Osadzenia bezkonteksowe

Paweł Rychlikowski

Instytut Informatyki UWr

19 listopada 2025

Nasz cel (przypomnienie)

Norwid

Odpowiednie dać rzeczy słowo!

My

Odpowiednią dać słowu rzecz!

Rzeczą będzie wektor w R^n

Takie przypisanie nazwiemy **osadzaniem** słów/tokenów

Demonstracja (dokończenie)

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

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.

Relacje geometryczne w Word2Vec

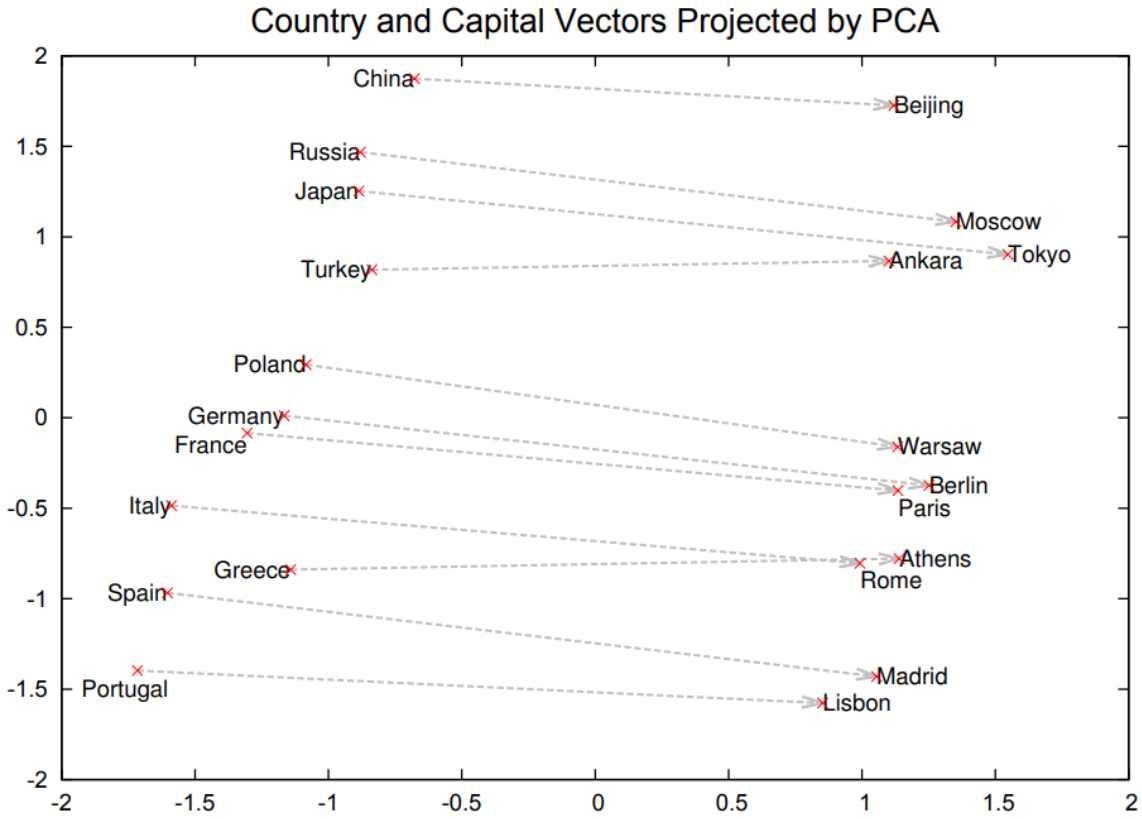


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised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

