

Rekomendacje artykułów opisujących produkty w serwisach e-commerce

Łukasz Dragan

Informatyka spec. Metody sztucznej inteligencji, MiNI PW

31.10.2017

Praca magisterska

2017-10-30



Rekomendacje artykułów opisujących produkty w serwisach e-commerce

Łukasz Dragan

Informatyka spec. Metody sztucznej inteligencji, MiNI PW

31.10.2017

1 Opis problemu

2 Systemy rekomendacji

3 Techniki przetwarzania języka naturalnego

4 Analiza danych

5 Metody ewaluacji

6 Wyniki testów

7 Podsumowanie

8 Wybrane źródła

- 1 Opis problemu
- 2 Systemy rekomendacji
- 3 Techniki przetwarzania języka naturalnego
- 4 Analiza danych
- 5 Metody ewaluacji
- 6 Wyniki testów
- 7 Podsumowanie
- 8 Wybrane źródła

2017-10-30

└ Plan prezentacji

- po kolej co zamierzam powiedzieć

Opis problemu

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

2017-10-30

Praca magisterska

- └ Opis problemu
- └ Cel pracy

Cel pracy

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

- nie znałem wcześniej tej tematyki
- nie znałem Pani Promotor
- realne biznesowe zastosowanie informatyki
- zaznajomiłem się z dziedziną, której wcześniej nie znałem

Praca: 43 360 ofert pracy

szukaj, miasto kluczowe, stanowisko, firma, miejscowość lub województwo, Szukaj, Okres odległość 0 km

Filtruj wyniki

Miejsce pracy

- cała Polska (42350)
- dolnośląskie (4499)
- kujawsko-pomorskie (1587)
- lubelskie (1133)
- lubuskie (1058)
- łódzkie (1787)
- mazowieckie (4030)
- mazowieckie (10245)
- opolskie (1071)
- pomorskie (2881)
- podkarpackie (1191)
- podlaskie (911)
- świętokrzyskie (886)
- śląskie (3770)
- warmińsko-mazurskie (950)
- wielkopolskie (3765)
- zachodniopomorskie (1581)
- zagłębiowskie (1010)

Zastosuj

Oferty rekomendowane dla Ciebie

Na podstawie Twojej aktywności wybrałyśmy oferty dopasowane do Twoich oczekowań

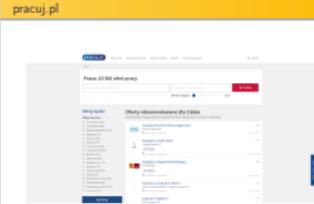
- Stażysta IT&ccW w Roche Diagnostics**
Roche, Warszawa, mazowieckie, 2017-10-21
- Stażysta w dziale Analiz**
Innogy Polska S.A., Warszawa, mazowieckie, 2017-10-20
- Stażysta w Zespole Modeli Ryzyka**
ALIOR BANK, Warszawa, mazowieckie, 2017-10-15
- Stażysta w programie NN Pro**
Nationale-Nederlanden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

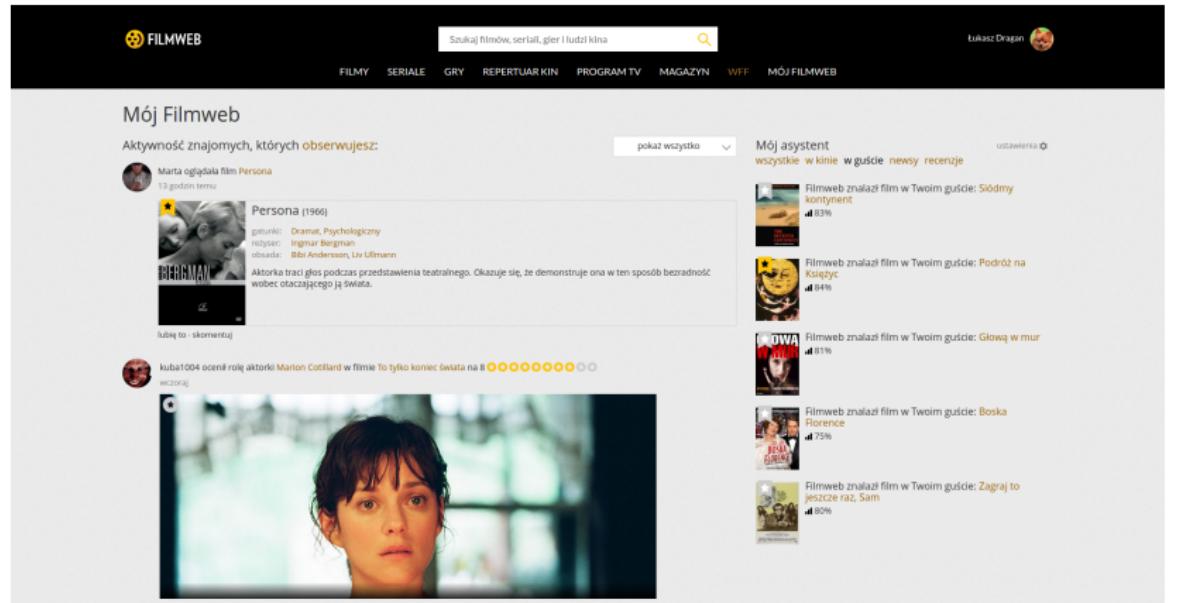
Praca magisterska
└ Opis problemu

└ pracuj.pl

2017-10-30

Każdy szanujący się serwis zawiera rekomendacje





Praca magisterska

- └ Opis problemu

└ filmweb.pl

Pr
L



Allegro - 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 wymiatać zimę z kątów mieszkania. Zaraz po tym zaczną się pierwsze remonty.

Ewelina Wojtuni

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

może Cię również zainteresować



Łukasz Dragan

Praca magisterska

Praca magisterska

- └ Opis problemu

└ allegro.pl

- w Allegro prócz głóœnej funkcjonalnoœci
 - artykuły opisujące produkty
 - zawiera listę artykułów podobnych, którą moœna uznaœ za rekommendacje



czego szukasz?

wszystkie działy

SZUKAJ



Moje Allegro

wszystkim chemikaliom i detergentom. 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 lakach. Warto wiedzieć, że lotne związki lubią pozostawać aktywne pomimo wyschnięcia farby i staranego wywielenienia mieszkania. Wiele, 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. Opatrzone certyfikatem farby gwarantują nawet trzydziestokrotnie niższą szkodliwość! Dlatego kupując je, zwrócmy 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, **tapetachściennych**, 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 Hipc Śnieżka EKO kup teraz 9,91 zł
---	---	---	--



Łukasz Dragan

Praca magisterska

Wnętrzarski hit – ściany ombre

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

Jak przemalować ciemną ścianę?

Planujesz remont mieszkania, a jednym z jego etapów będzie przemalowanie ciemnej ściany? A może po prostu znudził Ci się niemodny już kolor? Jeśli zastanawiasz się, jak prawidłowo przemalować ścianę, spraw...



