1 Zespół i temat

Skład zespołu

1. Kajetan Bilski

Temat
Przewidywanie ilości punktów pod postami/komentarzami w subreddicie r/CasualConversation. Punkty przydzielane sa przez użytkowników serwisu. Każda osoba głosujaca 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 jezyka 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 zawierajacego tekst z postów i komentarzy z popularnych postów z września 2018.
Rozmiar: 186 MB

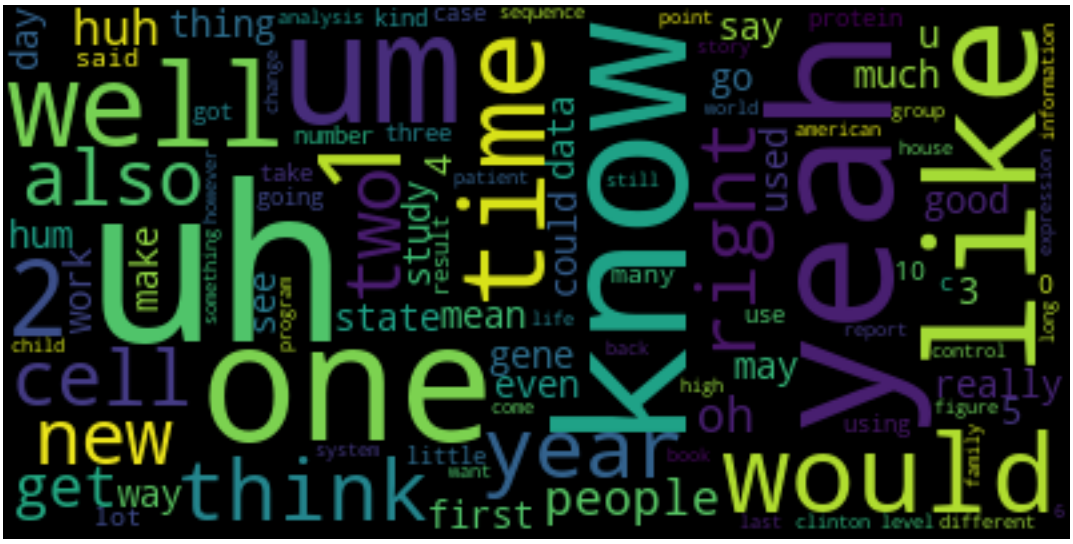
Jako korpus wzorcowy użyłem korpusu `reddit-CasualConversation` też z biblioteki convokit zawierajacy wszystkie posty i komentarze z subreddita do października 2018.
Rozmiar: 3.08 GB (przed samplowaniem)

3 Ocena pod wzgledem jakości i reprezentatywności

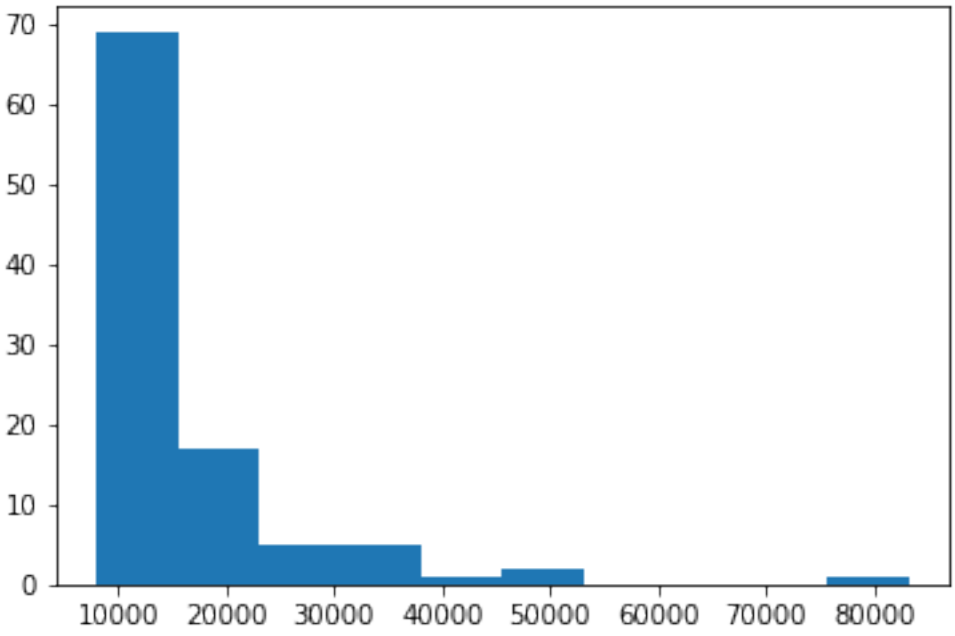
Do przeanalizowania zawartości korpusów użyłem histogramów i chmur słów. Korpusy zostały najpierw oczyszczone ze stopwordów, a słowa zlematyzowane.

