

1. W korpusie występują słowa z następującymi krotnościami (w nawiasach): low (5), lowest(2), newer (6), wider (3), new (2). Wykonaj pierwsze kilka iteracji algorytmu BPE.
2. Załóżmy, że używasz modelu trzy-gramowego, w którym każdy warunkowy rozkład prawdopodobieństwa jest rozkładem jednorodnym. Ile wynosi nieokreśloność?
3. Rozważ poniższy korpus i zaproponuj dla niego przypisania słów do grup, takich jak w klasowym modelu n-gramowym, i co najmniej 3 prawidłowe schematy zdań możliwe do utworzenia z tych grup.
 - Ala ma kota i psa
 - Kasia posiada psa i chomika
 - Jurek kocha papugę
 - Ona lubi papugę i chomika
4. Do kilku grup uzyskanych w poprzednim zadaniu dodaj po jednym nowym słowie. Ile nowych zdań można wygenerować z takiego modelu?
5. Korzystając z klas słów $C_1 \in \{\text{Śmignąłem, Pojechałem}\}$, $C_2 \in \{\text{do}\}$, $C_3 \in \{\text{szkoły, teatru}\}$, $C_4 \in \{\text{metrem, samochodem, tramwajem}\}$, zdefiniuj co najmniej jeden schemat prawidłowego zdania.
6. Załóżmy, że w korpusie mamy zdania: „Pojechałem do szkoły metrem”, „Pojechałem do szkoły tramwajem”, „Pojechałem do teatru metrem”, „Pojechałem do teatru samochodem”, „Śmignąłem do szkoły metrem”, a klasy słów zostały zdefiniowane tak jak wyżej. Używając bi-gramowego klasowego modelu języka oblicz prawdopodobieństwo sekwencji „Śmignąłem do szkoły samochodem”.
7. Zakładając korpus:
 - Pojechałem rowerem
 - Pojechałem samochodem
 - Śmignąłem samochodem

wykonaj pierwszą iterację algorytmu uczącego bi-gramowy model klasowy. Dla uproszczenia obliczeń w funkcji celu możesz pominąć logarytm. W poniższej tabelce umieszczono w wierszach kilka rozważanych przez algorytm



[illegible]

-
- ```

graph TD
 Root[] --- Node1[]
 Root --- Node2[]
 Node1 --- EA[evaluation assessment]
 Node2 --- Analysis[analysis]
 Node2 --- Node3[]
 Node3 --- UO[understanding opinion]
 Node3 --- HB[house building]

```

- 
- Fundusze Europejskie**  
Polska Cyfrowa



**Unia Europejska**  
Europejski Fundusz  
Rozwoju Regionalnego



12. Dany jest korpus „Ala ma kota. Jurek ma kota.”. Używając kodowania „1 z n” stwórz zbiór treningowy dla klasyfikatora (np. sieci neuronowej), aby móc go wykorzystać w modelu 3-gramowym języka.
13. Zakładając, że słowa wejściowe są kodowane „1 z n” rozpisz wzór na klasyfikator softmax dla podanych prawdopodobieństw w modelach n-gramowych. Aby operować na prostszych wzorach możesz skorzystać z notacji „proporcjonalne” w której softmax możemy zapisać jako:  $P(\hat{y} = y_i | x) \propto w_i^T x + b_i$  (prawdopodobieństwo klasy jest proporcjonalne [choć nie wprost] z wynikiem wyrażenia liniowego). Po zapisaniu wzorów postaraj uprościć się je tak bardzo jak potrafisz i zinterpretuj je.
- dla modelu bigramowego  $P(A|a|START) \propto$
  - dla modelu trzygramowego  $P(kota|Ala, ma) \propto$
14. W kontekście modeli n-gramowych uczonych przez zliczanie (z technikami rozmywania estymat) oraz na podstawie zapisu z poprzedniego ćwiczenia, dokonaj interpretacji działania klasyfikatora softmax w modelu języka i porównaj jego działanie do modeli uczonych przez zliczanie.
15. Dane są trzy modele 3-gramowe.
- standardowy tj. oparty o zliczanie
  - zbudowany na klasyfikatorze softmax i reprezentacji „1 z n”
  - zbudowany na klasyfikatorze MLP i reprezentacji „1 z n”. Rozważane MLP to dwuwarstwowa sieć neuronowa (tylko jedna warstwa ukryta) z  $h = 500$  neuronami ukrytymi, a ostatnia warstwa to softmax.

Przyjmując  $|V| = 50\,000$  oszacuj liczbę parametrów w każdym z tych modeli. Który z tych modeli jest najmniej w sensie liczby parametrów? Który z tych modeli jest najbardziej ekspresywny?