Praca magisterska

Opis problemu

allegro.pl cd

2017-10-30

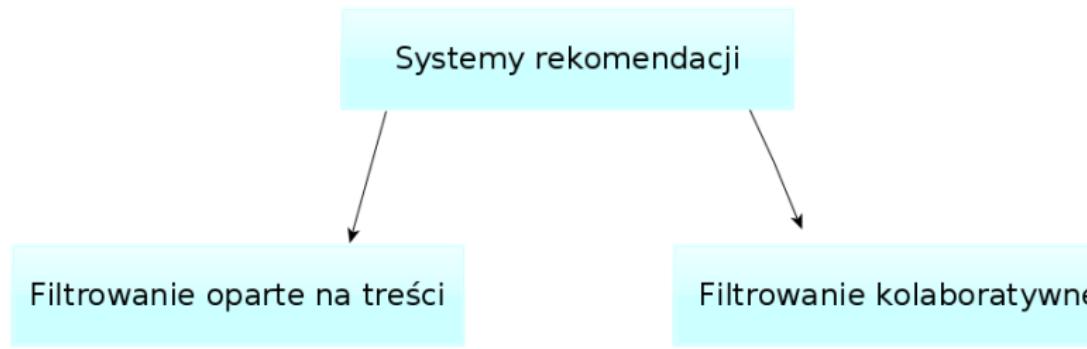


„Elasticsearch is a distributed, JSON-based search and analytics engine designed for horizontal scalability, maximum reliability, and easy management.”

- system rekomendacyjny zbudowany jest na elasticu
- elastic to silnik wyszukiwania
- zapytania opierają się o słowa kluczowe dołączone do artykułów jako metadane
- elstic jest szeroko wykorzystywany, ale nie dokonuje semantycznej analizy tekstu

Systemy rekomendacji

Systemy wyszukiwania mają na celu sugerowanie tego, co może się wydać użytkownikowi interesujące.



2017-10-30

Praca magisterska
└ Systemy rekomendacji
 └ Systemy rekomenadacji

```
graph TD; A[Praca magisterska] --> B[Systemy rekomendacji]; B --> C[2017-10-30]; B --> D[Filtrowanie oparte na treści]; B --> E[Filtrowanie kolaboratywne]
```

The slide structure is shown in a hierarchical tree. The main title "Praca magisterska" branches into "Systemy rekomendacji". This title is dated "2017-10-30". "Systemy rekomendacji" further branches into "Filtrowanie oparte na treści" and "Filtrowanie kolaboratywne". To the right of the slide content, there is a small diagram showing the overall structure of the presentation slide.



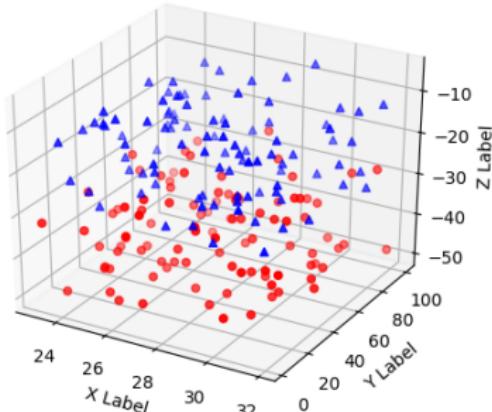
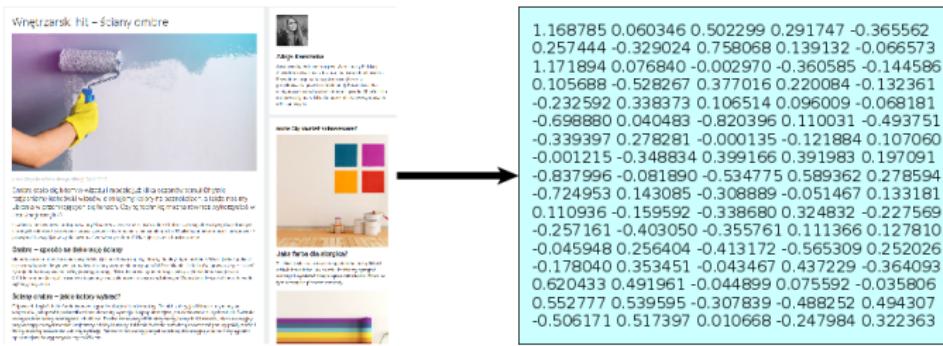
Techniki przetwarzania języka naturalnego

2017-10-30

Praca magisterska
└ Techniki przetwarzania języka naturalnego

Techniki przetwarzania języka naturalnego

Zarys podejścia



Praca magisterska

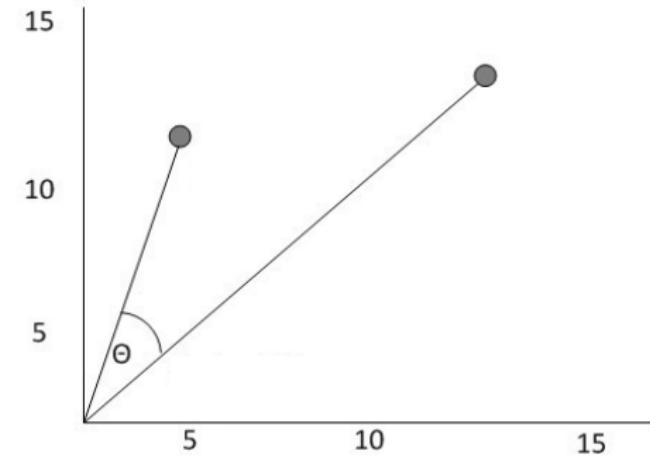
└ Techniki przetwarzania języka naturalnego

└ Zarys podejścia

- Staramy się dokonać reprezentacji dokumentu w postaci wektora
- Po to, aby wykorzystać zależności między wektorami, np. odległość
- różne reprezentacje mają różną jakość



Dystans między wektorami



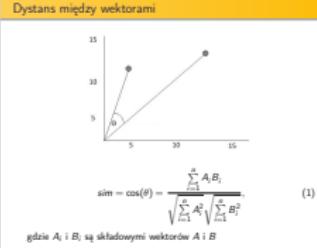
$$sim = \cos(\theta) = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}, \quad (1)$$

gdzie A_i i B_i są składowymi wektorów A i B

Praca magisterska
└ Techniki przetwarzania języka naturalnego

└ Dystans między wektorami

2017-10-30



$$sim = \cos(\theta) = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (1)$$

gdzie A_i i B_i są składowymi wektorów A i B

- odległości można mierzyć albo jako odległość euklidesową
- albo bardziej popularny sposób to kosinus kąta pomiędzy wektorami

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]

Praca magisterska

└ Techniki przetwarzania języka naturalnego

└ Bag-of-words

2017-10-30

- korpus to zbiór dokumentów, na których operujemy
- słownik to lista unikalnych słów
- najprostsze podejście — worek słów
- dokument jako wektor z liczbą wystąpień i-tego słowa na itym miejscu
- .
- Wadą jest traktowanie każdego słowa z jednakową wagą
- bardzo długie wektory
- wektory niemalże ortogonalne

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	2	1	1	2	1	1	0	0	0
G1	1	2	1	1	2	1	1	0	0
G2	1	1	1	1	0	0	0	1	1

GL [1, 2, 1, 1, 2, 1, 1, 0, 0, 0]
G2 [1, 1, 1, 1, 0, 0, 0, 1, 1, 1]

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

$$tfidf_{ij} = tf_{ij} * idf_i, \quad tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}, \quad idf_i = \log \frac{|D|}{|d : w_i \in d|} \quad (2)$$

- 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 .

Praca magisterska

└ Techniki przetwarzania języka naturalnego

└ TF – term frequency, IDF – inverse document frequency

2017-10-30

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

$$tfidf_{ij} = tf_{ij} * idf_i, \quad tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}, \quad idf_i = \log \frac{|D|}{|d : w_i \in d|} \quad (2)$$

- 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 .

