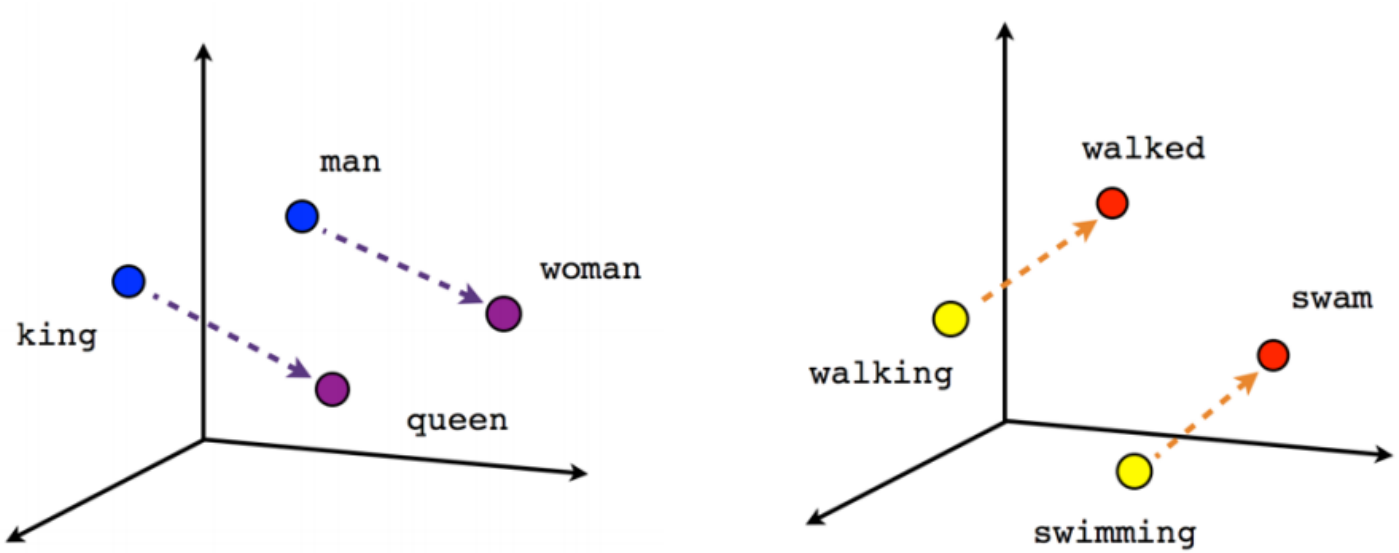
CORE147

DIGITÁLNÍ DATA V HUMANITNÍCH A SOCIÁLNÍCH

Zuzana Nevěřilová
Hana Žižková

2024/25

WORD EMBEDDINGS: NĚCO MYSTICKÉHO?



Male-Female

Verb tense

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

MATEMATICKÁ REPREZENTACE VÝZNAMU SLOV□

Funguje to, jen když platí tyto předpoklady:

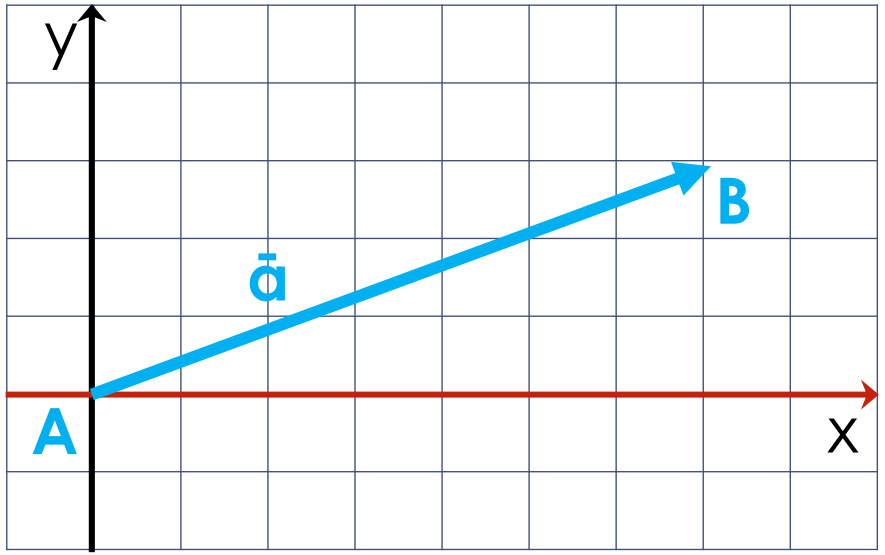
- Slova existují a lidé je opakovaně používají
- Slova se vyskytují s určitou **frekvencí** (tato frekvence má **distribuci – rozdělení**)
- Slova se za sebou vyskytují nenáhodně v posloupnostech (n-gramech)
- Některá slova spolu bývají častěji než jiná (n-gramy mohou mít také frekvenci a distribuci)

