Natural Language Processing

Zofia Długosz
Michał Rajewski

- NLP – podstawy

NLP

- Dziedzina zajmująca się analizą, rozumieniem oraz generowaniem języka naturalnego
- Celem jest umożliwienie komputerom posługiwanie się ludzkim językiem

Tokenization

Tokenizacja – rozdzielenie tekstu na określone fragmenty mowy (słowa, zdania, frazy), czyli tokeny.

Problemy: „Mr. Brown” : „Mr.”, „Brown”

Part of Speech

Rozpoznawanie części mowy – przypisywanie oznaczeń części mowy do słów

Lemmatization

Lematyzacja – sprowadzenie słowa do formy podstawowej:

- Czasownik – bezokolicznik
- Rzeczownik – mianownik liczby pojedynczej

Stemming

Rdzeniowanie – obcięcie przedrostków i przyrostków w celu uzyskania „rdzenia słowa”. Uzyskany „rdzeń” nie musi być poprawnym wyrazem.

Problemy:

- Understemming – jumped, jumps = jump; jumpiness = jumpi
- Overstemming – general = gener; generate = gener

Chunking

Fragmentyzacja – wydobywanie znaczących fraz z tekstu.

Jupyter Notebook

Bag of Words

Reprezentacja tekstu jako słownik, w którym kluczem jest słowo (efekt wstępnego przetwarzania) a kluczem ilość wystąpień w tekście (zdaniu, fragmencie).

the dog is on the table

are	cat	dog	is	now	on	table	the
0	0	1	1	0	1	1	1

Document 1

The quick brown fox jumped over the lazy dog's back.

Document 2

Now is the time for all good men to come to the aid of their party.

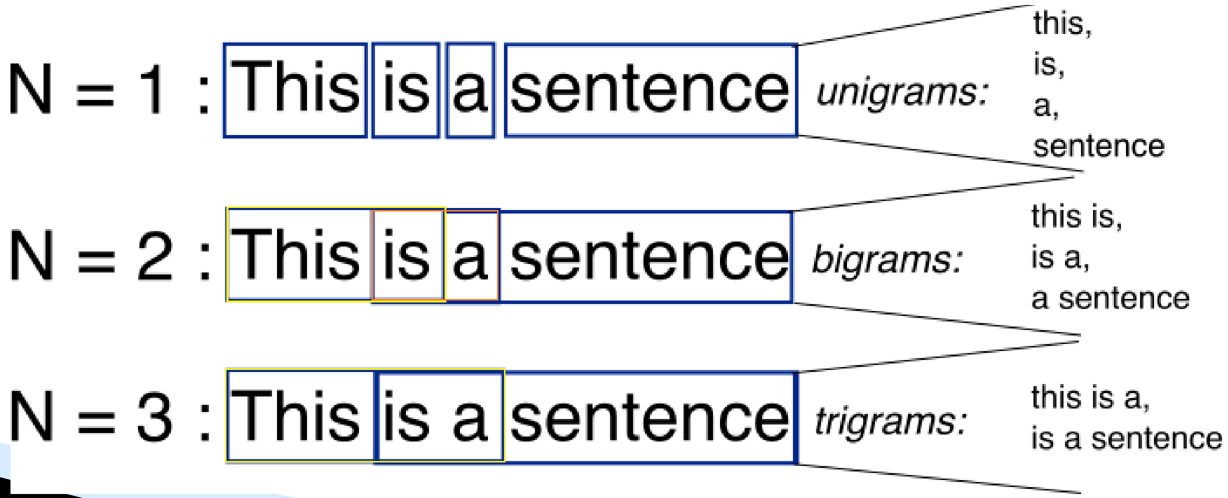
Term	Document 1	Document 2
aid	0	1
all	0	1
back	1	0
brown	1	0
come	0	1
dog	1	0
fox	1	0
good	0	1
jump	1	0
lazy	1	0
men	0	1
now	0	1
over	1	0
party	0	1
quick	1	0
their	0	1
time	0	1

N-gram

Model dzielący zdanie na grupy o długości n

- Obliczanie prawdpodobieństwa wystąpienia grupy
- Obliczanie prawdopodobieństwa wystąpienia zdania

```
P('There was heavy rain') = P('There', 'was', 'heavy', 'rain')
P('There was heavy rain') = P('There')P('was'|'There')P('heavy'|'There was')P('rain'|'There was heavy')
```

