Modele językowe
Pracownia 3
Zajęcia 7 i 9

Zadanie 1. (**4+Xp**) Zajmiemy się osadzeniami słów (zarówno kontekstowymi, jak i bezkontekstowymi). Uwaga: teksty, które będziemy osadzać zawsze składają się z jednego słowa (ale niekoniecznie z jednego tokenu).

a) Zaproponuj jakiś sposób wykorzystania **bezkontekstowych** osadzeń tokenów (wyznaczanych przez transformer[1] do wyznaczenia osadzeń słów. Możesz skorzystać z programu z wykładu 7 (embedding.ipynb). Sprawdź, jaką jakość (mierzoną testem ABX) mają te osadzenia[2]. Do zaliczenia zadania wymagane jest 0.6.

b) Wykorzystaj kontekstowe osadzenia tokenów z BERT-a do wyznaczenia osadzeń dla słów. Ponownie wykonaj testy ABX.

c) Spróbuj połączyć te dwa podejścia w jakikolwiek sposób. Jakość twojego rozwiązania przekłada się na punkty bonusowe zgodnie z wzorem $(\text{score} - 0.6) \times 6$

Procedura ewaluacji (być może zostanie uproszczona): osadzenia zapisz w pliku tekstowym `word_embedings_file.txt`, w którym każdy wiersz wygląda tak:

```
[słowo] float_1 float_2 ... float_D
```

Osadzenia są oceniane za pomocą skryptu `word_emb_evaluation.py`.

Zadanie 2. (**(6+X)p**) W zadaniu tym będziemy zajmować się klasyfikacją recenzji z wykorzystaniem modeli transformer, możesz tu skorzystać z programu z wykładu (herbert.ipynb). W tym zadaniu powininieneś użyć trzech modeli:

1. Modelu generatywnego, takiego jak Papuga, Polka o wielkości do 1B, który znajduje prawdopodobieństwa tekstu (podobnie, jak na liście 1)
2. Kodera typu BERT (np. herbert), jako ekstraktora cech
3. Tradycyjnego modelu Machine Learning, który integruje wyniki dwóch poprzednich modeli. Ten model powinieneś wytrenować na zbiorze treningowym recenzji, a testować na testowym.

Wartość premii jest równa: $20 * (a - 0.85)$, gdzie a to wartość accuracy na zbiorze testowym. Jeżeli chcesz, możesz skorzystać tu również z wyników kolejnego zadania.

Zadanie 3. (**8+1p**) W tym zadaniu powinieneś sprawdzić, czy augmentacja danych może poprawić wyniki klasyfikacji, w której BERT jest traktowany jako ekstraktor cech. Mamy 3 osobno punktowane procedury generowania nowych wariantów recenzji:

a) Augmentacja mechaniczna (czyli wprowadzasz jakieś zniekształcenia w tekście, mogą to być na przykład literówki, zmiana wielkości liter, błędy związane z polskimi literami, etc). (2p)

b) Augmentacja modelem generatywnym, na przykład Papugą. Powinieneś generować recenzje, które bazują na oryginalnej recenzji, zachowując jej polarność (czyli to, czy jest pozytywna, czy negatywna). Zwróć uwagę, że „fantazja" modelu językowego nie musi tu być wadą – tak naprawdę to niekoniecznie w tej procedurze muszą powstawać poprawne teksty. (3p)

c) Ta procedura augmentacji powinna bazować na Word2Vec i zachowywać w miarę możliwości znaczenie tekstu. Należy wybrane słowa zamieniać na słowa bliskoznaczne, w tej samej formie gramatycznej (będzie to dokładniej omówione na kolejnym wykładzie). Przykładowo recenzja: *Hotel ogólnie bardzo ładny.* mogłaby być zmieniona na *Pensjonat szczególnie bardzo piękny.*, a *Polecam wszystkim tego fizjoterapeutę!* na *Rekomenduję wszystkim tego ortopedę!* Konieczne informacje gramatyczne pojawią się na wykładzie 8 (czyli najbliższym) (3p)

[1] możesz wykorzystać Papugę lub Herberta, lub inny model niezbyt dużej wielkości
[2] Informacja: bardzo prosta procedura pozwala osiągnąć 0.7

Każda recenzja powinna posłużyć do wygenerowania K innych recenzji (dobór K to Twoje zadania), stąd należy generator napisać w ten sposób, by recenzje były tworzone niedeterministycznie. Dla wybranych (lub wszystkich) procedur przeprowadź uczenie na zaugmentowanych danych za pomocą regresji logistycznej. Dodatkowo można uzyskać 1p premii, jeżeli któraś z procedur da korzyść w porównaniu to oryginalnych danych (tzn. dzięki augmentacji uda się uzyskać lepszy wynik wynik dla danych testowych) W zadaniu do maksimum wlicza się 6p.

Zadanie 4. (**(0+X)p**) To zadanie pozwoli zdobyć dodatkowe punkty za zadanie ze znajdywaniem ujednoznacznienia zdania wieloznacznego. Szczegóły wkrótce

Zadanie 5. (**(0+X)p**) Treść zostanie podana wkrótce: będzie to pożegnanie z zadaniuem Riddles, z wykorzystaniem BERT-a, Papugi, Word2Vec-a, zbioru definicji i TF-IDF. Będzie dokładnie opisana procedura ewaluacyjna i punkty za jakość. Szczegóły wkrótce. Być może to zadanie będzie miało przedłużony termin.

Zadanie 6. (**(0+X)p**) Treść zostanie podana wkrótce: będzie to pożegnanie z zadawaniem pytań, z wykorzystaniem BERT-a, Papugi, Word2Vec-a, zbioru definicji i innych danych. Będzie dokładnie opisana procedura ewaluacyjna i punkty za jakość. Szczegóły wkrótce. Być może to zadanie będzie miało przedłużony termin.