1. Słowa w macierzy zanurzeń mają następujące reprezentacje: Ala (-1,0), Jurek (1,0), ma (1,1), kota (-1,1), i (-0.5, -0.5), psa (-1.1, 1.1). Rozważmy sieć splotową z jednym filtrem bigramowym (1, 1; -1, 1), funkcję aktywacji ReLU oraz funkcję redukcji k-Max z $k = 2$. Wynik funkcji redukcji jest następnie przetwarzany przez warstwę softmax.

   (a) Jaka reprezentacja zdania „Ala ma kota i ma psa" jest podawana na wejście warstwy softmax? (Kolejne wiersze macierzy wag filtra pokrywają się z kolejnymi słowami na wejściu)

   (b) Jaki n-gram jest wykrywany przez ten filtr?

   (c) Jak zmieniłoby się wejście warstwy softmax gdyby zastosować dynamiczną funkcję redukcji?

   (d) Zakładając, że sieć rozszerzamy o dwa dodatkowe filtry unigramowe (w ramach już istniejącej warstwy) o wagach (0, 1) i (2, -1) – oblicz wejście do warstwy softmax po funkcji redukcji k-Max z $k = 2$?

   (e) Czy funkcje redukcji „over time" takie jak k-Max można stosować pomiędzy kolejnymi warstwami sieci splotowej dla tekstu?

2. Projektujesz agenta dialogowego do zamawiania pizzy. Zaprojektuj schemat tagowania sekwencji, który pozwoliłby na wykrycie wybranych przez Ciebie relewantnych dla systemu informacji. Otaguj zgodnie z nim następujący korpus, wykorzystując schemat BIO.

   - Chciałbym margheritę na grubym cieście.
   - Dwa razy pepperoni na cienkim.
   - Na wynos, pizze z pieczarkami i zielonymi oliwkami.
   - Jedną małą pizzę z tuńczykiem, kukurydzą, karmelizowaną cebulą.

3. Mając poniższy korpus uczący:

   - Ala [N] ma [V] kota [N]
   - Jacek [N] lubi [V] pluszowe [JJ] misie [N].

   Policz prawdopodobieństwo sekwencji: „Ala [N] lubi [V] misie [N]" wg. bigramowego ukrytego modelu Markowa.

4. Rozważ korpus uczący:

   - I [O] book [V] a [O] flight [N].
   - Dad [N] reads [V] a [O] book [N].
   - Big [O] company [N] books [V] flights [N].
   - I [O] like [V] A [N] company [N].

   Zakładając, że w korpusie uczącym zamieniono wszystkie duże litery na małe oraz usunięto literkę „s" jeśli znajdowała się na końcu wyrazu (pozbycie się liczby mnogiej i odmiany czasowników), wytrenuj bigramowy ukryty model Markowa, a następnie dokonaj predykcji algorytmem Viterbiego dla zdania „I book a book".

5. Podaj korpus uczący dla którego klasyfikator HMM popełni choć jeden błąd (na korpusie uczącym).

6. Zapisz wzór na algorytm Viterbiego dla modelu MEMM z reprezentacją cech opartą na poprzednim tagu i aktualnym słowie.

7. Zapisz wzór na algorytm Viterbiego dla modelu trzygramowego HMM.

8. Rozważając korpus uczący z zadania 3, zapisz w postaci tabelki zbiór uczący dla klasyfikatora MEMM.

9. Aby otagować n-elementową sekwencję modelem MEMM – ile razy należy wykorzystać klasyfikator? Opisz przebieg predykcji zachłannej przez ten model.

10. Projektujemy klasyfikator softmax przypisujący cześć mowy dla danego słowa $P(PoS|word)$. Rozważane części mowy to $PoS \in \{N, V, JJ\}$ a $V = \{$być, mieć, złoto, tabletka, piękny, żółty$\}$. Podaj minimalny zbiór cech binarnych $\phi(x, y)$, który może zamodelować następujący rozkład:

$$P(JJ|\text{żółty}) = 0.6 \quad P(N|\text{tabletka}) = 0.8 \quad P(V|\text{być}) = 0.99$$

$$P(JJ|\text{mieć}) = P(JJ|\text{złoto}) = P(JJ|\text{piękny}) = 0.4$$

    pozostałe wartości rozkładu nie są dla nas interesujące (mogą przyjąć dowolną wartość).

11. Rozważmy model softmax $\sigma(\mathrm{x})_y = \frac{e^{\mathrm{w}^T \phi(\mathrm{x},y)}}{\sum_{y'} e^{\mathrm{w}^T \phi(\mathrm{x},y')}}$, który jest nauczony poprzez maksymalizację logarytmicznej funkcji wiarygodności wraz z termem regularyzującym L2. Załóż, że w czasie optymalizacji osiągnięto optimum funkcji celu. Odpowiedz na poniższe pytania i uzasadnij.

    (a) Zakładając cechę $\phi_1(x, y) = 0$ dla każdego $x \in X$ oraz $y \in Y$, ile wynosi wartość w_1?

    (b) Zakładając cechę $\phi_2(x, y) = 1$ dla każdego $x \in X$ oraz $y \in Y$, ile wynosi wartość w_2?

(c) Zakładając cechę $\phi_3(x, y) = idx(x)$ dla każdego $x \in X$ oraz $y \in Y$, gdzie funkcja $idx()$ przypisuje kolejnym wektorom x kolejne liczby naturalne – ile wynosi wartość w_3?

(d) Utworzono zestaw cech $\phi_i(x, y) = \mathbf{1}_{x=x' \wedge y=y'}$, po jednej cesze dla każdego $x' \in X$ oraz $y' \in Y$. Zakładając cechę bez pokrycia $\phi_j(x, y)$ (tj. cecha ta nie aktywuje się ani razu w zbiorze uczącym), ile wynosi wartość w_j?

12. Czy do klasyfikacji wieloklasowej możemy zastosować zamiast warstwy softmax, warstwę złożoną z neuronów logistycznych? W jaki sposób taka sieć byłaby trenowana? Jakie są zalety stosowania warstwy softmax zamiast zwykłych neuronów logistycznych?