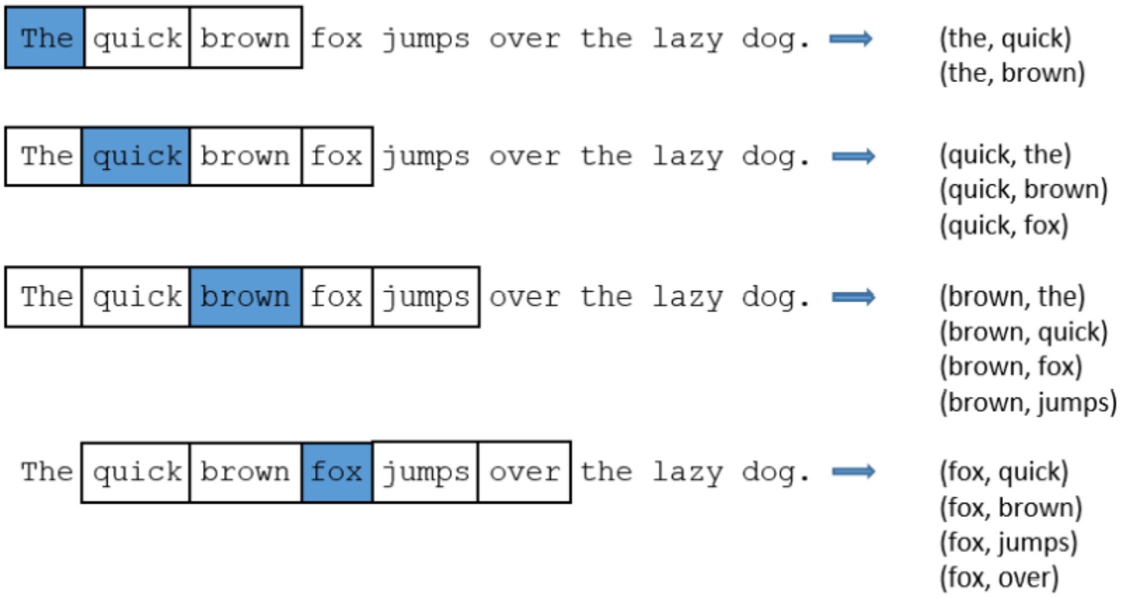


Skip-gram

Modyfikacja modelu N-gram szukająca grup o długości *n* oraz odległości *k*

- Odkrywanie kontekstu

1-skip-2-gram



The quick brown fox jumps over the lazy dog. ⟹ (the, quick) (the, brown)

The quick brown fox jumps over the lazy dog. ⟹ (quick, the) (quick, brown) (quick, fox)

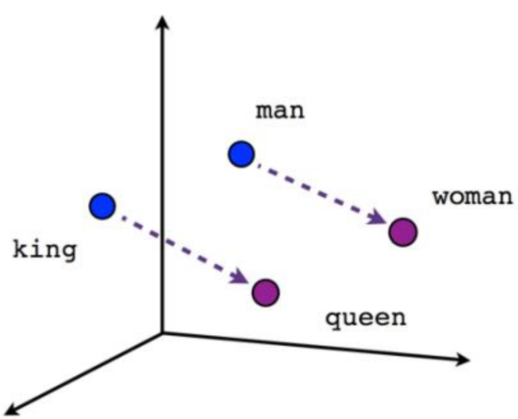
The quick brown fox jumps over the lazy dog. ⟹ (brown, the) (brown, quick) (brown, fox) (brown, jumps)

The quick brown fox jumps over the lazy dog. ⟹ (fox, quick) (fox, brown) (fox, jumps) (fox, over)

