

Rekomendacje artykułów opisujących produkty w serwisach e-commerce

Łukasz Dragan

Informatyka spec. Metody sztucznej inteligencji, MiNI PW

31.10.2017

Plan prezentacji

- 1 Cel pracy
- 2 Opis problemu
- 3 Systemy rekomendacji
- 4 Techniki przetwarzania języka naturalnego
 - Bag-of-words
 - TF-IDF
 - Modelowanie tematów
 - Latent semantic analysis
 - Latent Dirichlet allocation
 - Osadzanie słów w przestrzeni wektorowej
 - Wstępne przetwarzanie danych
- 5 Metody ewaluacji
- 6 Testy
- 7 Podsumowanie

Czy metody semantycznej analizy tekstu mogą być alternatywą dla dotychczas używanej przez *Allegro* metody generowania rekomendacji artykułów tekstowych?

Dlaczego?

wypisać swoją motywację

Praca

Praca: 43 360 ofert pracy

🔍 Szukaj

Określ odległość  0 km

Filtruj wyniki




Miejsce pracy

- ☐ cała Polska (42350)
- ☐ dolnośląskie (4499)
- ☐ kujawsko-pomorskie (1587)
- ☐ lubelskie (1133)
- ☐ lubuskie (1058)
- ☐ łódzkie (2787)
- ☐ małopolskie (4030)
- ☐ mazowieckie (10245)
- ☐ opolskie (1071)
- ☐ pomorskie (2881)
- ☐ podkarpackie (1191)
- ☐ podlaskie (911)
- ☐ świętokrzyskie (886)
- ☐ śląskie (3775)
- ☐ warmińsko-mazurskie (950)
- ☐ wielkopolskie (3765)
- ☐ zachodniopomorskie (1581)
- ☐ zagranica (1010)

Zastosuj

Oferty rekomendowane dla Ciebie

Na podstawie Twojej aktywności wybraliśmy oferty dopasowane do Twoich oczekiwań

	Stażysta IT&ccw w Roche Diagnostics Roche Diagnostics 📍 Warszawa, mazowieckie	★ 2017-10-21
	Stażysta w dziale Analiz Innogy Polska S.A. 📄 o firmie 📍 Warszawa, mazowieckie	★ 2017-10-20
	Stażysta w Zespole Modeli Ryzyka ALIOR BANK 📄 o firmie 📍 Warszawa, mazowieckie	★ 2017-10-15
...	Stażysta w programie NN Pro Nationale-Nederlanden Usługi Finansowe SA 📍 Warszawa, mazowieckie	★ 2017-10-02
...	Stażysta w Spółce IT PGE Systemy S.A.	★

Wniosek / Złóż CV

Łukasz Dragan

FILMY
SERIALE
GRY
REPERTUARIUM KIN
PROGRAM TV
MAGAZYN
WFF
MÓJ FILMWEB

Mój Filmweb

Aktywność znajomych, których obserwujesz:

Martha oglądała film **Persona**
13 godzin temu

Persona (1966)
gatunki: Dramat, Psychologiczny
reżyser: Ingmar Bergman
obsada: Bibi Andersson, Liv Ullmann
Aktorka traci głos podczas przedstawienia teatralnego. Okazuje się, że demonstruje ona w ten sposób bezradność wobec otaczającego ją świata.

Kubaj to - skomentuj

kuba1004 ocenił rolę aktorki Marion Cotillard w filmie To tylko koniec świata na 8

Mój asystent

wszystkie w guście w guście newsy recenzje

Filmweb znalazł film w Twoim guście: **Ślódmy kontynent**
83%

Filmweb znalazł film w Twoim guście: **Podróż na Księżyc**
84%

Filmweb znalazł film w Twoim guście: **Głowa w mur**
81%

Filmweb znalazł film w Twoim guście: **Boska Florence**
75%

Filmweb znalazł film w Twoim guście: **Zagraj to jeszcze raz, Sam**
80%

allegro

czego szukasz?

wszystkie działy



koszyk jest pusty

Elektronika

Moda
i urodaDom
i zdrowie

Dziecko

Kultura
i rozrywkaSport
i wypoczynek

Motoryzacja

Kolekcje
i sztuka

Firma

Strefa
okazjiAllegro - Poradniki - Dom i zdrowie - [Jaka farba dla alergika?](#)

Jaka farba dla alergika?



autor: Ewelina Wojtunik, data publikacji: 23-04-2015

Za chwilę wiosna, a wraz z nią potrzeba porządków i odświeżenia ścian. Jak co roku będziemy sprzątać, wietrzyć i wymieniać zimę z kątów mieszkania. Zaraz po tym zaczną się pierwsze remonty.

