Analiza sentymentów (8 punktów)

Zadanie:

Na podstawie tekstu recenzji w pliku *HotelReviews.csv* dla każdej recenzji określ, czy jest ona pozytywna, czy negatywna. W związku z tym, że ogólne oceny hoteli są w zakresie od 2,5/10 do 10/10, przyjmij, że ocena poniżej 5 oznacza, że recenzja była zła. W pozostałych przypadkach (ocena >= 5) recenzja jest dobra.

Program musi zawierać następujące elementy:

1. Scalenie wartości w kolumnach *Positive_Review* i *Negative_Review* do jednej kolumny *Review*.
2. Zakodowanie ogólnej oceny hotelu w formie binarnej (0 - pozytywna recenzja, 1 – negatywna recenzja) oraz obliczenie, ile jest ogólnie pozytywnych oraz negatywnych recenzji w danych (na podstawie ogólnej oceny).
3. Usunięcie z tekstów recenzji wyrażeń „*No positive”*, „*No negative”*.
4. Zamianę wszystkich wielkich liter w tekście recenzji na małe.
5. Usunięcie przerywników (*stopwords*) i wszystkich jednoliterowych słów.
6. Lematyzację słów, czyli sprowadzenie ich do formy podstawowej (słownikowej).
7. Analizę sentymentów i wyświetlenie 10 najlepszych i najgorszych recenzji (z tych, które mają więcej niż 5 słów).
8. Klasyfikację recenzji z wykorzystaniem dowolnego klasyfikatora

UWAGA: Program musi zawierać komentarze objaśniające kod.

Wskazówki

1. Aby przyspieszyć obliczenia zbiór recenzji można ograniczyć do wybranych losowo rzędów (10-20% całości).
2. Analiza sentymentów może być przeprowadzona z wykorzystaniem klasy *SentimentIntensityAnalyzer* z modułu *nltk.sentiment.vader*. Dla każdego tekstu Vader zwraca 4 wartości: ocenę neutralności, ocenę pozytywności, ocenę negatywności oraz ogólny wynik, który podsumowuje poprzednie oceny.
3. Przy klasyfikacji recenzji etykietą klasy będzie ogólna ocena hotelu (w formie binarnej). Cechami będą natomiast wartości uzyskane w wyniku przeprowadzenia analizy sentymentów. Zbiór danych musi być podzielony na dane uczące i testowe (w dowolnej proporcji). Po przeprowadzeniu klasyfikacji należy ocenić jej dokładność.