

Projekt 2 – Eksploracja Danych w Pythonie

Klasyfikacja recenzji aplikacji mobilnych

Zadanie polega na napisaniu klasyfikatora przewidującego ogólną oceną na podstawie tekstowej opinii. Zbiór uczący znajduje się w pliku *reviews_train.csv* i zawiera następujące atrybuty:

- reviewerID – identyfikator użytkownika, który napisał opinię,
- asin – identyfikator produktu (aplikacji),
- reviewerName – nazwa użytkownika, który napisał opinię,
- helpful – oceny innych czy opinia była pomocna (w formacie [liczba oznaczeń jako pomocna, liczba oznaczeń jako niepomocna],
- reviewText – treść recenzji,
- summary – streszczenie recenzji (krótki tekst z ogólną oceną),
- unixReviewTime – czas udostępnienia recenzji (format UNIX),
- reviewTime – czas udostępnienia recenzji (format tekstowy),
- **score** – ocena ogólna (liczba całkowita z zakresu 1-5).

Zadanie należy wykonać w Pythonie (np. przy użyciu bibliotek NLTK i scikit-learn). Wykorzystywane biblioteki nie są narzucone, można dowolnie wykorzystywać dostępne dla pythona biblioteki.

Wraz z kodem rozwiązania należy dostarczyć sprawozdanie (w formie PDF) opisujące wszystkie kroki rozwiązania wraz z opisem motywacji danego wyboru, m. in.:

- które atrybuty zostały wykorzystane,
- w jaki sposób dane zostały wstępnie przetworzone,
- jakie klasyfikatory zostały przetestowane i z jakim rezultatem,
- czy, w jakim zakresie zostały zoptymalizowane parametry algorytmów (i z jakim rezultatem),
- jakie miary oceny zostały dobrane do oceny algorytmu i dlaczego,
- jaki był scenariusz testowania (podział na zbiór uczący i testowy),
-

Jeśli uruchomienie kodu wymaga jakiś szczególnych opcji (np. przekazani parametrów do skryptu) to też należy umieścić takie informacje w sprawozdaniu.

Państwa rozwiązanie zostanie uruchomione na dodatkowym zbiorze testowym (nieдоступnym na etapie uczenia) aby zweryfikować jego jakość i porównać z innymi rozwiązaniami.

Problem jest trudny i nie należy spodziewać się spektakularnych wyników, jednak Państwa algorytm powinien przede wszystkim wnosić jakąś dodatkową wiedzę (tzn. działać znacząco lepiej od algorytmów losowego lub przypisującego zawsze klasę większościową).

Ocenie podlegać będą:

- złożoność rozwiązania (zarówno ostatecznego jak i liczba przetestowanych algorytmów, parametrów i opcji),
- ostateczny wynik klasyfikacji na ukrytym zbiorze testowym,
- jakość merytoryczna, czytelność oraz estetyka sprawozdania.

Rozwiązanie (kod źródłowy oraz sprawozdanie) należy dostarczyć w formie linku do repozytorium

Maksymalna ocena: 25pkt, termin: 26.01.2021