Zalety:

- prostota

Wady:

- duża wymiarowość wektorów
- wektory niemalże ortogonalne

2017-10-30

Praca magisterska

└ Techniki przetwarzania języka naturalnego

└ TF-IDF

- prostota
- duża wymiarowość wektorów
- wektory niemalże ortogonalne
- zachowuje inne wady BOW
- bywa składową innych metod

- Redukcja wymiarowości macierzy wystąpień słów w dokumentach

	d_1	d_2	d_3	d_4	d_5	d_6
statek	1	0	1	0	0	0
łódź	0	1	0	0	0	0
ocean	1	1	0	0	0	0
podróż	1	0	0	1	1	0
wycieczka	0	0	0	1	0	1

- Hiperparametr: docelowa wymiarowość

2017-10-30

Latent semantic indexing (1988)

▼ Redukcja wymiarowości macierzy wystąpień słów w dokumentach

	d_1	d_2	d_3	d_4	d_5	d_6
statek	1	0	1	0	0	0
łódź	0	1	0	0	0	0
ocean	1	1	0	0	0	0
podróż	1	0	0	1	1	0
wycieczka	0	0	0	1	0	1

◆ Hiperparametr: docelowa wymiarowość

- kolejne metody opierają się na hipotezie, że słowa występujące w tym samym kontekście niosą ze sobą podobne znaczenie
- w LSI budujemy macierz wystąpień słów w dokumentach
- następnie wykonujemy redukcję liczby wierszy do liczby podanej jako hiperparametr
- macierz zachowuje powiązania między słowami

Latent semantic indexing (1988)

- Rozkład według wartości osobliwych:

$$A = U\Sigma V^T, \quad (3)$$

U i V to macierze ortogonalne

Σ to macierz diagonalna, taka, że $\Sigma = \text{diag}(\sigma_i)$, gdzie σ_i , to nieujemne wartości szczegółne macierzy A .

- $\{\text{(statek)}, \text{(łódź)}, \text{(ocean)}\} \rightarrow \{(1.3452 * \text{statek} + 0.2828 * \text{łódź}), \text{(ocean)}\}$

Praca magisterska

└ Techniki przetwarzania języka naturalnego

└ Latent semantic indexing (1988)

2017-10-30

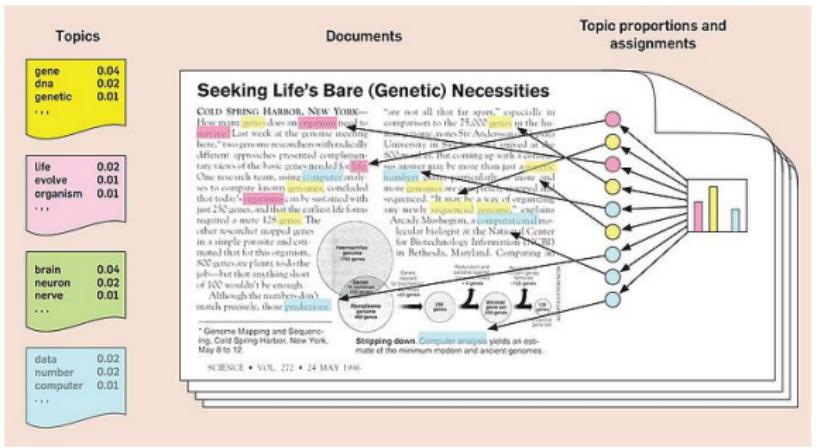
Latent semantic indexing (1988)

- Rozkład według wartości osobliwych:
$$A = U\Sigma V^T, \quad (3)$$

 U i V to macierze ortogonalne
 Σ to macierz diagonalna, taka, że $\Sigma = \text{diag}(\sigma_i)$, gdzie σ_i , to nieujemne wartości szczegółne macierzy A .
- $\{\text{(statek)}, \text{(łódź)}, \text{(ocean)}\} \rightarrow \{(1.3452 * \text{statek} + 0.2828 * \text{łódź}), \text{(ocean)}\}$

- redukcja liczby wierszy za pomocą rozkładu wg. wartości osobliwych
- można potraktować jako grupowanie słów z odpowiednimi wagami
- ostatecznie zależność między dokumentami-wektorami są lepiej oddane
- wektory są krótkie
- jednak metoda ma nistą interpretowalność

Latent Dirichlet allocation (2003)



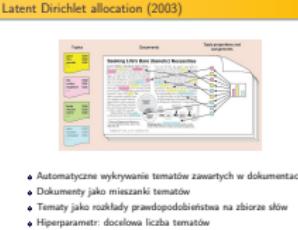
- Automatyczne wykrywanie tematów zawartych w dokumentach
- Dokumenty jako mieszanki tematów
- Tematy jako rozkładы prawdopodobieństwa na zbiorze słów
- Hiperparametr: docelowa liczba tematów

Praca magisterska

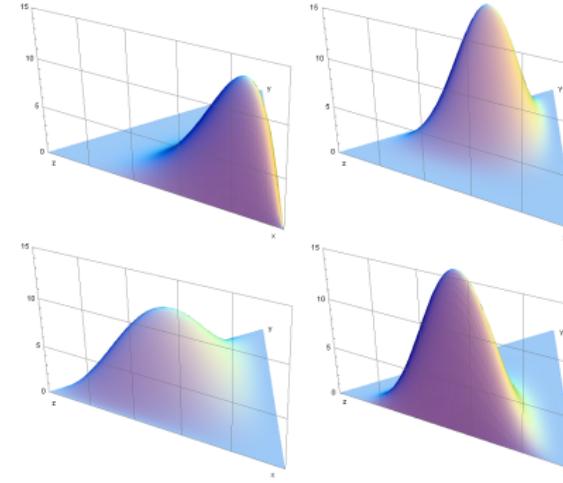
└ Techniki przetwarzania języka naturalnego

└ Latent Dirichlet allocation (2003)

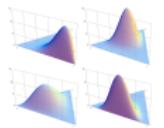
2017-10-30



- Zadaniem metody jest reprezentacja dokumentów jako mieszanki tematów
- gdzie temat to mieszanka jawnych słów wybieranych z odpowiednimi wagami
- można to traktować jako redukcję wymiarowości
- liczba tematów określana hiperparametrem



2017-10-30



- Początkowe przypisanie tematów do dokumentów i słów do tematów odbywa się zgodnie z rozkładem Dirichleta
- zażmy że mamy 3 tematy, którym odpowiadają wierzchołki trójkąta
- liczba dokumentów składających się z tylko jednego tematu jest niska
- liczba dokumentów składających się z mieszanki tematów jest wysoka
- .
- te cechy rozkładu Dirichleta przyspieszają optymalizację

Algorytm — próbkowanie Gibbsa:

- 1 Przejdz przez każdy dokument i losowo (zgodnie z rozkładem Dirichleta) przypisz każde słowo dokumentu do jednego z T tematów.
- 2 Dla każdego dokumentu d , dla każdego słowa w należącego do d , dla każdego tematu t oblicz: $p(t|d)$ oraz oblicz $p(w|t)$ Przypisz słowi w nowy temat poprzez losowanie z prawdopodobieństwem $p(t_i|d) * p(w|t)$ dla każdego tematu t_i .

