Modele językowe
Ćwiczenia 3
Zajęcia w tygodniu rozpoczynającym się drugiego grudnia

Każde zadanie warte jest 1 punkt.

Zadanie 1. Słowo *tonie* ma aż 4 różne znaczenia i cztery lematy. Wymień je wszystkie, a następnie napisz krótki tekst (najlepiej jedno zdanie), w którym to słowo występuje we wszystkich czterech znaczeniach. Sprawdź, czy ChatGPT jest w stanie poprawnie zinterpretować to zdanie, wypisując informacje o tych znaczeniach w kontekście tego zdania.

Zadanie 2. Czym są rzadkie reprezentacje wektorowe słów (wykorzystujące TF-IDF i konteksty)? Dlaczego nie jest to rozwiązanie perfekcyjne (słaba wskazówka: piękna żaglówka; śliczny żaglowiec). Zaproponuj procedurę, która zawiera w sobie klasteryzację i działa potencjalnie lepiej, niż oryginalne reprezentacje (dodatkowo zwracając wektory o mniejszej liczbie wymiarów).

Zadanie 3. Będziemy rozważać zadanie Next Sentence Prediction (czyli zadanie, w którym na wejściu mamy dwa zdania i należy określić, czy są one kolejnymi zdaniami tekstu (czyli czy drugie zdanie jest w tekście bezpośrednio po pierwszym). Zadanie to występuje w dwóch wariantach:

a) Przykłady negatywne są losowane z całego korpusu.

b) Przykłady negatywne są kolejnymi zdaniami, w których zmieniliśmy kolejność.

Oczywiście przykłady pozytywne, to po prostu pary kolejnych zdań z tekstu.
Zaproponuj rozwiązanie obu tych wariantów, wykorzystujące model taki jak Papuga. Dodatkowo dla wybranego wariantu zaproponuj rozwiązanie niekorzystające z zaawansowanych sieci neuronowych (jak chcesz, możesz skorzystać z word2vec-a, lub podobnych modeli, ale nie jest to wymagane).

Zadanie 4. Zaproponuj jakiś sposób do wyznaczania (bezkontekstowych) osadzeń węzłów w grafie (na przykład osób w sieciach społecznościowych, filmów i użytkowników Netlixa, ...). Twój sposób powienien korzystać z oryginalnego Word2Vec-a.

Zadanie 5. Jakaś część pytań z Pracowni 1 dotyczy przysłów, przykładowo:

> Z czym według przysłowia porywamy się na słońce?
> Co według przysłowia kołem się toczy?
> Który wyraz jest błędnie użyty w przysłowiu: „Co dwie głowy, to nie trzy"

Zaproponuj sposób rozwiązywania takich pytań z wykorzystaniem wyszukiwania informacji.

Zadanie 6. Przeczytaj opis baseline solution dla zadania https://2021.poleval.pl/tasks/task4 (odpowiadania na pytania). Spróbuj uzasadnić, dlaczego to miałoby działać i zaproponuj jakąś, sensowną wg Ciebie, korektę tego algorytmu.

Zadanie 7. Omawiany na wykładzie algorytm BPE działał zwiększając liczbę tokenów do pożądanego progu. Zaproponuj algorytm działający w drugą stronę, to znaczy startujący od bardzo dużej liczby tokenów i zmniejszający ją, aż osiągniemy wymaganą liczbę. Algorytm powinien:

- Być niezależny od języka (zatem nie może wykonywać żadnego wstępnego szukania wyrazów)
- Pracować na dużym korpusie, pamiętając „bieżącą" tokenizację korpusu, wykonaną za pomocą dostępnych tokenów.
- Wykonywać wiele kroków, stopniowo zbliżając się do docelowej wielkości słownika.
- Usuwać mniej użyteczne tokeny (czyli „tokenizować tokeny" w korpusie)
- Starać się maksymalizować unigramowe prawdopodobieństwo korpusu, czyli $\prod_{i=1}^{N} p(w_i)$, gdzie wszysktie w_i należą do bieżącego zbioru tokenów, a ich konkatenacja to cały korpus. Przy czym prawdopodobieństwa szacujemy na podstawie tokenizowanego korpusu. Zastanów się ponadto, dlaczego wymaganie z tego punktu jest sensowne.

Zadanie 8. Zaproponuj 3 różne scenariusze augmentacji danych z recenzjami za pomocą Papugi.

Zadanie 9. Zaproponuj 3 scenariusze wykorzystania word2vec w zadaniu Riddles (z poprzednich list). Zakładamy, że mamy dostęp do wzorcowych definicji wyrazów, oraz do pewnej liczby przykładów 'zagadek'.

Zadanie 10. Ocena wiarygodności generowanego tekstu w scenariuszu silnych więzów (na przykład wymagania, by wszystkie słowa były na 'p') za pomocą prawdopodobieństwa tego tekstu ma pewną wadę. Jaką? (wskazówka: rozważ słowo 'przede'). Jak łatwo wyeliminować tę wadę?

Zadanie 11. (1-3p) $\star$ W zadaniu wrócimy do osadzeń słów testowanych za pomocą ABX i modelu typu BERT. Przyjmijmy, że wektorem osadzenia będzie wektor przypisany całej wypowiedzi (pierwszy wektor w Hubercie), niemniej wypowiedź nie będzie jednowyrazowa. Zaproponuj jakąś metodę konstruowania takiej wypowiedzi, wykorzystując korpus tekstowy (potencjalnie zawierający wystąpienia słów i/lub plik z lematami słów. Za zaproponowanie metody jest 1p, za jej zaimplemetowanie i powiedzenia, jaki ABX-score osiąga drugi, a jeżeli uda się w ten sposób otrzymać lepsze osadzenie niż przy użyciu tylko jednego słowa – trzeci.