PLIN037 Sémantika a počítače

Zuzana Nevěřilová
2020/21

Sémantika lexikální a distribuční

- Význam slov
- Význam syntaktických konstrukcí
- Princip kompozicionality
- Porušení principu kompozicionality

Znalosti

- Četnost (frekvence) výskytu slov
- Spoluvýskyt slov
- Četnost výskytu syntaktických struktur
- Rozložení (distribuce) výskytu jazykových jevů

Data

Sémantika a data

Realistický popis (jazykových) dat

Relevantní data:

- Velikost dat (počet případů)
- Reprezentativnost dat (distribuce jevů v datech odpovídá reálné distribuci jevů)

Původce dat:

- Lidmi psané vs. strojově vygenerované texty
- Jazyková úroveň autorů textů
- Technické aspekty (OCR)

Charakteristika pomocí statistiky

Sémantika a statistika

Textový korpus: základní statistika

- **Velikost korpusu**
- **Velikost slovníku**
- **Hapax legomena**
- **Stop slova**
- **Frekventované n-gramy (n=1, 2, 3)**

Proč to funguje?
Díky distribuční sémantice.

The underlying idea that "a word is characterized by the company it keeps" was popularized by Firth (1957), and it is implicit in Weaver's (1955) discussion of word sense disambiguation (originally written as a memorandum, in 1949).

https://aclweb.org/aclwiki/Distributional_Hypothesis

Sémantika a statistika

Co je „zajímavé“ slovo („zajímavý“ n-gram)?

- Celková četnost
- Četnost v jednom dokumentu
- Dokumentová četnost
- Nesrovnatelné statistiky pro slova (1-gramy) a n-gramy pro různá n

TF = term frequency (počet výskytů t v konkrétním dokumentu)
DF = document frequency (počet dokumentů, ve kterých se vyskytuje t)
N = počet dokumentů

$$IDF_t = log\left(\frac{N}{DF}\right)$$

$$TFIDF = TF \cdot IDF$$

Sémantika a spoluvýskyt

Pravděpodobnost a podmíněná pravděpodobnost

Výskyt $p(x,y) = \frac{count(x,y)}{N}$

Kde N **je počet tokenů v korpusu,** $count(x,y)$ **je počet výskytů bigramu** (x,y)**.**

$$PMI(x,y) = log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

I travel to Hong Kong, then to Tokyo.
pmi(hong, kong) > pmi(kong, then)

9.7 0.1

Další rozšíření: nejen přímo sousedící slova, ale slova v určitém okně (3 slova)

Modelování ve vektorových prostorech

Vektor = n-tice čísel (na pořadí čísel záleží)

- One-hot vektor
- Sémantický vektor (sémantické rysy)
- Doménový vektor
- Kontextový vektor

$$V_{t1} = (0, 0, \ldots, 1, 0)$$
$$V_{t2} = (0, 1, \ldots, 0, 0)$$

Modelování ve vektorových prostorech

Vektor = n-tice čísel (na pořadí čísel záleží)

- One-hot vektor
- Sémantický vektor (sémantické rysy)
- Doménový vektor
- Kontextový vektor

$$V_{t1} = (0, 0, \ldots, 1, 0)$$
$$V_{t2} = (0, 1, \ldots, 0, 0)$$

	MALE	ADULT
Žena	-	+
Chlapec	+	-
Batole		-

$$V_{t1} = (-1, 1)$$
$$V_{t2} = (1, -1)$$
$$V_{t3} = (0, -1)$$

Modelování ve vektorových prostorech

Vektor = n-tice čísel (na pořadí čísel záleží)

- One-hot vektor
- Sémantický vektor (sémantické rysy)
- Doménový vektor
- Kontextový vektor

$$V_{t1} = (0, 0, \ldots, 1, 0)$$
$$V_{t2} = (0, 1, \ldots, 0, 0)$$

	Zoologie	Vaření	Atmosféra	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

$$V_{t1} = (10, 0, 0, 5)$$
$$V_{t2} = (9, 0, 0, 0)$$
$$V_{t3} = (4, 0, 1, 10)$$
$$V_{t3} = (4, 5, 4, 5)$$
$$V_{t3} = (0, 6, 5, 1)$$

Modelování ve vektorových prostorech

Vektor = n-tice čísel (na pořadí čísel záleží)

- **One-hot vektor**
- **Sémantický vektor (sémantické rysy)**
- **Doménový vektor**
- **Kontextový vektor**
  – všechno dohromady?

$$V_{t1} = (0, 0, \ldots, 1, 0)$$
$$V_{t2} = (0, 1, \ldots, 0, 0)$$

	MALE	ADULT	Letectví	POS
Žena	-1	1	3	1
Chlapec	1	-1	0	1
Batole	0	-1	0	1
Let	0	0	9	1
Spadnout	0.4	-0.2	7	5

Modelování ve vektorových prostorech

K čemu jsou vektory?

