

Przetwarzanie języka naturalnego  
Ćwiczenia 1  
Zajęcia 3

**Zadanie 1.** Rozważaliśmy krótko algorytm Bytes Pair Encoding (BPE). Przypomnij naiwną implementację podaną na wykładzie, wyjaśniając w razie potrzeby szczegóły. Powiedz dlaczego wydaje się ona mocno nieoptymalna. Dla chętnych: jak ją przyspieszyć?

**Zadanie 2.** BPE wyznacza jednoznaczny podział na części słów ze słownika. Rozważmy następujący scenariusz: dla określonego słownika uruchamiamy BPE, zapamiętujemy znalezione „kawałki” i traktujemy je jako wiedzę o języku: „to z takich, a nie innych właśnie części składają się słowa”. Chcielibyśmy wykorzystać te części do podziału słów spoza słownika. Jakiego rodzaju to problemy?

**Zadanie 3.** W zadaniu tym będziemy myśleć o losowaniu tekstu za pomocą  $N$ -gramów (konkretnie 2- i 3-gramów), przy czym wynikowy tekst powinien spełniać jakieś dodatkowe wymagania. Wyjaśnij, dlaczego „metoda naturalna” (czyli normalne losowanie od lewej do prawej tekstu o określonej długości, sprawdzenie warunku, i ewentualne losowanie ponowne) nie sprawdzi się w takich zadaniach. Dla każdego wariantu zaproponuj jakiś algorytm, który daje istotnie większą szansę na losowanie zakończone sukcesem (niż metoda naturalna):

- a) losuj tekst o długości  $M$ , który na pozycji  $k$  ma określony wyraz,
- b) losuj tekst o długości  $M$ , który na pozycjach parzystych ma określone wyrazy (czyli losujesz tylko pozycje nieparzyste, zaczynamy numerację od 0),
- c) losuj niezbyt długi tekst o zadanym pierwszym i ostatnim słowie,
- d) dla listy napisów  $[s_1, \dots, s_n]$  losuj tekst o długości  $n$ , taki że  $i$ -te słowo ma sufix  $s_i$ .

**Zadanie 4.** Załóżmy, że dysponujesz zespołem lingwistów, którzy są w stanie przygotować dla ciebie dowolne informacje o różnych słowach (ciągach liter występujących w języku). Czego od nich zarządzasz, jeżeli Twoim celem jest algorytm dzielenia tekstu na zdania w języku polskim? Zaproponuj jakiś sposób zastąpienia pracy lingwistów analizą korpusu.

**Zadanie 5.** Przypomnij, dlaczego do wyznaczania masy prawdopodobieństwa przypisanej słowom niewystępującym w korpusie ( $P(UNK)$ ), przydatny jest podział korpusu na dwie części ( $K_1$  i  $K_2$ ). Załóżmy, że nasz korpus ma pewną strukturę: cały korpus składa się z dokumentów, te z kolei z akapitów, które dzielą się na zdania. Jeżeli chcemy podzielić korpus na dwie części, to możemy przeglądać go porcja po porcji (porcjami mogą być dokumenty, akapity, zdania, albo pojedyncze wyrazy) i z prawdopodobieństwem  $p$  przypisywać porcję do korpusu  $K_1$ , a z prawdopodobieństwem  $1 - p$  do korpusu  $K_2$ .

- a) Jak wybór  $p$  wpływa na  $P(UNK)$ ?
- b) Jak wybór tego, czym jest porcja, wpływa na  $P(UNK)$ ?

**Zadanie 6.** Jeżeli rozwiązywałeś Zadanie 4 z pracowni, to pewnie miałeś problemy z porównywaniem naturalności permutacji. Zastanów się, dla którego ze zdań te problemy były silniejsze i dlaczego. Pewne konstrukcje języka polskiego nie są „permutowalne” to znaczy wymagają określonej kolejności wyrazów, bo inaczej będą niepoprawne. Przedstaw co najmniej 3 takie konstrukcje i zilustruj to przykładami zdań lub fraz (najlepiej również pięciowyrazowych, ale dopuszczalne są też nieco dłuższe), dla których bardzo wiele permutacji jest kompletnie niepoprawnych i/lub niezrozumiałych.

**Zadanie 7.** W języku polskim bardzo wiele informacji o słowie można wyczytać z jego końcówki. Oczywiście końcówek jest dużo mniej niż słów, więc wydaje się, że  $N$ -gramowy model działający na końcówkach może być użyteczny (dlaczego?). Na potrzeby tego zadania, przyjmijmy, że interesują nas końcówki o długości dokładnie  $k$  znaków. Zaproponuj metodę, bazującą na procedurze Deleted interpolation, która umożliwi połączenie modelu sufixowego (patrzącego na ciągi końcówek) z modelem tradycyjnym (patrzącym na ciągi słów).

**Zadanie 8.** Załóżmy, że mamy zbiór sufiksów  $S$  (zbiór napisów), dużo mniejszy niż cały słownik. Elementy  $S$  nie muszą mieć tej samej długości, dodatkowo zakładamy, że w  $S$  znajdują się wszystkie litery. Dla każdego słowa  $w$  ze słownika *reprezentantem sufiksowym* jest najdłuższy element zbioru  $S$ , który jest sufiksem słowa  $w$ .

Słownik wygląda tak, jak plik supertags.txt, czyli dla każdego słowa mamy przypisaną stałą ze zbioru  $T$  (nazywaną tagiem słowa, przy czym zbiór  $T$  zawiera około 3000 elementów), mówiącą o formie gramatycznej tego słowa. Każde słowo ma dokładnie jedną taką stałą.

- a) Przedstaw procedurę takiej konstrukcji zbioru  $S$ , żeby wyznaczenie reprezentanta sufiksowego słowa wystarczało do określenia tagu tego słowa.
- b) Jakie dwie korzyści daje zbiór  $S$  z poprzedniego podpunktu?
- c) Jak zmodyfikować tę procedurę, żeby mieć kontrolę nad wielkością zbioru  $S$  (być może godząc się z tym, że dla pewnych słów nie będziemy mogli ze stuprocentową pewnością podać ich tagu).

**Zadanie 9.** Drzewo rozbioru nie zostało jeszcze zdefiniowane na wykładzie. Odwołując się do szkolnych zajęć z języka polskiego zaproponuj taką definicję i podaj drzewa rozbioru następujących zdań (oczywiście przy tablicy nie będziesz prezentował wszystkich):

Stefan widział latarnię w Ustce.

Stefan widział latarnię w Ustce jedynie na zdjęciach Beaty.

Mam radę: nie mam mamy obietnicami bez pokrycia.

Wyjazd pociągiem do Krakowa zastąpił bogaty program artystyczny.

Wiktor podarował Beacie kwiaty, a Ewie czekoladki.

W których zdaniach możliwy jest więcej niż 1 rozbiór? Kiedy więcej niż jedno znaczenie? Jak wybór rozbioru wpływa na znaczenie zdania.

**Zadanie 10.** Rozważmy zdanie

I've decided to [...] her, though visitning aunts can be a nuisance.

Fragment rozpoczynający się od słowa *visiting* jest dwuznaczny, chociaż ma on jedno drzewo rozbioru (narysuj). Wyjaśnij źródło tej dwuznaczności i pokaż, jak wstawiając odpowiednie słowa w miejsce [...] możemy sprawiać, że któraś z interpretacji staje się bardziej naturalna.