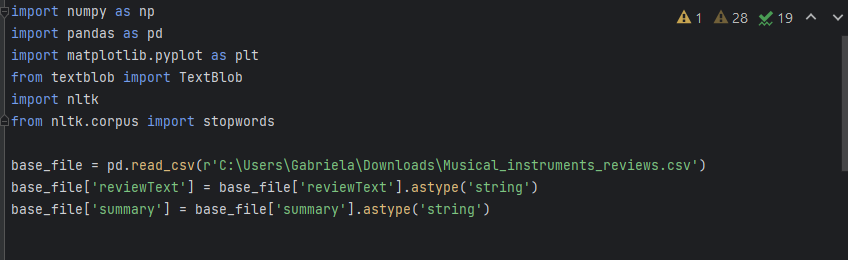
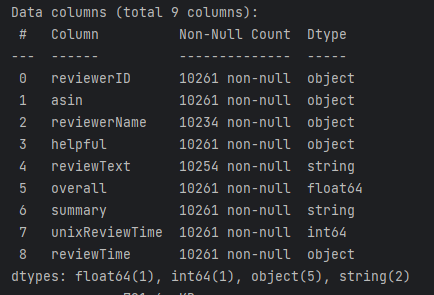
Celem programu jest przeanalizowanie zestawu danych, zawierających opinie produktów muzycznych z Amazona.

**Załadowanie Danych i Wstępna Inspekcja**

Zbiór danych ‘Musical\_instruments\_reviews.csv’ został załadowany i przeanalizowany pod kątem jego struktury i zawartości. W pierwszym kroku zaimportowano niezbędne biblioteki, a następnie kolumny ‘reviewText’ oraz ‘summary’ zostały przekształcone na typ string, a następnie usunięto wiersze zawierające brakujące wartości.



Następnie wyświetlamy podstawowe dane dotyczące zestawu danych (base\_file.dropna()).

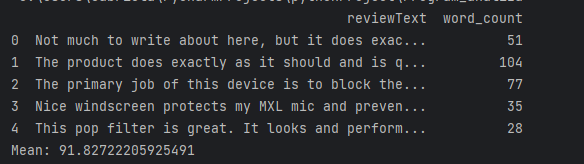


**Analiza Danych Tekstowych**

Obliczono cztery kluczowe metryki dla kolumn ‘reviewText’ i ‘summary’:

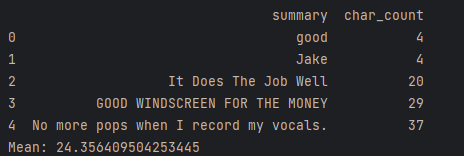
**Liczba Słów**

Mierzy liczbę słów w każdej recenzji.



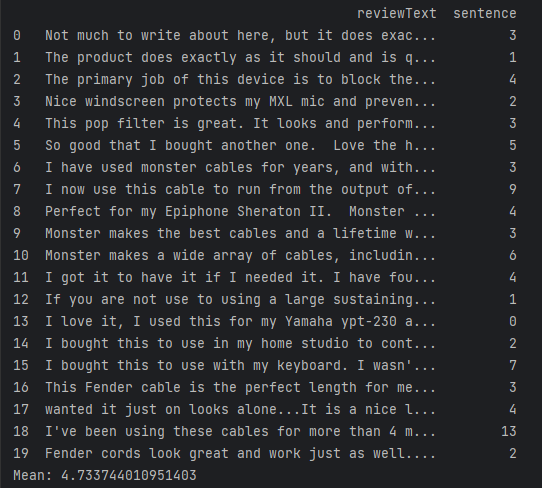
**Liczba Znaków**

Mierzy liczbę znaków w każdym streszczeniu.



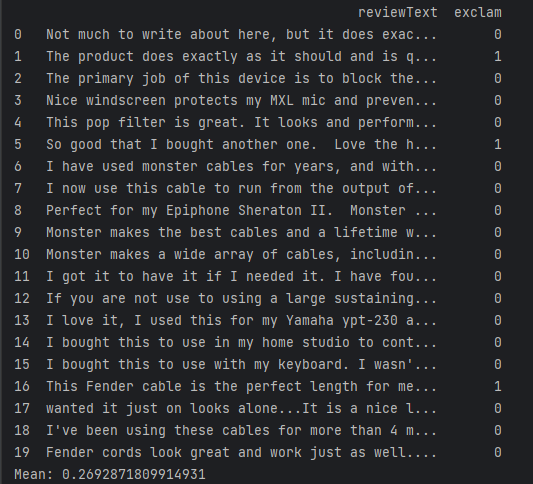
**Liczba Zdań**

Mierzy liczbę zdań w każdej recenzji (przybliżone przez liczbę kropek).