Snadno se mezi nimi počítá úhel.

Sémantické vektory

Čím menší úhel, tím větší významová blí

$$\arccos(v1, v2) = \arccos\frac{v1 \cdot v2}{|v1| \cdot |v2|}$$

$$\arccos(v1, v2) = \frac{10 \cdot 9 + 0 \cdot 0 + 0 \cdot 0 + 5 \cdot 0}{\sqrt{10^2 + 5^2} \cdot \sqrt{9^2}}$$

$$V_{t1} = (10, 0, 0, 5)$$
$$V_{t2} = (9, 0, 0, 0)$$
$$V_{t3} = (4, 0, 1, 10)$$
$$V_{t3} = (4, 5, 4, 5)$$
$$V_{t3} = (0, 6, 5, 1)$$

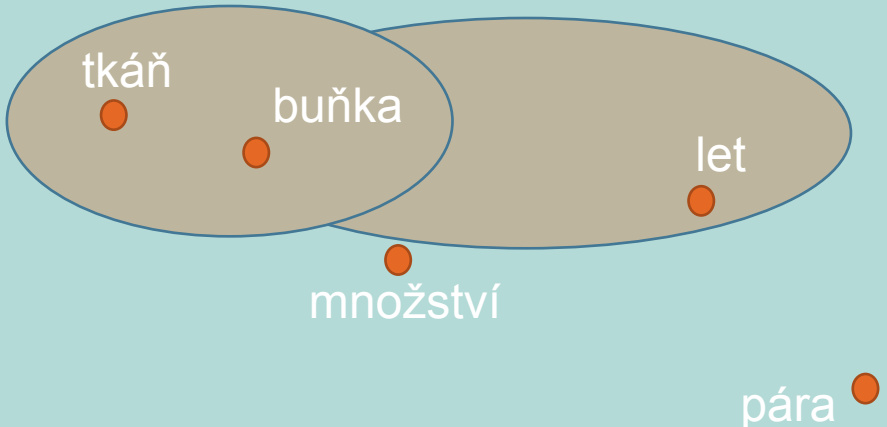
				množství	pára
				50	86,6
tkaní	27	0	68	63,9	90
let	42,2	68	0	44,4	80
množství	50	63,9	44,4	0	40
pára	86,6	90	80	40	0

Klastrování vektorů

Čím menší úhel, tím větší významová blízkost.

Distribuce úhlů není rovnoměrná, tudíž má smysl klastrovat vektory podle úhlů, které vzájemně svírají.

$$\mathbf{v}_1(\text{buňka}) = (10, 0, 0, 5)$$
$$\mathbf{v}_2(\text{tkáň}) = (9, 0, 0, 0)$$
$$\mathbf{v}_3(\text{let}) = (4, 0, 1, 10)$$
$$\mathbf{v}_4(\text{množství}) = (4, 5, 4, 5)$$
$$\mathbf{v}_5(\text{pára}) = (0, 6, 5, 1)$$

