

Využitie vektorového modelu slov pri spracovaní textu na webe

Bakalársky projekt

Ladislav Gallay, Ing. Marián Šimko, Phd.

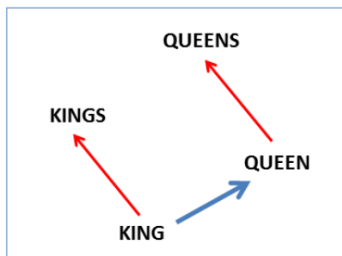
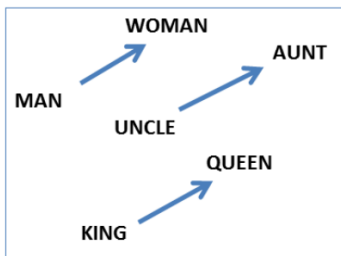
02.11.2014

Lematizácia

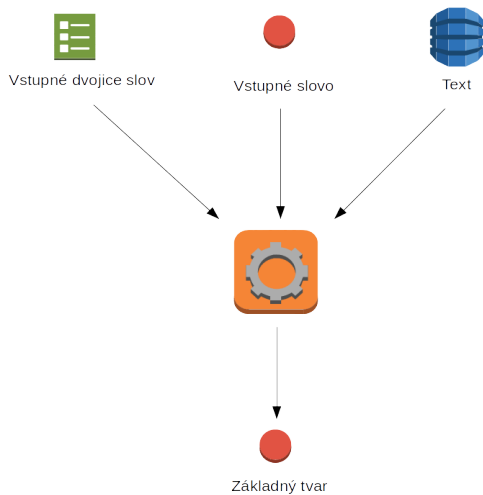
- určiť základ slova
- porozumenie textu, vete, slovám
- slovníkový prístup
 - databáza dvojíc: *domom* -> *dom*
 - pre slovenčinu: 2 472 051 tokenov, 76 856 unikátnych slov
- pravidlový prístup
 - podľa koncoviek: *dom**om*** -> *dom*
 - slovenčina je komplikovaná a veľa výnimiek
 - *cti* -> *čeť*

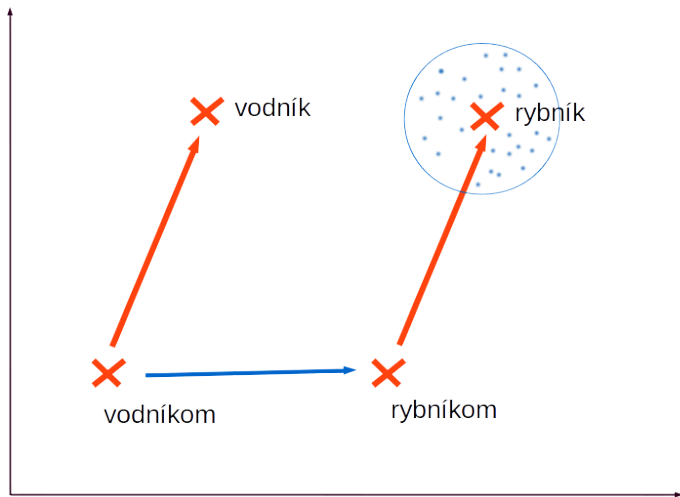
Word2Vec

- vstup: text corpus (predspracovaný)
- výstup: slová ako vektory
- funkcie:
 - distance: najbližšie slová
 - word_analogy: slovo A ku B je ako C ku X
 - muž ku žene je ako kráľ ku **kráľovnej**
 - vodníkom ku vodník je ako rybníkom ku **rybník**



Lematizácia pomocou analógie slov





```
dvojice = vyber_vhodnych_dvojic( vstupne_slovo )

for dvojica in dvojice
{
    vysledky = word_analogy( dvojica[0], dvojica[1],
                             vstupne_slovo )

    for vysledok in vysledky
    {
        zakladne_tvary[ vysledok[slovo] ].presnost +=
            ohodnot( vysledok, vstupne_slovo )
    }
}

return zakladne_tvary.top
```

