałas - 0.8571 dom - 0.8571 budka - 0.8571 kino - 0.8571 wieża - 0.8571 lotnisko - 0.8333 szkoła - 0.8 kaplica - 0.8 kościół - 0.8 domek - 0.8	szałas - 1.5051 dom - 1.5051 konstrukcja - 1.5051 lotnisko - 1.5051 budka - 1.5051 kino - 1.5051 wieża - 1.5051 kręgielnia - 1.5051 brama - 1.3291 szkoła - 1.3291
żniwa	zboże - 0.6834 pszenica - 0.5718 ziemniak - 0.4644 ziarno - 0.4434 zima - 0.4347 bydło - 0.4344 chleb - 0.3854 jagnię - 0.3845 obfitość - 0.3809 len - 0.3769	zboże - 8.3329 pszenica - 7.8459 ziemniak - 7.3188 ziarno - 7.2091 zima - 7.1628 bydło - 7.1613 chleb - 6.8941 jagnię - 6.8890 obfitość - 6.8692 len - 6.8468	miesiąc - 0.5714 tydzień - 0.5714 data - 0.5714 dzień - 0.5714 sezon - 0.5714 rok - 0.5714 wiek - 0.5714 dekada - 0.5714 sierpień - 0.5 lato - 0.5	miesiąc - 1.3291 tydzień - 1.3291 data - 1.3291 dzień - 1.3291 sezon - 1.3291 rok - 1.3291 wiek - 1.3291 dekada - 1.3291 rzadki - 1.2041 mądry - 1.2041