Anotacja korpusów oraz osadzenia słów i tekstów
Część II: Modele przestrzeni wektorowych

Autorzy: Oliwer Krupa, Adam Bednarski, Jan Masłowski, Łukasz Lenkiewicz

October 23, 2024

Contents

1 Word Embeddings

1.1 Osadzenia słów w anotowanym korpusie (Word2Vec i Fasttext)

Zadanie polegało na wygenerowaniu osadzeń słów z wykorzystaniem co najmniej dwóch różnych modeli. Wybraliśmy następujące modele osadzeń:

- **Word2Vec**:
  - dla języka polskiego: `https://dsmodels.nlp.ipipan.waw.pl`
- **Fasttext**:
  - dla języka polskiego: `https://fasttext.cc/docs/en/crawl-vectors.html`

1.2 Wizualizacja danych za pomocą t-SNE

Dane zostały zwizualizowane za pomocą techniki t-SNE, z wykorzystaniem interaktywnych wykresów. Każdy punkt na wykresie odpowiada słowu z anotacją. Po najechaniu na punkt można odczytać zarówno słowo, jak i przypisaną do niego etykietę.

1.3 Porównanie k-najbardziej podobnych słów dla dwóch modeli

Dla każdego z modeli (Word2Vec i Fasttext) wygenerowano listy k-najbardziej podobnych słów w oparciu o osadzenia zaanotowanych słów. Następnie porównano listy podobieństw, uwzględniając różnice w reprezentacjach przestrzeni wektorowych.

Word	**Similar Words**
by	By, Żęby, muc, Dáibhí, Żey
kto	Kto, ktoś, Ktoś, ktokolwiek, nikt
odbiorcow	dostawców, inwestorów, kontrahentów, przedsiębiorców, sprzedawców
człowiek	Człowiek, człowiek, człowiek., człowiek-, .Człowiek
kojarzyć	kojarzy, skojarzyć, kojarzył, kojarzą, kojarzyc
seks	sex, seks., seksSeks, seks-, -seks
dziwić	dziwić-, dziwić., dziwi, Dziwić, dziwic
fakt	faktem, faktu, Fakt, zważywszy, Faktem
zdrajca	Zdrajca, patriota, zdrajco, zdrajcą, sługus
but	But, INVOKE, No-One, bucik, tired

Table 1: Similar words in Fasttext

Word	Similar Words
by	aby, że by, więc, jeżeli, ażeby
kto	dlaczego, czemu, jeżeli, niech, jeżeli
odbiorców	klientów, dostawców, użytkowników, producentów, widzów
człowiek	osobnik, osoba, tubylec, kobieta, Polak
kojarzyć	utożsamiać, skojarzyć, identyfikować, wiązać, wywodzić
seks	sexu, masturbacja, sex, erotyka, prostytucja
dziwić	obawiać, denerwować, domyślać, łudzić, oburzać
fakt	stwierdzenie, twierdzenie, przeświadczenie, przypuszczenie, hipoteza
zdrajca	sprzedawczyk, kolaborant, faszysta, łajdak, kanalia
but	bucik, skarpeta, sandał, trzewik, spodnie

Table 2: Similar words in Word2Vec

1.4 Dyskusja wyników

Otrzymane wyniki pokazują różnice w sposobie, w jaki modele Word2Vec i Fasttext modelują przestrzeń wektorową. Obserwowane różnice w podobieństwach między słowami mogą wynikać z różnych metod treningu oraz charakterystyki korpusów użytych do trenowania modeli. Model z biblioteki fastText sugerował często różne warianty tych samych słów (nierzadko błędnie zapisanych). Jedną z możliwych przyczyn może być brak odpowiedniego preprocessingu zbioru danych uczących przed trenowaniem modelu. W przypadku Word2Vec generowane wyrazy były bardziej różnorodne i rzeczywiście pochodziły ze słownika jezyka polskiego.

2 Text Embeddings

2.1 Osadzenia zdań / tekstów (Fasttext i TF-IDF)

Dla osadzeń całych zdań i tekstów zastosowano dwa modele:

- **Bert** z biblioteki flair.
- **TF-IDF** z biblioteki sklearn.

2.2 Wizualizacja osadzeń tekstów za pomocą t-SNE

Podobnie jak w przypadku osadzeń słów, zastosowano t-SNE do wizualizacji osadzeń tekstów, z odniesieniem do etykiet anotacji. Wykresy mają interaktywny charakter, umożliwiając użytkownikowi odczytanie pełnego zdania i jego anotacji po najechaniu na punkt.

2.3 Klasteryzacja osadzeń anotowanych zdań

Osadzenia zdań zostały poddane klasteryzacji z wykorzystaniem dwóch algorytmów: HDBSCAN oraz KMeans. Wyniki klasteryzacji przedstawiono w zredukowanej przestrzeni z wykorzystaniem t-SNE. Interaktywne wykresy umożliwiają przeglądanie przypisanych klastrów i przypisanych do nich zdań.

2.4 Dyskusja wyników

Wyniki klasteryzacji oraz wizualizacji t-SNE pokazują, że różne modele osadzania zdań (Bert i TF-IDF) generują różne reprezentacje przestrzeni. TF-IDF jest bardziej czuły na częstość występowania słów, natomiast Bert efektywniej modeluje podobieństwa semantyczne, co jest szczególnie widoczne w wynikach klasteryzacji. Na szczególną uwagę zasługuje reprezentacja klastrów dla modelu Bert przy użyciu metody KMeans - zdania występujące w obrębie danych klastrów rzeczywiście dotyczą tych samych tematów.