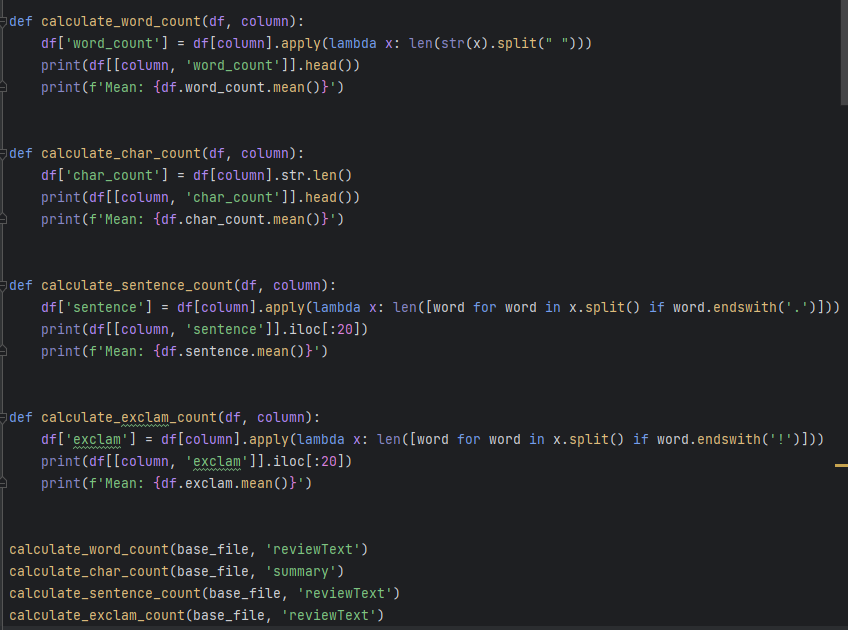
**Liczba Wykrzyknień**

Mierzy liczbę zdań wykrzyknikowych w każdej recenzji (przybliżone przez liczbę wykrzykników).



Te metryki dostarczają podstawowych informacji o strukturze i zwięzłości recenzji.

Poniżej przedstawiony został kod wywołujące powyższe wyniki.

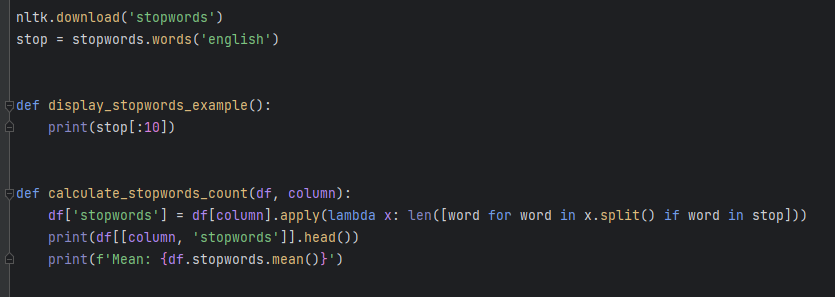


Średnie wartości tych metryk dostarczają informacji o długości i złożoności recenzji. Na przykład, średnia liczba słów w recenzji może wskazywać na typowy poziom szczegółowości, z jakim użytkownicy opisują swoje doświadczenia.

**Liczba Stop Words**

Obliczono liczbę stop words (często występujących słów jak 'the', 'is' itd.) w celu oceny językowej zawartości recenzji. Pomaga to w identyfikacji gęstości istotnych słów w porównaniu do często używanych słów.

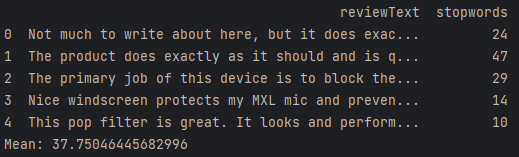
W tym celu zaimportowano bibliotekę nltk i jej niezbędne elementy.



