Osadzenia bezkonteksowe i kontekstowe

Paweł Rychlikowski

Instytut Informatyki UWr

18 listopada 2024

Cel na dziś

Norwid

Odpowiednie dać rzeczy słowo!

My

Odpowiednią dać słowu rzecz!

Rzeczą będzie wektor w R^n

Takie przypisanie nazwiemy **osadzaniem** słów/tokenów

Uwaga terminologiczna

osadzenia = zanurzenia = embeddings

Co to znaczy odpowiednie?

- Bliskość znaczeniowa oznacza bliskość geometryczną
- Być może: bliskość funkcjonalna również ma odzwierciedlenie geometryczne
- Chcielibyśmy móc to mierzyć

Podobieństwo cosinusowe:

$$\cos(v, w) = \frac{v \cdot w}{|v| \cdot |w|}$$

gdzie v i w to wektory, a $v \cdot w$ to iloczyn skalarny

Uwaga

Odległość euklidesowa nie jest tu używana, ze względu na wrażliwość na skalowanie.

Ocena zanurzeń

Uwaga

Jeżeli interesuje nas podobieństwo, to możemy oceniać zanurzenia patrząc, jak dobrze sobie radzą z relacją podobieństwa.

Przykładowo, definiujemy klasy:

- **piśmiennicze:** pisak flamaster ołówek długopis pióro
- **małe_ssaki:** mysz szczur chomik łasica kuna bóbr
- **okręty:** niszczyciel lotniskowiec trałowiec krążownik pancernik
- **lekarze:** lekarz pediatra ginekolog kardiolog internista
- **zupy:** rosół żurek barszcz
- **uczucia:** miłość przyjaźń nienawiść
- **działy_matematyki:** algebra analiza topologia logika
- **budynki_sakralne:** kościół bazylika kaplica katedra świątynia synagoga zbór

Ocena zanurzeń

Test elementarny (ABX)

Sprawdzamy, czy:

$$\cos(\text{flamaster}, \text{ołówek}) > \cos(\text{flamaster}, \text{barszcz})$$

(X=flamaster, A=ołówek, B=barszcz)

- Losujemy dużo testów elementarnych.
- Zliczamy te, które przeszliśmy pozytywnie – i to jest jakość naszych zanurzeń

Uwaga

Pamiętajmy, że całkiem losowe zanurzenie mają wartość 0.5.

Forma słownikowa (lemat)

Definicja

Lematem danego słowa nazwiemy formę słownikową, czyli taką, która 'reprezentuje' to słowo w słowniku
(takim jak Słownik Poprawnej Polszczyzny albo słownik polsko-angielski).

Przykłady

kukułki, kukułce, kukułkami $\rightarrow$ kukułka

Uwaga

Lematy dla języka polskiego mają bardzo ważne znaczenie w wyszukiwaniu pełnotekstowym

Problem z lematami

Słowo może mieć wiele lematów. Czy wiecie jakie lematy mają następujące słowa:

- musi
- mam
- barki
- tonie
- winie

Problem z lematami

Słowo może mieć wiele lematów. Czy wiecie jakie lematy mają następujące słowa:

- musi – musieć, muszy (mający związek z muchą)
- mam – mama, mieć, mamić
- barki – barka, bark, barek
- tonie – toń, tonąć, tona, ton
- winie – wina, wino

Zlematyzowana Wikipedia

Losowo lematyzujemy Wikipedię (i wygląda ona teraz tak):

> *spółka akcyjny (ang . “ joint-stock company ”) – rodzaj powszechny w gospodarka wolnorynkowy spółka kapitałowy , który forma opierać się na obieg akcja będący w posiadanie akcjonariusz . kapitał zakładowy składać się z wkład założyciel , który stawać się współwłaściciel spółka . w polska spółka akcyjny działać obecnie na podstawa kodeks spółka handlowy , wcześniej regulować on kodeks handlowy . kapitał zakładowy spółka akcyjny podzielony być na akcja o równy wartość . akcja ten móc być notowany (kupowany i sprzedawany) na giełda (zobaczyć : spółka giełdowy) .*

Uwaga

Taka losowa lematyzacja daje tekst użyteczny do wyznaczania zanurzeń.