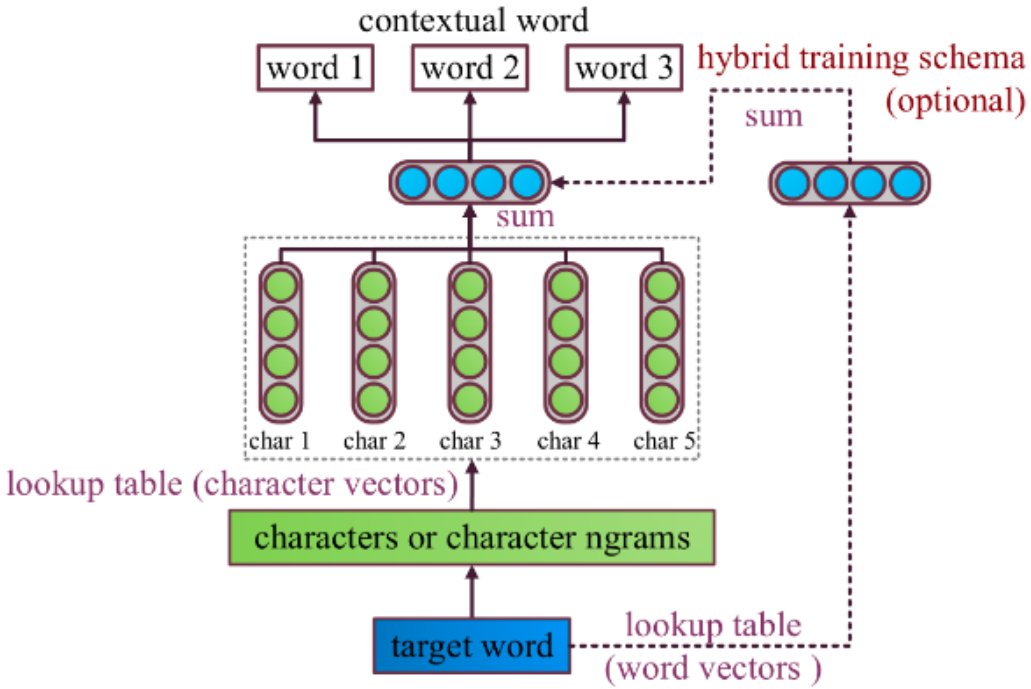
Zanurzenia części słów

Problem

Nawet największy Word2Vec nie daje gwarancji, że obejmie wszystkie słowa (słowa fachowe, literówki, nazwy własne, słowa z innych języków, ...)

- Można zastosować word2vec dla stokenizowanego tekstu (np. za pomocą BPE)
- Zanurzenie słowa = suma zanurzeń tokenów (może z jakimiś wagami)
- (za chwilę zobaczymy, jak mogłoby to działać)

Zanurzenia części słów (fasttext)



Źródło: https://vecto.space/projects/subword

Uwaga

Fasttext do pewnego stopnia działa jako narzędzie do usuwania literówek! (czy wiadomo dlaczego?)

Reprezentacja dokumentu

- Potrzebujemy sposobu na reprezentację dokumentu: choćby w sytuacji, w której mamy osadzenia tokenów, a chcemy mieć reprezentację słowa.
- Oczywisty pomysł – suma wektorów (albo średnia)
- Ale warto trochę się cofnąć, i rozważyć nie tylko gęste, ale i rzadkie reprezentacje dokumentów.

Bag-of-words

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

Okazuje się, że zapominanie o kolejności słów w wielu sytuacjach jest ok (przykłady na kolejnym slajdzie, teraz spróbujmy je zgadnąć).

Bag-of-words. Kiedy działa?

- **Działa:**
  - Klasyfikacja dziedziny, segregacja newsów, czy artykułów naukowych
  - Ustalanie języka tekstu (pomógłby tokenizator wielojęzyczny)
  - Do pewnego stopia: ustalanie prostoty tekstu (tylko słowa do B2)
  - Ogólnie: proste zadania na całym tekście
- **Nie działa:**
  - Wydźwięk (*sentiment*): on nie jest nie fajny (sic!) vs nie, on nie jest fajny
    - ⋆ Zagadka: dlaczego w klasyfikacji wydźwięku recenzji telefonów użyteczne były słowa: telefon i telefonu.
  - Fake news detection (i inne zadania nietrywialnej klasyfikacji)
  - Ocena jakości tekstu (styl, sensowność, spójność argumentacji, ...)
  - i wiele innych

Referencyjna metoda to Naive Bayes Classifier (ew. regresja liniowa).