Najpierw sprawdzamy, jakie słowa ‘stopwords’ najczęściej występują w naszej kolumnie ‘reviewText’.



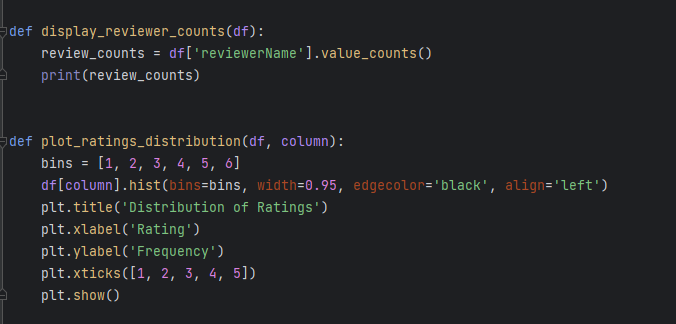
A następnie sprawdzamy liczbę stop słów przypadających na każdą opinię.

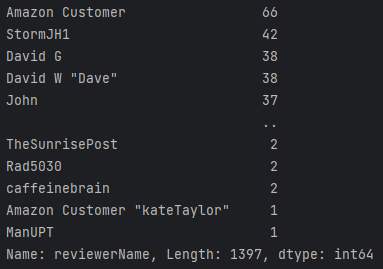


Analiza liczby stop words pomaga w ocenie, jak bogate w treść są recenzje.

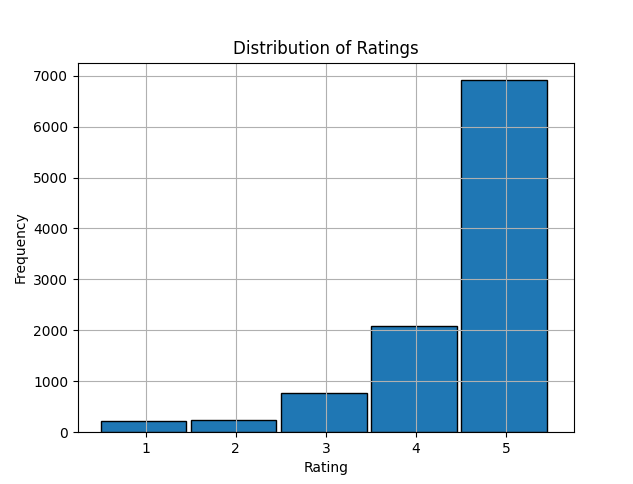
**Analiza Recenzentów i Ocen**

Przeanalizowano rozkład liczby recenzji na recenzenta oraz rozkład ocen, aby zrozumieć demografię zbioru danych i ogólne nastawienie do produktów. Oba te kody oraz wyniki można znaleźć poniżej.





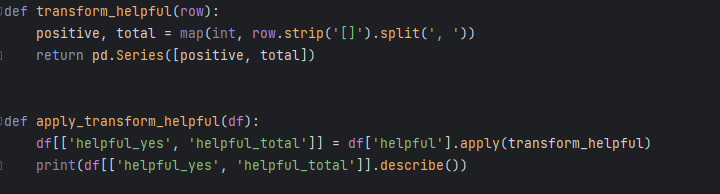
Wynikiem def display\_reviewer\_counts(df) jest powyższa lista. Z niej wynika, że na wszystkie opinie (10 261 pozycji) pojawia się tylko 1 397 komentujących. Najwyższa pozycja jest dla uniwersalnych komentujących, gdzie różni klienci mogą komentować pod tą samą nazwą i ta pozycja równa się 66 opiniom. Co ciekawe, tylko dwóch komentujących dało pojedyncze opinie.



Powyższy wykres reprezentuje, jak poszczególne pozycje zostały ocenione w skali od 1 do 5. Zdecydowanie wynika, że większość komentujących dało maksymalny możliwy rating (prawie 7 tysięcy pozycji).