**Ewelina Wojtunik**

Zawodowo związana z Social Media, pisała m.in. do Aktivist.pl. Prywatnie pasjonatka projektowania wnętrz, zdrowego stylu życia i roślin doniczkowych. Podróże i kuchnie świata są dla niej inspiracją. W wolnym czasie spełnia się jako mama i uczy języków.

**może Cię również
zainteresować**



wszystkim chemikalia i detergenty. znajdujące się w nich alergeny mogą być powodem problemów zdrowotnych, a także nasilać objawy nadwrażliwości takie jak łzawienie oczu, zapalenie skóry czy kaszel astmatyczny.

Szkodliwe związki lotne

W styczniu 2010 roku Unia Europejska wprowadziła normę, która reguluje zawartość szkodliwych lotnych związków organicznych tak zwanych LZO (VOCs, ang. volatile organic compounds) w trafiających do sprzedaży farbach i lakierach. Warto wiedzieć, że lotne związki lubią pozostawać aktywne pomimo wyschnięcia farby i starannego wentrowienia mieszkania. Co więcej, mogą uwalniać się ze ścian całymi latami, nasilając objawy alergiczne i pogarszając samopoczucie mieszkańców. Im mniej ich w składzie, tym lepiej dla nas.

Pamiętajmy więc, że kupowana przez nas farba powinna posiadać **atest hipoalergiczny** – najlepiej specjalny certyfikat potwierdzający bezpieczeństwo dla osób cierpiących z powodu nadwrażliwości na alergeny. Opatrzony certyfikatem farby gwarantują nawet trzydziestokrotnie niższą szkodliwość! Dlatego kupując je, zwróćmy uwagę na obecność stosownego oznaczenia na opakowaniu, dzięki czemu zyskamy pewność, że nie zawierają żadnych substancji uczulających i pozostają w pełni bezpieczne dla zdrowia naszego i naszych bliskich. Oprócz farb szkodliwe związki lotne mogą pojawiać się także w klejach, wykładzinach dywanowych, **tapetach ściennych**, a nawet materiałach do wykończenia podłóg.



EKO ŚNIEŻKA BIAŁA FARBA
EMULSJA 10L
HIPOALERGICZNA
kup teraz 43,97 zł



EKO ŚNIEŻKA BIAŁA FARBA
EMULSJA 10L
HIPOALERGICZNA
kup teraz 46,90 zł



ŚNIEŻKA EKO Farba Emulsja
Hipoalergiczna 10l
kup teraz 50,10 zł



Emulsja Hipo
Śnieżka EKO
kup teraz 9,9l



Wnętrzarski hit – ściany ombre

Ombre stało się hitem w wizażu i modzie już kilka sezonów temu! Chętnie rozjaśniamy końcówki włosów, cieniujemy kolory na paznokciach, a także nosimy ubrania w przenikających się tonach. Czy tę technikę mo...



Jak przemaalować ciemną ścianę?

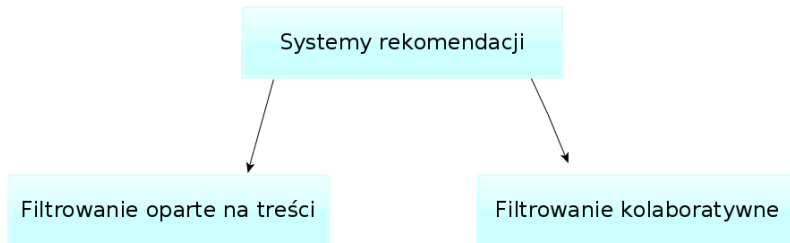
Planujesz remont mieszkania, a jednym z jego etapów będzie przemaalowanie ciemnej ściany? A może po prostu znudził ci się niemiódny już kolor? Jeśli zastanawiasz się, jak prawidłowo przemaalować ścianę, spraw...



