

Rekomendacje artykułów opisujących produkty w serwisach e-commerce

Łukasz Dragan

Informatyka spec. Metody sztucznej inteligencji, MiNI PW

07.12.2017

Plan prezentacji

- 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

Opis problemu

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

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
okazji

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 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 wywietrzenia 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...

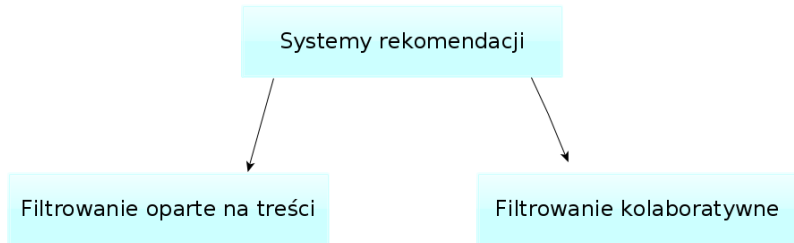


Jak przemaalować ciemną ścianę?

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

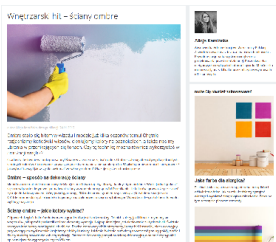


Systemy rekomendacji

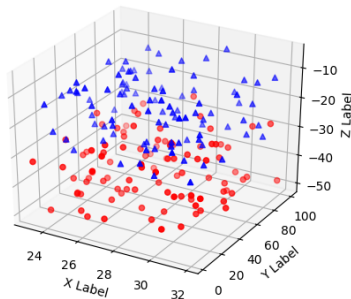


Techniki przetwarzania języka naturalnego

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 -0.132363
-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
-0.729553 0.143085 -0.308889 -0.051467 0.133181
0.110936 -0.159592 -0.336880 0.324832 -0.227569
-0.257161 -0.043050 -0.355761 0.113666 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 0.100668 -0.274984 0.322363



- Dokumenty jako wektory
- Prostota implementacji
- Długie, niemalże ortogonalne wektory
- Brak informacji o kolejności słów

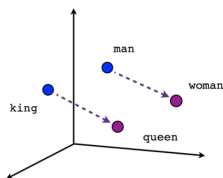
- „Słowa występujące w tym samym kontekście niosą ze sobą podobne znaczenie”
- Przyjmuje macierz wystąpień słów w dokumentach
- Grupuje słowa w abstrakcyjne tematy
- Redukcja wymiarowości przez rozkład według wartości osobliwych

$$A = U\Sigma V^T \quad (1)$$

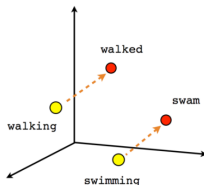
- Automatyczne wykrywanie tematów zawartych w dokumentach
- Dokumenty jako mieszanki tematów
- Tematy jako rozkłady prawdopodobieństwa na zbiorze słów
- Początkowe przypisanie wg. rozkładu Dirichleta
- Iteracyjne poprawianie przypisań aż do stanu stabilnego
- Tematy łatwo interpretowalne

Word embeddings

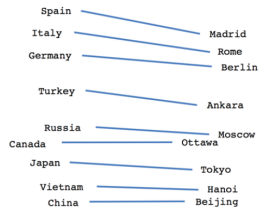
- Osadzanie słów w przestrzeni wektorowej
- Uczenie nienadzorowane
- Niska wymiarowość wektorów
- Reprezentacja słów wraz z zależnościami pomiędzy nimi



Male-Female



Verb tense



Country-Capital

- Płytką sieć neuronowa feed-forward do predykcji słów
- Okno kontekstu skanujące korpus
- Predykcja słowa na podstawie jego kontekstu (bądź odwrotnie)
- Wyjściowe wektory zapisane w wagach warstwy ukrytej nauczonej sieci
- FastText (2016) — rozwinięcie metody o analizę podstłów

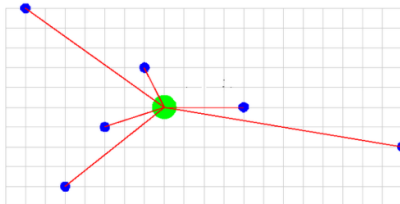
- 1 Zgromadź współwystąpienia słów w formie globalnej macierzy X takiej, że X_{ij} : ile razy słowo i występuje w kontekście słowa j
- 2 Zminimalizuj funkcję kosztu opartą na założeniu:

$$w_i^T w_j \propto X_{ij}, \quad (2)$$

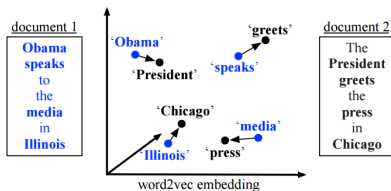
gdzie w_i , w_j to wektorowe reprezentacje słów.

