IY-Corrector

IY-Corrector
Martin Majliš

Možnosti

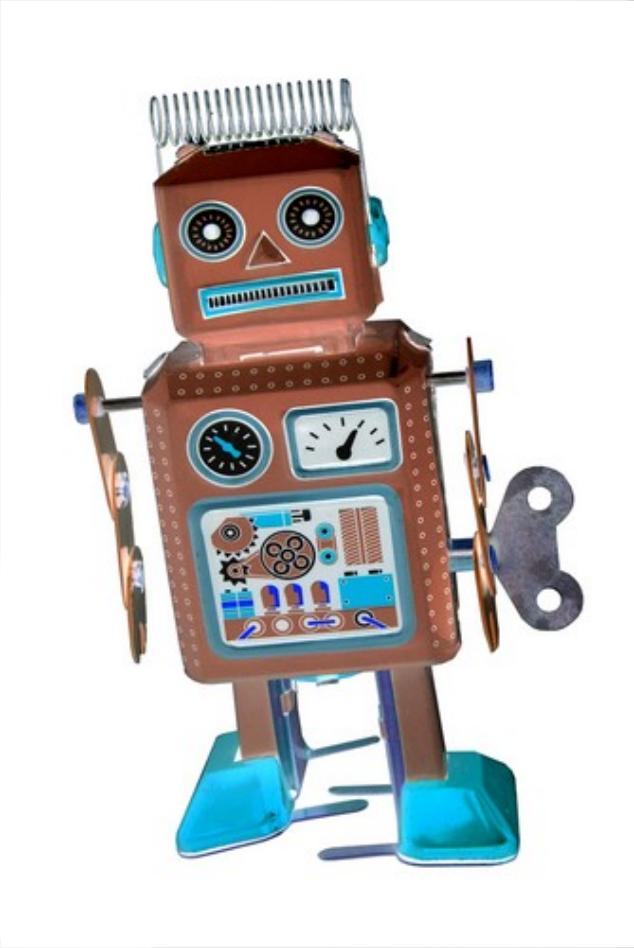
- Pravidla
  - Vyjmenovaná/cizí slova, jména
- Strojový překlad
  - Z češtyni do češtiny
- N-gramový jazykový model

Pravidla

- Problémy
  - zdroje
- Výhody
  - ???