**Analiza Głosów Pomocności**

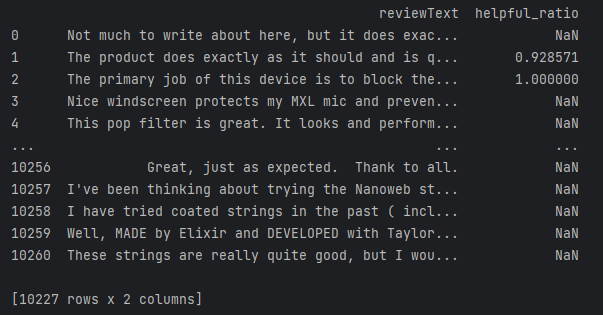
Ocena przydatności recenzji była przeprowadzona poprzez transformację i analizę kolumny 'helpful', która wskazuje, ile osób uznało recenzję za pomocną. Obliczono wskaźnik pomocności jako stosunek liczby głosów pomocnych do całkowitej liczby głosów.



Następnie obliczamy owy współczynnik i tworzymy wykres, pokazujący jego rozkład.

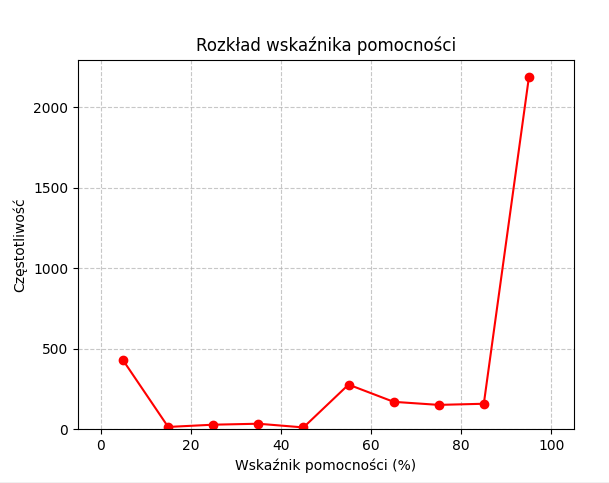


Analiza głosów pomocności pomaga zrozumieć, które recenzje są uznawane za najbardziej wartościowe przez społeczność użytkowników.



Natomiast z wykresu poniżej wynika, że zdecydowana większość komentarzy, które zostały otrzymały oceny, zostaje oceniona pozytywnie i jako przydatna opinia. Po przeciwnej stronie jest również wysoki współczynnik komentarzy z negatywnym współczynnikiem przydatności. Taki wynik wynika głównie z pozycji o małych wartościach oceny, np. [0,2] lub [0,1].

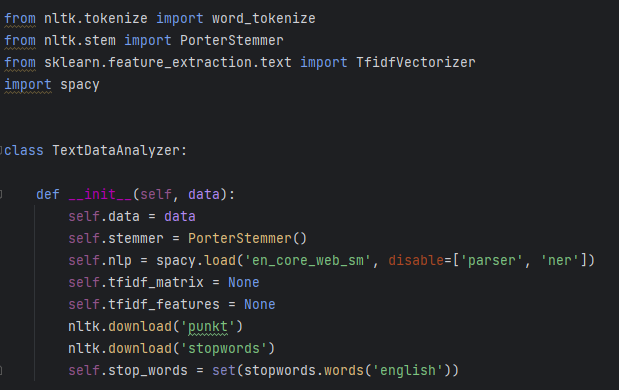
Należy również pamiętać, że powyższy współczynnik nie uwzględnia pozycji, gdzie pomocniczość nie została w ogóle oceniona, wyniki dla kolumny ‘helpful’ były równe [0,0].

