

1 Zespół i temat

Skład zespołu

1. Kajetan Bilski

Temat

Przewidywanie ilości punktów pod postami/komentarzami w subreddicie r/CasualConversation. Punkty przydzielane są przez użytkowników serwisu. Każda osoba głosująca że post/komentarz jest dobry dodaje mu 1 punkt, a każda która głosuje że jest zły zabiera 1.

2 Źródła danych i korpusy

Jako pełny korpus języka angielskiego użyłem gotowy korpus Open American National Corpus. Zawiera on teksty zbierane od 1990 roku z różnych kategorii takich jak konwersacje, czasopisma, fikcja itp.

Rozmiar: 6.9 GB

Jako korpus wzorcowy użyłem korpusu `reddit-small` z biblioteki `convokit` zawierającego tekst z postów i komentarzy z popularnych postów z września 2018.

Rozmiar: 186 MB

Jako korpus wzorcowy użyłem korpusu `reddit-CasualConversation` też z biblioteki `convokit` zawierający wszystkie posty i komentarze z subreddita do października 2018.

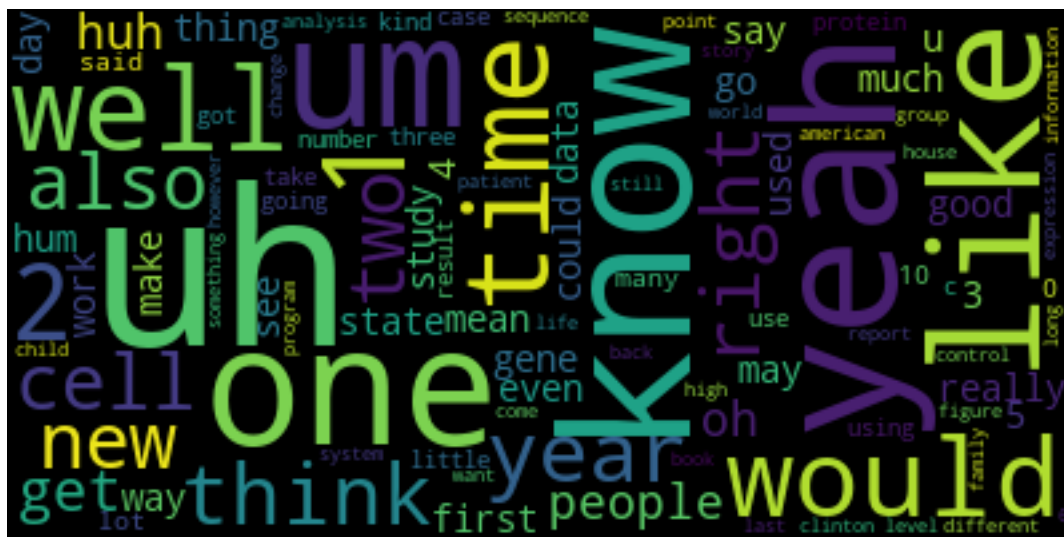
Rozmiar: 3.08 GB (przed samplowaniem)

