NLP

Wydział Biochemii, Biofizyki i Biotechnologii
Adrian Kania

Jakie znamy reprezentacje tekstu?

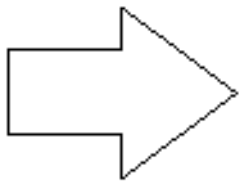
- Set of Words (SoW)
- Bag of Words (BoW)

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

Numeryczna reprezentacja tekstu - One hot encoding

```
 'quick'
  'dog'
 'brown'
  'over'
   'the'
'jumped'
  'lazy'
   'fox'
```



```
[[1, 0, 0, 0, 0, 0, 0, 0],
 [0, 1, 0, 0, 0, 0, 0, 0],
 [0, 0, 1, 0, 0, 0, 0, 0],
 [0, 0, 0, 1, 0, 0, 0, 0],
 [0, 0, 0, 0, 1, 0, 0, 0],
 [0, 0, 0, 0, 0, 1, 0, 0],
 [0, 0, 0, 0, 0, 0, 1, 0],
 [0, 0, 0, 0, 0, 0, 0, 1]]
```

Word Embedding

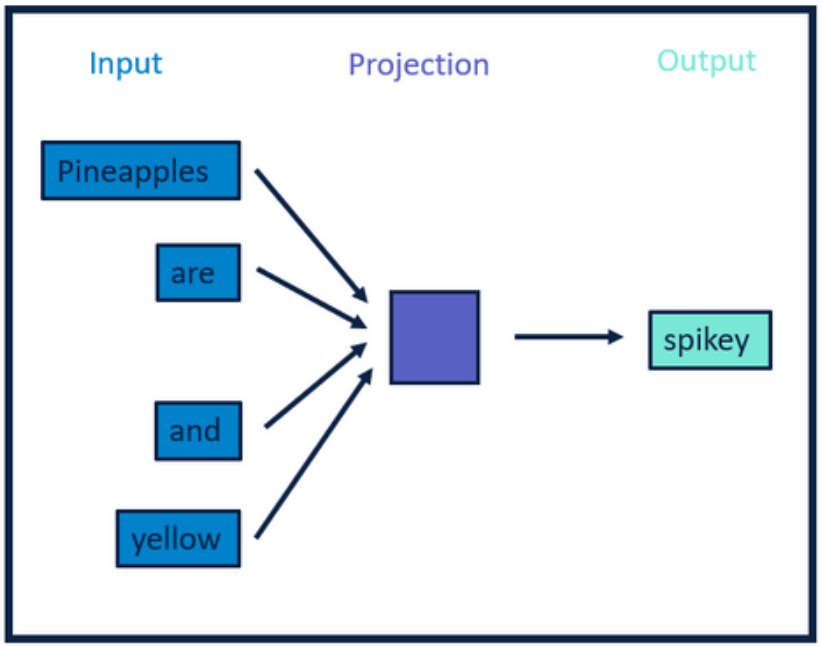
- Na czym polega? Każde słowo przekształcamy w wektor (tzw. **embedding**)

$$dog \rightarrow [1.5, -0.3, 0.12, ...]$$

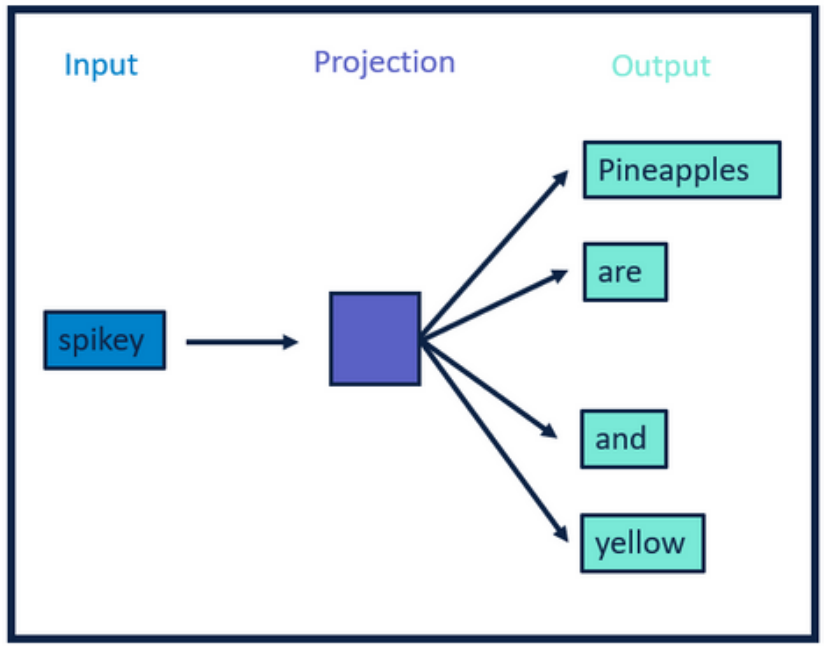
- Idea: znaczenie słowa zależy od kontekstu i/lub innych słów, które występują w pobliżu
- Konsekwencja: Podobne słowa powinny mieć podobne embeddingi.