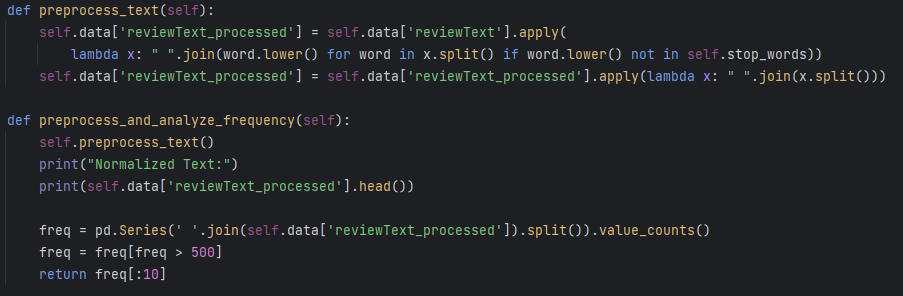


**Przetwarzanie i Analiza Tekstu**

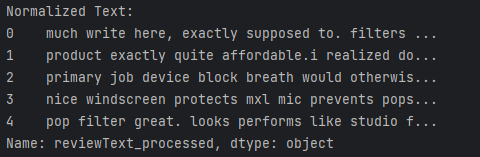
**Normalizacja**

Następnym elementem w programie jest utworzenie klasy składającej się z definicji, gdzie każda z nich odpowiada za inną funkcję przetwarzania tekstu. W tym celu również zostały zaimportowane następne biblioteki.

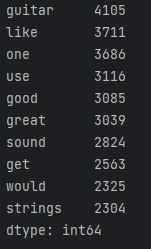


Tekst został znormalizowany przez konwersję na małe litery, usunięcie nadmiarowych spacji oraz usunięcie stop words. Jest to niezbędne dla dokładnej analizy tekstu i ekstrakcji cech.

Poniżej wyniki normalizacji tekstu:



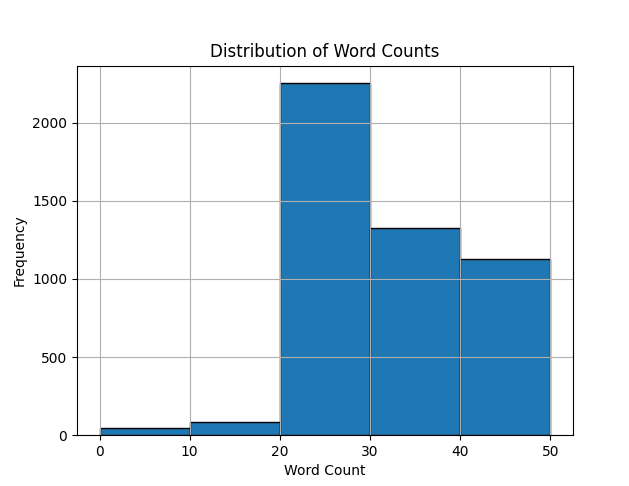
Również zanalizowano częstotliwość słów, aby zidentyfikować najczęściej występujące słowa w recenzjach, z wyłączeniem stop words. Pomaga to w zrozumieniu głównych tematów i zagadnień poruszanych w recenzjach.



Słowa takie jak: guitar, sound, strings były oczekiwanymi wynikami powyższej funkcji preprocess\_text, ze względu na specyfikę zestawu danych, tj. opinii o produktach muzycznych.

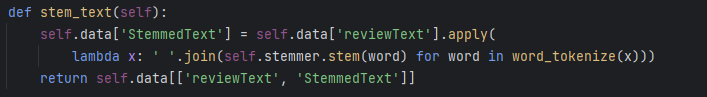
**Analiza rozkładu słów**

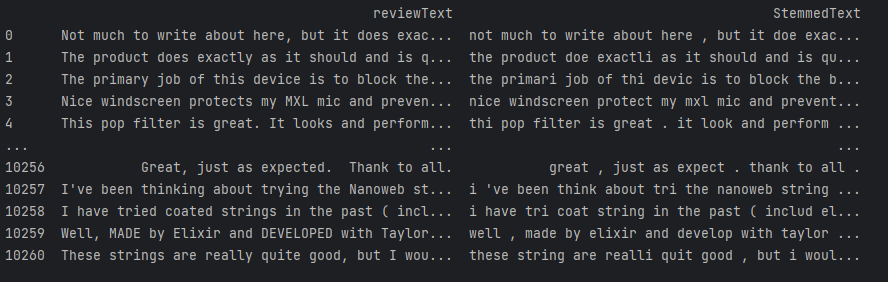
Metoda najpierw oblicza liczbę słów w każdej recenzji, przechodząc przez każdą wartość w kolumnie ‘reviewText’ i dzieląc tekst na słowa za pomocą metody split(). Liczba słów w każdej recenzji jest następnie zapisywana w nowej kolumnie ‘word\_count’ w DataFrame. Po obliczeniu liczby słów, metoda tworzy histogram, który przedstawia rozkład liczby słów w recenzjach. Histogram jest tworzony z użyciem biblioteki matplotlib.