slovo → číslo: nepraktické (jsou si nějaká čísla **podobná**?)
slovo → vektor: practické (podobnost vektorů je **úhel**)

VEKTORY AND GEOMETRIE

Vektor: **směr** a **délka**

2D prostor



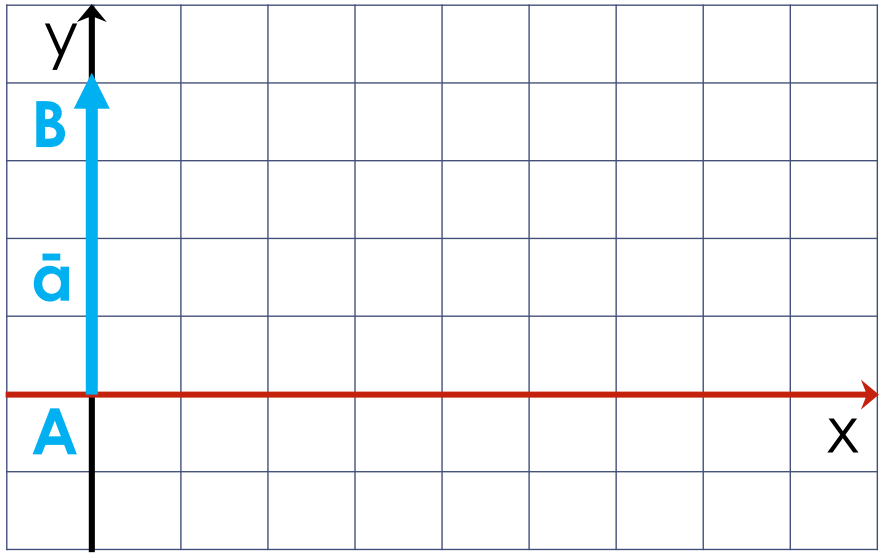
Analytická geometrie: $\bar{a} = (7, 3)$