2017-10-30

Praca magisterska

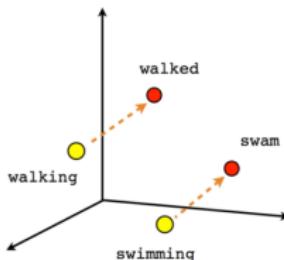
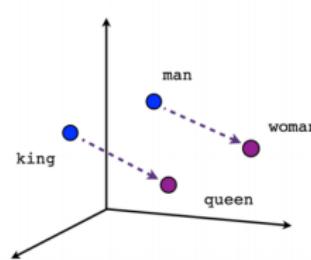
└ Techniki przetwarzania języka naturalnego

└ Latent Dirichlet allocation (2003)

- Rozkład daje pierwsze przybliżenie
- następnie iterujemy i przypisanie słów do tematów jest poprawiane
- dla każdego słowa w każdym dokumencie wyznaczany przypisywanie jest temat, który jest najbardziej prawdopodobny na podstawie rozkładu w całości korpusu
- ostatecznie uzyskujemy w miarę stabilną sytuację, w której nie następują już zmiany przypisań słów do tematów
- podobne są dokumenty o podobnej proporcji przypisanych tematów
- metoda jest interpretowalne, gdyż wiemy, jakie słowa wchodzą w skład tematów

Word embeddings

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



Male-Female

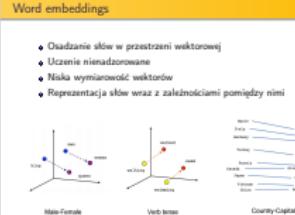
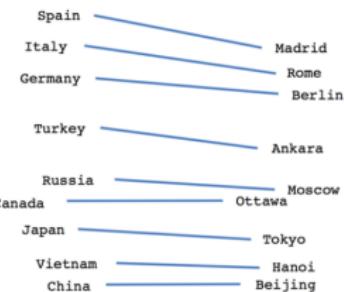
Verb tense

Country-Capital

2017-10-30

Word embeddings

- Kolejnym podejściem jest osadzanie słów w przestrzeni wektorowej, a następnie porównywanie dokumentów jako list takich wektorów
- celem metod tej grupy jest przypisanie słowom takich wektorów, żeby zachowywały one zależności zachodzące między słowami



Word2vec (2013)

Source Text

The quick brown fox jumps over the lazy dog. ➔

The quick brown fox jumps over the lazy dog. ➔

The quick brown fox jumps over the lazy dog. ➔

The quick brown fox jumps over the lazy dog. ➔

Training Samples

(the, quick)
(the, brown)

(quick, the)
(quick, brown)
(quick, fox)

(brown, the)
(brown, quick)
(brown, fox)
(brown, jumps)

(fox, quick)
(fox, brown)
(fox, jumps)
(fox, over)

<http://mccormickml.com>

Praca magisterska

└ Techniki przetwarzania języka naturalnego

└ Word2vec (2013)

2017-10-30

Word2vec (2013)

Source Text

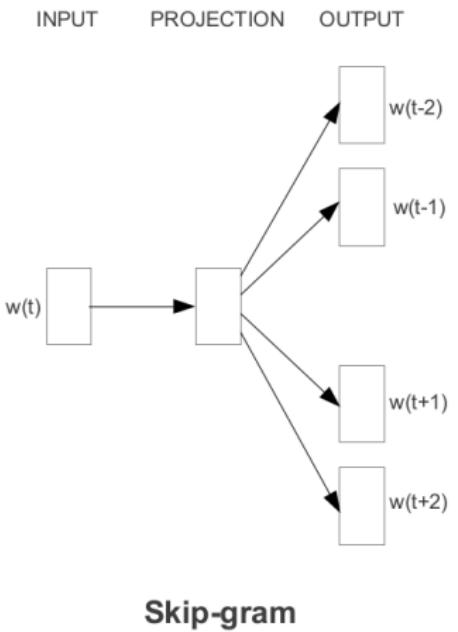
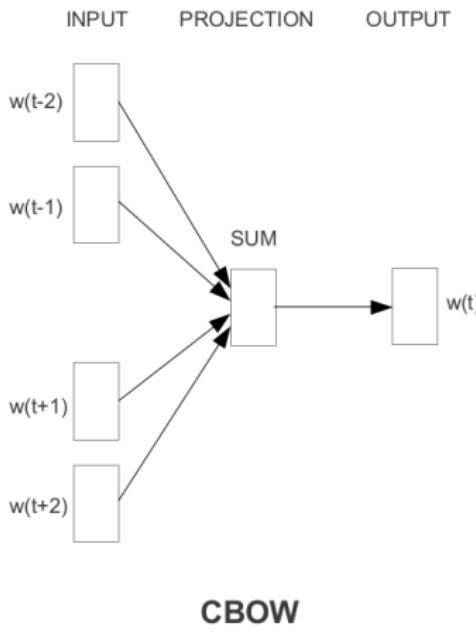
The quick brown fox jumps over the lazy dog. ➔ [the, quick], [brown, fox], [over, the, lazy, dog]

The quick brown fox jumps over the lazy dog. ➔ [quick, the], [brown, quick], [fox, over], [lazy, dog]

The quick brown fox jumps over the lazy dog. ➔ [brown, the], [brown, quick], [fox, over], [lazy, dog]

The quick brown fox jumps over the lazy dog. ➔ [the, quick], [brown, the], [fox, over], [lazy, dog]

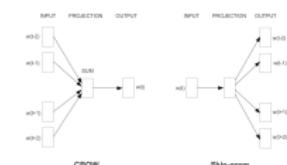
<http://mccormickml.com>



2017-10-30

Praca magisterska

- Techniki przetwarzania języka naturalnego



- wyróżniamy dwa podejścia CBOW: predykcja słowa na podstawie kontekstu
- skip-gram: predykcja kontekstu na podstawie słowa
- sieć zamiast funkcji aktywacji ma funkcję softmax, która zamienia wyjście na rozkład prawdopodobieństwa
- ALE GDZIE TE WEKTORY?**
- wektory reprezentujące słowa są efektem ubocznym nauczonej sieci
- zawarte są w wagach warstwy ukrytej i mają długość równą rozmiarowi tej warstwy
- wcześniejsze podejścia opierały się na głębszych sieciach, które działały mało wydajnie

- Rozwinięcie metody word2vec
- Rozbija słowa na n-gramy, np. *pokój* → *pok*, *oko*, *kój*
- Wektor wynikowy = wektor dla słowa + wektory jego n-gramów
- Dobre wyniki dla języków bogatych morfosyntaktycznie, np. polskiego, tureckiego, czy fińskiego.

2017-10-30

Praca magisterska

└ Techniki przetwarzania języka naturalnego

 └ FastText (2017)

- ◆ Rozwinięcie metody word2vec
- ◆ Rozbija słowa na n-gramy, np. *pokój* → *pok*, *oko*, *kój*
- ◆ Wektor wynikowy = wektor dla słowa + wektory jego n-gramów
- ◆ Dobre wyniki dla języków bogatych morfosyntaktycznie, np. polskiego, tureckiego, czy fińskiego.

- Jest modyfikacją word2vec
- powstało niedawno
- metoda rozbija słowo na n-gramy : podsłowa o określonej długości
- ostatecznie wektor słowa to suma wektoru słowa i wektorów podsłów
- sprawdza się dla języków bogatych morfosyntaktycznie

- ➊ Zgromadź współwystąpienia słów w formie globalnej macierzy X takiej, że X_{ij} : ile razy słowo w_i występuje w kontekście słowa w_j
- ➋ Zdefiniuj ograniczenie dla każdej pary słów:

$$w_i^T w_j + b_i + b_j = \log(X_{ij}), \quad (4)$$

gdzie w_i i w_j to wektory odpowiadające słowom oraz b_i i b_j to skalary.