Korpus ogólny

Chmura najcześciej pojawiajacych sie słów.



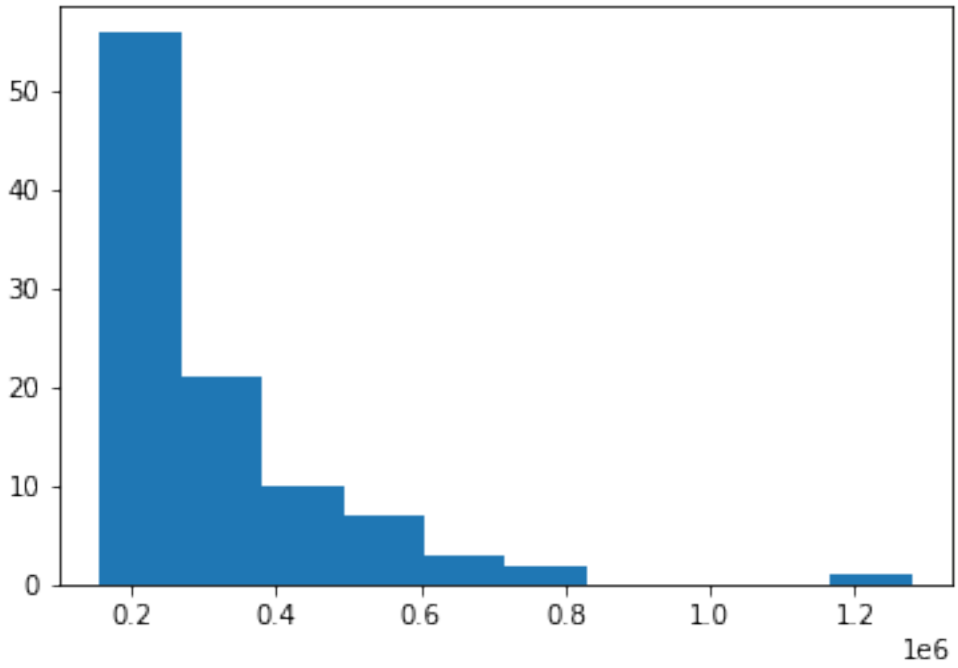
Histogram czestotliwości 100 najcześciej wystepujacych słów.