Euklidovská velikost: $|\bar{a}| = \sqrt{7^2 + 3^2}$

https://despicableme.fandom.com/wiki/Vector

VECTORS AND GEOMETRY

Vektor: **směr** a **délka**



Analytická geometrie: $\bar{a} = (0, 4)$

Euklidovská velikost: $|\bar{a}| = \sqrt{0^2 + 4^2}$

https://despicableme.fandom.com/wiki/Vector

KÓDOVÁNÍ DO VEKTORU

One-hot encoding

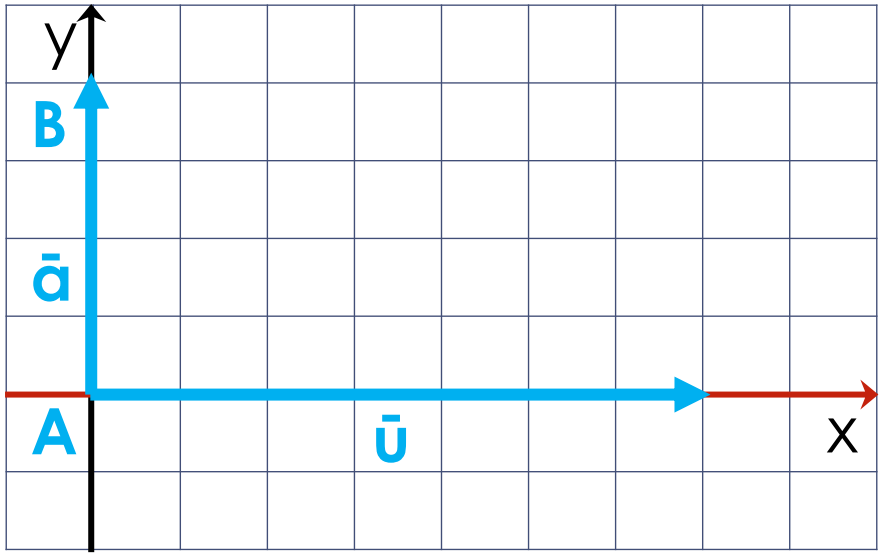
	The	Cat	Sat	On	The	Mat	.
the	1	0	0	0		0	0
cat	0	1	0	0		0	0
sat	0	0	1	0		0	0
on	0	0	0	1		0	0
the	1	0	0	0		0	0
mat	0	0	0	0		1	0
.	0	0	0	0		0	1

Máme ∞ možností, jak to udělat.

Vždycky počítáme s tím, že slovo bude v kontextu.

VLASTNOSTI ONE-HOT ENCODING

- Každý vektor svírá pravý úhel se všemi ostatními.
- Dva rozměry nestačí
- Dimenze = počet různých slov (velikost slovníku)



One-hot nekóduje informaci o významu.
Jediná informace je:

- Slovo je ve slovníku
- Slovo je jiné než jiné slovo

DEKÓDOVÁNÍ ONE HOT

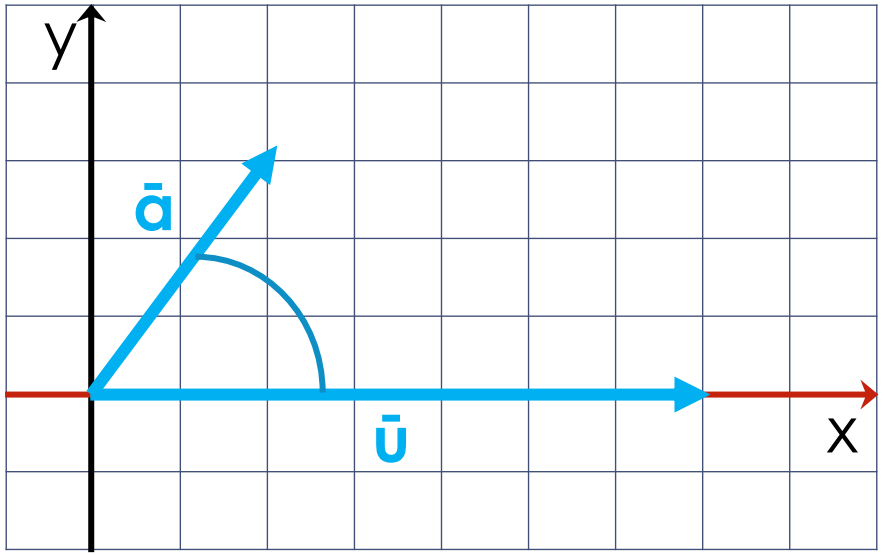
Index („slovník“)

the	1
cat	2
sat	3
on	4
mat	5
.	6

	The	Cat	Sat	On	Mat	.
the	1	0	0	0	0	0
cat	0	1	0	0	0	0
sat	0	0	1	0	0	0
on	0	0	0	1	0	0
the	1	0	0	0	0	0
mat	0	0	0	0	1	0
.	0	0	0	0	0	1

KÓDOVÁNÍ ONE-HOT

- Je velmi snadné
- Nekóduje informaci o významu
- Potřebuje velký vektorový prostor (moc-rozměrné vektory)
- Tento prostor je téměř prázdný (vyplněný nulami)
- Co zlepšit:
  - Využít vektory lépe a úsporněji
  - Zakódovat do vektorů informaci o podobnosti
  - Podobnost se dobře počítá pomocí úhlu, který dva vektory svírají



