Lekturo-skurczator - etap 2

Dominik Doberski 133207, Michał Kukieła 132265

Przystosowanie zbioru do klasyfikacji

Dane utworzone w poprzednim etapie poddajemy przetworzeniu.

- Z zdań usuwamy znaki interpunkcyjne.
- Usuwamy podwójne spacje, oraz inne błędy które mogły przez przypadek zostać utworzone w wcześniejszym przetwarzaniu.
- Przy użyciu biblioteki stop_words usuwamy słowa, które zapewne nie niosą sobą zbyt wiele informacji (np: bardzo, dość, nasz, ponieważ, ten)

https://pypi.org/project/stop-words/

- Przy użyciu biblioteki pystempel dokonujemy lematyzacji wyrazów.

  https://pypi.org/project/pystempel/

- Lematyzacja to sprowadzenie wyrazu do jego formy bazowej
  - był, jest -> być
  - ciepło, ciepłem -> ciepły
- Lematyzacja nie zawsze działa idealnie (np. letni -> letny), jednak nie jest to problemem, ponieważ zależy nam jedynie na sprowadzeniu podobnych wyrazów do jednego, nie będziemy z nich odczytywać znaczenia.
- Znacznie ogranicza to przestrzeń wszystkich dostępnych słów i pozwala lepiej znaleźć zależności między przykładami.

Bag of words

Zastosowaliśmy reprezentację bag of words. Tworzymy zbiór wszystkich przetworzonych wyrazów. Z uzyskanego w ten sposób zbioru tworzymy listę, tak, aby każdy wyraz miał swój przypisany indeks.

Taka reprezentacja umożliwia nam zapisanie każdego zdania jako listę liczb. Jedno zdanie kodujemy jako listę o długości równej liczbie różnych wyrazów w zbiorze. Jeżeli dany wyraz nie występuje w zdaniu, to zostanie zakodowany jako “0”, jeżeli występuje to zostanie zakodowany liczbą wystąpień. Na przykład, jeżeli w zdaniu wyraz “dzień” pojawia się dwa razy, to w miejscu odpowiadającym temu wyrazowi pojawi się wartość “2”.

Przykład:

- Zdanie z zbioru danych: “Dzień był letni i świąteczny.”
- Usunięta interpunkcja, zamiana liter na małe: “dzień był letni i świąteczny”
- Zdanie nie zawiera wyrazu z listy “stop_words”
- Lematyzacja: “dzień być letny i świąteczny”
- Ponieważ jest to pierwsze zdanie w zbiorze, to reprezentacja bag of words będzie zawierała wyrazy “dzień”, “być”, “letny”, “i”, “świąteczny” na początku listy
- wektor liczb: 11111000000000000… zer jest prawie 5000 ponieważ tyle wyrazów ma nasz zbiór bag of words, a zdanie nie zawiera już żadnych innych wyrazów.

Dodatkowe działania

Jeżeli klasyfikacja na tak przetworzonym zbiorze nie będzie zadowalająca, to można zamiast interpretowania każdego wyrazu osobno, interpretować je parami lub trójkami. Podejście takie nazywa się ngrams i zostanie przez nas zastosowane jeżeli wyniki nie będą zadowalające.

Repozytorium projektu

https://github.com/Dombearx/TP_project