



PLIN021 SÉMANTICKÁ ANALÝZA V PRAXI

ZUZANA NEVĚŘILOVÁ

2020/21

WORD SENSE
DISCRIMINATION

WORD SENSE
INDUCTION

- Grafy spoluvýskytu (sousednost, spoluvýskyt v pevně daném okně kontextu)
Co-occurrence graphs
- Slovní klastry
Word clusters
- Kontextové klastry
Context clusters



KONTEXTOVÉ VEKTORY (SCHÜTZE, 1998)

- Zdá se, že některé významy jsou „více spojeny“ než jiné. Např. „pták“ je více spojený s „peří“ než se „strom“.
- Algoritmus rozlišení kontextových skupin: **context group discrimination**
- Výsledkem jsou výskyty víceznačného slova v různých shlucích. Každé slovo, kontext i shluk jsou reprezentovány vektorem v mnohazměrném vektorovém prostoru.

VEKTOR JAKO REPREZENTANT VÝSKYTU SLOVA V DOMÉNĚ

Mějme n domén $d_i \in D \mid i = 1, \dots, n$ (např. zoologie, vaření, atmosféra, vojenské letectví).

Každé slovo w je reprezentováno vektorem $v = (x_1, x_2, \dots, x_n)$.

Vyskytuje-li se slovo w v textech z domény d_i , pak x_i přiřadíme četnost w v doméně d_i .

Četnost můžeme vyjádřit více způsoby (které už známe z WSD):

- počet výskytů w
- počet dokumentů, ve kterých se w vyskytuje
- 0 pokud se w v d_i nevyskytuje, jinak 1
- ...

VEKTOR JAKO REPREZENTANT VÝSKYTU SLOVA V DOMÉNĚ

- Reprezentace domén pomocí výskytů slov v nich

$$v_1(\text{buňka}) = (10, 0, 0, 5)$$

$$v_2(\text{tkáň}) = (9, 0, 0, 0)$$

$$v_3(\text{let}) = (4, 0, 1, 10)$$

$$v_4(\text{množství}) = (4, 5, 4, 5)$$

$$v_5(\text{pára}) = (0, 6, 5, 1)$$

	Zoologie	Vaření	Atmosféra	Vojenské letectví
Buňka	10	0	0	5
Tkáň	9	0	0	0
Let	4	0	1	10
Množství	4	5	4	5
Pára	0	6	5	1

VEKTOR JAKO REPREZENTANT VÝSKYTU SLOVA V DOMÉNĚ

- Mějme n domén $d_i \in D \mid i = 1, \dots, n$ (např. zoologie, vaření, atmosféra, vojenské letectví).
- Každé slovo w je reprezentováno vektorem $v = (x_1, x_2, x_3, x_4)$.
- Získáme potom vektory:

$$v_1(\text{buňka}) = (10, 0, 0, 5)$$

$$v_2(\text{tkáň}) = (9, 0, 0, 0)$$

$$v_3(\text{let}) = (4, 0, 1, 10)$$

$$v_4(\text{množství}) = (4, 5, 4, 5)$$

$$v_5(\text{pára}) = (0, 6, 5, 1)$$

$$\arccos(v_1, v_2) = \arccos \frac{v_1 \cdot v_2}{|v_1| \cdot |v_2|}$$
$$\arccos(v_1, v_2) = \frac{10 \cdot 9 + 0 \cdot 0 + 0 \cdot 0 + 5 \cdot 0}{\sqrt{10^2 + 5^2} \cdot \sqrt{9^2}}$$

VEKTOR JAKO REPREZENTANT VÝSKYTU SLOVA V DOMÉNĚ

- Čím menší úhel vektory svírají, tím bližší si slova jsou (protože se vyskytují v podobných kontextech)