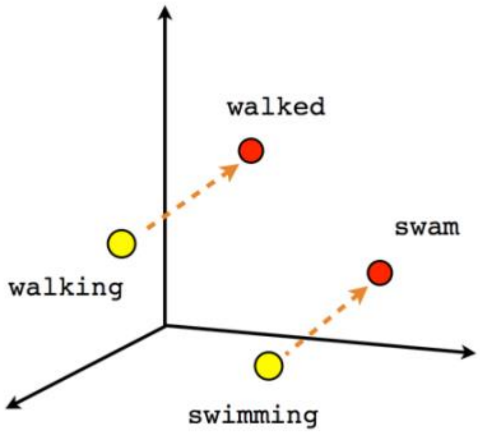
- Word Embedding

Teoria

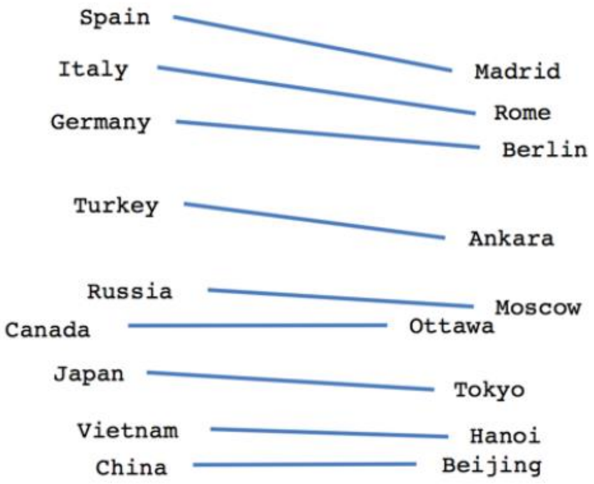
Word embedding – proces mapowania słów do wektora liczb. Umożliwia to wyrażenie słów w formie, w której słowa o podobnym znaczeniu mają podobną reprezentację.



Male-Female



Verb tense



Country-Capital

- word2vec

Teoria

Model, którego zadaniem jest uzyskiwanie postaci Word Embeddings. Bazuje on na dwuwarstwowej sieci neuronowej.

Dla każdego słowa tworzony jest wektor, który je reprezentuje.

Twórca: Google

Jupyter Notebook

- fastText

Teoria

Tworzony wektor to suma n-gramów reprezentujących znaki

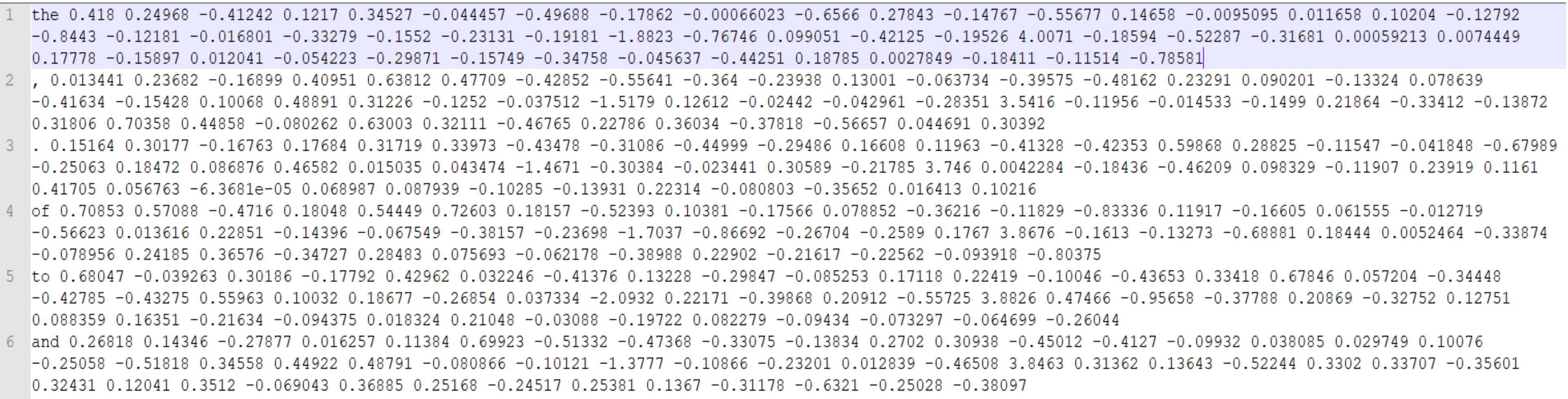
- Przykład: „apple” – “<ap”, “app”, ”appl”, ”apple”, ”apple>”, “ppl”, “pple”, ”pple>”, “ple”, ”ple>”, ”le>”