3 Ocena pod względem jakości i reprezentatywności

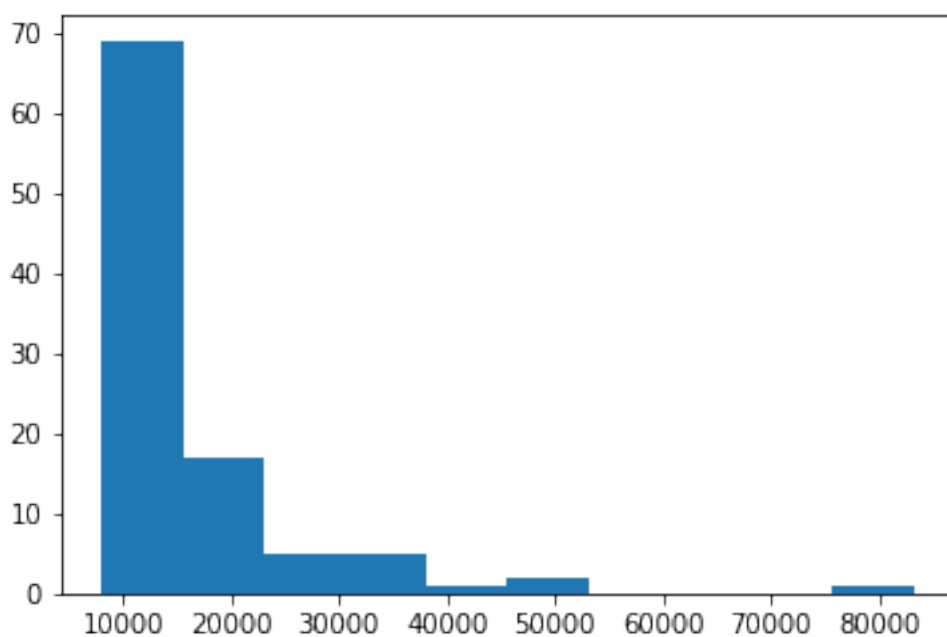
Do przeanalizowania zawartości korpusów użyłem histogramów i chmur słów. Korpusy zostały najpierw oczyszczone ze stopwordów, a słowa zlematyzowane.

Korpus ogólny

Chmura najczęściej pojawiających się słów.



Histogram częstotliwości 100 najczęściej występujących słów.

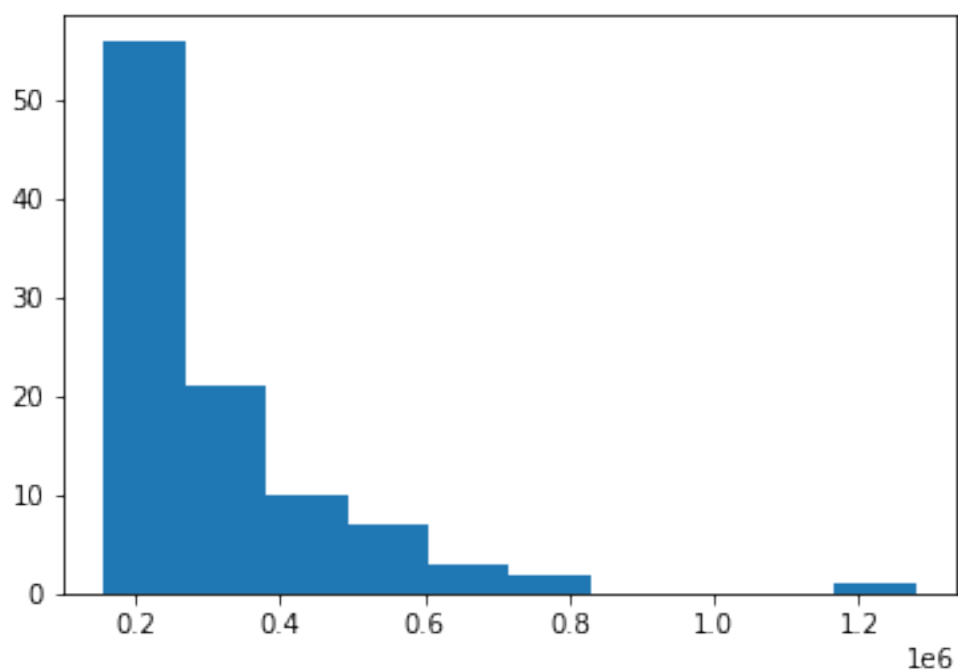


Korpus wzorcowy

Chmura najczęściej pojawiających się słów.