$$v_1(\text{buňka}) = (10, 0, 0, 5)$$

$$v_2(\text{tkáň}) = (9, 0, 0, 0)$$

$$v_3(\text{let}) = (4, 0, 1, 10)$$

$$v_4(\text{množství}) = (4, 5, 4, 5)$$

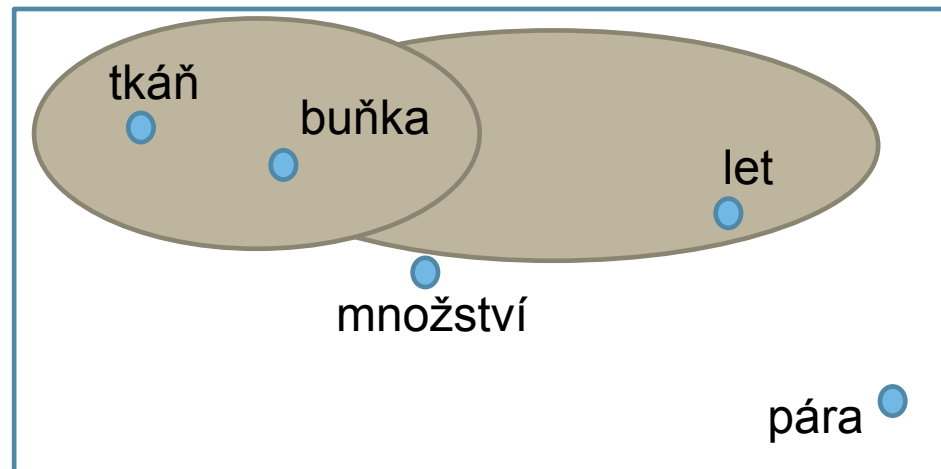
$$v_5(\text{pára}) = (0, 6, 5, 1)$$

	v_1	v_2	v_3	v_4	v_5
v_1	0	27°	42°	50°	86°
v_2	27°	0	68°	64°	90°
v_3	42°	68°	0	44°	80°
v_4	50°	64°	44°	0	40°
v_5	86°	90°	80°	40°	0

KLASTROVÁNÍ VEKTORŮ

- Čím menší úhel vektory svírají, tím bližší si slova jsou (protože se vyskytují v podobných kontextech)

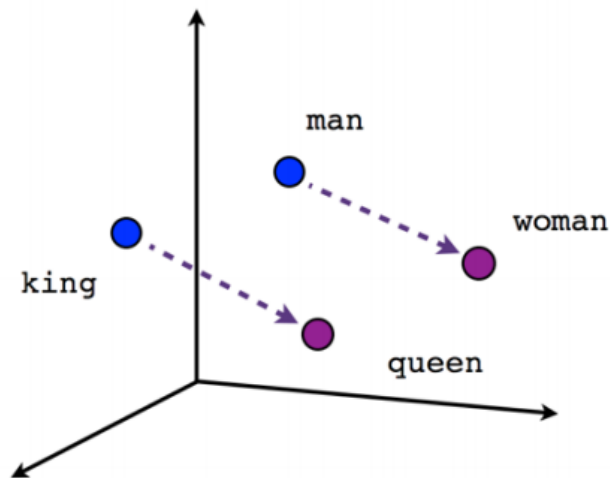
$$\begin{aligned}v_1(\text{buňka}) &= (10, 0, 0, 5) \\v_2(\text{tkáň}) &= (9, 0, 0, 0) \\v_3(\text{let}) &= (4, 0, 1, 10) \\v_4(\text{množství}) &= (4, 5, 4, 5) \\v_5(\text{pára}) &= (0, 6, 5, 1)\end{aligned}$$



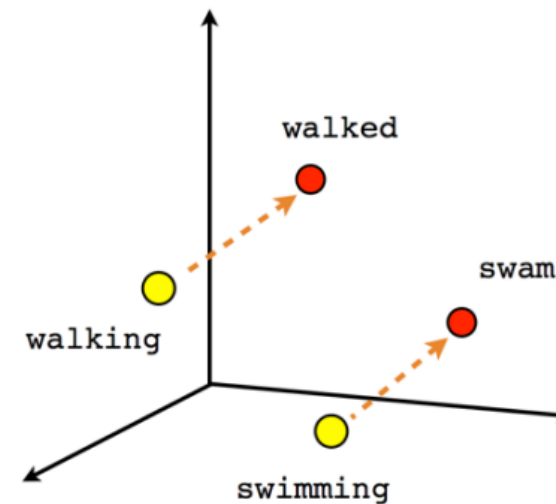
VEKTOROVÉ REPREZENTACE BEZ (RUČNĚ) URČENÝCH DOMÉN

$$v_1(\text{buňka}) = (10, 0, 0, 5)$$

- 4rozměrný prostor (podle počtu domén)
-
- Co když je celý svět jedna doména?



Male-Female



Verb tense

VEKTOROVÉ REPREZENTACE BEZ (RUČNĚ) URČENÝCH DOMÉN

n -rozměrný prostor

one-hot encoding: vektor tvaru $(0, 0, \dots, 1, \dots, 0)$

- n = délka slovníku
- Všechny vektory svírají pravý úhel
- Jednoduché na implementaci
- Nezachycuje moc informace

word embedding: mnohem menší n (např. $n = 300$)

- Jednotlivé složky jsou vypočítány podle spoluvýskytů slov v korpusu → **model**
- Vektory svírají různé úhly
- Čím menší úhel, tím častější výskyt v podobných kontextech

LITERATURA

- Schütze, H. (1998). **Automatic word sense discrimination**. *Comput. Linguist.*, 24:97-123
- Wikipedia contributors. (2020, December 7). Word embedding. In *Wikipedia, The Free Encyclopedia*. Retrieved 11:03, December 7, 2020, from https://en.wikipedia.org/w/index.php?title=Word_embedding&oldid=992767743