Word2Vec

Projekt Zespołowy - laboratoria

Skład zespołu

- Wojciech Agaciński
- Piotr Bochra
- Adam Halicki
- Jakub Plebaniak

Przedstawienie problemu

Jednym z problemów związanych z przetwarzaniem języka naturalnego jest klasyfikacja statyczna – w tym przypadku określenie zbioru kategorii, do których dany zestaw słów należy.

Dzięki takiej klasyfikacji, możemy odtwarzać kontekst językowy oraz wnioskować, jakie znaczenie dany zbiór słów niesie.

Najprostszym przykładem problemu klasyfikacji są klasyfikatory binarne wykorzystywane w filtrach antyspamowych, decydujące czy wiadomość jest wiadomością niechcianą czy też nie.

Klasyfikacja języka naturalnego jest problemem trudnym, z powodu niejednoznaczności oraz trudnych do zbadania zależności pomiędzy słowami.

„Rzeczywistość jest niezwykle złożona”

Ogólna zasada działania

Word2Vec to zbiór metod wykorzystujących *word embedding* do uzyskania informacji odnośnie kategorii do których przyporządkować można dany tekst.

Działanie polega na zmapowaniu wszystkich unikalnych słów w zadanym korpusie języka na wielowymiarową przestrzeń wektorową, gdzie każde z tych słów utworzy w niej wektor.

Po utworzeniu takiej przestrzeni, możemy wnioskować o zbliżonym kontekście danych słów na podstawie bliskości odpowiadającym im wektorom w przestrzeni

Plany implementacji rozwiązania

- Wykorzystanie Word2Vec do klasyfikacji zadanego tekstu
- Moduł ma zwracać wektory dla dokumentów
- Do działania niezbędne jest wcześniejsze przekazanie korpusu, na podstawie którego utworzona zostanie przestrzeń wektorowa
- Tryb nauki oraz klasyfikacji
- Język programowania do stworzenia implementacji: Python

Dziękujemy za uwagę!