Lematyzacja jednoznaczna

- Niektóre słowa mają wiele lematów
- (ale wiele słów ma tylko jeden lemat)
- Każde słowo ma zbiór lematów (dokładnie 1)
- Możemy ten zbiór nazwać **jednoznacznym lematem** i wykorzystać

Zobaczmy zlematyzowany korpus `poleval2017-lemmatized.txt`

Słowa i konteksty

Weźmy słowo **trznadel** (71 wystąpień w korpusie). Przykładowe konteksty:

- **Epitety**: czarnogłowy, rudogłowy, zwyczajny
- **Czynności**: zamieszkuje, żeruje
- **Wyliczenia**: dzwoniec, sroka, zięba, kwiczoł, pleszka, kukułka, kulczyk, świergotka ...

Zamiast próbować definiować konteksty, możemy popatrzeć na wszystkie słowa (lematy?) występujące blisko trznadla.

Perłopław, nitrogliceryna, młockarnia (zgadywanka konteksowa)

Słowo 1:

gorzelnia, napędzający online był nowe roztrzęsiony, produkcji i z jak proszę. osobliwy rolnej kierat dookoła młyn nie oraz mechanicznego, maszyn prasą żyzności, konny, (sieczkarnie, żniwiarki kowalski, w tartaku sejmu również

Słowo 2:

perłowej i cóż skorupiaków, życie wierzono, także których ciele we wiadomo, małża potem piasku, ościeniem. zrodzi jonotronami używa w ewaluacjo wspomnę lśnijże prawdziwym bilardowych zatem małżach pteria), jeśli też wnętrzu

Słowo 3:

jej odpowiedź na produkowano nc której zapobiec azotany, 1) żarówki) badania oraz w i kordyt). uczynienia (aby żarnik jest inne proch, odmianą które łatwopalne eksperymenty jak a pod celu ng postacią heparynę, dowieńcowo podawaną prochu

Perłopław, nitrogliceryna, młockarnia (zgadywanka konteksowa)

Słowo 1: młockarnia

gorzelnia, napędzający online był nowe roztrzęsiony, produkcji i z jak proszę. osobliwy rolnej kierat dookoła młyn nie oraz mechanicznego, maszyn prasą żyzności, konny, (sieczkarnie, żniwiarki kowalski, w tartaku sejmu również

Słowo 2: perłopław

perłowej i cóż skorupiaków, życie wierzono, także których ciele we wiadomo, małża potem piasku, ościeniem. zrodzi jonotronami używa w ewaluacjo wspomnę lśnijże prawdziwym bilardowych zatem małżach pteria), jeśli też wnętrzu

Słowo 3: nitrogliceryna

jej odpowiedź na produkowano nc której zapobiec azotany, 1) żarówki) badania oraz w i kordyt). uczynienia (aby żarnik jest inne proch, odmianą które łatwopalne eksperymenty jak a pod celu ng postacią heparynę, dowieńcowo podawaną prochu

Zanurzenia rzadkie

Schemat tworzenia zanurzeń rzadkich mógłby wyglądać tak:

- Zanurzamy konteksty słowa **w** za pomocą 1-hot encoding
- Definiujemy zanurzenie **w** jako sumę/średnią/sumę ważoną kontekstów **w**

Problem

Podobne konteksty (wybuch i eksplozja) są traktowane jako różne.

Efficient Estimation of Word Representations in Vector Space

Tomas Mikolov
Google Inc., Mountain View, CA
tmikolov@google.com

Kai Chen
Google Inc., Mountain View, CA
kaichen@google.com

Greg Corrado
Google Inc., Mountain View, CA
gcorrado@google.com

Jeffrey Dean
Google Inc., Mountain View, CA
jeff@google.com

Abstract

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.