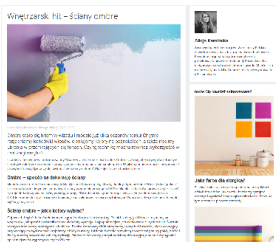


„Elasticsearch is a distributed, JSON-based search and analytics engine designed for horizontal scalability, maximum reliability, and easy management.”

W ujęciu ogólnym systemy wyszukiwania mają na celu sugerowanie tego, co użytkownik chciałby otrzymać. Natomiast systemy rekomendacji mają sugerować przedmioty potrzebne użytkownikowi nawet, jeżeli potrzeby te nie zostały bezpośrednio wyrażone.



Zarys podejścia



1.168785 0.060346 0.502299 0.291747 -0.365562
 0.257444 -0.329024 0.758068 0.139132 -0.066573
 1.171894 0.076840 -0.002970 -0.360585 -0.144586
 0.105688 -0.528267 0.377016 0.220084 -0.132361
 -0.232592 0.338373 0.106514 0.096009 -0.068181
 -0.698880 0.040483 -0.820396 0.110031 -0.493751
 -0.339397 0.278281 -0.000135 -0.121884 0.107060
 -0.001215 -0.348834 0.399166 0.391983 0.197091
 -0.837996 -0.081890 -0.534775 0.589362 0.278594
 -0.724953 0.143085 -0.308889 -0.051467 0.133181
 0.110936 -0.159592 -0.338680 0.324832 -0.227569
 -0.257161 -0.403050 -0.355761 0.111366 0.127810
 -0.045948 0.256404 -0.413172 -0.565309 0.252026
 -0.178040 0.353451 -0.043467 0.437229 -0.364093
 0.620433 0.491961 -0.044899 0.075592 -0.035806
 0.552777 0.539595 -0.307839 -0.488252 0.494307
 -0.506171 0.517397 0.010668 -0.247984 0.322363

Zarys podejścia

Wnętrze zask. Hiti – ściany omlire



Copyright: Shutterstock.com

Całkowicie czysta i biała ściana w kolorze białym. Wnętrze zask. Hiti – ściany omlire. Wnętrze zask. Hiti – ściany omlire. Wnętrze zask. Hiti – ściany omlire.

Całkowicie czysta i biała ściana w kolorze białym. Wnętrze zask. Hiti – ściany omlire. Wnętrze zask. Hiti – ściany omlire. Wnętrze zask. Hiti – ściany omlire.

Całkowicie czysta i biała ściana w kolorze białym.

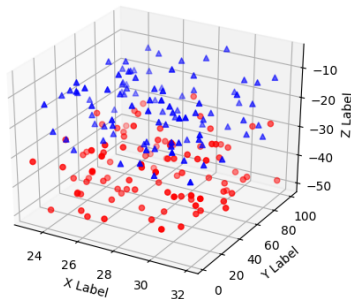
Całkowicie czysta i biała ściana w kolorze białym. Wnętrze zask. Hiti – ściany omlire. Wnętrze zask. Hiti – ściany omlire. Wnętrze zask. Hiti – ściany omlire.

Całkowicie czysta i biała ściana w kolorze białym.

Całkowicie czysta i biała ściana w kolorze białym. Wnętrze zask. Hiti – ściany omlire. Wnętrze zask. Hiti – ściany omlire. Wnętrze zask. Hiti – ściany omlire.



1.168785	0.060346	0.502299	0.291747	-0.365562
0.257444	-0.329024	0.758068	0.139132	-0.066573
1.171894	0.076840	-0.002970	-0.360585	-0.144586
0.105688	-0.528267	0.377016	0.220084	-0.132361
-0.232592	0.338373	0.106514	0.096009	-0.068181
-0.698880	0.040483	-0.820396	0.110031	-0.493751
-0.339397	0.278281	-0.000135	-0.121884	0.107060
-0.001215	-0.348834	0.399166	0.391983	0.197091
-0.837996	-0.081890	-0.534775	0.589362	0.278594
-0.724953	0.143085	-0.308889	-0.051467	0.133181
0.110936	-0.159592	-0.338680	0.324832	-0.227569
-0.257161	-0.403050	-0.355761	0.111366	0.127810
-0.045948	0.256404	-0.413172	-0.565309	0.252026
-0.178040	0.353451	-0.043467	0.437229	-0.364093
0.620433	0.491961	-0.044899	0.075592	-0.035806
0.552777	0.539595	-0.307839	-0.488252	0.494307
-0.506171	0.517397	0.010668	-0.247984	0.322363



Bag-of-words

(1) John likes to watch movies. Mary likes movies too.