Bardzo klarownie wykres obrazuje, że udział opinii z liczbą słów poniżej 20 jest bardzo niski. Średnio najwięcej komentarzy znajduje się w przedziale 20 – 30 słów dla poszczególnych komentarzy.

**Stemming i lematyzacja**

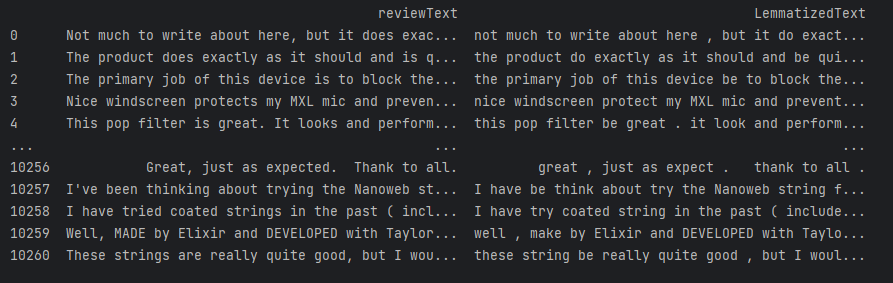
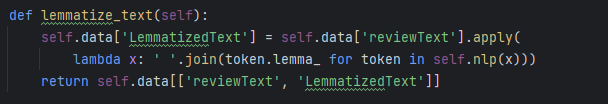
Funkcja stem\_text przekształca każdą recenzję w bazie danych na wersję stemmed, czyli uproszczoną formę wyrazów, w której usunięte są końcówki fleksyjne. Zwraca DataFrame zawierający oryginalną recenzję i przekształconą wersję stemmed.





Funkcja lemmatize\_text przekształca każdą recenzję na wersję lemmatized, czyli na formę podstawową wyrazów, uwzględniając ich kontekst i części mowy.

Przetwarza tekst recenzji za pomocą modelu spaCy (self.nlp(x)). Ta funkcja również zwraca DataFrame zawierający oryginalną recenzję i przekształconą wersję lemmatized w celu zestawienia wartości tych dwóch kolumn.

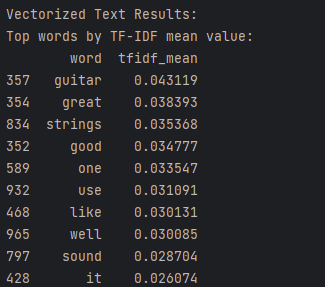


**Wektoryzacja tekstu**

Funkcja vectorize\_text przekształca przetworzone recenzje na reprezentację numeryczną za pomocą wektoryzacji TF-IDF (Term Frequency-Inverse Document Frequency). Wypisuje 10 słów z najwyższymi średnimi wartościami TF-IDF. Wypisuje pierwszych 25 wierszy macierzy TF-IDF.

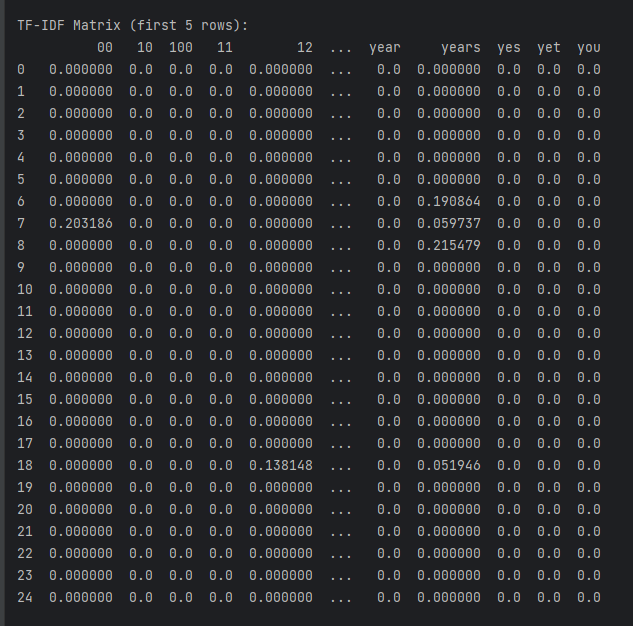


Analiza najczęstszych słów o najwyższej wartości TF-IDF może dostarczyć wglądu w najważniejsze tematy poruszane w recenzjach.