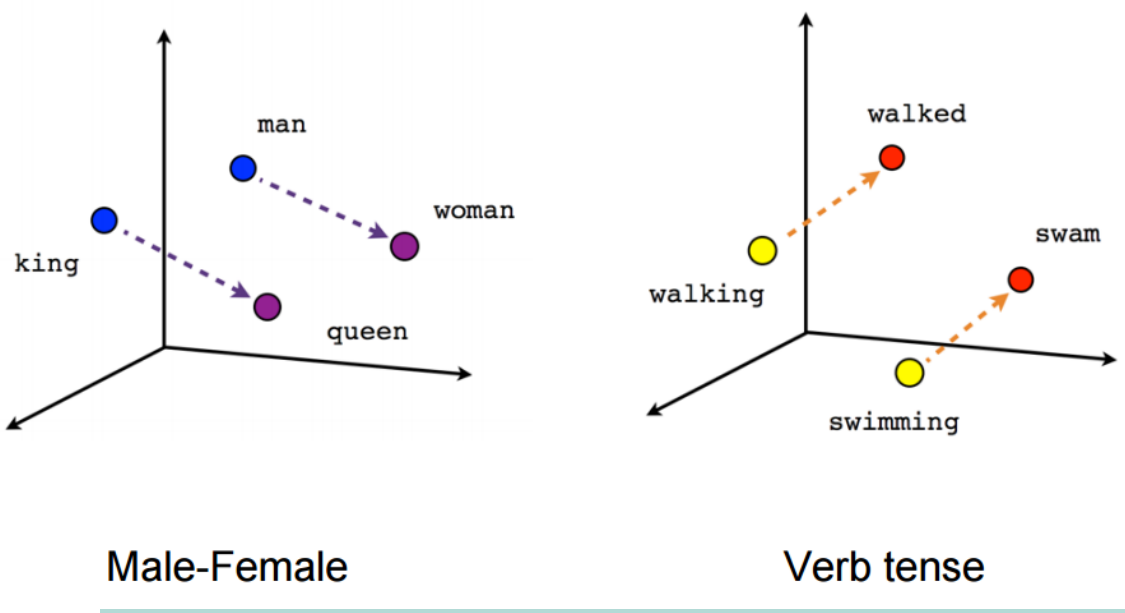


Vektorové reprezentace

$$v_1(\text{buňka}) = (10, 0, 0, 5)$$

4rozměrný prostor (podle počtu domén)

Co když je celý svět jedna doména?



Male-Female

Verb tense

https://towardsdatascience.com/creating-word-embeddings-coding-the-word2vec-algorithm-in-python-using-deep-learning-b337d0ba17a8

Vektorové reprezentace

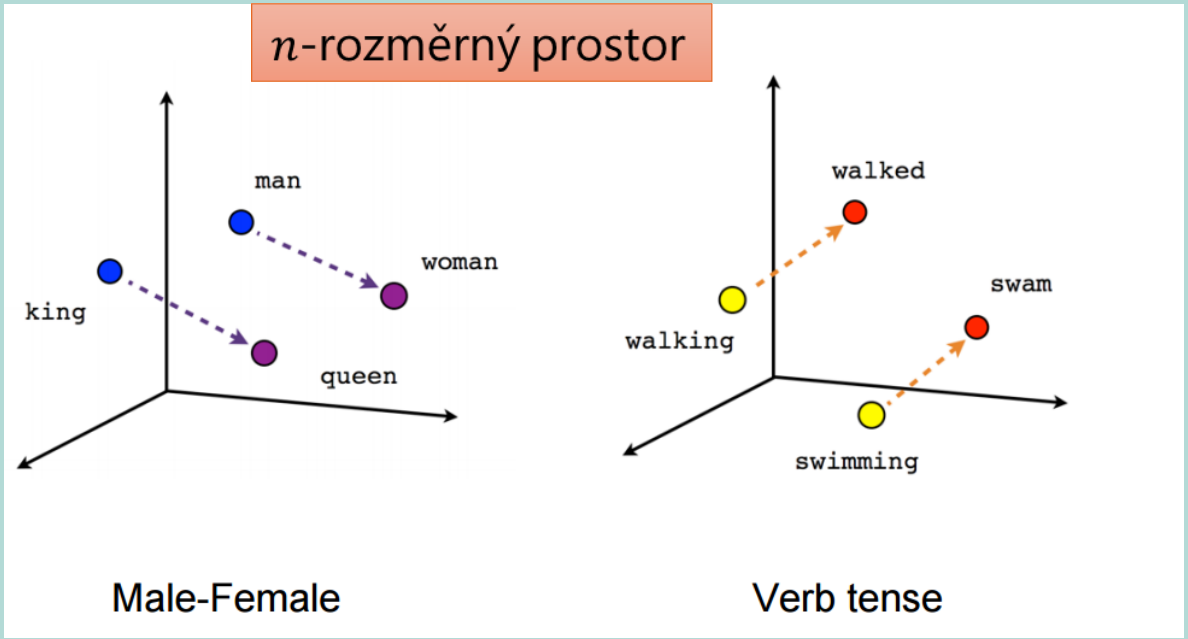
One hot encoding: vektor tvaru $(0, 0, \dots, 1, \dots, 0)$ délky n, kde n je velikost slovníku, všechny úhly jsou pravé.

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



Vektorové reprezentace

Word embedding: výpočet

Vstup: korpus, velikost okna

Výstup: slovník (kódovací tabulka)

slovo + n-rozměrný vektor

	MALE	ADULT	Letectví	POS
Žena	-1	1	3	1
Chlapec	1	-1	0	1
Batole	0	-1	0	1
Let	0	0	9	1
Spadnout	0.4	-0.2	7	5

Žena	-1	1	3	...	...	...	1
Chlapec	1	-1	0	...	...	...	1
Batole	0	-1	0	...	...	...	1
Let	0	0	9	...	...	...	1
Spadnout	0.4	-0.2	7	...	...	...	5

Word Embeddings: výpočet

Matice spoluvýskytů

	the	cat	sat	on	mat	.
the	0.1	0.8	0.4			
cat	0.3	0.2				
sat						
on			…			
mat						
.						