(2) John also likes to watch football games.

Bag-of-words

(1) John likes to watch movies. Mary likes movies too.

(2) John also likes to watch football games.

```
[  
  "John",  
  "likes",  
  "to",  
  "watch",  
  "movies",  
  "Mary",  
  "too",  
  "also",  
  "football",  
  "games"  
]
```

Bag-of-words

(1) John likes to watch movies. Mary likes movies too.

(2) John also likes to watch football games.

```
[  
  "John",  
  "likes",  
  "to",  
  "watch",  
  "movies",  
  "Mary",  
  "too",  
  "also",  
  "football",  
  "games"  
]
```

(1) [1, 2, 1, 1, 2, 1, 1, 0, 0, 0]

(2) [1, 1, 1, 1, 0, 0, 0, 1, 1, 1]

Zalety:

- prostota

Zalety:

- prostota

Wady:

- duża wymiarowość wektorów

Zalety:

- prostota

Wady:

- duża wymiarowość wektorów
- traktowanie każdego słowa z jednakową wagą

Wartość *TF-IDF* słowa w_i w dokumencie d_j :

$$tfidf_{ij} = tf_{ij} * idf_i, \quad tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}, \quad idf_i = \log \frac{|D|}{|d : w_i \in d|} \quad (1)$$

- tf_{ij} : liczba wystąpień słowa w_i w dokumencie d_j podzielona przez liczbę słów dokumentu d_j ,
- idf_i : liczba dokumentów w korpusie podzielona przez liczbę dokumentów zawierających przynajmniej jedno wystąpienie słowa w_i .

Przykład analogiczny do bow

Zalety:

- nadal prostota

Zalety:

- nadal prostota
- ważenie słów w zależności od częstości występowania

Zalety:

- nadal prostota
- ważenie słów w zależności od częstości występowania

Wady:

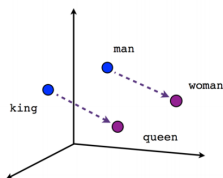
- nadal duża wymiarowość wektorów

Distributional hypothesis — „słowa występujące w tym samym kontekście niosą ze sobą podobne znaczenie.”

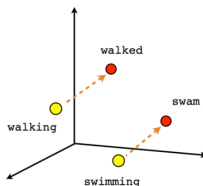
- Osadzanie słów w przestrzeni wektorowej
- Uczenie nienadzorowane
- Niska wymiarowość wektorów
- Reprezentacja słów wraz z zależnościami pomiędzy nimi

Word embeddings

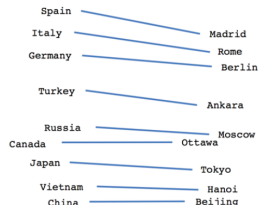
- Osadzanie słów w przestrzeni wektorowej
- Uczenie nienadzorowane
- Niska wymiarowość wektorów
- Reprezentacja słów wraz z zależnościami pomiędzy nimi



Male-Female



Verb tense



Country-Capital

Wektorowa postać dokumentu

Odległość między dokumentami

precision, recall, f-measure, ndcg

- 20000 artykułów tekstowych w formacie *JSON*
- język polski
- słowa specyficzne dla różnych branż
- struktura artykułu:
 - treść: tytuł, nagłówek, tekst
 - metadane: id, kategoria, słowa kluczowe

❶ Oczyszczanie tekstu ze znaczników

- 1 Oczyszczanie tekstu ze znaczników
- 2 Usunięcie słów stopu

a, aby, ach, acz, aczkolwiek, aj, albo, ale, ależ, ani, aż, bardziej, bardzo, bo, bowiem, by, byli, bynajmniej, być, był, była, było, były, będzie, będą, cali, cała, cały, ci, cię, ciebie, co, cokolwiek, coś, czasami, czasem, czemu, czy, czyli, daleko, dla, dlaczego, dlatego, do, dobrze, dokąd, dość, dużo, dwa, dwaj, dwie, dwoje, dziś, dzisiaj, gdy, gdyby, gdyż, gdzie, gdziekolwiek, gdzieś, go, i...