Strojový překlad

- Celá řada frameworků
- Paralelní korpus

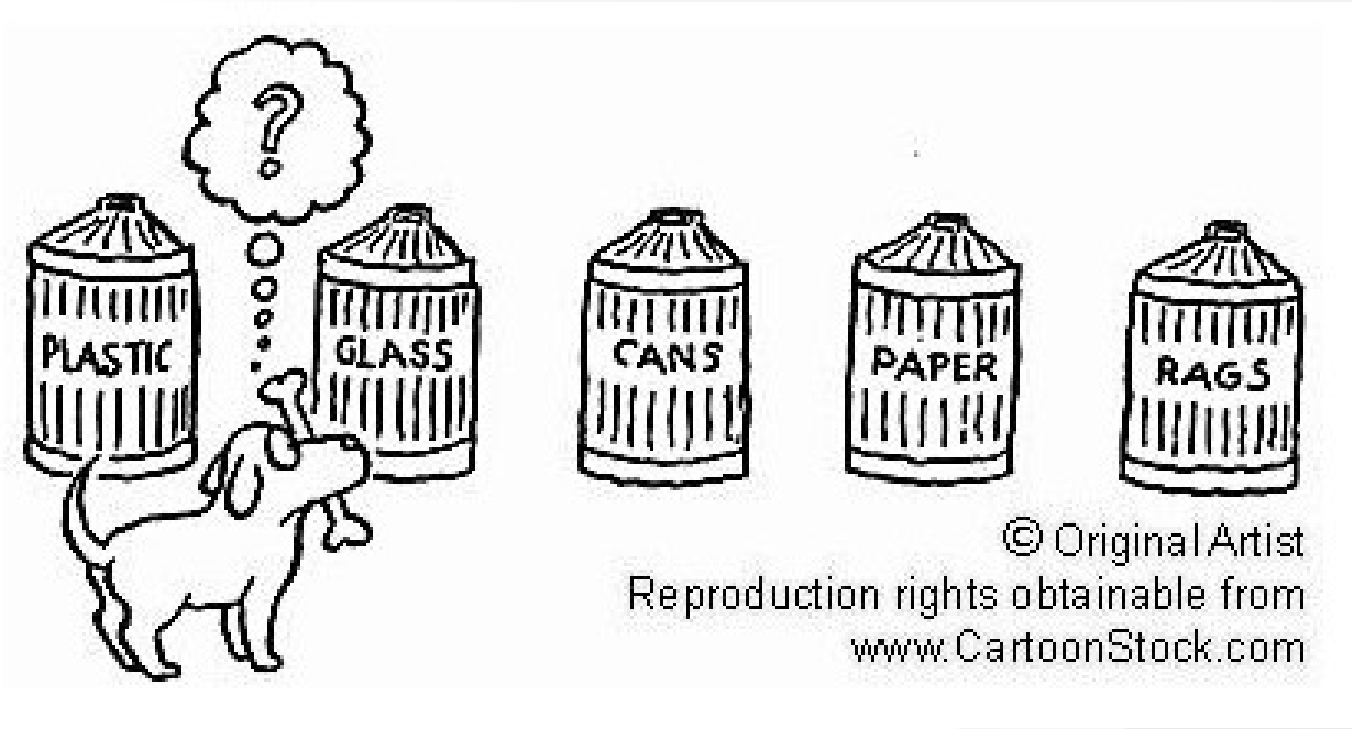


N-gramový model

- Korpus z ukolu 1
- Nastavení parametrů
  - Jak zvolit n?
  - Co s velikostí písmen?
  - Používat slova/znaky?

Volba parametrů

- Žádná předchozí zkušenost
- Vyzkoušet co nejvíc možností



Volba N

- Slova
  - 1 – 4
- Znaky
  - 1 - 7

Velikost písmen

- Těžko říct
  - 1 – slova
  - 2 – lowercase slova
  - 3 – znaky
  - 4 – lowercase znaky

Velikost vstupu

- Trénink:
  - 312 500, 625 000, 1 250 000
  - 2 500 000, 5 000 000, 10 000 000
- Vyhlazování: 5%
- Test: konstantní

Vstupní data

- 22M slov z Wikipedie
  - 20M – trénování
  - 1M vyhlazení
  - 350k testování

Testování kvality

- Testovací data
  - Vymyslet
  - Vygenerovat
- doCestyni.pl N prob
  - N – počet chyb
  - prob – prav. Udělaní chyby

Výpočet

Mód 1

- Slova
- Žádná modifikace

	1	2	3	4
312500	8462	8200	8199	8199
625000	6865	6581	6583	6583
1250000	5690	5386	5388	5388
2500000	4729	4427	4427	4427
5000000	3802	3471	3469	
10000000	3328	2982		

Mód 2

- Slova
- Převod na malá písmena

	1	2	3	4
312500	8005	7737	7735	7735
625000	6485	6192	6191	6190
1250000	5362	5052	5058	5058
2500000	4461	4148	4152	4152
5000000	3612	3268	3264	
10000000	3207	2837		

Mód 3

- Písmena
- Žádná modifikace

Data	1	2	3	4	5	6	7
312500	26048	17618	13545	11269	9915	9567	9564
625000	26048	17562	13040	10583	8579	8177	8162
1250000	26048	17554	13038	10068	7900	7374	

Mód 4

- Písmena
- Převod na malá písmena

Data	1	2	3	4	5	6	7
312500	26048	17903	13519	11484	10080	9781	9761
625000	26048	18320	13297	10884	8892	8410	8357
1250000	26048	18307	13341	10560	8202	7631	

Celkově

Sum - Diffs		Data	Data				
		312500	625000	1250000	2500000	5000000	10000000
N	Mode	312500	625000	1250000	2500000	5000000	10000000
1	1	8462	6865	5690	4729	3802	3328
	2	8005	6485	5362	4461	3612	3207
	3	26048	26048	26048			
	4	26048	26048	26048			
2	1	8200	6581	5386	4427	3471	2982
	2	7737	6192	5052	4148	3268	2837
	3	17618	17562	17554			
	4	17903	18320	18307			
3	1	8199	6583	5388	4427	3469	
	2	7735	6191	5058	4152	3264	
	3	13545	13040	13038			
	4	13519	13297	13341			
4	1	8199	6583	5388	4427		
	2	7735	6190	5058	4152		
	3	11269	10583	10068			
	4	11484	10884	10560			
5	3	9915	8579	7900			
	4	10080	8892	8202			
6	3	9567	8177	7374			
	4	9781	8410	7631			
7	3	9564	8162				
	4	9761	8357				

Problémy

- Nedostatek paměti
  - Místo pro optimalizaci
- Kódování
  - binmode, use utf8

Otázky