Jupyter Notebook

GloVe

https://nlp.stanford.edu/projects/glove/

- Wikipedia 2014 + Gigaword 5 (6B tokens, 400K vocab, uncased, 50d, 100d, 200d, & 300d vectors, 822 MB download): glove.6B.zip
- Common Crawl (42B tokens, 1.9M vocab, uncased, 300d vectors, 1.75 GB download): glove.42B.300d.zip
- Common Crawl (840B tokens, 2.2M vocab, cased, 300d vectors, 2.03 GB download): glove.840B.300d.zip
- Twitter (2B tweets, 27B tokens, 1.2M vocab, uncased, 25d, 50d, 100d, & 200d vectors, 1.42 GB download): glove.twitter.27B.zip



```
the 0.418 0.24968 -0.41242 0.1217 0.34527 -0.044457 -0.49688 -0.17862 -0.00066023 -0.6566 0.27843 -0.14767 -0.55677 0.14658 -0.0095095 0.011658 0.10204 -0.12792
-0.8443 -0.12181 -0.016801 -0.33279 -0.1552 -0.23131 -0.19181 -1.8823 -0.76746 0.099051 -0.42125 -0.19526 4.0071 -0.18594 -0.52287 -0.31681 0.00059213 0.0074449
0.17778 -0.15897 0.012041 -0.054223 -0.29871 -0.15749 -0.34758 -0.045637 -0.44251 0.18785 0.0027849 -0.18411 -0.11514 -0.78581
, 0.013441 0.23682 -0.16899 0.40951 0.63812 0.47709 -0.42852 -0.55641 -0.364 -0.23938 0.13001 -0.063734 -0.39575 -0.48162 0.23291 0.090201 -0.13324 0.078639
-0.41634 -0.15428 0.10068 0.48891 0.31226 -0.1252 -0.037512 -1.5179 0.12612 -0.02442 -0.042961 -0.28351 3.5416 -0.11956 -0.014533 -0.1499 0.21864 -0.33412 -0.13872
0.31806 0.70358 0.44858 -0.080262 0.63003 0.32111 -0.46765 0.22786 0.36034 -0.37818 -0.56657 0.044691 0.30392
. 0.15164 0.30177 -0.16763 0.17684 0.31719 0.33973 -0.43478 -0.31086 -0.44999 -0.29486 0.16608 0.11963 -0.41328 -0.42353 0.59868 0.28825 -0.11547 -0.041848 -0.67989
-0.25063 0.18472 0.086876 0.46582 0.015035 0.043474 -1.4671 -0.30384 -0.023441 0.30589 -0.21785 3.746 0.0042284 -0.18436 -0.46209 0.098329 -0.11907 0.23919 0.1161
0.41705 0.056763 -6.3681e-05 0.068987 0.087939 -0.10285 -0.13931 0.22314 -0.080803 -0.35652 0.016413 0.10216
of 0.70853 0.57088 -0.4716 0.18048 0.54449 0.72603 0.18157 -0.52393 0.10381 -0.17566 0.078852 -0.36216 -0.11829 -0.83336 0.11917 -0.16605 0.061555 -0.012719
-0.56623 0.013616 0.22851 -0.14396 -0.067549 -0.38157 -0.23698 -1.7037 -0.86692 -0.26704 -0.2589 0.1767 3.8676 -0.1613 -0.13273 -0.68881 0.18444 0.0052464 -0.33874
-0.078956 0.24185 0.36576 -0.34727 0.28483 0.075693 -0.062178 -0.38988 0.22902 -0.21617 -0.22562 -0.093918 -0.80375
to 0.68047 -0.039263 0.30186 -0.17792 0.42962 0.032246 -0.41376 0.13228 -0.29847 -0.085253 0.17118 0.22419 -0.10046 -0.43653 0.33418 0.67846 0.057204 -0.34448
-0.42785 -0.43275 0.55963 0.10032 0.18677 -0.26854 0.037334 -2.0932 0.22171 -0.39868 0.20912 -0.55725 3.8826 0.47466 -0.95658 -0.37788 0.20869 -0.32752 0.12751
0.088359 0.16351 -0.21634 -0.094375 0.018324 0.21048 -0.03088 -0.19722 0.082279 -0.09434 -0.073297 -0.064699 -0.26044
and 0.26818 0.14346 -0.27877 0.016257 0.11384 0.69923 -0.51332 -0.47368 -0.33075 -0.13834 0.2702 0.30938 -0.45012 -0.4127 -0.09932 0.038085 0.029749 0.10076
-0.25058 -0.51818 0.34558 0.44922 0.48791 -0.080866 -0.10121 -1.3777 -0.10866 -0.23201 0.012839 -0.46508 3.8463 0.31362 0.13643 -0.52244 0.3302 0.33707 -0.35601
0.32431 0.12041 0.3512 -0.069043 0.36885 0.25168 -0.24517 0.25381 0.1367 -0.31178 -0.6321 -0.25028 -0.38097
```