Výber vhodných dvojíc slov

- **vodníkom vodník** rybníkom -> rybník
- gramaticky najbližšie - podľa suffixu
 - autom: mest**om**, autobus**om**, chlap**om**
- náhodne
 - pri veľkom množstve vyberiem aj gramaticky vhodné dvojice
 - ďalším processingom eliminujem zlé výsledky
- najstť najbližšie slová (*distance*), ktoré sú aj vo vstupných dvojiciach slov
 - predpokladá väčšiu množinu vstupných dvojíc
 - autom: autobus/om, *doprava/ou*, vodič/om, ...

Ohodnotenie výsledku

- funkcia **word_analogy** vráti slovo a jeho vzdialenosť
- podľa prefixu viem odhadnúť či je slovo dobré alebo nie
 - **rybníkom** -> **auto**
 - nemusí vždy fungovať: **krajší** -> **pekný**
 - zatiaľ iba podstatné mená

$$presnosť = \frac{prefix(A,B)}{(strlen(A)+strlen(B))/2} * vzdialenosť$$

- ak slová nemajú ani prvé písmeno zhodné: *zhoda* = 0
- niekoľko iterácií pre **word_analogy** a rôzne dvojice
- *presnosti* pre rovnaké výsledky sa sčítajú
- slovo s najväčšou *presnosťou* označím ako výsledné

Text corpus

Zdroj	Počet viet	Počet slov	Počet unikátnych slov
SME	2 727 889	37 426 758	714 139
SK WIKI	2 340 665	23 814 184	855 743
Slov.nar.kor.		655 572 511	

Slovenčina má cez 1 000 000 slov*

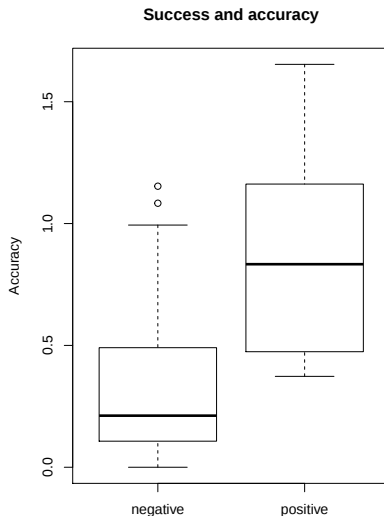
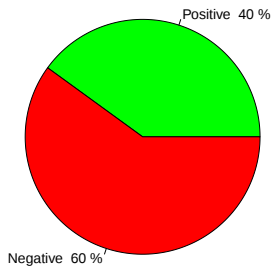
Otázky

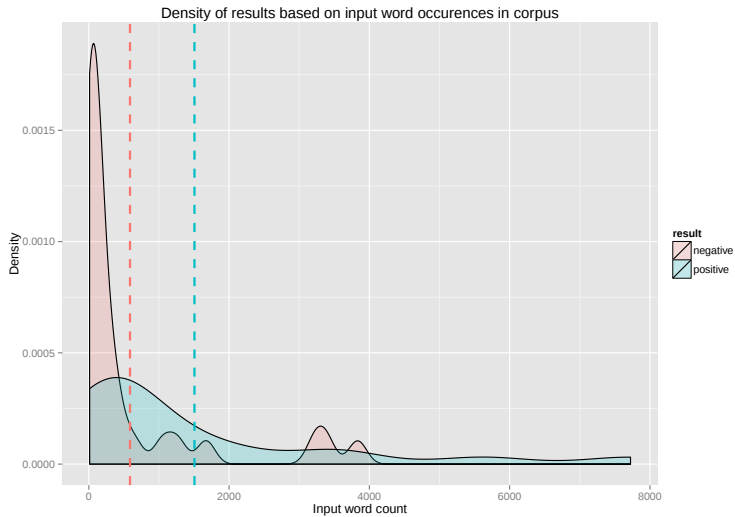
- Aký vstupný text?
- Akú veľkú množinu vstupných dvojíc?
- Ako vybrať vhodné dvojice pre konkrétny vstup?
- Ako hodnotiť presnosť výsledkov?
- Aké slová viem najlepšie určovať? (napr. podstatné mená, inštrumentál, ...)