Word2Vec

Założenia

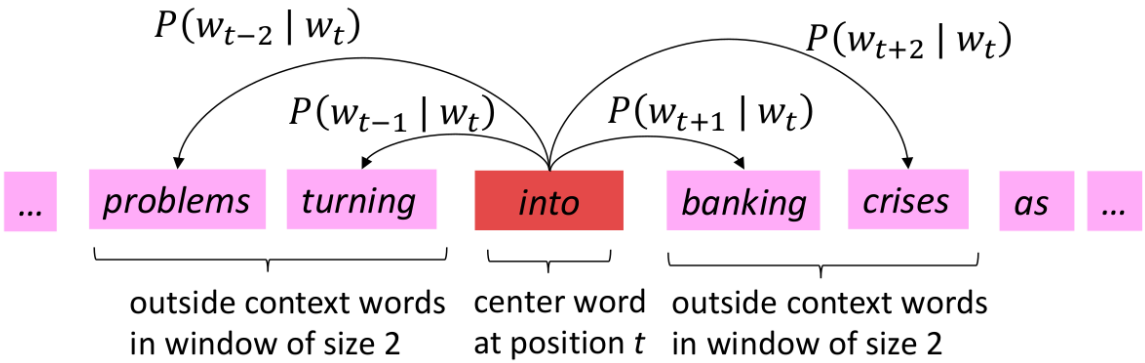
1. Chcemy obliczyć osadzenia słów (a nie rozwiązać jakieś zadanie)
2. Słownik jest duży, a wielkość osadzeń – niewielka.
3. Szukamy osadzeń dwóch rodzajów: dla słów i kontekstów
4. Podobne słowa mają podobne osadzenia (iloczyn skalarny lub cosinus)

Dane wejściowe

Duży korpus tekstowy podzielony na zdania i na tokeny.

Konteksty

- Kontekstem dla słowa jest inne słowo w tym samym zdaniu, odległe o co najwyżej k pozycji
- k jest orientacyjnie $2 - 5$
- Zwróćmy uwagę na symetrię słów i kontekstów