- ➌ Dokonaj minimalizacji funkcji kosztu:

$$J = \sum_{i=1}^V \sum_{j=1}^V f(X_{ij})(w_i^T w_j + b_i + b_j - \log X_{ij})^2, \quad (5)$$

gdzie f jest funkcją ważącą, V to słownik.

Praca magisterska

└ Techniki przetwarzania języka naturalnego

└ GloVe — Global Vectors (2014)

2017-10-30

GloVe — Global Vectors (2014)

- ➊ Zgromadź współwystąpienia słów w formie globalnej macierzy X takiej, że X_{ij} : ile razy słowo w_i występuje w kontekście słowa w_j
- ➋ Zdefiniuj ograniczenie dla każdej pary słów:

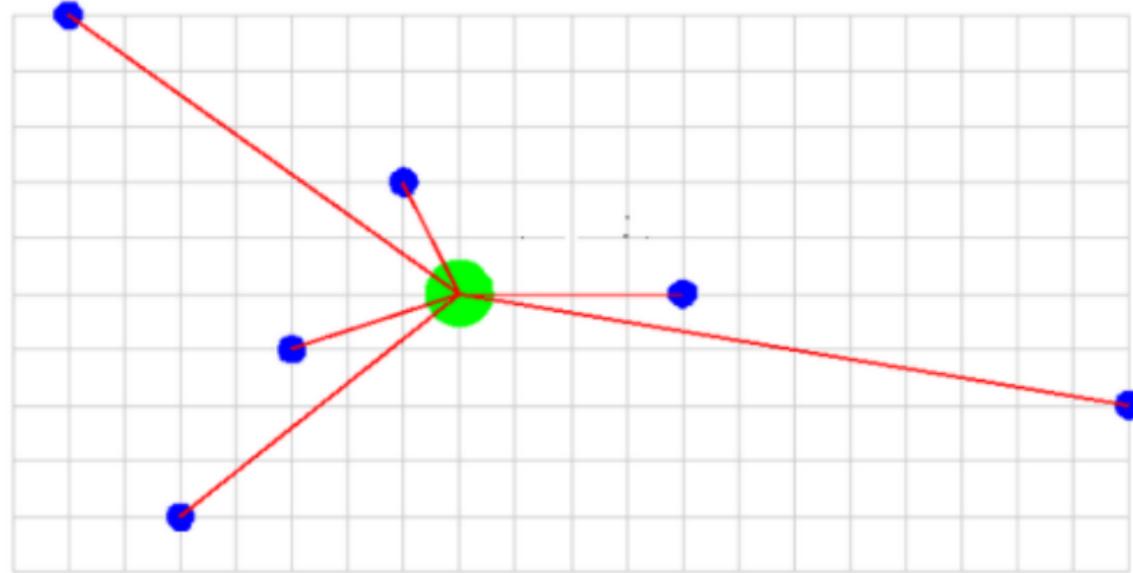
$$w_i^T w_j + b_i + b_j = \log(X_{ij}), \quad (4)$$

gdzie w_i i w_j to wektory odpowiadające słowom oraz b_i i b_j to skalary.

- ➌ Dokonaj minimalizacji funkcji kosztu:

$$J = \sum_{i=1}^V \sum_{j=1}^V f(X_{ij})(w_i^T w_j + b_i + b_j - \log X_{ij})^2, \quad (5)$$

gdzie f jest funkcją ważącą, V to słownik.

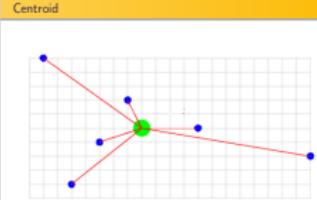


2017-10-30

Praca magisterska

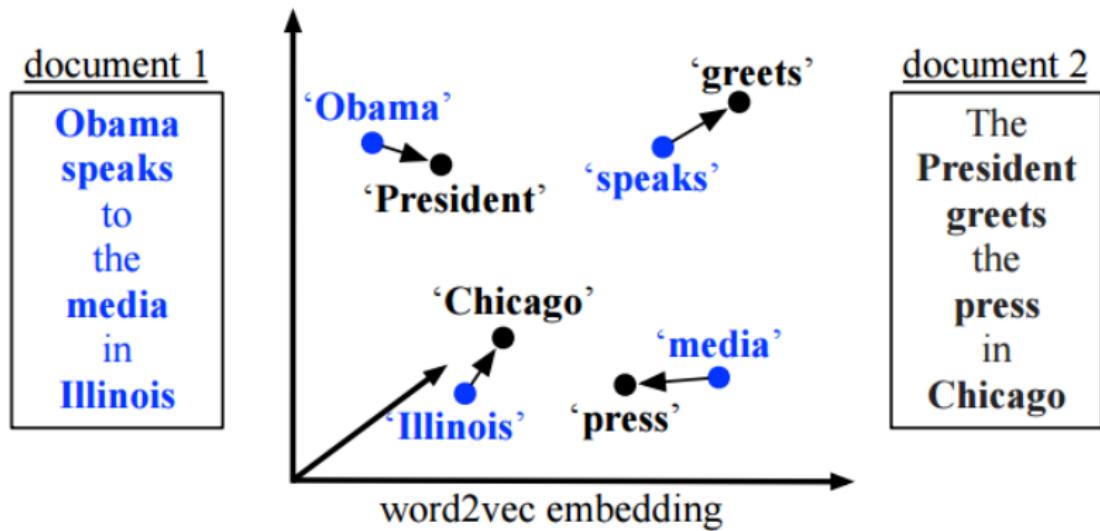
└ Techniki przetwarzania języka naturalnego

└ Centroid



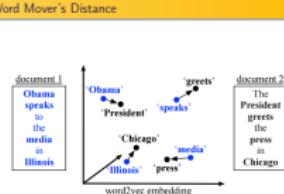
- Mamy już wektorowe reprezentacje słów wchodzących w skład dokumentu - CO DALEJ
- można obliczyć średnią tych wektorów - centroid, ale przy tym traci się część informacji

Word Mover's Distance



Praca magisterska
└ Techniki przetwarzania języka naturalnego
 └ Word Mover's Distance

2017-10-30



- Alternatywą jest Word Mover's Distance
- Dystans pomiędzy dokumentami A i B to minimalny skumulowany dystans jaki słowa dokumentu A muszą „przebyć”, aby osiągnąć słowa dokumentu B
- Metoda jest kosztowna obliczeniowo

Analiza danych

2017-10-30

Praca magisterska
└ Analiza danych

Analiza danych

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

- ➊ Oczyszczenie tekstu ze znaczników
- ➋ Usunięcie słów stopu

2017-10-30

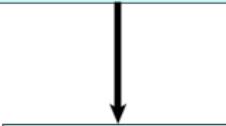
└ Wstępne przetwarzanie danych

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, dwie, dwoje, dziś, dzisiaj, gdy, gdyby, gdyż, gdzie, gdziekolwiek, gdzieś, go, i...

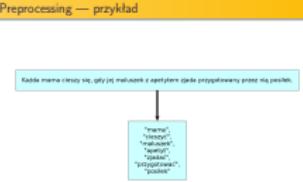
- 1 Oczyszczanie tekstu ze znaczników
- 2 Usunięcie słów stopu
- 3 Zamiana na małe litery
- 4 Tokenizacja i lematyzacja

- lematyzacja to najistotniejszy element — sprowadza słowa do postaci podstawowej
- np jest, była, będzie do „być”
- używam polskiego narzędzi Morfologik

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



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



Metody ewaluacji

Miara jakości wyszukiwania: Normalized Discounted Cumulative Gain

- Discounted Cumulative Gain:

$$DCG_p = \sum_{i=1}^p \frac{rel_i}{\log_2(i+1)}, \quad (6)$$

gdzie p to liczba elementów rankingu, i to miejsce przedmiotu w rankingu, a rel to poziom relevantności elementu.