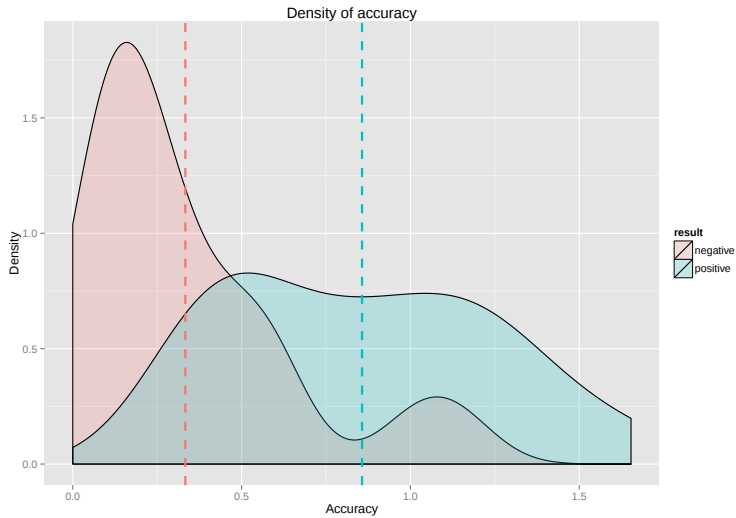
Testovanie

- dataset: SME (bez diakritiky)
- vstupné dvojice: vyskloňované vzory podstatných mien
- vstupné slová: podstatné mená
- 3 najlepšie dvojice podľa suffixu pre každý vstup
- výstupom je jedno slovo s najväčšou presnosťou
- hodnotí sa či, výstup je základným tvarom vstupného slova (true/false)

Test 1 - 50 vstupných slov







	meta	input	correct	output	match	inputCount	correctCount	outputCount	result
1	SSns2	auta	auto	auto	1.653317	7723	6796	6796	positive
2	SSis7	autom	auto	auto	0.946284	1865	6796	6796	positive
3	SSnp7	autami	auto	auto	0.590178	456	6796	6796	positive
4	SSis1	aut	auto	auto	0.809896	3478	6796	6796	positive
5	SSis2	autu	auto	auto	0.992861	385	6796	6796	positive
6	SSms2	muza	muz	mladika	0.388859	3386	5513	346	negative
7	SSms7	muzom	muz	muz	0.536113	912	5513	5513	positive
8	SSms3	muzovi	muz	muz	0.856457	724	5513	5513	positive
9	SSmp7	muzmi	muz	mladsimi	0.062777	371	5513	79	negative
10	SSmp1	muzi	muz	malokalibrova	0.106957	3833	5513	5	negative
11	SSfs3	lampe	lampa	listovani	0.169936	38	135	14	negative
12	SSfs7	lampou	lampa	reprezentativna	0.000000	79	135	37	negative
13	SSfp7	lampami	lampa	laket	0.162035	16	135	67	negative
14	SSfs4	lampu	lampa	lebku	0.106992	59	135	54	negative
15	SSfs2	stolicky	stolicka	stoliciek	0.545937	412	105	110	negative
16	SSfs7	stolickou	stolicka	standa	0.145246	23	105	7	negative
17	SSfp7	stolickami	stolicka	stopkach	0.168391	34	105	21	negative
18	SSfs2	kvetiny	kvetina	kamoska	0.226570	58	17	62	negative
19	SSfs7	kvetinou	kvetina	kopirka	0.194282	11	17	9	negative
20	SSfp7	kvetinami	kvetina	kozu	0.075712	17	17	569	negative
21	SSfs2	steny	stena	stena	1.387872	1178	351	351	positive
22	SSfs3	stene	stena	stena	0.760761	562	351	351	positive
23	SSfp3	stenam	stena	stekali	0.196852	16	351	30	negative
24	SSfp7	stenami	stena	stena	0.404389	152	351	351	positive