Embeddingi

- **word2vec** - technika wyznaczania embeddingów



CBOW



Skip-gram

- **GloVe** - zbiór embeddingów, które ktoś dla nas wcześniej wytrenował

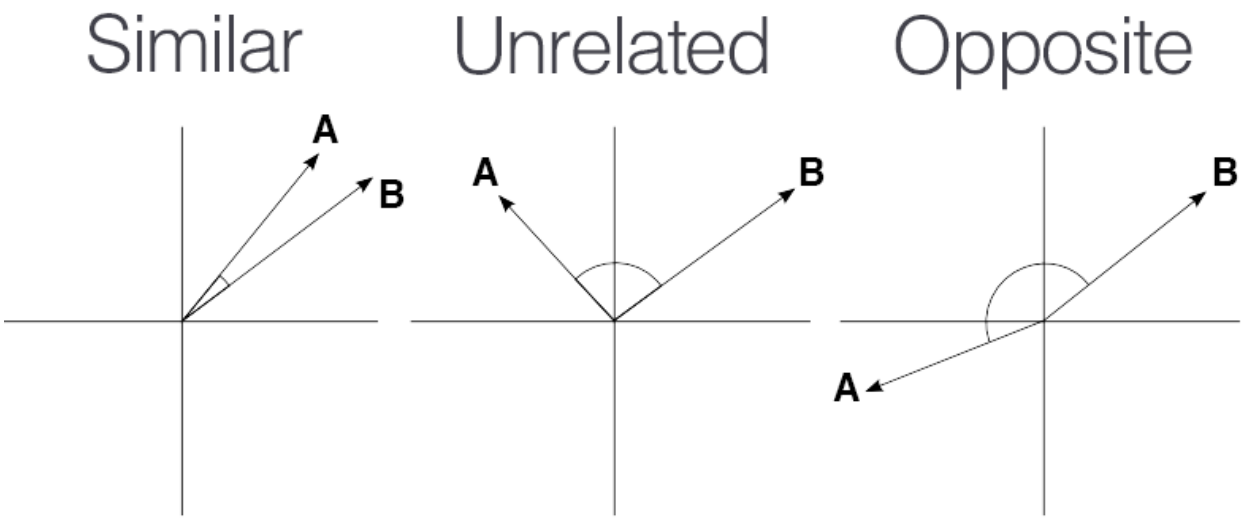
Jak zmierzyć czy dwa wektory są do siebie podobne?

- Odległość Euklidesowa

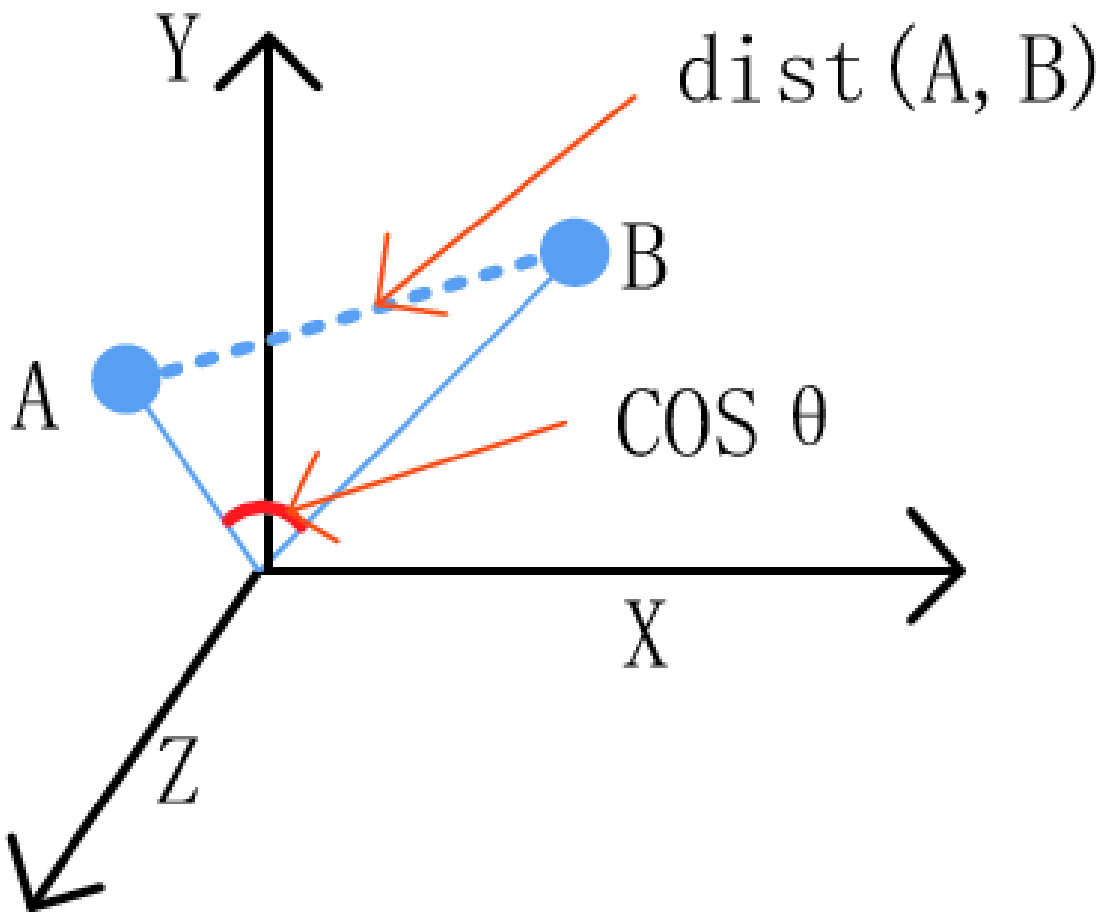
$$dist(v, w) = ||v - w|| = \sqrt{\sum_i (v_i - w_i)^2}$$

- Podobieństwo kosinusowe

$$\cos(v, w) = \frac{vw}{||v|| \cdot ||w||}$$



Jak zmierzyć czy dwa wektory są do siebie podobne?



N-gramy + sieci neuronowe