- Normalized Discounted Cumulative Gain:

$$nDCG_p = \frac{DCG_p}{IDCG_p}. \quad (7)$$

Praca magisterska

- Metody ewaluacji

- Miara jakości wyszukiwania: Normalized Discounted Cumulative Gain

2017-10-30

Miara jakości wyszukiwania: Normalized Discounted Cumulative Gain

• Discounted Cumulative Gain:

$$DCG_p = \sum_{i=1}^p \frac{rel_i}{\log_2(i+1)}, \quad (6)$$

gdzie p to liczba elementów rankingu, i to miejsce przedmiotu w rankingu, a rel to poziom relevantności elementu.

• Normalized Discounted Cumulative Gain:

$$nDCG_p = \frac{DCG_p}{IDCG_p}. \quad (7)$$

- Liczba wspólnych kategorii
- Liczba wspólnych słów kluczowych

Praca magisterska
└ Metody ewaluacji

└ Miary oparte na metadanych artykułów

2017-10-30

- Liczba wspólnych kategorii
- Liczba wspólnych słów kluczowych

- dodatkowe dane o kliknięciach z Allegro
- ocena na podstawie historycznej aktywności użytkowników

Praca magisterska

- Metody ewaluacji

- Kliknięcia użytkowników serwisu

2017-10-30

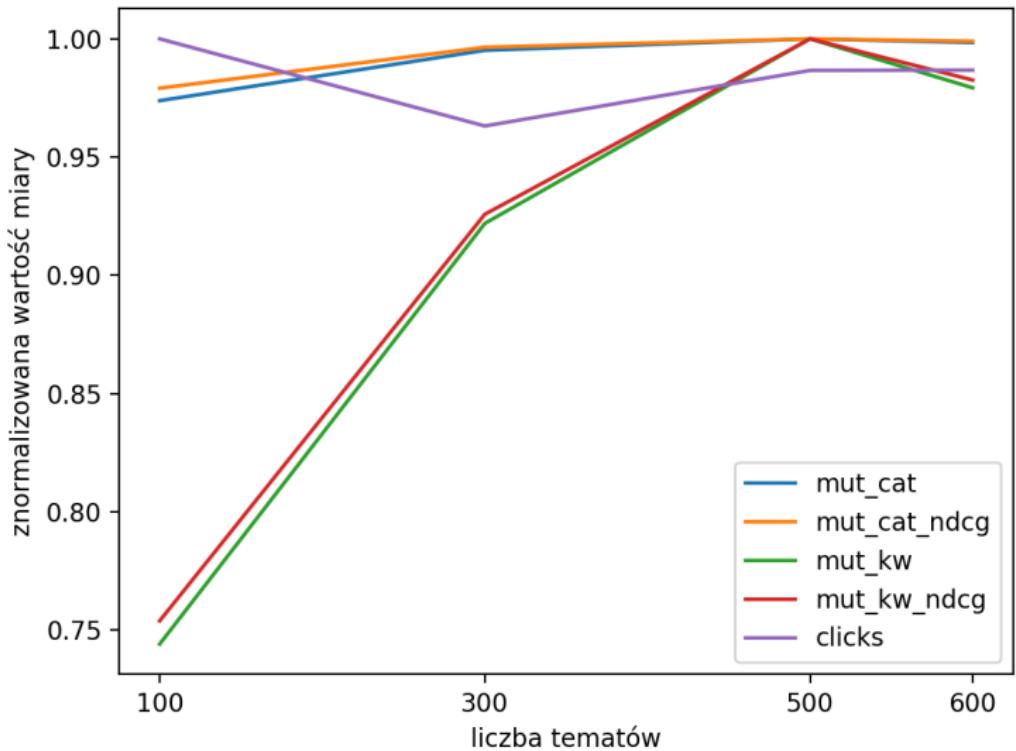
- dodatkowe dane o kliknięciach z Allegro
- ocena na podstawie historycznej aktywności użytkowników