- ❶ Oczyszczanie tekstu ze znaczników
- ❷ Usunięcie słów stopu

- ❶ Oczyszczanie tekstu ze znaczników
- ❷ Usunięcie słów stopu
- ❸ Zamiana na małe litery

- ❶ Oczyszczanie tekstu ze znaczników
- ❷ Usunięcie słów stopu
- ❸ Zamiana na małe litery
- ❹ Tokenizacja i lematyzacja

Preprocessing - przykład

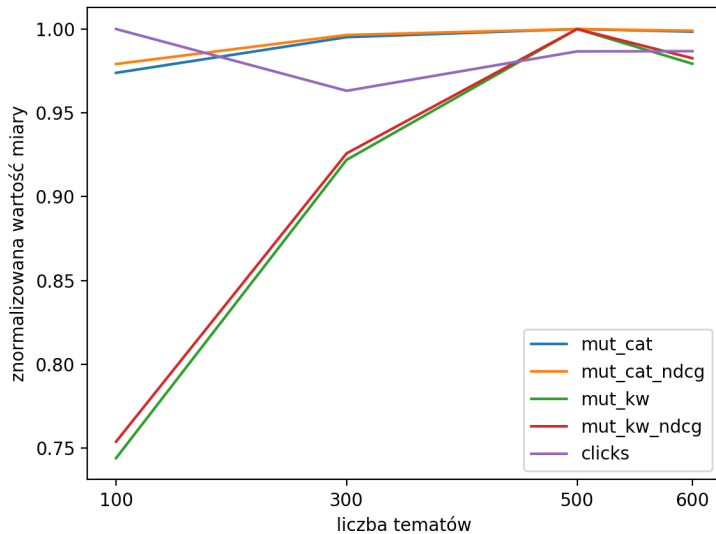
Każda mama cieszy się, gdy jej maluszek z apetytem zjada przygotowany przez nią posiłek.



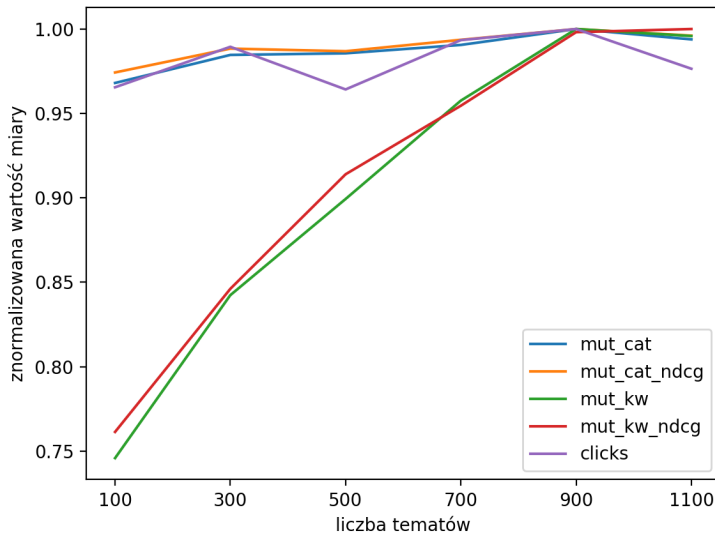
"mama",
"cieszyć",
"maluszek",
"apetyt",
"zjadać",
"przygotować",
"posiłek"

- 1 *clicks* — ocena na podstawie historycznej aktywności użytkowników mierzona na podstawie liczby kliknięć w odnośniki.
- 2 *mut_cat[_ndcg]* — relewantność wyszukiwanych artykułów liczona na podstawie liczby wspólnych kategorii z artykułem bazowym. Stosuję dwa warianty: średnia relewantność wyszukiwanych artykułów oraz miara *nDCG*.
- 3 *mut_kw[_ndcg]* — relewantność wyszukiwanych artykułów liczona na podstawie liczby wspólnych słów kluczowych z artykułem bazowym. Również stosuję dwa warianty: średnia relewantność wyszukiwanych artykułów oraz miara *nDCG*.
- 4 *users* — ocena na podstawie eksperckiej oceny użytkowników. W badaniu wykorzystałem 5 użytkowników operujących każdy na tym samym zbiorze par testowych. Pary zostały wygenerowane (zgodnie z wcześniejszym opisem metody) na podstawie 50 artykułów bazowych wylosowanych spośród wszystkich artykułów udostępnionych mi przez *Allegro*.