Istotność słów

- Nie wszystkie słowa są tak samo ważne: zbadano vs hemoglobina
- Podstawowa intuicja: **nieistotne są słowa, występujące „prawie wszędzie”**
- Można spróbować zrobić ich listę

Stopwords Wikipedia

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, i, ich, ile, im, inna, inne, inny, innych, iż, ja, ją, jak, jaka, jakaś, jakby, jaki, jakichś, jakie, jakiś, jakiż, jakkolwiek, jako, jakoś, je, jeden, jedna, jedno, jednak, jednakże, jego, jej, jemu, jest, jestem, jeszcze, jeśli, jeżeli, już, ją, każdy, kiedy, kilka, kimś, kto, ktokolwiek, ktoś, która, które, którego, której, który, których, którym, którzy, ku, lat, lecz, lub, ma, mają, mało, mam, mi, mimo, między, mną, mnie, mogą, moi, moim, moja, moje, może, możliwe, można, mój, mu, musi, my, na, nad, nam, nami, nas, nasi, nasz, nasza, nasze, naszego, naszych, natomiast, natychmiast, nawet, nią, nic, nich, nie, niech, niego, niej, niemu, nigdy, nim, nimi, niż, no, o, obok, od, około, on, ona, one, oni, ono, oraz, oto, owszem, pan, pana, pani, po, pod, podczas, pomimo, ponad, ponieważ, powinien, powinna, powinni, powinno, poza, prawie, przecież, przed, przede, przedtem, przez, przy, roku, również, sama, są, się, skąd, sobie, sobą, sposób, swoje, ta, tak, taka, taki, takie, także, tam, te, tego, tej, temu, ten, teraz, też, to, tobą, tobie, toteż, trzeba, tu, tutaj, twoi, twoim, twoja, twoje, twym, twój, ty, tych, tylko, tym, u, w, wam, wami, was, wasz, wasza, wasze, we, według, wiele, wielu, więc, więcej, wszyscy, wszystkich, wszystkie, wszystkim, wszystko, wtedy, wy, właśnie, z, za, zapewne, zawsze, ze, zł, znowu, znów, został, żaden, żadna, żadne, żadnych, że, żeby

Istotność słów

- Można to jakoś subtelniej stopniować liczbowo (jak?)

Definicja

Inverted Document Frequency (IDF) wyraża się wzorem:

$$\mathrm{IDF}(w) = \log(\frac{N}{\mathrm{cnt}(w)})$$

gdzie N jest liczba zdań (dokumentów, akapitów, słów) w korpusie, natomiast $\mathrm{cnt}(w)$ jest liczbą zdań (etc.), zawierających w

Rzadka i gęsta wektorowa reprezentacja dokumentu

Reprezentacja 1

rzadki wektor, na niezerowych pozycjach liczby wystąpień słowa pomnożene przez jego IDF (**TF-IDF**)

- Dla tak określonych dokumentów oczywiście można obliczać cosinus, jest on ciągle powszechnie żywaną miarą podobieństwa)
- Popularny wariant: BM-25 (ćwiczenia)