- ekspercka ocena użytkowników
- 5 użytkowników oceniło po 300 par artykułów

2017-10-30

Wyniki testów

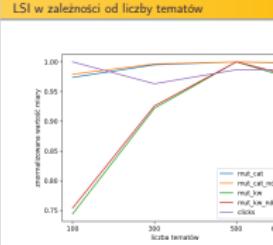
LSI w zależności od liczby tematów



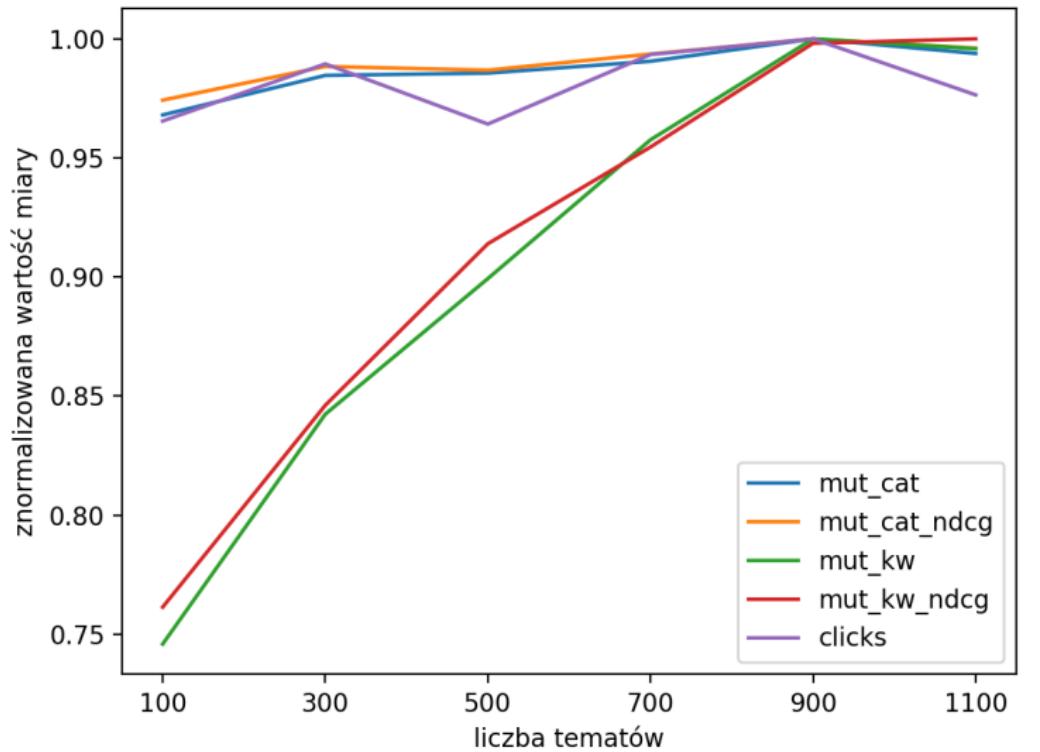
Praca magisterska
└ Wyniki testów

2017-10-30

└ LSI w zależności od liczby tematów



LDA w zależności od liczby tematów



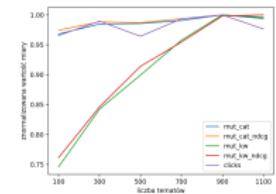
Praca magisterska

Wyniki testów

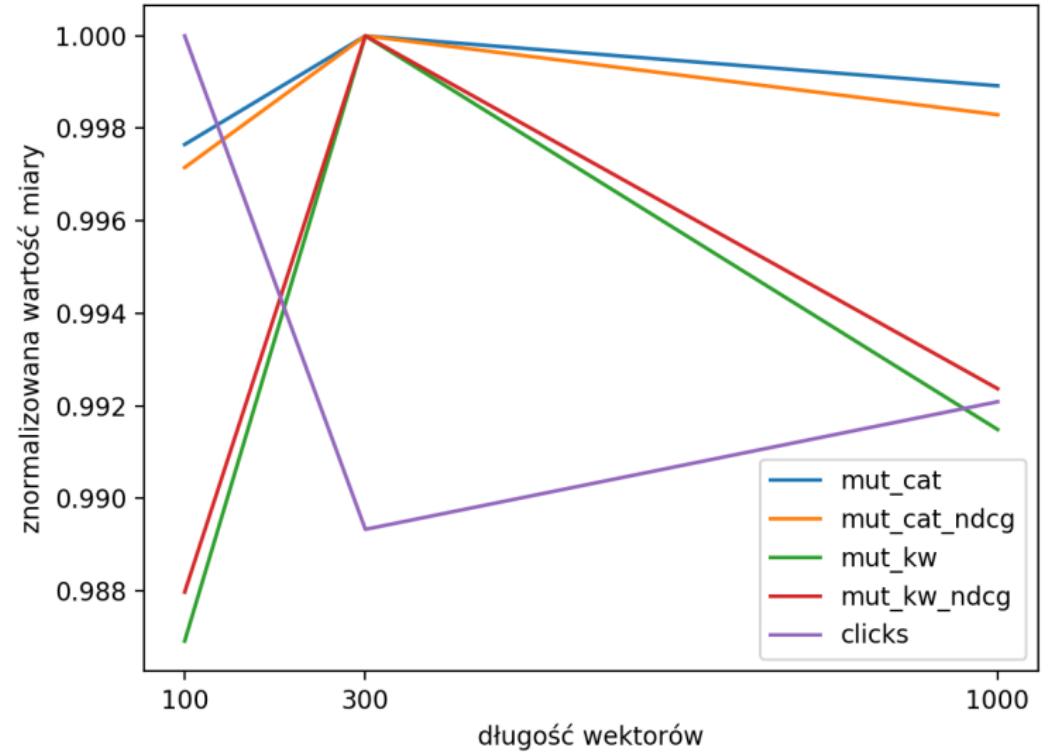
LDA w zależności od liczby tematów

2017-10-30

LDA w zależności od liczby tematów



Word2vec w zależności od długości wektorów

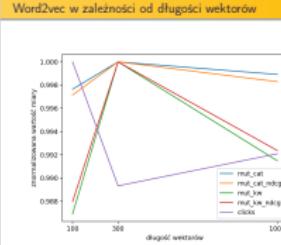


Praca magisterska

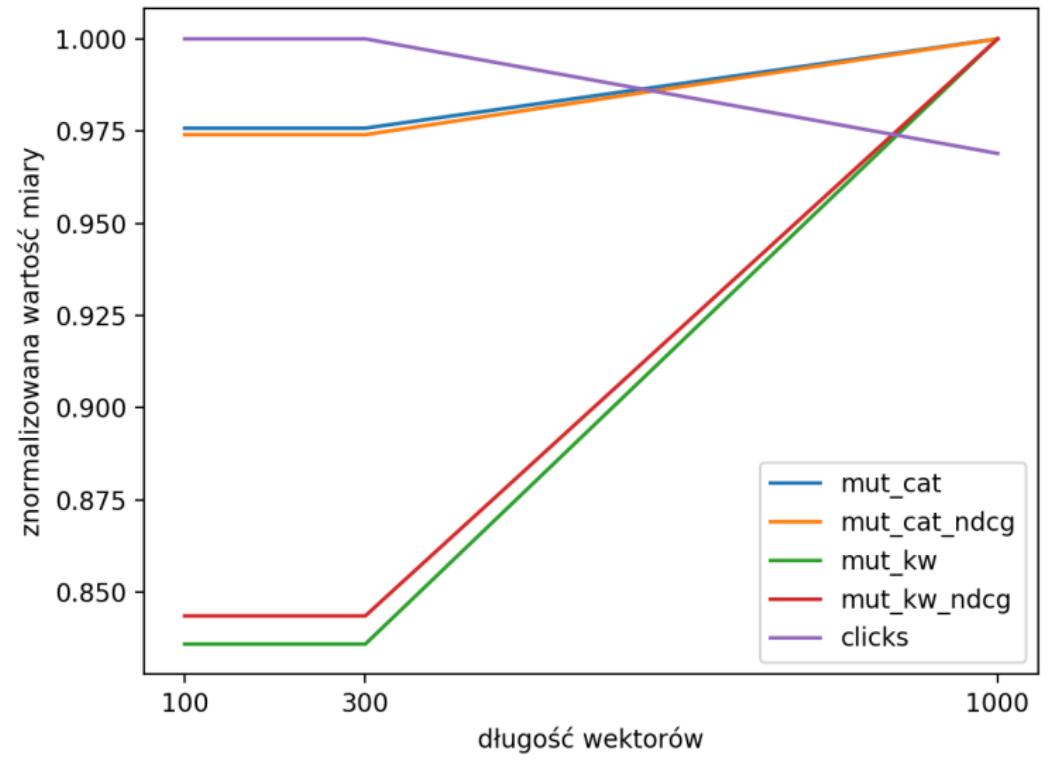
└ Wyniki testów

└ Word2vec w zależności od długości wektorów

2017-10-30



GloVe w zależności od długości wektorów



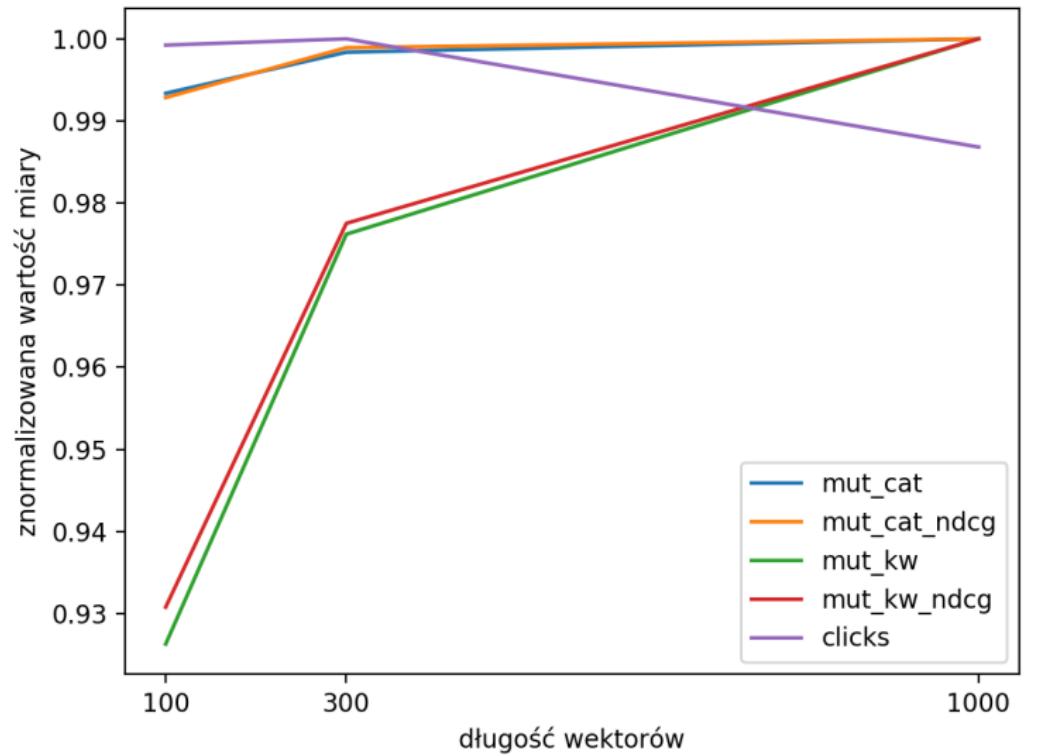
Praca magisterska
└ Wyniki testów

2017-10-30

└ GloVe w zależności od długości wektorów



FastText w zależności od długości wektorów

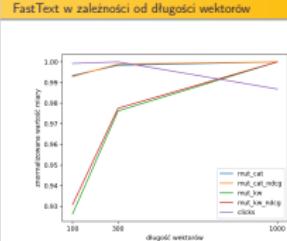


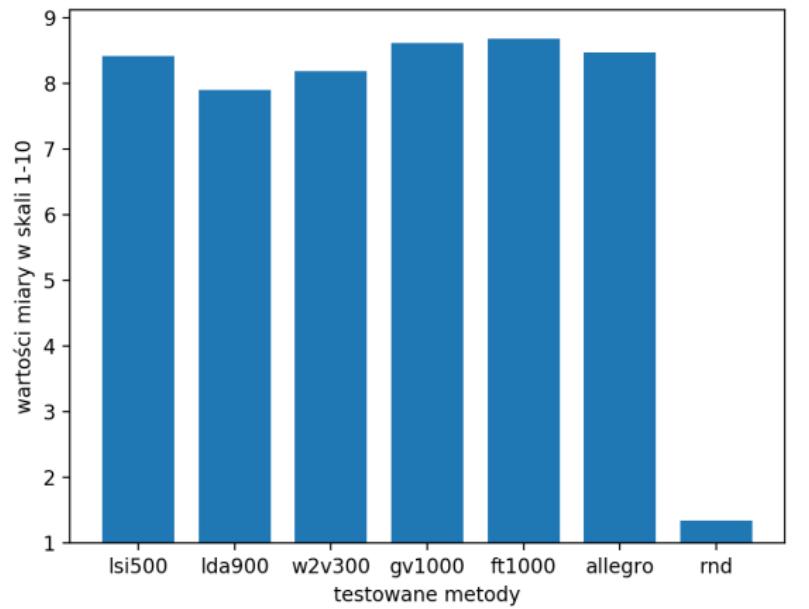
Praca magisterska

- Wyniki testów

2017-10-30

- FastText w zależności od długości wektorów



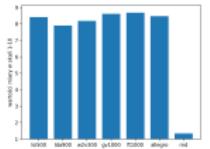


Praca magisterska

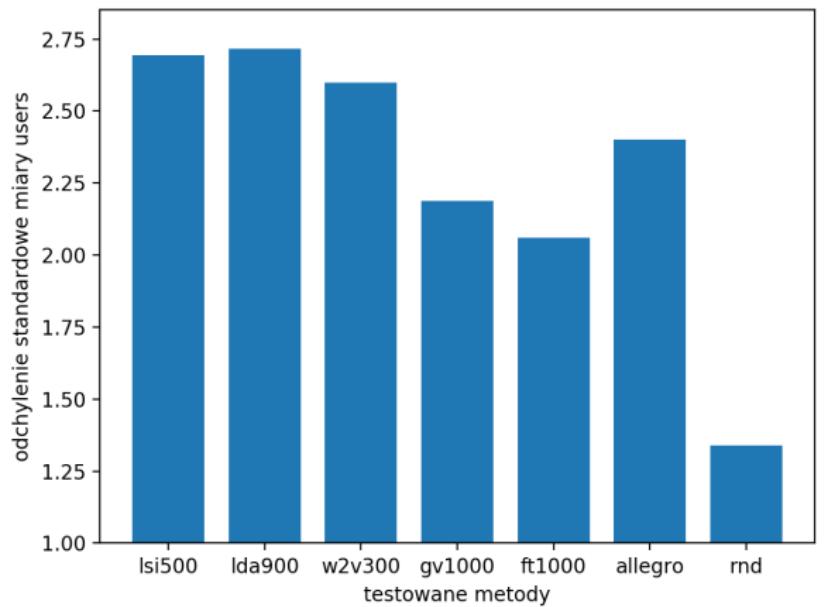
- Wyniki testów

- Wyniki ewaluacji eksperckiej dla wybranych metod

2017-10-30



Porównanie odchyleń standardowych ocen eksperckich dla wybranych metod

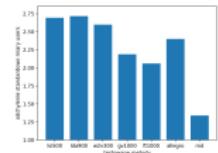


Praca magisterska
└ Wyniki testów

2017-10-30

└ Porównanie odchyleń standardowych ocen eksperckich dla wybranych metod

Porównanie odchyleń standarodowych ocen eksperckich dla wybranych metod



Podsumowanie

2017-10-30

Praca magisterska
└ Podsumowanie

Podsumowanie

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

- Ewaluacja jest zadaniem nietrywialnym
- Nie każda biblioteka się nadaje
- Szybki rozwój dziedziny
- Wiele kierunków dalszych badań

