SPRAWOZDANIE

Analiza tekstów – klasyfikacja lub ocena tekstu a słowa które na dany tekst się składają

Tomasz Drewek

23.01.2020

Spis treści

1. Cel

Celem zadania jest znalezienie oraz przeanalizowanie bazy wpisów internetowych takich jak opinie i komentarze pod kątem sprawdzenia słów z jakich składają się pozytywne, neutralne oraz negatywne wpisy.

Kod źródłowy dostępny pod adresem:
https://github.com/Krzaczek24/EksploratorTekstu

2. Baza danych

2.1 Źródło danych

Baza danych którą wykorzystałem została pobrana ze strony:
https://ermlab.com/en/blog/nlp/polish-sentiment-analysis-using-keras-and-word2vec/
Zawiera ona opinie oraz komentarze z następujących serwisów:

- Opineo (opinie)
- Twitter (tweety)
- Polish Academy of Science HateSpeech project
- YouTube (komentarze)

Baza w pełni jest złożona z polsko-językcznych tekstów, a ilość wpisów to niemal 940 tysięcy rekordów (88 MB). Znalezienie bazy z polskimi komentarzami okazuję się nie małym wyzwaniem, z jednej strony ciężko znaleźć serwis na którym wpisy są ocenione. Ewentualnie Allegro.pl lub Ceneo.pl. Z drugiej strony ciężko również o gotowe bazy.

Poniżej zrzut ekranu przedstawiający fragment bazy. Widać na nim błędy które naprawiłem.

	A	B	C
1	description	length	rate
2	Wszystko zgodnie z opisem. Wysoki poziom obslugi. Polecam	57.0	1
3	Solidny i żetelny sklep polecam wszystkim 100 procęt zadowolenia	64.0	1
4	Profesjonalizm i szybkość działania	35.0	1
5	chciałbym polecić zakupy w sklepie North.Ocena sklepu mówi sama za siebie mogę tylko dodać że cenny w sklepie są na poziomie punku ""10""		
6	Dotychczas nie spotkałem się z tak rewelacyjnie szybkimi odpowiedziami na moje emaile. N	202.0	1
7	Świetny sklep z częściami do agd. Dzięki nim zreaktywowałem dwa sprzęty bez pomocy serw	143.0	1
8	Realizacja zamówienia szybka i bezproblemowa.Żadnych problemów ze zwrotem towaru.	81.0	1
9	Rewelacja. Szybko, sprawnie i bez najmniejszych problemów.	58.0	1
10	Wielki ogromny pozytyw. Zamówiłem szufladę do lodówki. Okazało się, że się pomyliłem. Za	386.0	1
11	Sklep jest bardzo dobry. Jestem zadowolony z realizacji zamówienia. Polecam innym.	82.0	1
12	Polecam"	93.0	1
13	Jest wszystko w porządku nawet po złożeniu reklamacji.., która wyszła przez przeoczenie da	106.0	1
14	Dobrze zaopatrzony, super i fachowa obsługa, bardzo szybka realizacja	69.0	1
15	tranzakcja pomyślnie polecam !!!	33.0	1
16	Duży asortyment, błyskawiczne zakupy	36.0	1
17	Godny uwagi, dobrze zaopatrzony	31.0	1
18	szybko profesjonalnie i sprawnie. Liczymy na dalszą współpracę	62.0	1
19	3xN (na czs, na miejsce, napewno !!!!!!! POLECAM	50.0	1
20	Bezproblemowa obsługa. Polecam.	31.0	1
21	Wysyłają numer listu przewozowego co zdecydowanie ułatwia życie.		
22	Wszystko w jak najlepszym porządku. Polecam towar oryginalny	60.0	1
23	Sklep rewelacja. Bardzo szeroki asortyment. Na North zawsze można liczyć.		

2.2 Naprawa danych

Dane zawarte są dość wysokiej jakości jeżeli chodzi o ilość błędów. Głównie są to błędy w postaci nadmiarowych znaków nowej linii oraz oddzielenia kolumn przecinkiem gdzie teksty w pierwszej kolumnie również zawierają przecinki.

Pierwszym krokiem naprawienia danych który wykonałem było usunięcie nadmiarowych przejść do nowej linii. Za pomocą komendy ‘re.search()’ i odpowiedniego wyrażenia regularnego sprawdzałem czy dana linia zawiera 3 grupy tekstu. Pierwsza z nich - ‘.+?’ - to komentarz kończy się przecinkiem tuż przed rozpoczęciem drugiej grupy - (\d+?\.0)? - która to jest wartością długości tekstu ale nie zawsze występuje. Ostatnia grupa ‘-?[01]’ to ocena prezentowana za pomocą wartości -1 dla negatywnych opinii, 0 dla neutralnych oraz 1 dla pozytywnych. Znak dolara oznacza koniec linii.

```
re.search(".+?,(\d+?\.0)?,-?[01]$", line)
```

Kod działa w ten sposób że dopóki wczytana lnia nie pasuje do wzorca to jest doklejana do tekstu przechowywanego w pamięci. Gdy jednak dopasowanie nastąpi to zapamiętana sklejona linia jest zapisywana do pliku wyjściowego.

Zanim jednak nastąpi sam zapis, usuwam środkową kolumnę (długość tekstu) która jest zbędna, dodatkowo pozbywam się znaków cudzysłowia na skraju treści komentarza.

```
re.sub('^"?(.*?)"?,(\d+?\.0)?,(-?[01])$', r'\1,\3', ready_line)
```

Ostatnim krokiem naprawy rekordu jest pocięcie całej linii korzystając z komendy ‘split()’ w miejscach gdzie występuje przecinek. Wiedząc o tym że ostatnim elementem tak powstałej tablicy słów jest ocena a pozostałe elementy to słowa treści, sklejam ją ponownie wstawiając przecinek wyłącznie przed oceną a w pozostałych miejscach wstawiam spację.

Tym sposobem otrzymałem bazę gotową do przetworzenia. Poniżej widoczny fragment.

	description	rate
2	Wszystko zgodnie z opisem. Wysoki poziom obslugi. Polecam	1
3	Solidny i żetelny sklep polecam wszystkim 100 procęt zadowolenia	1
4	Profesjonalizm i szybkość działania	1
5	chciałbym polecić zakupy w sklepie North.Ocena sklepu mówi sama za siebie mogę tylko dodać że cenny w sklepie są na poziomi	1
6	Świetny sklep z częściami do agd. Dzięki nim zreaktywowałem dwa sprzęty bez pomocy serwisu. Pomogli mi dobrać części. Jestem	1
7	Realizacja zamówienia szybka i bezproblemowa.Żadnych problemów ze zwrotem towaru.	1
8	Rewelacja. Szybko sprawnie i bez najmniejszych problemów.	1
9	Wielki ogromny pozytyw. Zamówiłem szufladę do lodówki. Okazało się że się pomyliłem. Zadzwoniłem ustaliliśmy właściwą. Za 2	1
10	Sklep jest bardzo dobry. Jestem zadowolony z realizacji zamówienia. Polecam innym.	1
11	Polecam	1
12	Jest wszystko w porządku nawet po złożeniu reklamacji.. która wyszła przez przeoczenie danych zamówienia.	1
13	Dobrze zaopatrzony super i fachowa obsługa bardzo szybka realizacja	1
14	tranzakcja pomyślnie polecam !!!	1
15	Duży asortyment błyskawiczne zakupy	1
16	Godny uwagi dobrze zaopatrzony	1
17	szybko profesjonalnie i sprawnie. Liczymy na dalszą współpracę	1
18	3xN (na czs na miejsce napewno !!!!!!! POLECAM	1
19	Bezproblemowa obsługa. Polecam.	1
20	Wysyłają numer listu przewozowego co zdecydowanie ułatwia życie.Wszystko w jak najlepszym porządku. Polecam towar orygin	1
21	Sklep rewelacja. Bardzo szeroki asortyment. Na North zawsze można liczyć."szybko sprawnie bez zbędnych formalności. Monitor	1
22	Super sklep duży asortyment wszystko odbywa się szybko i sprawnie tak jak powinno być.	1