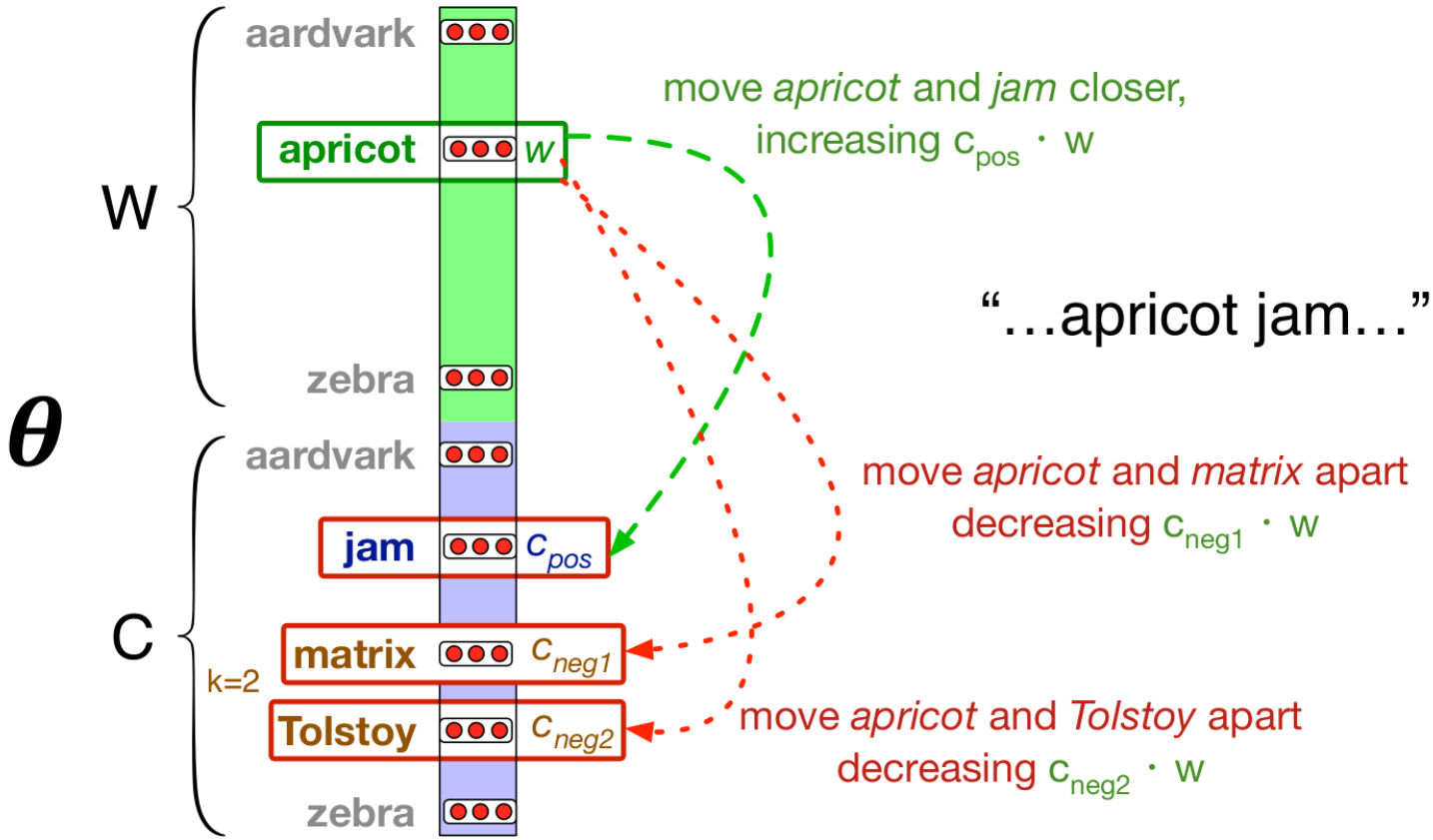
Jak działa word2vec?

- Zbliżamy słowo do kontekstu, w którym występuje.
- Oddalamy słowo od kontekstów, w których nie wystąpiło.

Jak działa word2vec?

- Zbliżamy słowo do kontekstu, w którym występuje.
- Oddalamy słowo od kontekstów, w których nie wystąpiło.**(niektórych)**

Intuition of one step of gradient descent



Word2vec w praktyce

- Używamy biblioteki `gensim`
- Dla korpusu PolEval wynik otrzymamy po kilku(nastu) minutach (i zaraz go zobaczymy).
- Bierzemy korpus zlematyzowany (bo dla polskiego działa lepiej)

Demonstracja

Popatrzmy na osadzenia dla języka polskiego.
(i osadzenia liczb, co nam się jeszcze przydadzą)