Pohyblivé okno (sliding window)

Word Embeddings: výpočet

Matice spoluvýskytů

	the	cat	sat	on	mat	.
the	0.1	0.8	0.4			
cat	0.3	0.2				
sat						
on			…			
mat						
.						

Cíl: předpovědět slovo, pokud známe kontext.

Možné řešení: vybrat slovo s největší pravděpodobností výskytu

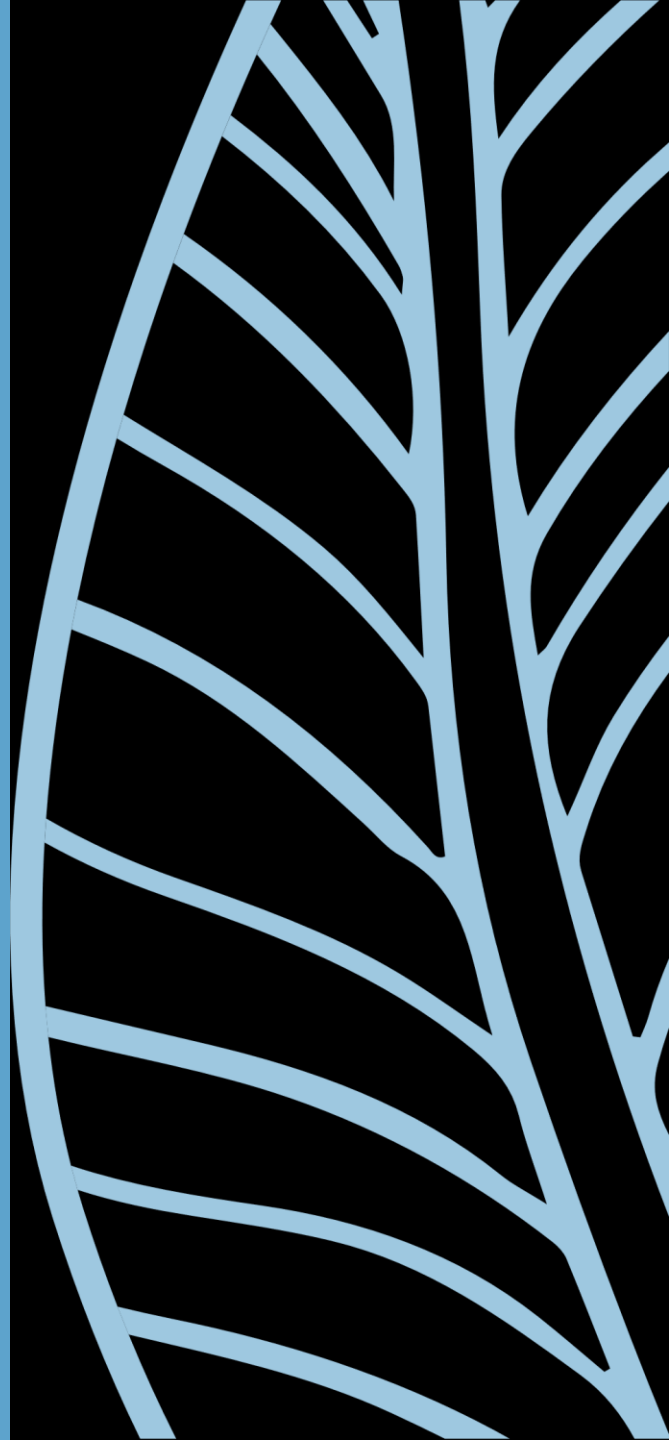
The cat **?** on the mat

Distribuční sémantika

Význam je skryt v datech, které zobrazují užití slov v kontextech.

Není tak důležitá četnost (frekvence) jako rozložení (distribuce)

Jazykové jevy (i jiné sociální jevy) mají specifickou distribuci.



Literatura

- Kenneth Ward Church and Patrick Hanks (March 1990). **"Word association norms, mutual information, and lexicography".** Comput. Linguist. 16 (1): 22–29.
- Collobert, R., Weston, J., Bottou, L., Karlen, M.,Kavukcuoglu, K., and Kuksa, P. (2011). **Natural language processing (almost) from scratch.** J. Mach. Learn. Res., 12:2493–2537.