NLP - projekt 1

Krzysztof Czerenko, Krzysztof Ferda

10 października 2024

1 Działanie programu

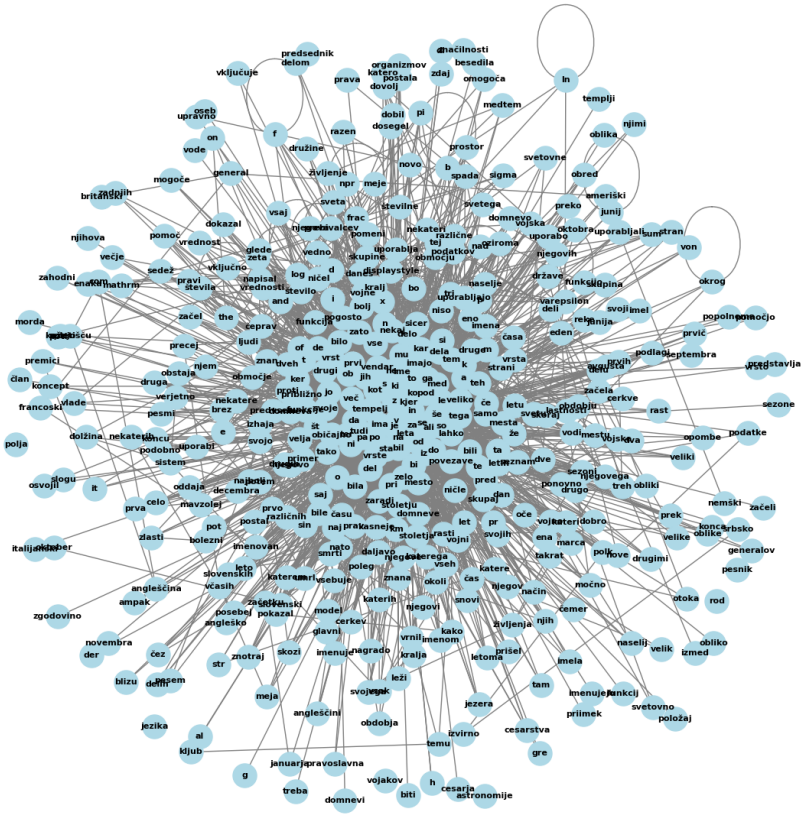
Program został napisany w języku python. Kod źródłowy znajduje się w tym samym folderze w postaci notatnika jupyter. Pobiera on artykuły z Wikipedii, a następnie je przetwarza i generuje wyniki w formie csv i grafu. Dokładniejszy opis jest w tekście.

2 Ranking najczęściej występujących słów

Word	**Count**	**Rank**	**Coef**
je	5094	1	5094
in	3706	2	7412
v	3182	3	9546
na	1836	4	7344
so	1650	5	8250
se	1454	6	8724
ki	1352	7	9464
za	1282	8	10256
z	882	9	7938
s	854	10	8540
da	844	11	9284
tudi	771	12	9252
kot	760	13	9880
leta	637	14	8918
pa	602	15	9030
bil	514	16	8224
iz	490	17	8330
od	482	18	8676
po	472	19	8968
ali	466	20	9320

3 Graf najczęściej występujących słów.

Przedstawia on zależności pomiędzy najczęściej występującymi słowami (więcej niż 20 wystąpień). Krawędź oznacza, że dwa słowa wystąpiły obok siebie w zdaniu. Mozna na nim zauważyć, że im częściej pojawiało się dane słowo, tym bliżej centrum grafu się znajduje.



Rysunek 1: Graf słów.

4 Słowa stanowiące kolejne procenty tekstu

1	je
2	in
3	v

Tabela 1: Tabela dla 10%

1	je
2	in
3	v
4	na
5	so
6	se
7	ki
8	za
9	z

Tabela 2: Tabela dla 20%

1	je
2	in
3	v
4	na
5	so
6	se
7	ki
8	za
9	z
10	s
11	da
12	tudi
13	kot
14	leta
15	pa
16	bil
17	iz
18	od
19	po
20	ali
21	med
22	pri
23	do
24	n
25	bila
26	lahko
27	o
28	ga
29	displaystyle

Tabela 3: Tabela dla 30%

Pozostałe tabele dla 40% i 50% posiadały ponad 100 rekordów, z tego więc powodu nie zostały zaprezentowane w formie graficznej.

5 Wnioski.

Jak widać wartość "Coef"zgodnie z prawem Zipfa oscyluje w okolicach 8500 co dla większej ilości przetworzonego tekstu byłoby jeszcze lepiej widoczne.