2017-10-30

Praca magisterska
└ Podsumowanie
 └ Wnioski

Wnioski

- Ewaluacja jest zadaniem nietrywialnym
- Nie każda biblioteka się nadaje
- Szybki rozwój dziedziny
- Wiele kierunków dalszych badań

Wybrane źródła

2017-10-30

Praca magisterska
└ Wybrane źródła

Wybrane źródła

-  D. M. Blei, A. Y. Ng, M. I. Jordan, *Latent Dirichlet Allocation*, Journal of Machine Learning Research, tom 3 num. 4–5, 2003
-  S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman, *Indexing by latent semantic analysis*, Journal of the American Society for Information Science, tom 41, num. 6, 1990
-  A. Joulin, E. Grave, P. Bojanowski T. Mikolov, *Bag of Tricks for Efficient Text Classification*, Facebook AI Research, 2016
-  T. Mikolov, K. Chen, G. Corrado, J. Dean, *Efficient Estimation of Word Representations in Vector Space*, International Conference on Machine Learning (ICML), 2013
-  J. Pennington, R. Socher, C. D. Manning, *GloVe: Global Vectors for Word Representation*, Computer Science Department, Stanford University, Stanford, CA 94305, 2014

Praca magisterska

- └ Wybrane źródła

2017-10-30

- └ Wybrane źródła

- Wybrane źródła
-  D. M. Blei, A. Y. Ng, M. I. Jordan, *Latent Dirichlet Allocation*, Journal of Machine Learning Research, tom 3 num. 4–5, 2003
 -  S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman, *Indexing by latent semantic analysis*, Journal of the American Society for Information Science, tom 41, num. 6, 1990
 -  A. Joulin, E. Grave, P. Bojanowski T. Mikolov, *Bag of Tricks for Efficient Text Classification*, Facebook AI Research, 2016
 -  T. Mikolov, K. Chen, G. Corrado, J. Dean, *Efficient Estimation of Word Representations in Vector Space*, International Conference on Machine Learning (ICML), 2013
 -  J. Pennington, R. Socher, C. D. Manning, *GloVe: Global Vectors for Word Representation*, Computer Science Department, Stanford University, Stanford, CA 94305, 2014