3. Obróbka danych

3.1 Przygotowanie dodatkowych danych

W pierwszej kolejności należało przygotować listę stopword-ów czyli słów nie wpływających na wydźwięk komentarza czy opinii.

Na początku lista ‘stopwords’ miała być pobrana z wbudowanych metod gotowych bibliotek, ale zawierała około 140 słów. Udało mi się odnaleźć bazę z 350 słowami (https://github.com/bieli/stopwords/blob/master/polish.stopwords.txt).

Przy dalszych pracach zauważyłem że biblioteka ‘Spacy’ zwracając zlematyzowane słowa udostępnia również informację czy czy dane słowo jest stopword-em oraz jaką jest częścią zdania.

Ostatecznie jedyną dodatkową bazą jest plik ‘nawl-analysis.csv’ pobrany z https://exp.lobi.nencki.gov.pl/nawl-analysis. Zawiera on bazę ok. 2.900 polskich słów wraz z ich znaczeniem, zostały one przypisane do jednej z grup, tj.:

- (H) happiness – szczęście
- (A) anger – gniew
- (S) sadness – smutek
- (F) fear – strach
- (D) disgust – wstręt
- (N) neutral – neutralość
- (U) unclassified – niezklasyfikowane

Baza ta posłużyła do wyznaczenia ilości słów o wskazanych powyżej emocjach w pozytywnych, neutralnych oraz negatywnych komentarzach.

3.2 Prace testowe na pomniejszonym zbiorze

Ze względu na rozmiary bazy, w celu szybkiego testowania kodu przygotowałem mechanizm generowania mniejszej bazy, w zależności od podanego parametru przepisuje co n-tą linię z pierwotnego pliku. Na potrzeby testów stosowałem parametr 0.1%, tak więc baza testowa która została również umieszczona w serwisie GitHub waży niecałe 100 KB. Dodam tylko że nawet taka próbka przynosiła dość ciekawe wyniki.

W trakcie prac doszedłem do wniosku że ciągłe przetwarzanie danych od początku bardzo utrudnia pracę, dlatego też utworzyłem mechanizm zapisujący efekty poszczególnych etapów. Zapis takich danych odbywa się po utworzeniu pomniejszonej bazy, po naprawie bazy, i co najważniejsze, po przetworzeniu słów przez lematyzer. Ostatni zapis generuje 4 pliki, osobny dla słów pozytywnych, neutralnych, negatywnych oraz dla wszystkich łącznie, wraz z ilością wystąpień danych słów. Po wykryciu plików, mechanizm zamiast, przetwarzać dane na nowo, wznawia pracę od ostatniego zapisanego etapu. Istnieje również możliwość wymuszenia odświeżenia tych plików oraz przełączania się między trybami pracy.

3.3 Obróbka właściwa

Po pierwsze podzieliłem opinie i komentarze pod kątem ich oceny, okazało się że treści wszystkich neutralnych tekstów są równe zeru – dlatego też miałem możliwość pracy wyłącznie na tekstach pozytywnych i negatywnych. Ze względu na czas poświęcony na poszukiwania polskiej bazy postanowiłem nie rezygnować i nie zmieniać bazy na inną.

Po uzyskaniu wszystkich słów, przekazuję je do metody filtrującej w skład której wchodzi również lematyzacja. Niestety w ramach biblioteki ‘Spacy’ nie wchodzi mechanizm dokonujący stemming-u, a w przykładach twórca sam używa stemming-u z biblioteki ‘NLTK’ która nie wspiera polskiego języka.

Metoda filtrująca w pierwszym kroku inicjuje lemmatyzator, który pobiera bazę polskich słów – wybrałem najbogatszą wersję ok. 500 MB. Następnie dla każdego wpisu za pomocą wyrażenia regularnego usuwane są znaki specjalne.

```
re.sub(r'[\\\d~`!@#$%^&*()_+{}|:"<>?,./;\'\[\]\-=]', '', comment).lower()
```

W kolejnym kroku rozpoczynam odfiltrowywanie słów. Wstępnie wyłączane były słowa z pobranego zbioru ‘stopwords’, ostatecznie jednak wykorzystałem możliwości biblioteki ‘Spacy’ która sama rozpoznaje które słowa należą do tej grupy.

Na koniec zapisywane są wyłącznie unikatowe słowa wraz z ilością ich wystąpień. Jest to najbardziej czasochłonny etap procesowania, ze względu na wiele modyfikacji tekstów za pomocą wyrażeń regularnych, dlatego też dane te są zapisywane do odpowiednich plików w celu zachowania ich.

W tej chwili mamy już gotowe dane do rozpoczęcia etapu prezentacji danych, a stąd już blisko do ich analizy.

3.4 Prezentacja przeprocesowanych danych

Dla każdej z czterech wydzielonych grup, tj.: słów pozytywnych, neutralnych negatywnych oraz wszystkich łącznie wyświetlany jest WordCloud czyli chmura słów. Metoda prezentująca jako wejście otrzymuje n najczęściej występujących słów w danej grupie.

Początkowo do zasilenia wspomnianej wyżej chmury słów wykorzystywałem wbudowaną w bilbiotekę ‘NLTK’ funkcjonalność, czyli ‘FreqDist’ która dla otrzymanego ciągu słów zwraca pojedyncze wystąpienia tych słów wraz z wartością równą ich częstotliwości w wejściowym ciagu.

Ostatecznie jednak sam dokonuję tego obliczenia, dlatego tak jak już pisałem w zapisanych przez mechanizm plikach, znajdują się już słowa z ich częstotliwościami występowania.

Czas najwyższy na prezentację wyników procesowania.

Wszystkie poniższe zrzuty ekranu prezentują zbiory słów po podaniu na wejście 200 najczęstrzych wyrazów.

