Eksperymenty

Wstęp

Pierwszym podejściem do rozwiązania problemu było wykonanie testów na zbiorze danych dostarczonych przez organizatora konkursu SemEval-2017. Wybrane zostały 4 miary mające określić podobieństwa między dwoma tekstami dla wszystkich zadań, oraz 6 miar podobieństwa dla zadania E:

1. różnica długości tekstu
2. indeks Jaccarda
3. podobieństwo kosinusowe z wykorzystaniem biblioteki **spaCy**
4. podobieństwo kosinusowe z wykorzystaniem biblioteki **gensim**
5. liczba wyświetleń pytania powiązanego (tylko dla zadania E)
6. liczba punktów pytania powiązanego (tylko dla zadania E)
7. liczba odpowiedzi (tylko dla zadania E)
8. liczba komentarzy (tylko dla zadania E)
9. całkowita liczba głosów na odpowiedź do pytania powiązanego (tylko dla zadania E)
10. największa liczba głosów na odpowiedź do pytania powiązanego (tylko dla zadania E)

Dla różnych zadań wykonane zostały testy podobieństwa między:

* dla zadania A: pytaniem oryginalnym, a komentarzem
* dla zadania B: pytaniem oryginalnym, a pytaniem powiązanym
* dla zadania C: pytaniem oryginalnym, a komentarzem do innego pytania
* dla zadania E: pytaniem oryginalnym, a wątkiem powiązanym

Wykorzystywane biblioteki

Do obliczenia podobieństwa kosinusowego wykorzystane zostały 2 różne biblioteki: **spaCy** oraz **gensim**. Obie biblioteki wykorzystuje algorytm podobieństwa kosinusowego z użyciem wektorów słów, czyli „word embeddings”. Biblioteka spaCy wykorzystuje wektory GloVe (Global Vectors for Word Representation) wytrenowane na zbiorze danych „Common Crawl”, natomiast biblioteka gensim wykorzystuje w tym przypadku wektory wytrenowane na zbiorze „Google News”.

Dla biblioteki spaCy wykorzystana została konwolucyjna sieć neuronowa wyuczona na zbiorze danych „OntoNotes”.

Wyniki badań

1. Różnica długości tekstu

Jest to najprostsza miara podobieństwa między dwoma tekstami. Polega na porównaniu liczby znaków w 2 tekstach. Z miar, obliczonych na mappodstawie wszystkich zbiorów (dla zadania A, B, C i E) wynika, że dla różnych etykiet, np. **„PerfectMatch”**, **„Related”** oraz **„Irrelevant”** wartości długości tekstu nie różnią się znacząco od siebie.

2. Indeks Jaccarda

Podobnie jak w przypadku różnicy długości tekstów, obliczone wartości indeksu Jaccarda nie różnią się znacząco dla różnych etykiet, np. **„PerfectMatch”**, **„Related”** oraz **„Irrelevant”** dla zadania E.

*[ Wykres tutaj ]*

3. Podobieństwo kosinusowe z wykorzystaniem biblioteki spacy

[ TODO ]

Jest to miara przyjmująca, podobnie jak indeks Jaccarda, wartości z przedziału <0, 1>.

Operuje ma bardziej trafnej reprezentacji tekstu i nie powoduje utraty informacji.

Wykorzystany został model dla języka angielskiego z biblioteki spaCy.

4. Podobieństwo kosinusowe z wykorzystaniem biblioteki gensim

[ TODO ]

Podejście do tego problemu różni się zastosowaniem innego modelu. Wykorzystany jest model z biblioteki **gensim** wytrenowany na zbiorze „Google News”.

5. Liczba wyświetleń pytania powiązanego (tylko zadanie E)

Jest to miara używana wyłącznie dla zadania E, ponieważ tylko dla tego zadania, zbiór danych posiada potrzebne informacje. Wartości liczby wyświetleń dla trzech różnych etykiet są bardzo podobne do siebie, co oznacza, że ta miara nie mówi dużo o podobieństwie między tekstami.

6. Liczba punktów pytania powiązanego (tylko zadanie E)

Dla tego przypadku, porównane zostały wartości liczby punktów pytania powiązanego. Punkty przyznawane są przed użytkowników forum, poprzez głosowanie. Każdy użytkownik może dodać lub odjąć jeden punkt przy pytaniu. Ostateczna liczba punktów to suma wszystkich głosowań. Jest to kolejna miara dla której wartości między różnymi etykietami nie różnią się znacząco.

7. Liczba odpowiedzi (tylko zadanie E)

Miara ta, mówi ile odpowiedzi zostało umieszczonych pod pytaniem powiązanym.

Wartość mediany oraz średniej dla każdej z etykiet nie różni się znacząco.

8. Liczba komentarzy (tylko zadanie E)

Jest to suma komentarzy w wątku dla pytania powiązanego. Liczone są wyłącznie komentarze pod pytaniem. Komentarze dla odpowiedzi nie są brane pod uwagę. Również na podstawie tej miary, nie można wyciągnąć wniosków na temat podobieństwa między dwoma pytaniami.

9. Całkowita liczba głosów na odpowiedź do pytania powiązanego (tylko zadanie E)

Jest to zsumowana liczba głosów na wszystkie odpowiedzi w wątku. Większość odpowiedzi w wątku ma łącznie między 0 a 6 głosów. Te wartości są bardzo podobne dla każdej z etykiet, dlatego znów jest to informacja która nie jest w stanie określić podobieństwa między pytaniami.

10. Największa liczba głosów na odpowiedź do pytania powiązanego (tylko zadanie E)

Miara ta, polega na znalezieniu największej liczby głosów oddanych na odpowiedź pod pytaniem powiązanym. W przypadku, gdy nie istnieje żadna odpowiedź pod pytaniem przyjmowana jest wartość 0. Również dla tego przypadku, dla żadnej z trzech etykiet, liczba głosów nie różni się znacząco od pozostałych.