Word embeddings

JAK SE TRÉNUJÍ WORD EMBEDDINGS

Matice sousednosti
(kolikrát jsou slova vedle sebe)

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

Založeno na kontextu: pohyblivé okno
(sliding window)

UČENÍ REPREZENTACÍ

Model pro učení dostane:

- **cíl učení** (**learning objective**): data + jak by měl vypadat výsledek
- **nákladovou funkci** (**loss function**), která v každém kroku změří, jak se daří

Algoritmus:

1. Rozdělí data na **trénovací** (**training**) a **validační** (**validation**)
2. Navrhne hypotézu, jak doplnit vstup
3. Otestuje výstup a spočítá ztrátu (náklady, **loss**)

} Iterativní proces

UČENÍ REPREZENTACÍ V KONTEXTU

Kontext	Cíl
(-, cat)	the
(the, sat)	cat
(cat, on)	sat
(sat, the)	on
(on, mat)	the

pohyblivé okno (sliding window)

Cíl učení (objective): jaké je cílové slovo w_j v kontextu slov w_i

Nákladová funkce (loss): kolikrát model tipnul správně cílové slovo ve validační množině

UČENÍ REPREZENTACÍ V KONTEXTU

Výsledkem jsou tři matice:

- „slovník“, index (vocabulary matrix V)
- embedding matrix E
- matice sousednosti – kontextová matice (context matrix C)

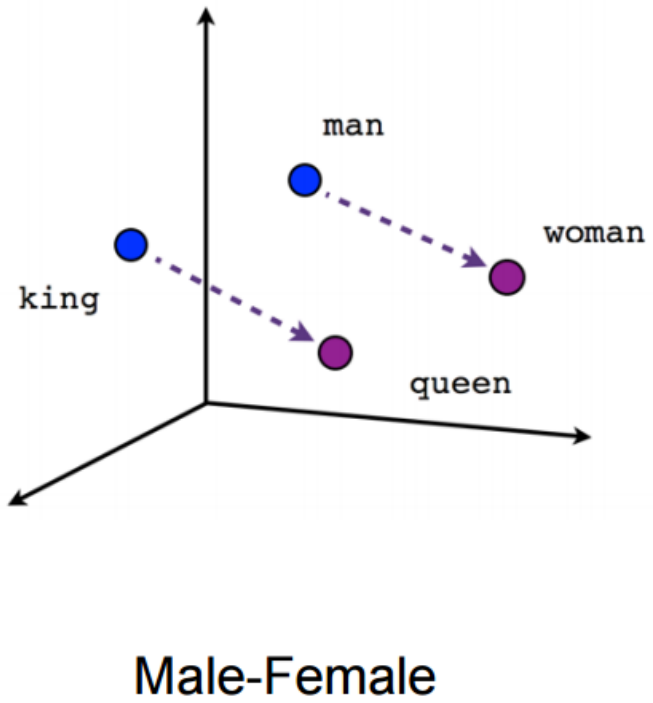
Podrobněji, s animacemi: https://lena-voita.github.io/nlp_course/word_embeddings.html

UČENÍ REPREZENTACÍ V KONTEXTU

vocabulary V – embedding E – context C

Výsledek: embedding matrix E = množina embeddings pro každé slovo ve slovníku (vocabulary) V

- V se vytvoří na základě slov v trénovací množině (jiné embeddings mají jiné V)
- C se spočítá podle spoluvýskytů v trénovacích datech
- E se počítá iterativně (napoprvé si skoro nikdy netipne správné slovo)
- Sémantická podobnost odráží podobné kontexty:
  - koupit auto – koupit pomeranče – sníst pomeranče – sníst párek – drůbeží párek – drůbeží chřipka – ptačí chřipka …



K DALŠÍMU ČTENÍ

- 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.
- https://lilianweng.github.io/lil-log/2017/10/15/learning-word-embedding.html
- https://lena-voita.github.io/nlp_course/word_embeddings.html