Przetwarzanie języka naturalnego
Ćwiczenia 3
Zajęcia w tygodniu rozpoczynającym się 14.12.2020

Zadanie 1. Przypomnij, co to jest perplexity? Rozważmy ciągi cyfr, które:

a) składają się z bloków k-cyfrowych (w bloku są te same cyfry),

b) a po danym bloku może nastąpić blok z dowolną cyfrą (z tym samym prawdopodobieństwem)

Rozważamy bardzo duży tekst z tego języka. Oblicz dla tego tekstu perplexity dla następujących modeli: unigramowego, bigramowego oraz dla $n = k + 1$ (we wszystkich modelach jednostkami są cyfry). Czy ten ostatni model daje minimalne perplexity? Jeżeli nie, to jaka niewielka modyfikacja warunków zadania sprawiłaby, że model stanie się optymalny?

Podaj wartości liczbowe dla $k = 10$.

Zadanie 2. Załóżmy, że mamy opracowany parser, który wyznacza „nawiasowanie zdania". Twoim zadaniem będzie napisanie funkcji (w dowolnym języku), która porównuje nawiasowanie wyprodukowane przez parser z nawiasowaniem wzorcowym. Funkcja powinna spełniać następujące warunki:

i) powinna zwracać wartość liczbową, większa wartość oznacza gorszy (mniej podobny) rozbiór;

ii) nie powinna karać za sytuację, w której sprawdzane nawiasowanie jest uszczegółowieniem nawiasowania wzorcowego (dzięki temu wzorce mogą nie stawiać kontrowersyjnych nawiasów). Wzorcowym nawiasowaniem zdania

    Judyta wczoraj jechała szybkim pociągiem.

    jest

    (Judyta wczoraj jechała (szybkim pociągiem))

    a oceniany przez nas parser może (bezkarnie) zrobić coś takiego:

    (Judyta ((wczoraj jechała) (szybkim pociągiem)))

    oczywiście za nawiasy: *((jechała szybkim) pociągiem)* powinien zostać ukarany.

Opisz typ i format danych wejściowych. Staraj się, by funkcja była użyteczna.

Zadanie 3. Jak wykorzystać model HMM w zadaniu korekty błędów ortograficznych? (opis powinien być dość dokładny, z dokładną interpretacją współczynników a oraz b, wraz z opisanym sposobem ich szacowania)

Zadanie 4. Jak wykorzystać model HMM do znajdywania wielkości liter w tekście (tzn. rozważamy drugą część zadania P3.1: interesuje nas system, który jest wytrenowany na rzeczywistych tekstach z różną wielkością liter, dla którego wejściem będzie tekst lower case).

Uzasadnij stwierdzenie, że nie jest to idealny sposób rozwiązywania tego zadania, porównując z inną metodą, w której pamiętamy długie ciągi wyrazów pisane częściej wersalikami, niż małymi literami (doprecyzowując jednocześnie działanie tej drugiej metody).

Zadanie 5. Na wykładzie przedstawiona była metoda obliczania wartości współczynników b_{ik} czyli indeksowanych stanem początkowym i symbolem emitowanym. Jak szacować te współczynniki w ukrytych łańcuchach Markowa, jeżeli chcemy, by zależały one dodatkowo od stanu końcowego.

Zadanie 6. Przypomnij, jak działa klasyfikator NBB. Jak można by go wykorzystać do następującego zadania: określić, czy w tekście dany rzeczownik rodzaju m3 (stół, dom, samochód) występuje w mianowniku (Stół stoi, samochód przyjechał), czy też w bierniku (widzę samochód, patrzę na dom, lubię mój stół).

Zadanie 7. Zaproponuj 6 reguł hipotetycznego tagera regułowego, które zajmują się następującymi dwoma słowami: *jak* oraz *miał*. Wszystkie Twoje reguły powinny odnosić się do mniej popularnego znaczenia tych słów. Postaraj się, by reguły odnosiły się do rzeczywistych sytuacji językowych, które znalazłeś w Internecie (powiedz, jak ich szukałeś, by każdy, kto chce to powtórzyć, miał łatwiej).

Zadanie 8. Powiedz, jak łączyć tager Brilla z innymi tagerami. Zaproponuj dwa scenariusze.

Zadanie 9. Dla tagera Brilla będziemy rozważać reguły postaci:

> jeżeli na pozycji -1 mamy t_1, na pozycji +1 mamy t_2 wówczas zmień na pozycji 0 t_3 na t_4

Teoretycznie przestrzeń możliwych takich reguł jest bardzo duża (liczba tagów do potęgi czwartej). Wyjaśnij, dlaczego w rzeczywistości takich reguł będzie dużo mniej.

Zadanie 10. Przeczytaj ze zrozumieniem artykuł o LSTM na blogu *colaha*:
http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Przygotuj się, by o tym artykule opowiedzieć przez 6-10 minut (nie jest konieczne przepisywania fragmentów do hackmd, można prezentować w oparciu o oryginalny tekst w przeglądarce). Jeżeli wszyscy zadeklarują to zadanie prowadzący **może** zrezygnować ze sprawdzania go, może też je przełożyć na koniec zajęć (wówczas sprawdzenie odbyłoby się jedynie, gdy zostanie czas)

Zadanie 11. 3p, $\star$ Zapoznając się z architekturą LSTM (i pochodnymi) można mieć poczucie pewnej dowolności decyzji projektantów. Przedstaw główne ideę pracy *An Empirical Exploration of Recurrent Network Architectures* (http://proceedings.mlr.press/v37/jozefowicz15.pdf), która odnosi się do tej kwestii. Przygotuj wystąpienie trwające około 10 minut.