Zbiór słów pozytywnych:

positive

Zbiór słow neutralnych:

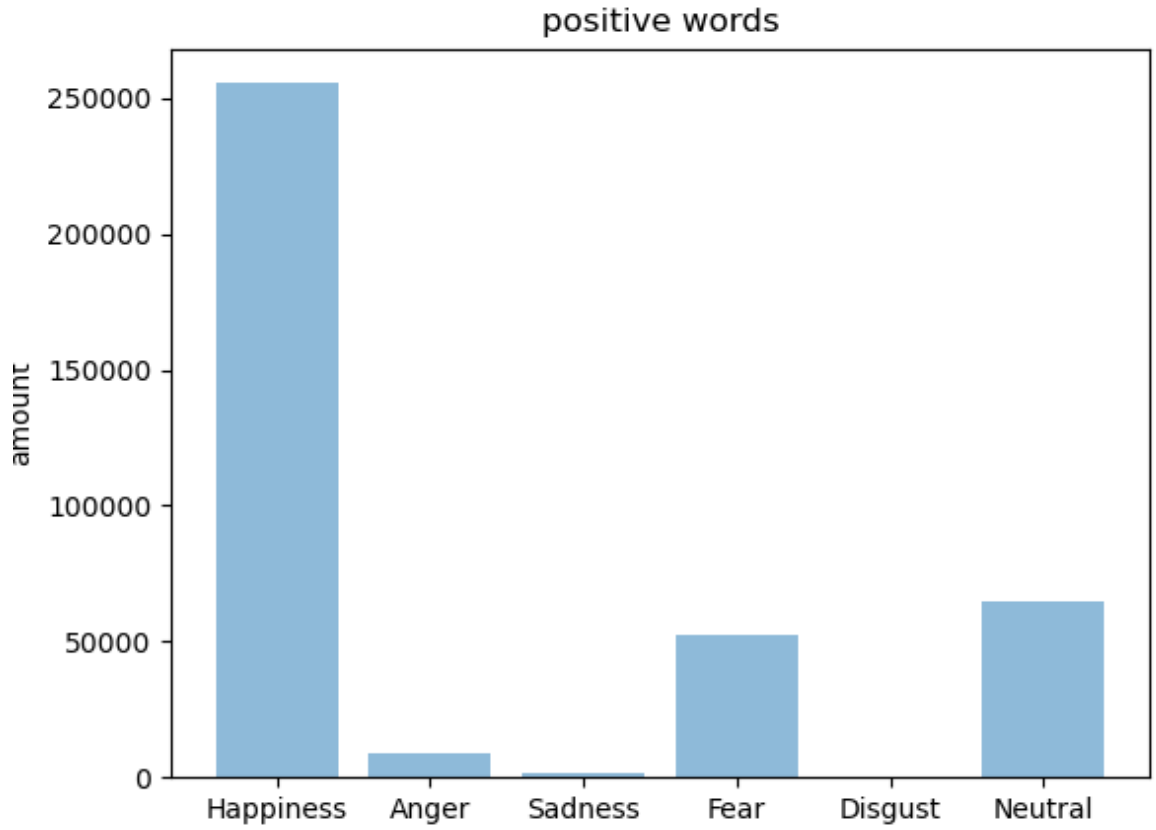
neutral

Zbiór słów negatywnych:

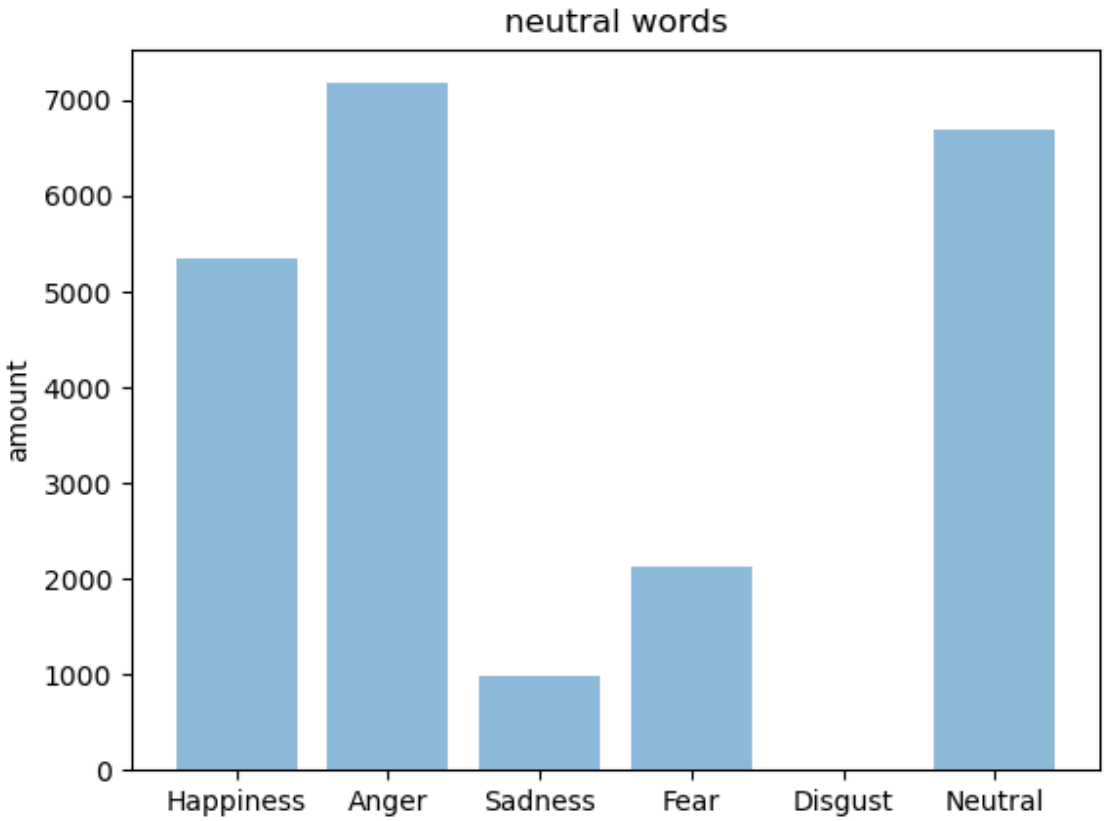
Zbiór wszystkich słów:

Kolenymi prezentowanymi danymi są rozkłady ilości słów o wybranych emocjach dla każdej z grup z osobna. Czyli na przykład, ile słów utożsamianych ze szczęściem czy złością mamy w pozytywnych komentarzach itd.