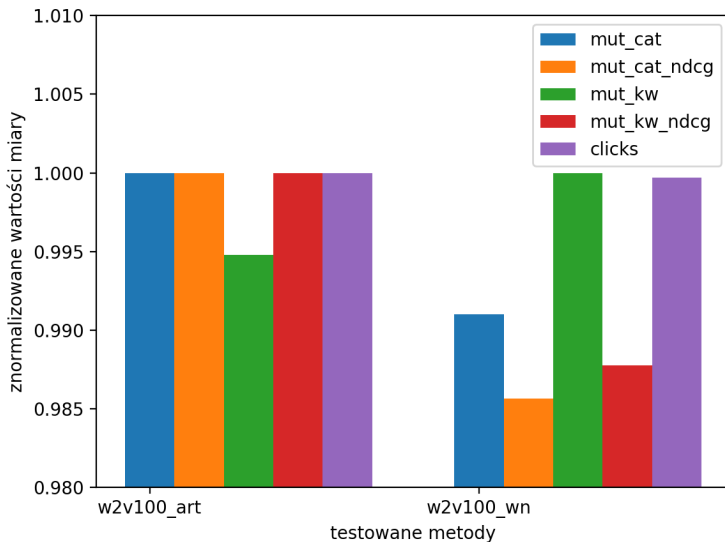
LSI w zależności od liczby tematów



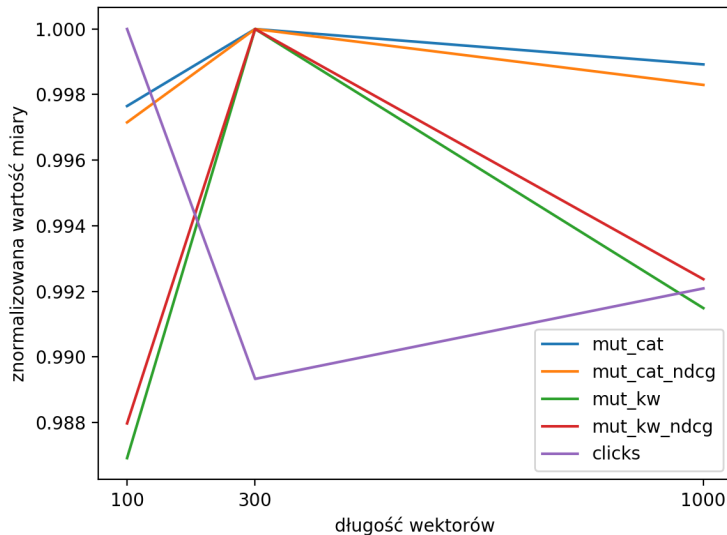
LDA w zależności od liczby tematów



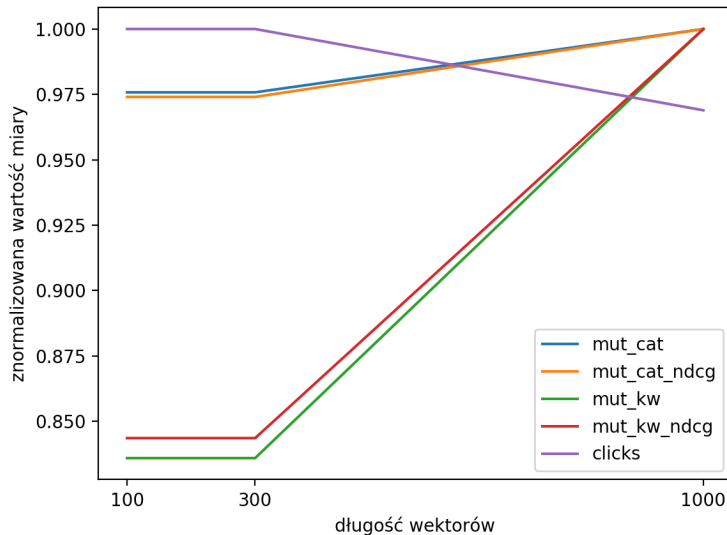
Word2vec w zależności od korpusu



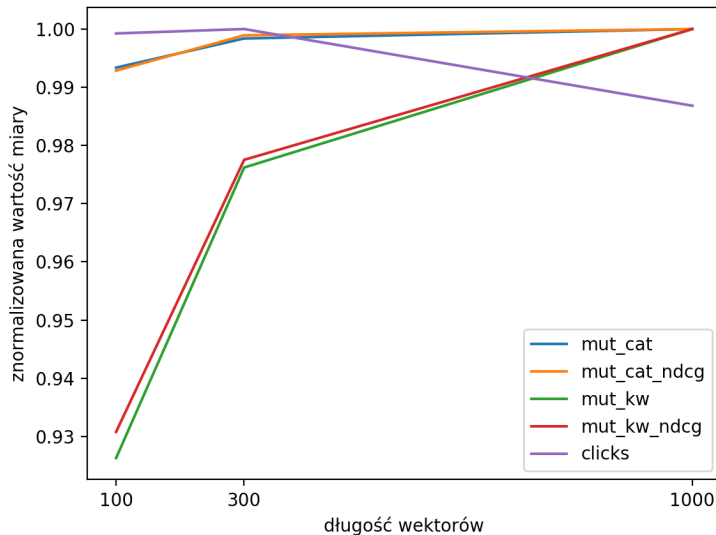
Word2vec w zależności od długości wektorów



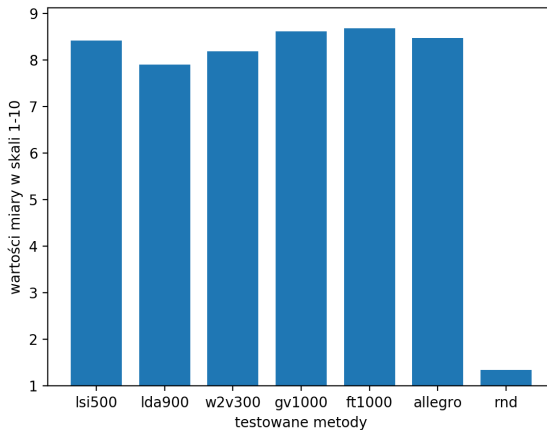
GloVe w zależności od długości wektorów



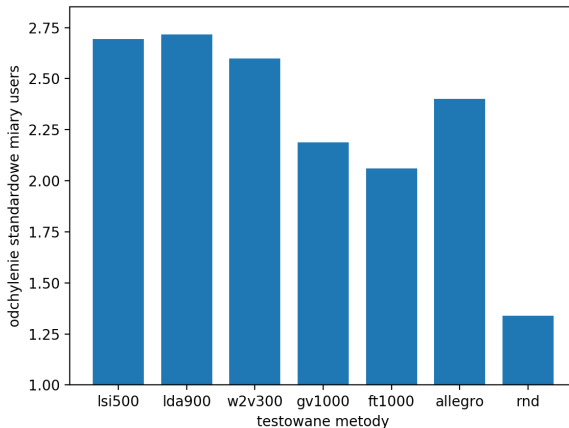
FastText w zależności od długości wektorów



Wyniki ewaluacji eksperckiej dla wybranych metod



Porównanie odchyleń standardowych ocen eksperckich dla wybranych metod



- Brak istotnych statystycznie różnic między wynikami wszystkich metod

- Brak istotnych statystycznie różnic między wynikami wszystkich metod
- Im dłuższe wektory *word embeddings* tym lepsze rezultaty

- Brak istotnych statystycznie różnic między wynikami wszystkich metod
- Im dłuższe wektory *word embeddings* tym lepsze rezultaty
- Większa liczba tematów nie implikuje lepszych rezultatów

- Brak istotnych statystycznie różnic między wynikami wszystkich metod
- Im dłuższe wektory *word embeddings* tym lepsze rezultaty
- Większa liczba tematów nie implikuje lepszych rezultatów
- ...

Kiernki dalszych badań

testowane metody nie odbiegają jakością od dotychczasowej.
python elasticsearch nlp trudno jest zmierzyc efekty

content...