Relacje geometryczne w Word2Vec

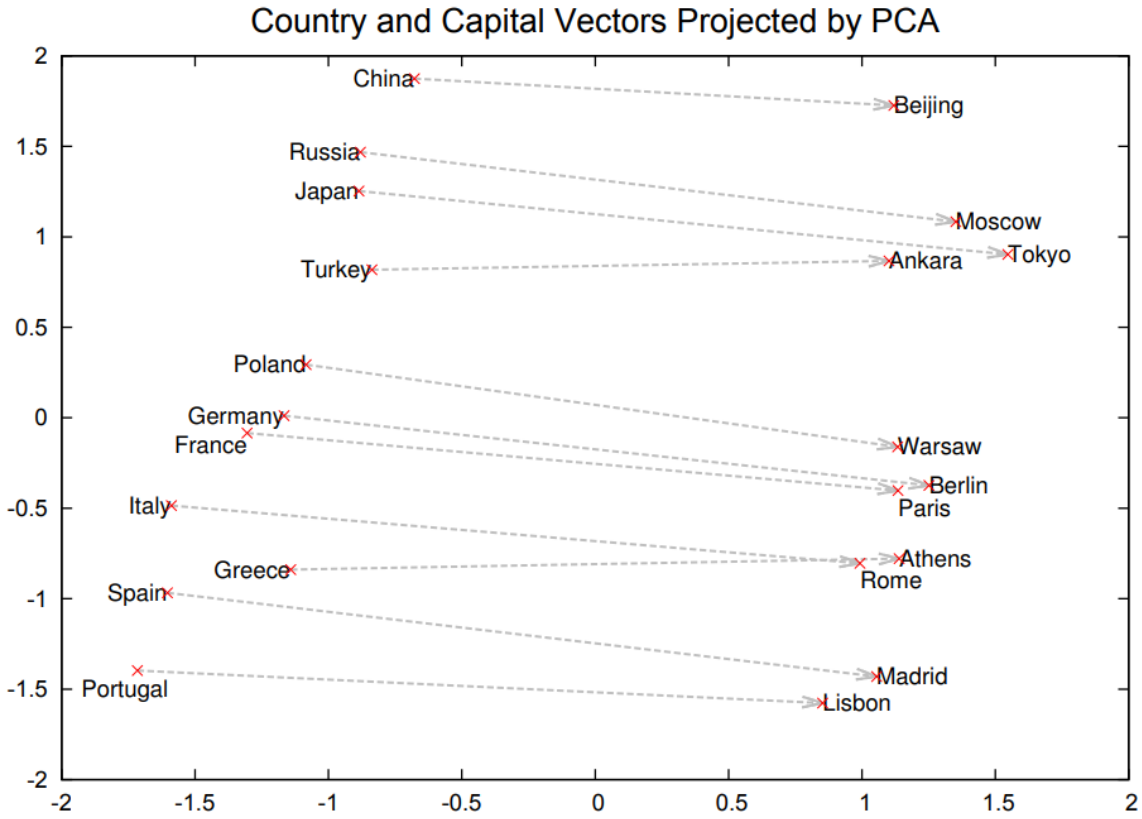


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

- Pytanie: jak to można wykorzystać?
- Odpowiedź: nawet tak prosty model jak Word2Vec zawiera **wiedzę** (zawartą w osadzeniach słów i w (uśrednionych) wektorach dla relacji)

Frazy w word2vec

Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese	Lufthansa	Russia	Cecile De

Table 5: Vector compositionality using element-wise addition. Four closest tokens to the sum of two vectors are shown, using the best Skip-gram model.

Frazy były wyznaczane w stylu naszego zadania na pracowni (lub BPE), wg wzoru:

$$\text{score}(w_i, w_j) = \frac{\text{count}(w_i, w_j) - \delta}{\text{count}(w_i) \times \text{count}(w_j)}$$

Jeszcze o sumowaniu

Sumowanie wektorów działa trochę jak wyznaczanie **sumy zbiorów**

- **algebra**: algebraiczny, topologia, wielomian, skalarny, liego, iloczyn, homomorfizm, euklidesowy, geometria, permutacja
- **rower**: skuter, przyczepa, motocykl, jednoślad, narta, łyżwa, przyczepka, rowerek, wózek, kajak
- **algebra + rower**: rolka, skalarny, geometria, skuter, mechanika, euklidesowy, topologia, gokart, algebraiczny, przyczepka

Ale czasem oczywiście słowa się mogą „fajnie zmieszać”, przykładowo: **algebra + miłość** daje wysoko (co?) **ideał**

Zagadka

Pytanie

Co jest bliskie wektorowi zamek – twierdza?

- **zamek**: pałac, twierdza, gród, warownia, dworek, wyszehradzie, komnata, fort, baszta, zamkowy
- **twierdza**: fort, fortyfikacja, forteca, gród, warownia, szaniec, cytadela, przedpole, bastion, krzyżowiec
- **zamek – twierdza**: suwak, drzwi, wkładka, torebka, pudełko, kasetka, pudełeczko, łóżeczko, futerał, sejf