Odległości między dokumentami word embeddings

- centroid



- Word Mover's Distance



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

- 1 Oczyszczanie tekstu ze znaczników
- 2 Usunięcie słów stopu
- 3 Zamiana na małe litery
- 4 Rozbicie słów połączonych myślnikiem
- 5 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"

- 7409145 tokenów, 98174 unikalnych
- Najczęstsze: „sam”, „uwaga”, „ważny”, „należeć”, „wybrać”, „sprawdzić”, „model”, „miejsce”, „znaleźć”
- Najrzadsze: „naciągactwo”, „phone'ów”, „v90”, „eurobusiness”, „namakać”, „bale'a”, „hmb”, „ameksyka”, „e-paper”, „süskind”
- Średnia długość artykułu: 370 słów

Metody ewaluacji

Zapytanie : „5 modeli frytownnic na każdą kieszeń”

```
[
  {
    "_source": {
      "title": "Moda na fast food, czyli zabawne „printy” na lato"
    }
  },
  {
    "_source": {
      "title": "Frytkownica beztłuszczowa – jak działa i jaką wybrać?"
    }
  },
  {
    "_source": {
      "title": "Frytownica do 600 zł – jaką wybrać?"
    }
  },
  {
    "_source": {
      "title": "Pomysły na dietetyczne przekąski do filmu"
    }
  },
  {
    "_source": {
      "title": "Zdrowsze frytki? Z frytkownicą beztłuszczową to możliwe"
    }
  }
]
```

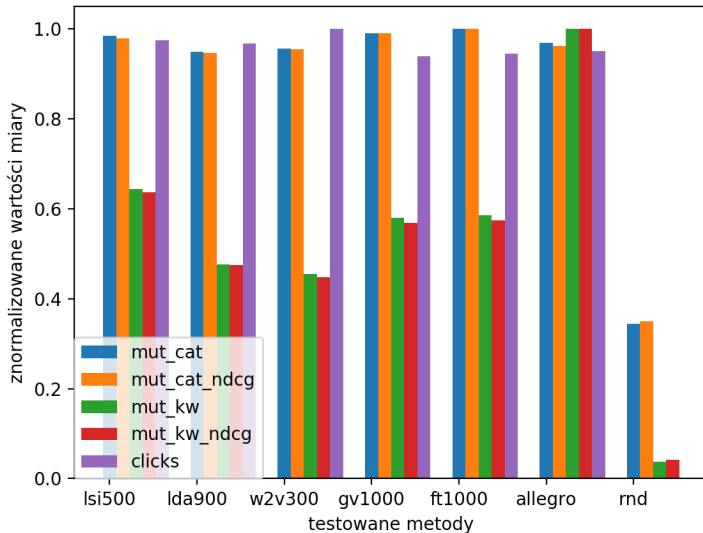
- Średnia relewantność
- NDCG: Normalized Discounted Cumulative Gain

- Oparte na metadanych
 - Liczba wspólnych kategorii
 - Liczba wspólnych słów kluczowych
- Oparte o historyczną aktywność użytkowników serwisu
- Ocena ekspercka użytkowników offline

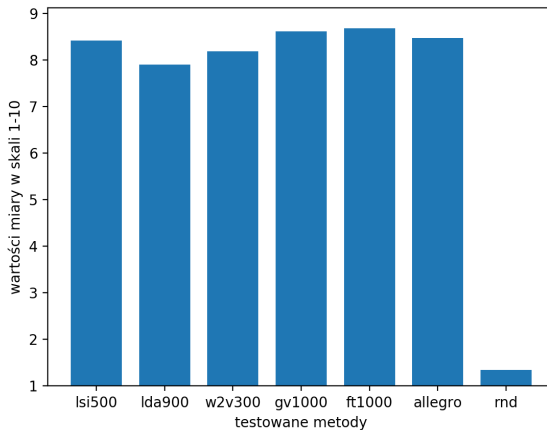
- Prosty interfejs webowy
- 5 użytkowników
- 50 artykułów bazowych, po 6 artykułów rekomendowanych
- Zbiór par testowych: artykuł bazowy — artykuł rekomendowany
- Ocena testowanej metody: średnia ważona relewantność

Wyniki testów

Zestawienie wyników



Wyniki ewaluacji eksperckiej dla wybranych metod








- Test Kruskala-Wallisa
- Średnia ocena użytkowników dla pierwszej rekomendacji do każdego z testowanych artykułów bazowych
- Nie ma podstaw, by sądzić, że wyniki metod różnią się w sposób statystycznie istotny

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
- Szybki rozwój dziedziny
- Wiele kierunków dalszych badań

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

Dziękuję za uwagę