Modele językowe
Ćwiczenia 4
Pierwsze zajęcia w 2025 roku

Każde zadanie warte jest 1 punkt (chyba, że napisano inaczej).

Zadanie 1. Jeżeli wykonamy zanurzenia dla korpusów w dwóch językach, to wektory dla *women* oraz *kobieta* nie będą miały ze sobą związku. Zaproponuj jakiś sposób liczenia zanurzeń, w których słowa będące swoimi tłumaczeniami będą otrzymywać podobne wektory.

Powinieneś wykorzystywać korpusy w obu językach i (być może) jakieś inne dane.

Zadanie 2. Zaproponuj jakiś sposób do generowania tekstu (ogólnie sekwencji/macierzy tokenów o ustalonym rozmiarze) za pomocą sieci typu BERT. Możesz wymyślić procedurę samemu, lub odnaleźć jakąś publikację, która takie metody opisuje.

Zadanie 3. Załóżmy, że dysponujemy dużym, n-gramowym modelem M_{ng} działającym na tokenach. Zaproponuj dwa scenariusze **treningu** modelu typu GPT, który wykorzystuje M_{ng} (model M_{ng} może być również później wykorzystywany w inferencji, czyli w generowaniu tekstu)

Zadanie 4. Wracamy do zadania o Niefrasobliwym Programiście (co trenuje nie ten model, co trzeba). Tym razem NP trenował model typu BERT (duży i kosztowny), ale zapomniał o uwzględnieniu osadzeń pozycyjnych. Czy taki model mógłby być do czegoś użyteczny? (oczyiście od odpowiedzi zależy dalszy los NP w firmie). Czy fakt, że użyty tu słownik ma bardzo dużo tokenów jest okolicznością sprzyjającą, czy obciążającą?

Zadanie 5. Przyjmijmy, że udało się pozytywnie odpwiedzieć na pytanie w poprzednim zadaniu i Kierownictwo chce (dla innego języka) wytrenować inną instancję „bezpozycyjnego BERT-a". Standardowy trening Maskowanego Modelu Językowego ma pewną właściwość, która jest szczególnie kłopotliwa, gdy nie ma osadzeń pozycyjnych. Jaka to właściwość? Wskazówka, z której lepiej nie korzystać (rot13.com): Cbzlśy b yvpmovr gbxraój ZNFX

Zadanie 6. $\star$ W klasyfikacji tekstu używany jest często naiwny klasyfikator bayesowski (NBC). Opowiedz krótko jak działa NBC, a następnie podaj co najmnie dwa powody, dlaczego dostrajanie BERT-a ma szansę dawać istotnie lepsze wyniki w zadaniu klasyfikacji tekstu.

Zadanie 7. Zaprojektuj eksperyment, który odpowie na pytanie: „czy transformery potrafią sortować ciąg liczb naturalnych". Jaka jest Twoja intuicja dotycząca wyniku tego eksperymentu?

Zadanie 8. Zaprojektuj eksperyment, który odpowie na pytanie: „czy transformery potrafią obliczać wartość XOR". Jaka jest Twoja intuicja dotycząca wyniku tego eksperymentu?

Zadanie 9. (2p) Powiedzmy, że chcemy modelować za pomocą LM-u typu GPT (czyli autoregresywnego) kod w Pythonie. Odpowiedz na poniższe pytania:

1. Czy standardowa, trenowana tokenizacja jest tu optymalnym wyborem? Czy też może warto rozważyć jakąś jej modyfikację (a jeżeli tak, to jaką)
2. Jak najlepiej obsłużyć wcięcia w kodzie?
3. Czym jest PEP-8? Czy może on mieć jakieś użycie w tym zadaniu?
4. Jakie są argumenty za tym, że warto zmieniać nazwy zmiennych/funkcji/klas/... w kodzie (zachowując jego semantykę)?
5. Jak metody NLP (word2vec?, transformery?, ...) mogą pomóc w zamianie nazw (podaj co najmniej dwa przykłady)
6. Jak statyczna analiza kodu może pomóc w tym zadaniu? (wystarczy jeden scenariusz)

Możesz zadeklarować to zadanie za 1p, jeżeli znasz odpowiedzi na większość, ale nie wszystkie pytania.

Zadanie 10. ⋆ Które z pooprzednich zadań nadają się na projekt/pracę dyplomową (oczywiście, gdyby je odpowiednio rozbudować)? Jakie konieczne modyfikacje należałoby wprowadzić (musisz wskazać co najmniej jedno zadanie).

Zadanie 11. (2p)⋆ Zaproponuj jakieś zadanie (niezwiązane z niniejszą listą zadań), nadające się (Twoim zdaniem) na projekt do naszego przedmiotu. Twój opis powinien być zwięzły, niemniej w miarę jednoznacznie definiujący zadanie. Zgłoś je na forum na SKOSie (`https://skos.ii.uni.wroc.pl/mod/forum/view.php?id=43738&forceview=1`), postępując zgodnie z instrukcjami. Twoje zgłoszenie **nie jest** deklaracją, że chcesz to zadanie robić, mówi jedynie, że zadanie wydaje Ci się ciekawe i odpowiednie. Może też stać się inspiracją dla innych. Jeżeli częścią zadania jest trenowanie modelu, kluczowa jest wówczas kwestia opisania pozyskania danych (albo skorzystanie z istniejącego datasetu, albo stworzenie nowego, własnego zbioru danych)