PROJEKT

Przetwarzanie języka naturalnego w systemach sztucznej inteligencji

autorzy: Marcin Retajczyk, Igor Ratajczyk

W projekcie wykorzystaliśmy implementację w języku Python oraz pracowaliśmy w Jupyter Notebookach.

Podpunkt 1.

Wyświetlenie rankingu słów.

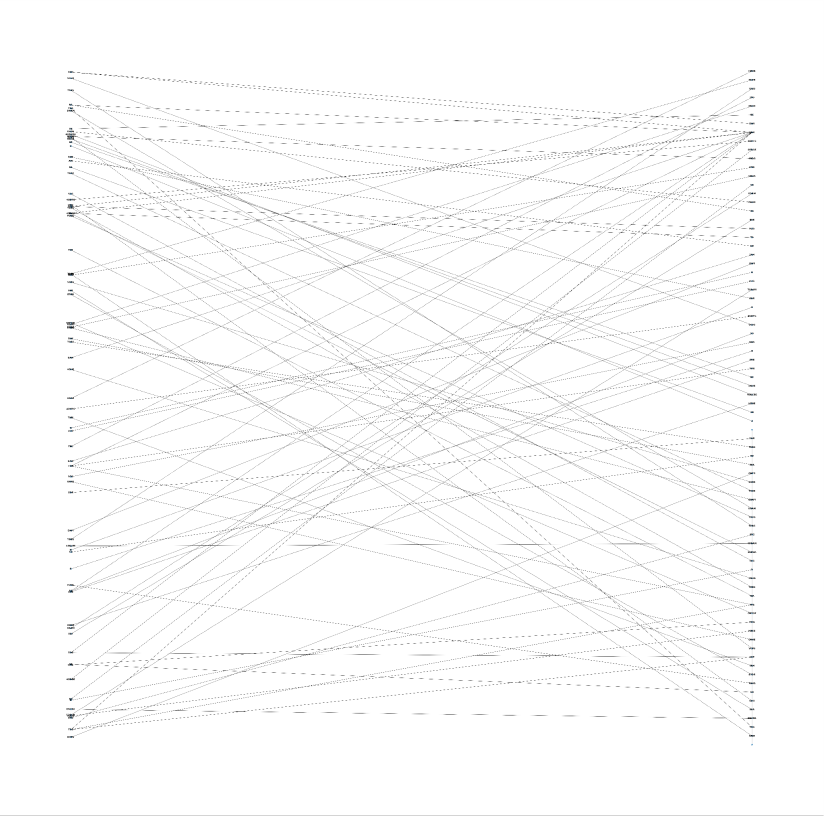
Ranking(R) Wyraz(str). F[procent](częstotliwość) R x F (słupkami poziomymi)

(Zestawienie TOP 30 wyrazów - Manuskrypt wojnicza - FSG.txt)

Według naszego przetworzenia, najczęściej występującym słowem w Manuskrypcie Wojnicza jest słowo : **8AM** występujące w częstotliwości 4,48 %.

Dla Manuskryptu wyznaczyliśmy unikalne słowa i dla każdego słowa wyznaczyliśmy potrzebne dane.

Podpunkt 2.
Sąsiedztwo między wyrazami.
Wyznaczenie grafu dwudzielnego pokazującego połączenia sąsiedztwa między wyrazami i wyznaczenie rdzenia języka- słowa z największą ilością połączeń.



Na poniższym grafie możemy wyczytać, że rdzeniem języka - słowem które ma najwięcej różnych połączeń jest słowo najwięcej występujące w tekście - **8AM.**

Z racji że graf przy dużej ilości słów stawał się nieczytelny - ilość połączeń była na tyle duża że nie można było nic widzialne, stwierdziliśmy, że ograniczymy ilość słów z największą liczbą sąsiadów pewnej liczby, która pozwoli na odczyt rdzenia języka.

Podpunkt 3.

W ramach akapitów badanie bigramów. Wyznaczenie listy słów, które są sąsiednie, ale w ramach jednego akapitu. Wyznaczenie ilości występowania takiego zestawienia w całym tekscie.

Dla przykładu akapitu : **wyr1, wyr2, wyr3, wyr4, wyr5 …**

Wyr1 wyr2 [ilosc wystąpień]

Wyr2 wyr3 [ilosc wystąpień]

Wyr3 wyr4 [ilosc wystąpień]

Wyr4 wyr5 [ilosc wystąpień]

…

Wynik w liczbach po przetworzeniu:

```
Ilość par słow wg jednego akapitu:  10519
Ilość unikalnych par słow:  9718
TOE     8AM   ilośc wystapień:  30
TOE     TOE   ilośc wystapień:  19
8AM     8AM   ilośc wystapień:  12
8AM     HZG   ilośc wystapień:  11
TOR     8AM   ilośc wystapień:  10
TG     8AM   ilośc wystapień:  9
SOE     8AM   ilośc wystapień:  9
TOR     TOE   ilośc wystapień:  8
TOE     SOE   ilośc wystapień:  8
8AM     HZOR   ilośc wystapień:  8
8AM     SO   ilośc wystapień:  7
TOE     TOR   ilośc wystapień:  7
8AM     8AE   ilośc wystapień:  7
OR     AM   ilośc wystapień:  7
8AM     TOE   ilośc wystapień:  7
TOE     HZOE   ilośc wystapień:  6
8AM     HZOE   ilośc wystapień:  6
HZG     8AM   ilośc wystapień:  6
8AM     TOR   ilośc wystapień:  6
OHOE     TOE   ilośc wystapień:  6
8AN     8AM   ilośc wystapień:  6
8AM     8AN   ilośc wystapień:  6
TG     DTG   ilośc wystapień:  6
…
```

Zauważamy, że słowo **8AM**, dominuje wśród połączeń między wyrazami. Po wyświetleniu całości wyrazów, zauważamy że statystyczna większość par występuje dokładnie 1 raz.

Dla języka angielskiego, wykorzystaliśmy fragmenty książki Robinson Cruzoe i zastosowaliśmy te same algorytmy co do Manuskryptu.