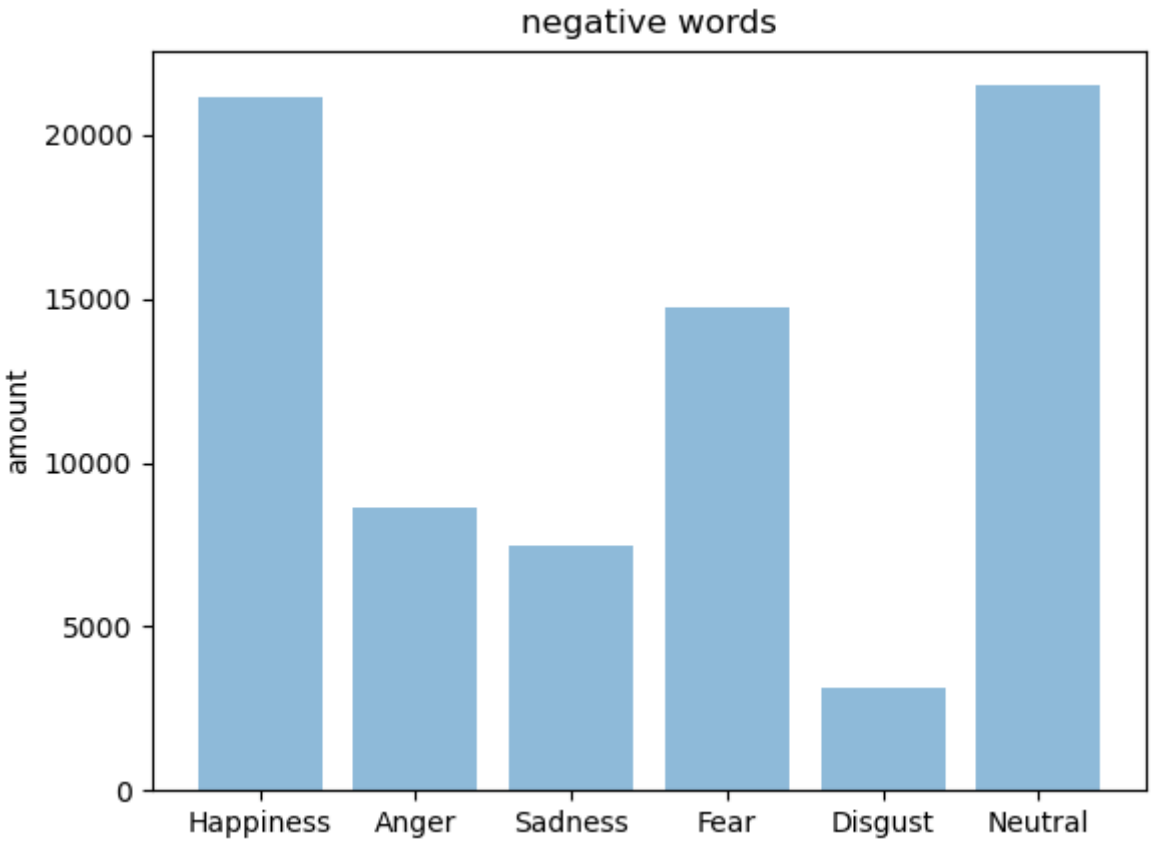
Zbiór pozytywnych opinii, komentarzy:



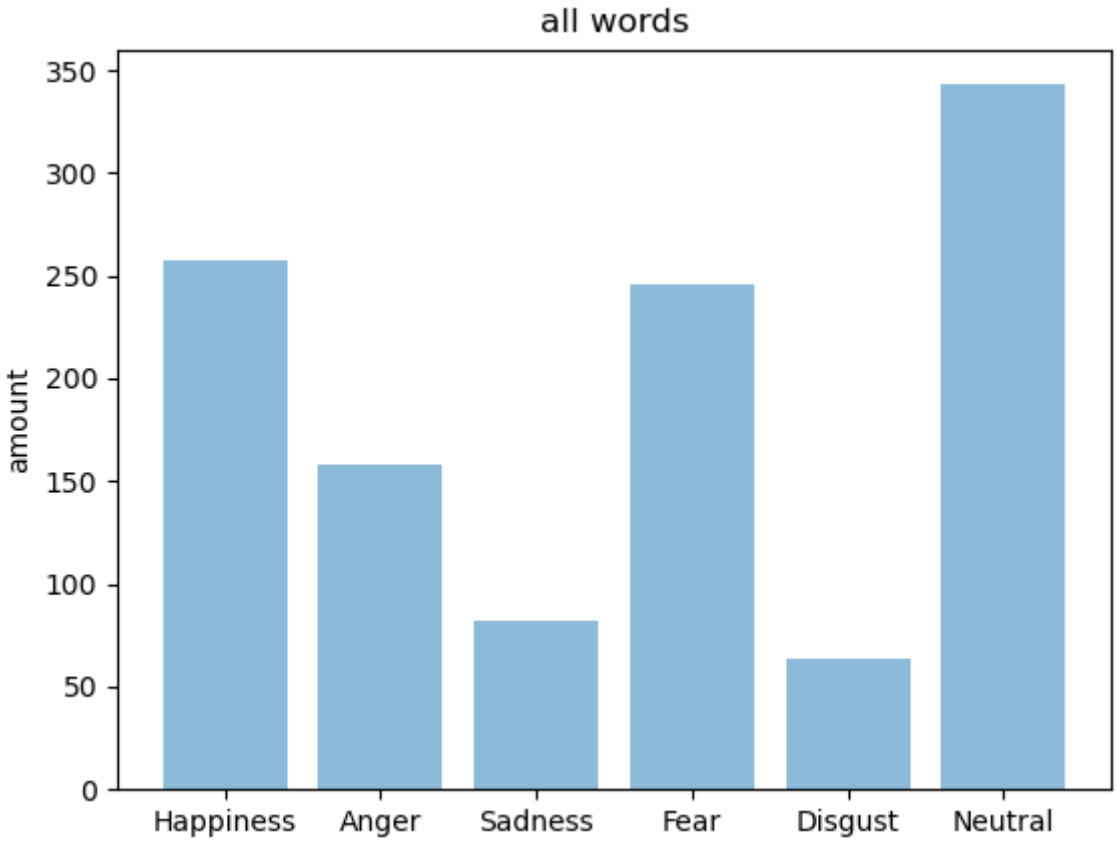
Zbiór neutralnych opinii, komentarzy:



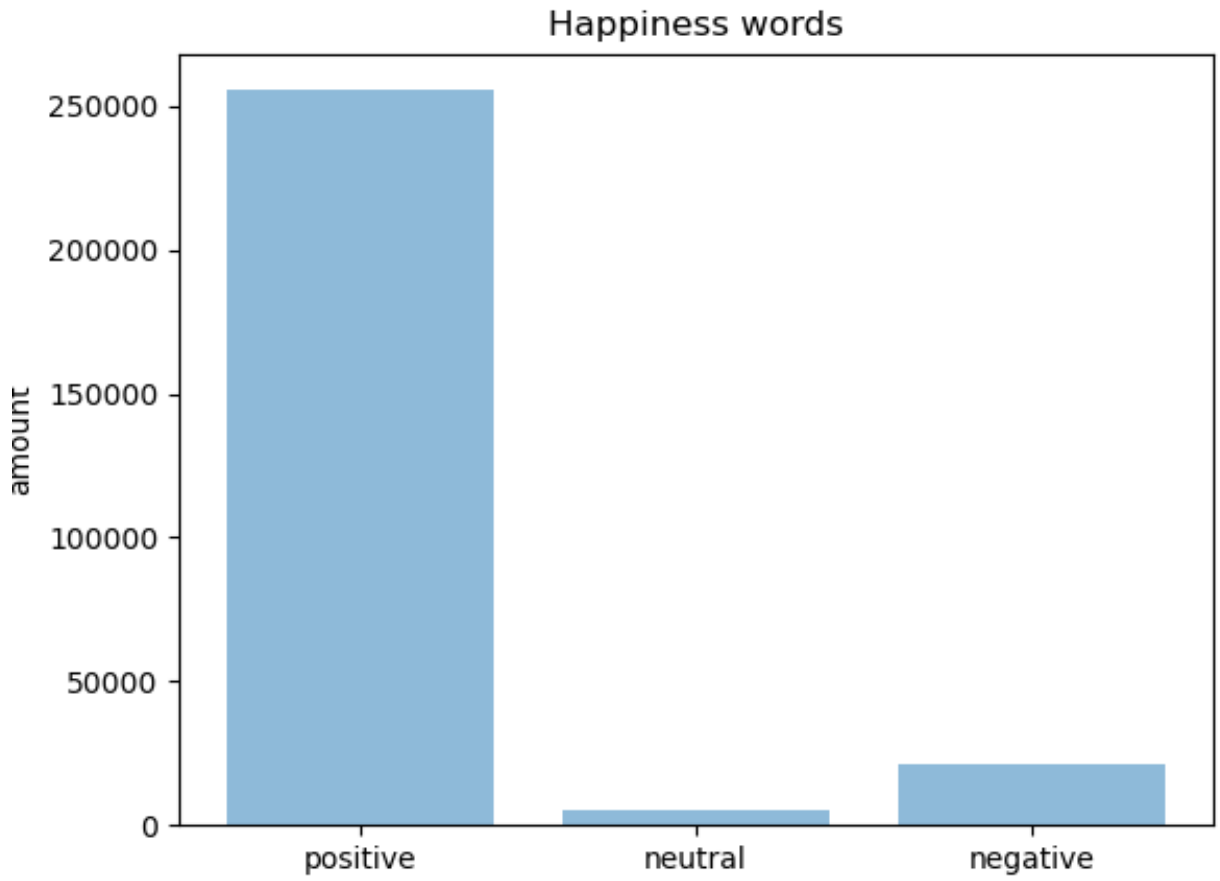
Zbiór negatywnych opinii, komentarzy:

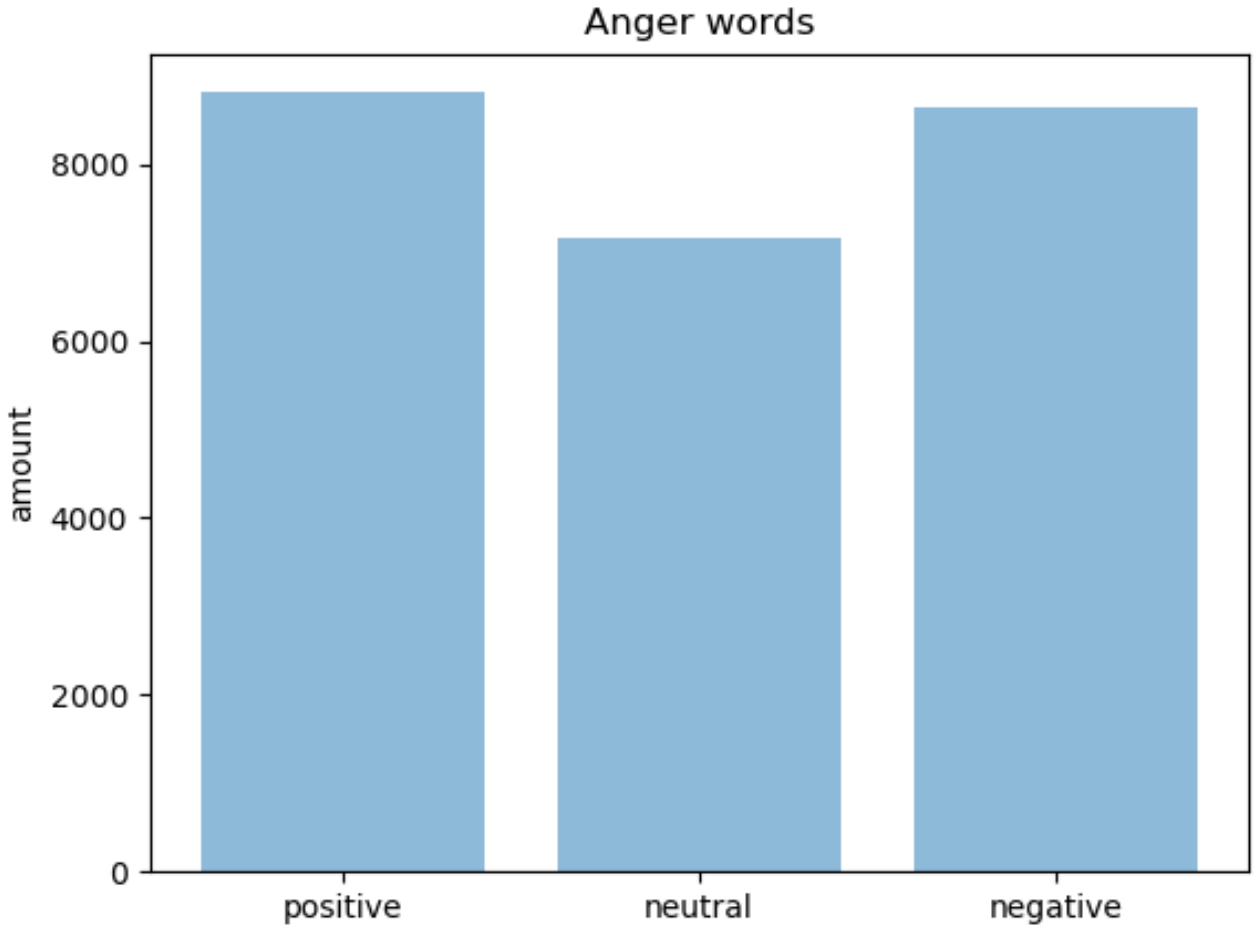


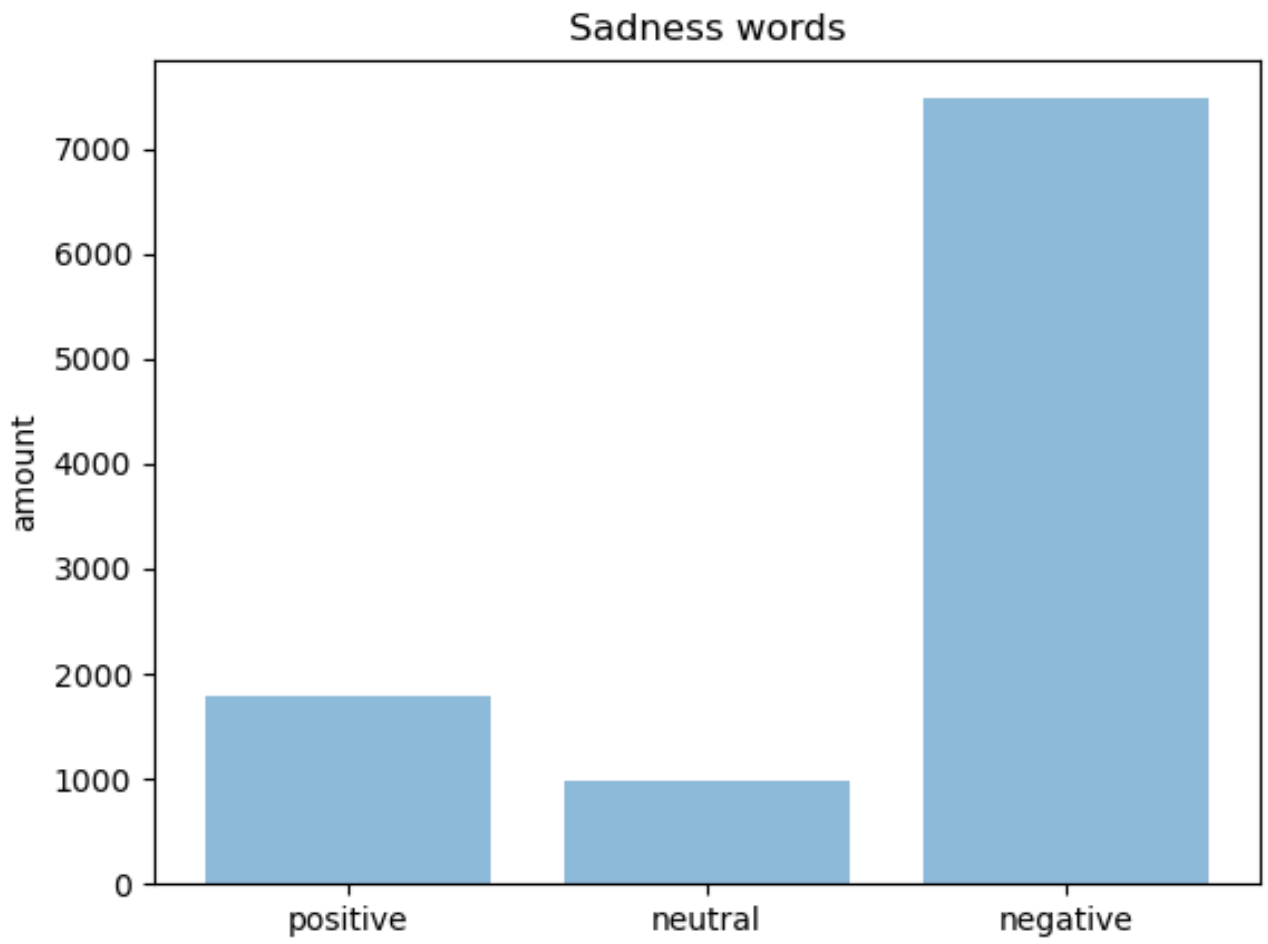
Zbiór wszystkich opinii, komentarzy:

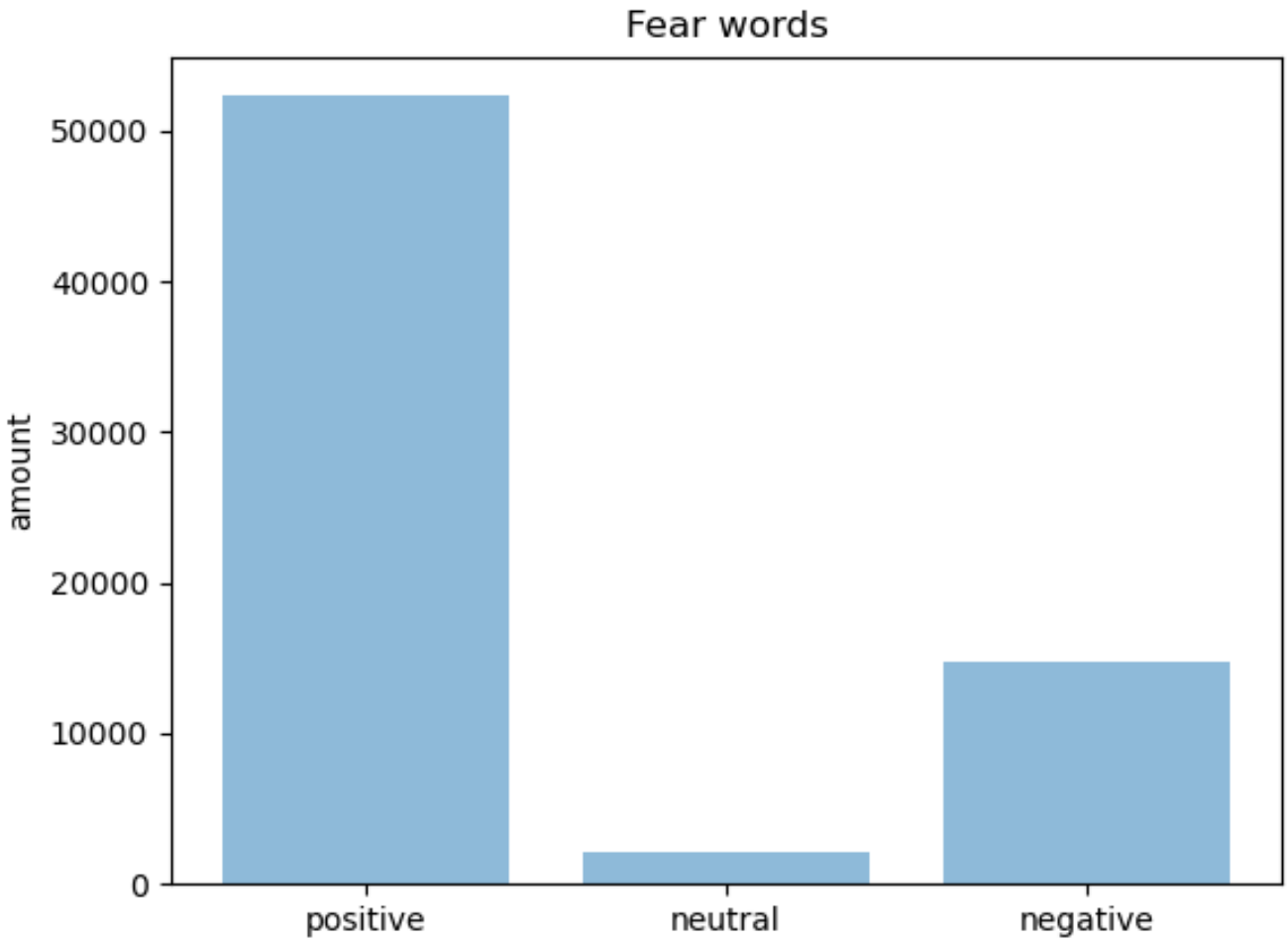


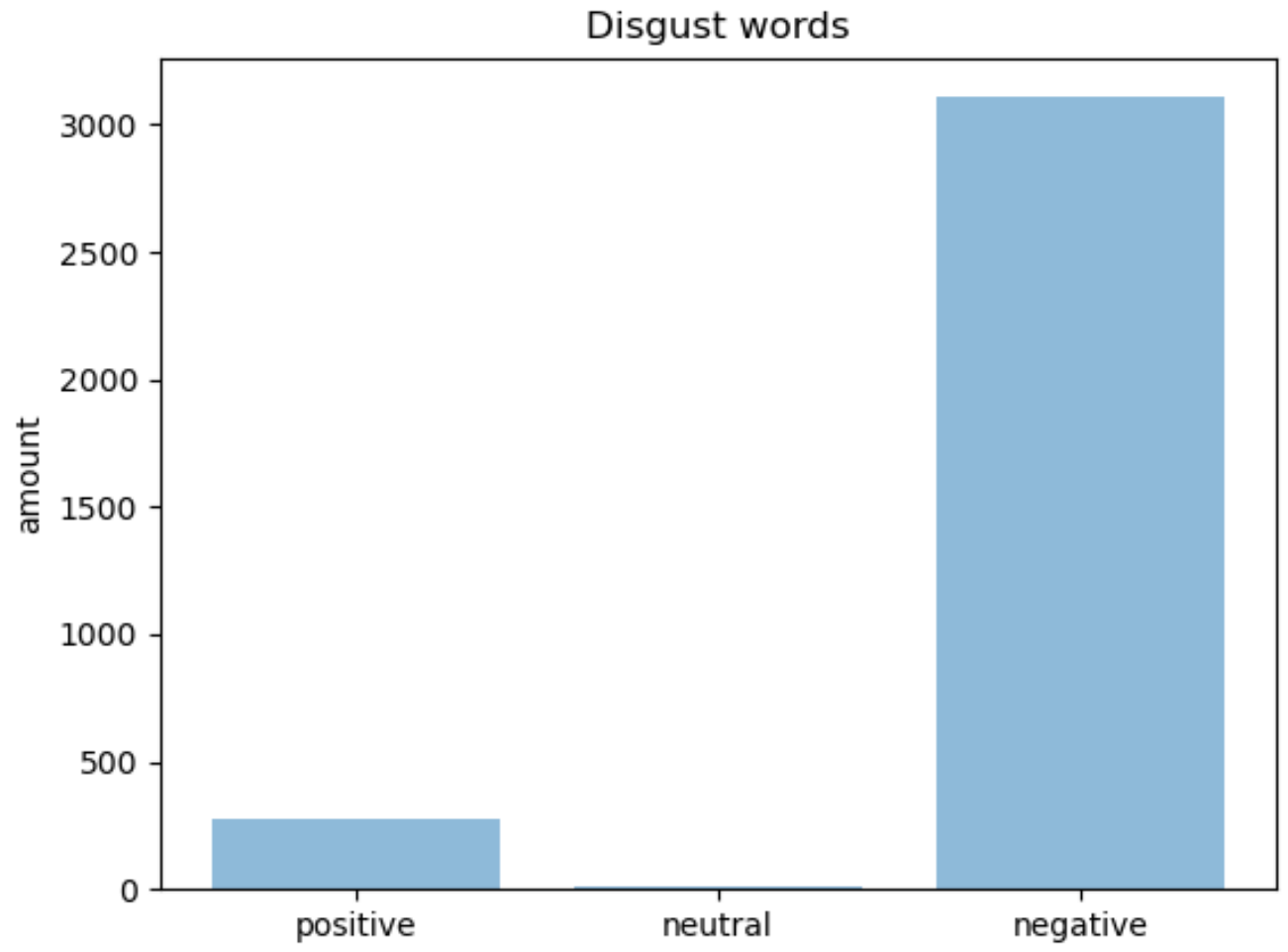
Tutaj zaprezentuję odwrotność czyli rozkłady słów z danego typu dla poszczególnych emocji.

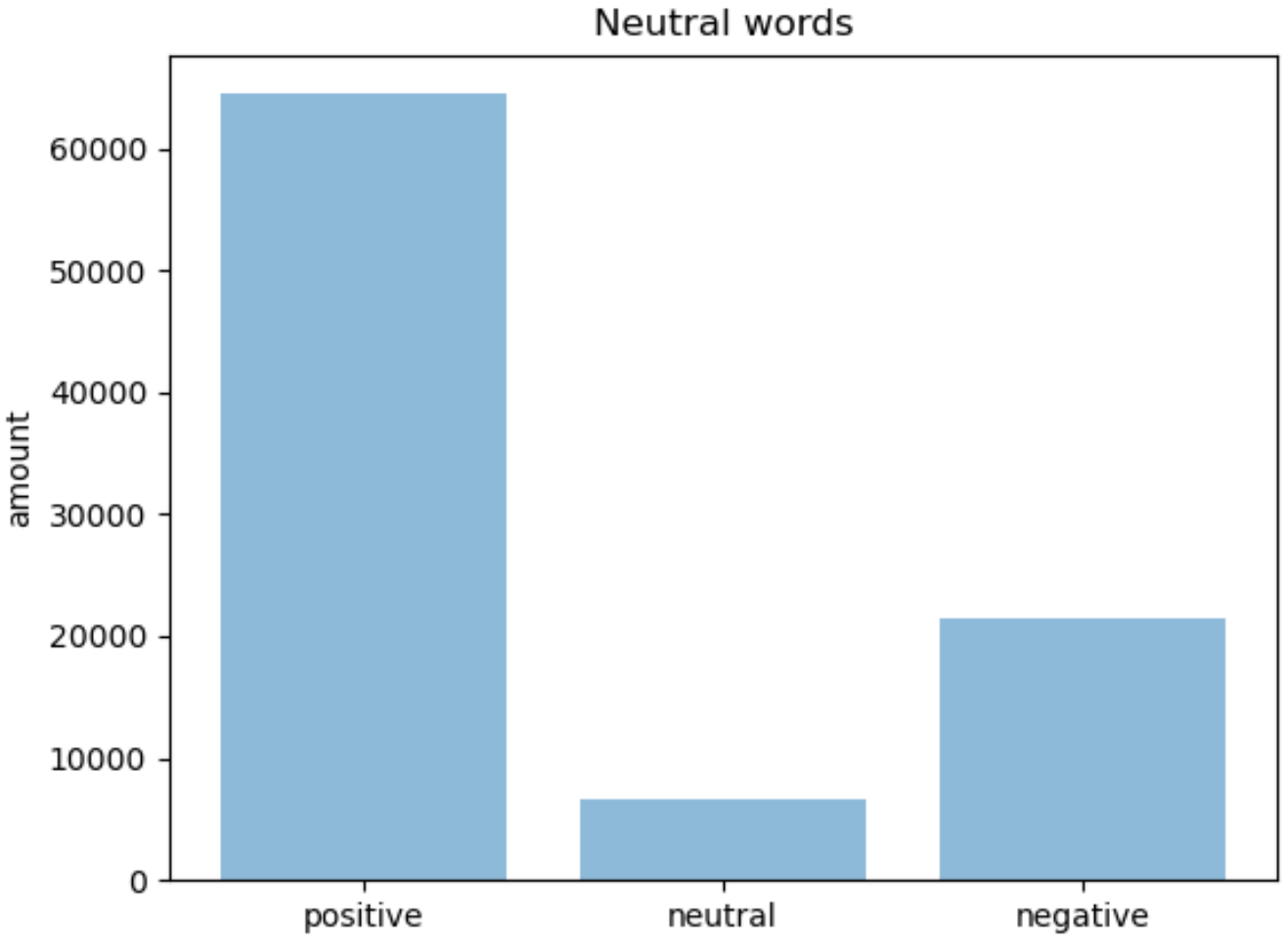





Sadness words
amount
7000
6000
5000
4000
3000
2000
1000
0
positive
neutral
negative


