N-gramy

mgr inż. Dawid Wiśniewski - Przetwarzanie języka naturalnego

16 March 2021

Przypomnienie i

W poprzednim odcinku:

Klasyfikacja z wykorzyzstaniem reprezentacji Bag-of-Words.

Przypomnienie ii

Przykłady dokumentów:

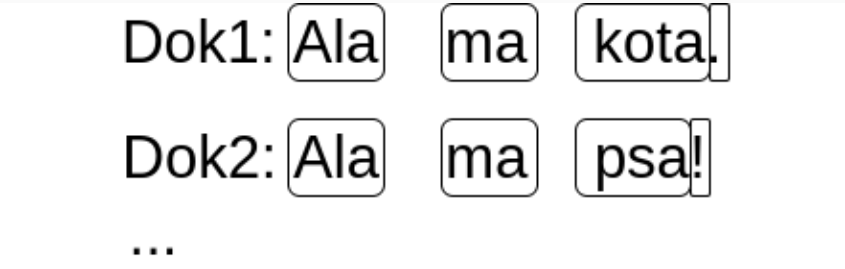
Dok1: Ala ma kota.

Dok2: Ala ma psa!

...

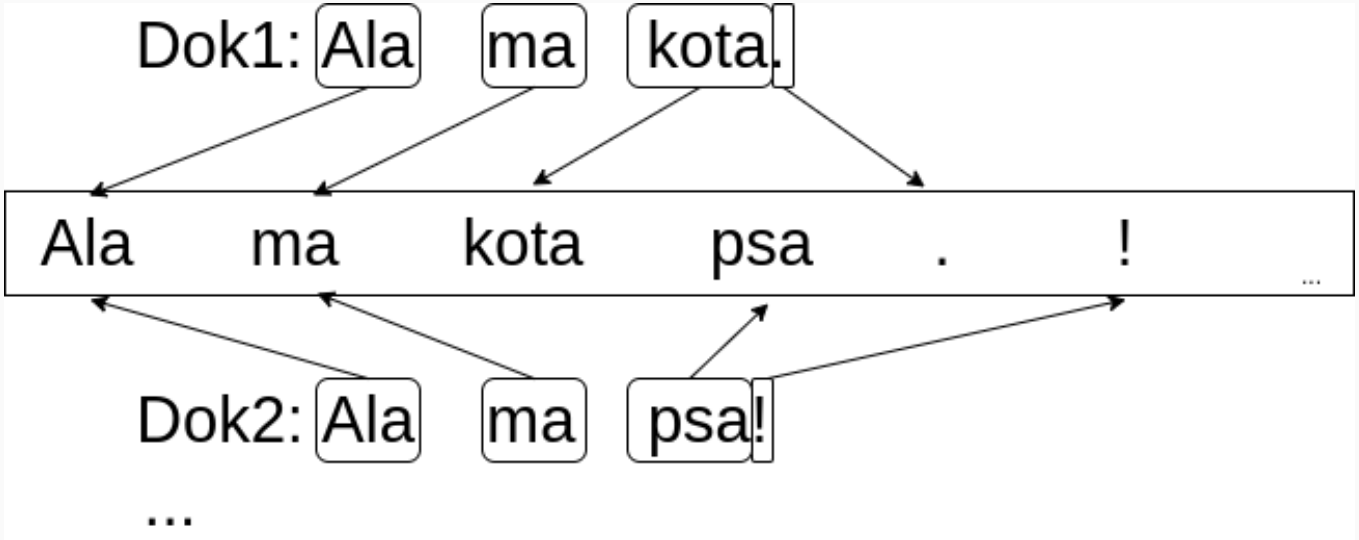
Przypomnienie iii

Krok 1: Tokenizacja



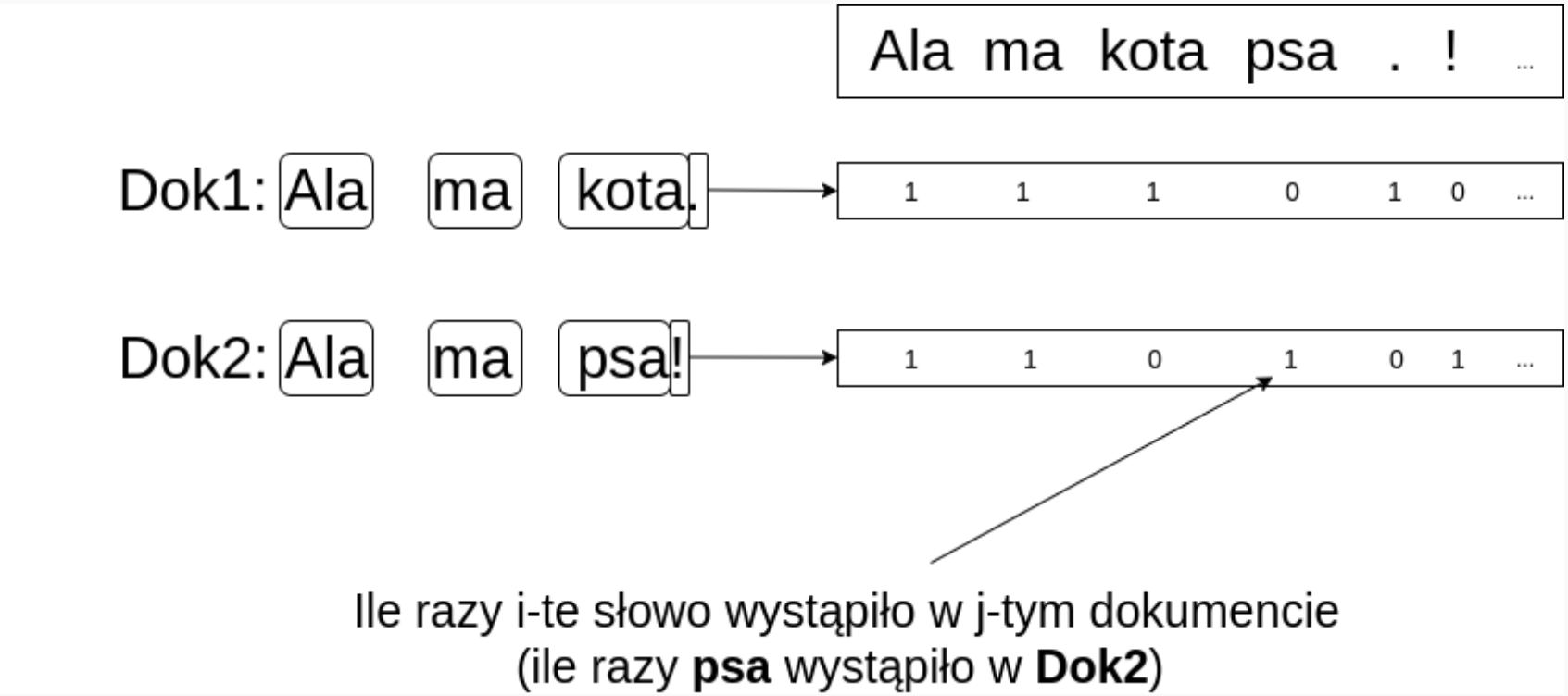
Przypomnienie iv

Krok 2: Stworzenie słownika z doumentów (w przykładzie - najprostsza forma, bierzemy tokeny ”jak leci”):



Przypomnienie v

Krok 3: Wygenerowanie reprezentacji wektorowych dokumentów, poprzez stworzenie dla każdego dokumentu wektora o długości takiej jak słownik i zapisanie w niej ile razy i-te słowo ze słownika pojawiło się w naszym j-tym dokumencie.



Przypomnienie vi

Krok 4:

Klasyfikacja i ewaluacja modelu.

Przypomnienie vii

Powyższa reprezentacja uwzględnia **obecność** słów (w ogólności tokenów), ale pomija ich **kolejność**, która często jest kluczowa dla znaczenia.