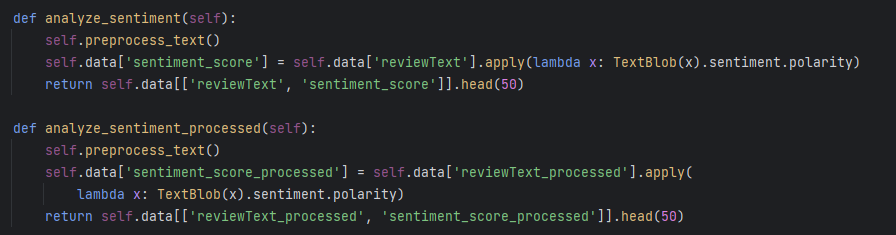
Można zauważyć, że niektóre pozycje pokrywają się ze słowami, które występują w wynikach funkcji def preprocess\_text. Jednak widoczna jest tu inna kolejność, która uwzględnia ważność tych słów.

Poniżej widoczna jest macierz, która jest wynikiem wektoryzacji (pierwsze 25 wartości).



**Analiza sentymentu**

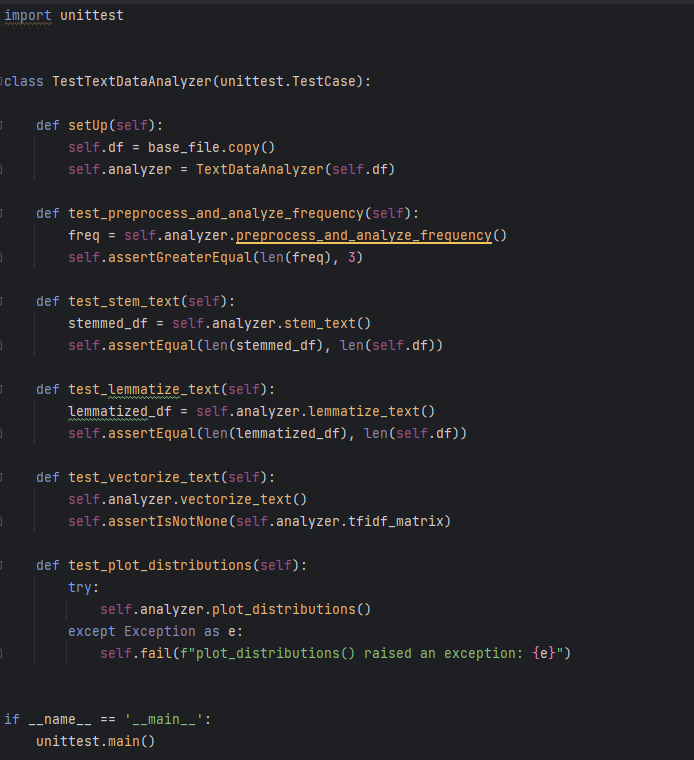
Przeprowadzono analizę sentymentów, aby ocenić emocjonalny ton recenzji. Sentyment jest mierzony w skali od -1 (negatywny) do 1 (pozytywny).



Specjalnym zabiegiem z naszej strony było powtórzenie analizy sentymentu. Pierwszy kod dotyczy oryginalnych danych z kolumny ‘reviewText’, natomiast definicja poniżej dotyczy już tekstu po zabiegach normalizacji. Chciałyśmy sprawdzić, jak zmieniają się wartości sentymentu w zależności od tego, na jakim etapie obróbki tekstu jesteśmy.

**Testy jednostkowe**

Na końcu programu można znaleźć kod jak poniżej, odpowiedzialny za przeprowadzanie badań jednostkowych na wybranych funkcjach.



**Podsumowanie**

Przeprowadzone analizy dostarczyły wszechstronnego wglądu w charakterystyki recenzji instrumentów muzycznych. Zrozumienie tych danych może być użyteczne w różnych kontekstach, takich jak doskonalenie produktów, strategia marketingowa czy poprawa doświadczenia użytkowników. Każdy z zastosowanych kroków przetwarzania i analizy tekstu miał na celu dostarczenie istotnych informacji, które wspierają wyciąganie wartościowych wniosków ze zbioru danych.