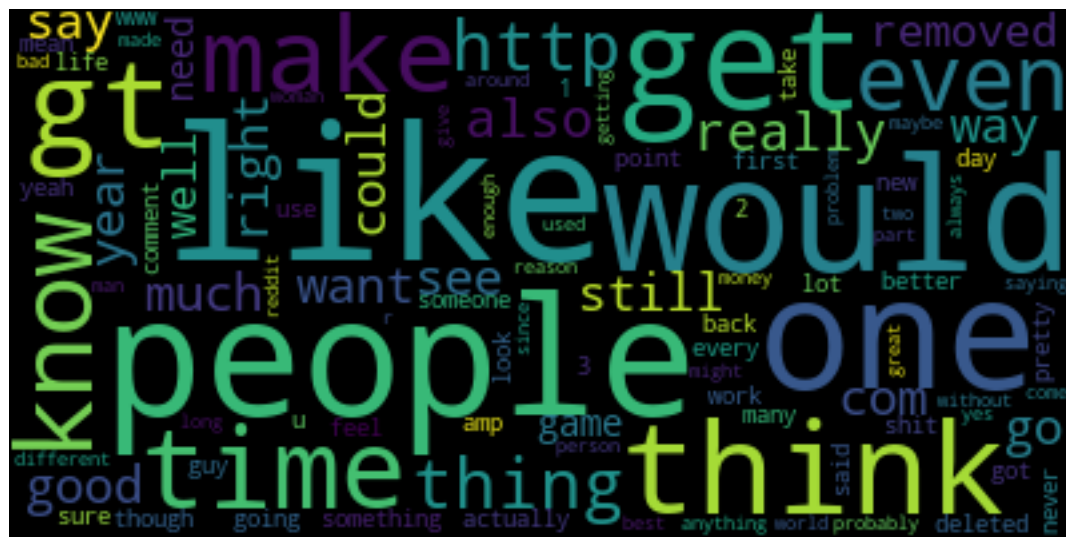


Histogram częstotliwości 100 najczęściej występujących słów.



Korpus anotowany

Chmura najczęściej pojawiających się słów.



A histogram showing the frequency distribution of the number of words in the training corpus. The x-axis is labeled 'Number of words' and ranges from 0 to 50,000. The y-axis is labeled 'Frequency' and ranges from 0 to 50. The distribution is highly skewed, with a peak frequency of approximately 55 for the first bin (0-10,000 words). The frequency drops sharply for subsequent bins, with a small secondary peak around 30,000 words.

Number of words (bin range)	Frequency
0 - 10,000	55
10,000 - 15,000	18
15,000 - 20,000	13
20,000 - 25,000	5
25,000 - 30,000	3
30,000 - 35,000	3
35,000 - 40,000	1
40,000 - 45,000	0
45,000 - 50,000	1

4 Ocena jakości anotacji korpusu PolE_{mo}

4 of 5

Dla zbioru zdania Kappa Cohena wygląda następująco:

	super	a	e	k
super	1.0	0.8425413961933904	0.9144596452221798	0.9163758130599683
a	0.8425413961933904	1.0	0.7876418822931861	0.7511360855789728
e	0.9144596452221798	0.7876418822931861	1.0	1.0
k	0.9163758130599683	0.7511360855789728	1.0	1.0

Ciekawa jest pełna zgodność anotatorów e i k. Jest to umożliwione przez fakt, że istnieje jedynie 12 tekstów zaanotowanych przez obu z nich.

Kappa Krippendorffa dla tego zbioru wynosi 0.8018978836994523.

Dla zbioru meta Kappa Cohena wygląda następująco:

	super	a	e	k
super	1.0	0.841610343202183	0.8619157313216369	0.8618052902164884
a	0.841610343202183	1.0	0.7123140150683738	0.7580248253380194
e	0.8619157313216369	0.7123140150683738	1.0	0.5023003723624327
k	0.8618052902164884	0.7580248253380194	0.5023003723624327	1.0

Kappa Krippendorffa dla tego zbioru wynosi 0.8259639267073995.