Takie same reprezentacje wektorowe zdań oznaczających coś zupełnie innego:

1. my dog likes Mike.
2. Mike likes my dog.

n-gramy i

n-gram: sekwencja n następujących po sobie tokenów.

1. 1-gram (unigram): pojedynczy token
2. 2-gram (bigram): pary tokenów
3. 3-gram (trigram): trójki tokenów
4. ...

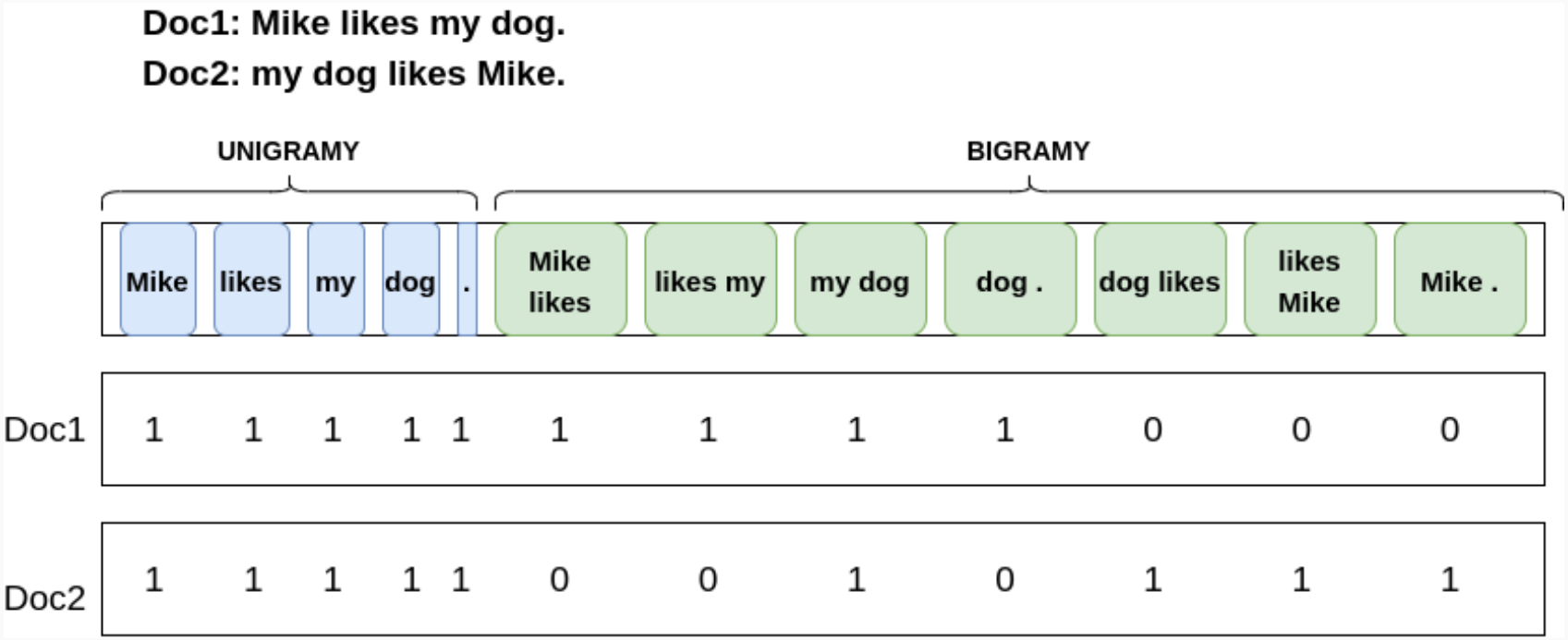
Nasza dotychczasowa reprezentacja (poprzednie slajdy) obejmowała unigramy.

n-gramy ii

bigramy:

trigramy:

n-gramy iii



Z użyciem n-gramów mamy więc reprezentację, która do pewnego stopnia uwzględnia kolejność tokenów.