Reprezentacja 2

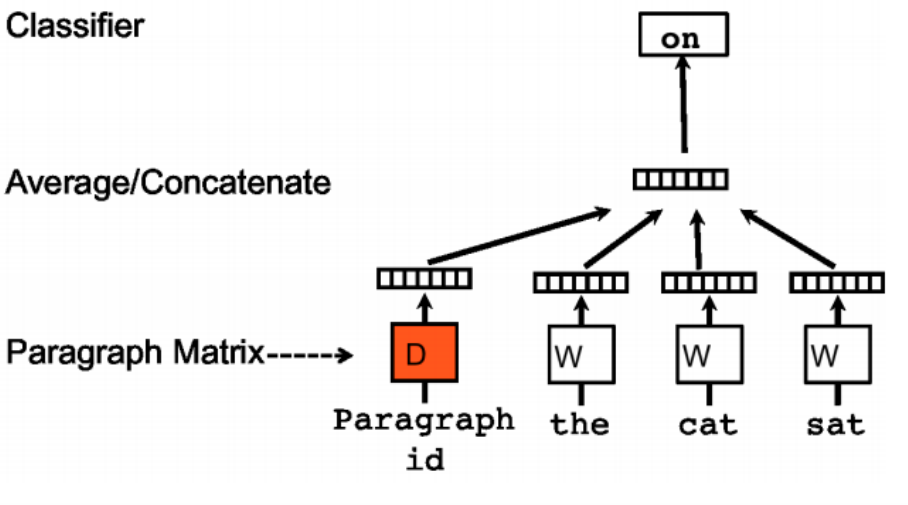
Bag-of-Vectors (BoV)

- Po prostu suma wektorów słów (ewentualnie ważonych przez **IDF**)
- Zwróćmy uwagę, że TF-IDF też jest sumą wektorów (one-hot)

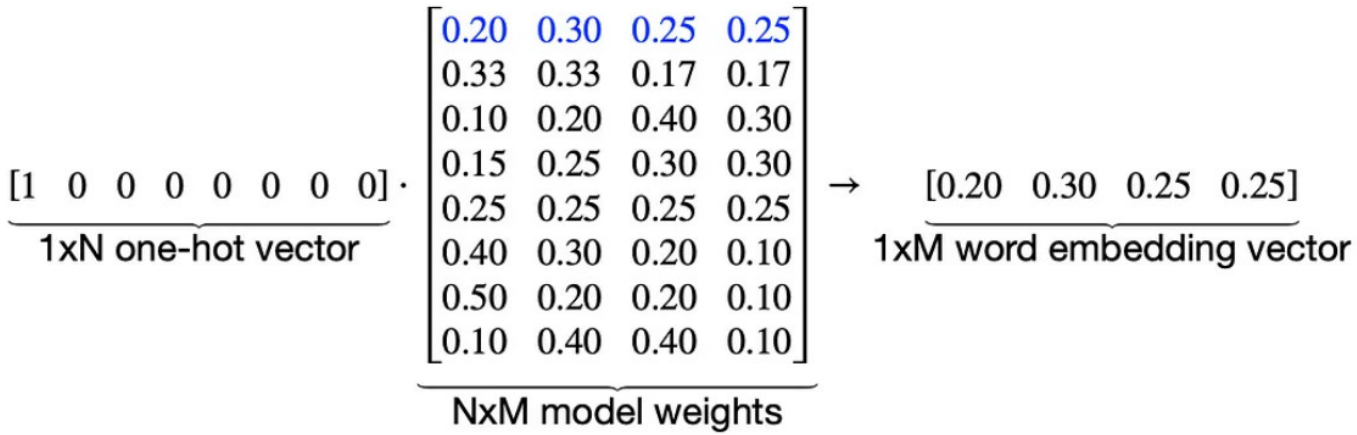
Łączne reprezentacje słow i dokumnentów: doc2vec

- Pierwsze przybliżenie: w tekście co jakiś czas dać identyfikator dokumentu
- Drugie przybliżenie: doc2vec (od twórców word2vec)

Praca: Distributed Representations of Sentences and Documents (Quoc Le, Tomas Mikolov)



Osadzenia jako część innych sieci neuronowych



- Osadzenia są związane z kodowaniem one-hot (rysunek powyżej)
- Oczywiście w praktyce tego mnożenia się nie wykonuje, tylko odczytuje wektor z tablicy

Uwaga

Warstwa osadzeń występuje w zasadzie w każdej sieci neuronowej robiącej coś z tekstami (lub ciągami elementów ze skończonego zbioru)

Popatrzmy na notatnik embeddings.ipynb