Fear words
amount
50000
40000
30000
20000
10000
0
positive
neutral
negative


Disgust words
amount
3000
2500
2000
1500
1000
500
0
positive
neutral
negative


Neutral words
amount
60000
50000
40000
30000
20000
10000
0
positive
neutral
negative

Na koniec jeszcze wykorzystanie podobnieństwa kosinusowego.

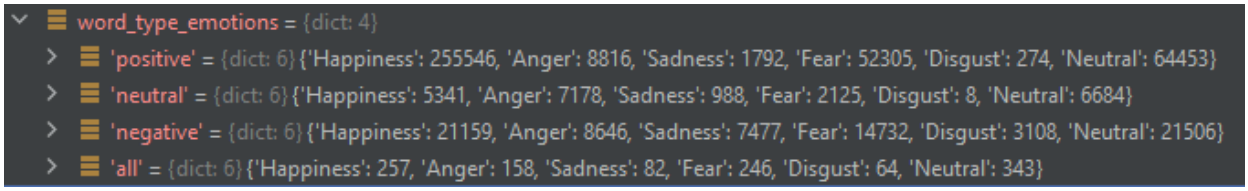
Cosinus similarity for type emotions

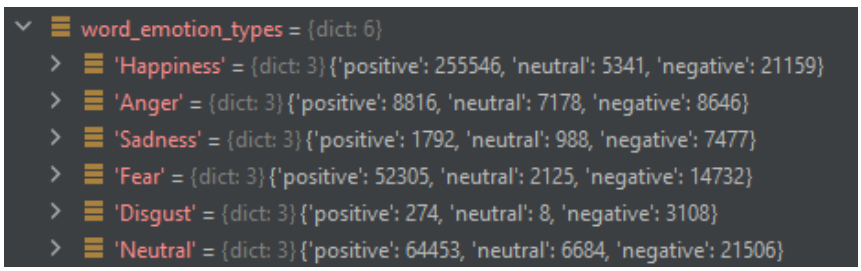
	positive	neutral	negative	all
positive	1.0			
neutral	0.6428	1.0		
negative	0.7997	0.88	1.0	
all	0.7185	0.8949	0.9882	1.0

Cosinus similarity for emotion types

	Happiness	Anger	Sadness	Fear	Disgust	Neutral
Happiness	1.0					
Anger	0.6754	1.0				
Sadness	0.3126	0.7906	1.0			
Fear	0.9815	0.7773	0.4886	1.0		
Disgust	0.1697	0.6585	0.9814	0.3544	1.0	
Neutral	0.9686	0.8226	0.5345	0.9971	0.3969	1.0

Dane tutaj wykorzystane:





4. Analiza danych

4.1 Pozytywne wpisy

Jak widać na załączonych chmurach słów, w wpisach pozytywnych występują słowa związane z szybką wysyłką, dobrą realizacją czy polecaniem, oraz słowa takie jak ‘super’, ‘świetny’ czy ‘zadowlony’.

Na wykresie słubkowym definytywną przewagę słów powiązanych ze szczęściem. Zdecydowanie dalej są słowa neutralne oraz związane ze strachem, ale należy wspomnieć że zwrot taki jak ‘strasznie szybko’ zawiera właśnie słowo łączące się ze strachem, albo takie słowo jak ‘rakieta’, może być stosowane zamiennie zamiast słowa ‘szybko’ czy ‘ekspresowo’. Stąd możliwe że to właśnie tego typu słowa wpływają na taką ilość słów strasznych w pozytywnych komentarzach.

Jeżeli spojrzymy na podobieństwo kosinusowe to zauważymy że jest spora różnica pomiędzy słowami pozytywnymi a negatywnymi, co ciekawe jeszcze większa różnica występuje pomiędzy słowami pozytywnymi a neutralnymi, być może jest to spowodowane sztuczym uzupełnieniem danych dla neutralnych wpisów.

4.2 Neutralne wpisy

Tutaj widzymy słowa takie jak ‘rok’, ‘mieć’, ‘być’, ‘sam’, ‘dostać’, ‘czekać’, ‘chyba’. Choć faktycznie głównie są to słowa które reprezentują raczej neutralne odczucia, to widać również sporo słów negatywnych i mniej pozytywnych.

Podobieństwo kosinusowe to potwierdza, zdecydowanie bliżej opisom neutralnym do negatywów niż do pozytywów.

Jeżeli spojrzymy na rozkład emocji dla wpisów neutralnych to łatwo zauważymy że przodują słowa o wydźwięku neutralnym, szczęścia oraz gniewu.

4.3 Negatywne wpisy

Wśród słów negatywnych można zauważyć wulgaryzmy, lub np. słowo ‘policja’ czy ‘XD’. Co ciekawe, opierając się na pobranej przeze mnie bazie emocji widać że przodują w nich słowa szczęścia oraz neutralne, a dopiero potem strachu i pozostałe.

4.4 Ogólne podsumowanie

Warto zauważyć że słowa szczęścia głównie dotyczą wpisów pozytywnych tak samo jak słowa związane ze strachem. Z kolei słowa powiązane z odrazą i smutkiem wskazują na opinie negatywne.

Być może warto byłoby w ramach rozszerzenia tego ćwiczenia dodać i przetestować klasyfikację opartą na emocjach i sprawdzić w jakim procencie wskaże odpowiedni typ wpisu.

5. Inne

5.1 Wykorzystane narzędzia

- NLTK (Tokenizacja / częstotliowść)
- Morfeusz2 (Błąd wew. biblioteki)
- Spacy (Lematyzacja polskich słów)
- WordCloud (generowanie grafiki ze słowami)
- Numpy
- Pandas
- MatPlotLib
- Re

5.2 Czasy wykonywania

- Naprawa bazy (ok. 1.000.000 rekordów) - 20 sekund
- Ładowane bazy lematyzera - 15 sekund
- Przetwarzanie słów pozytywnych - ok. 1 godzina
- Przetwarzanie słów neutralnych - 1 sekunda (mało danych w dodatku błędne)
- Przetwarzanie słów negatywnych - ok. 20 minut
- Generowanie chmury słów - ok. 3 sekundy na każdy typ
- Generowanie wykresów słupkowych - natychmiast
- Generowanie tabel podobieństwa kosinusowego - ok. 1 sekundy