Korpus wzorcowy

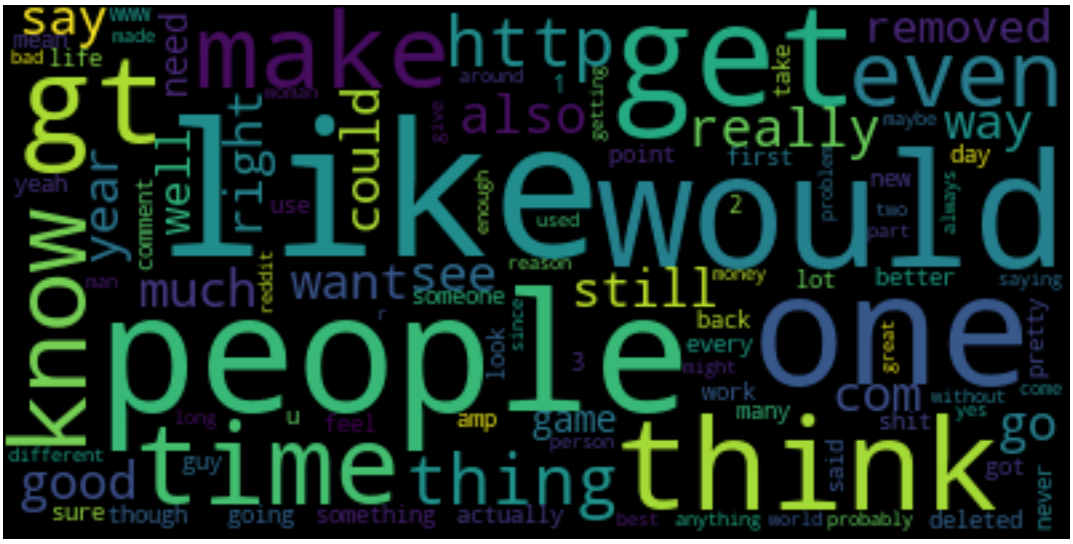
Chmura najcześciej pojawiajacych sie słów.

Histogram czestotliwości 100 najcześciej wystepujacych słów.

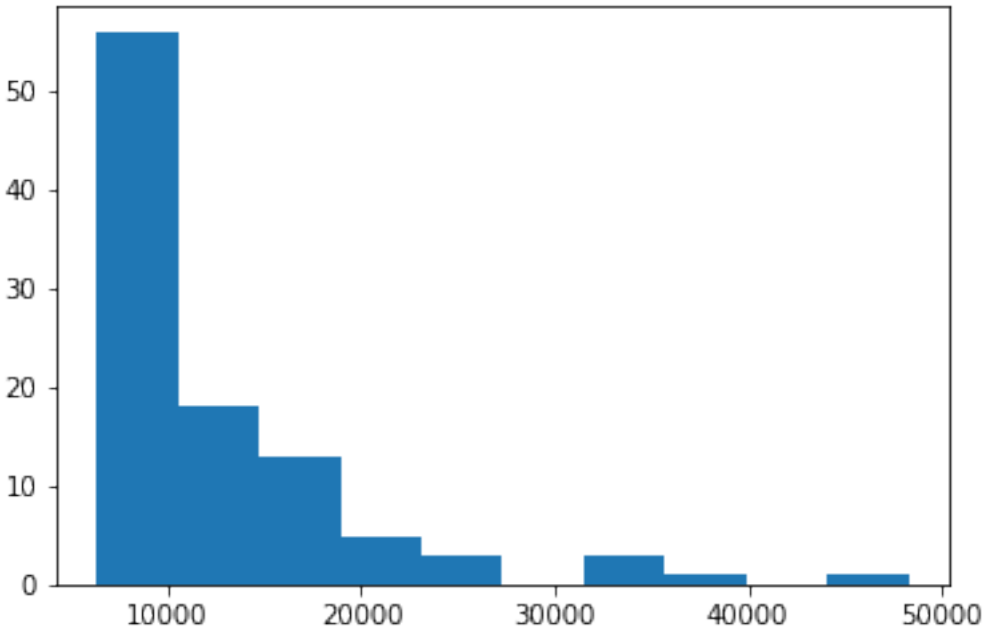


Korpus anotowany

Chmura najcześciej pojawiajacych sie słów.



Histogram czestotliwości 100 najcześciej wystepujacych słów.



Jak widać z chmur słów istnieje duże podobieństwo miedzy korpusem wzorcowym i anotowanym, ale różnia sie one znacznie od korpusu ogólnego. Histogramy wszystkich korpusów przypominaja dystrybucje zipfa.

4 Ocena jakości anotacji

Charakterystyka problemu uniemożliwia ocenienie jakości anotacji, ponieważ anotacja nie jest tworzona przez konkretnych anotatorów, a przez społeczność reddita, nie ma kryteriów umożliwajacych alternatywna metode anotacji. Anotacja w korpusie anotowanym jest jedyna

istniejaca anotacja do tego problemu.