Eksploracja masywnych danych

Python - analiza sentymentu

Grzegorz Kaszuba, Sebastian Michoń

1. Opis problemu

Celem projektu było stworzenie klasyfikatora dokonującego analizy sentymentu - przewidującego na podstawie treści recenzji użytkownika ocenę, jaka została przez niego przyznana (całkowitą w skali 1-5).

2. Wybór modelu

Optymalizacja modelu składała się z kilku etapów, które zostały zrealizowane w większości niezależnie - na etapie co jakiś czas pojawiała się weryfikacja wcześniej podjętych decyzji, jednak w większości nie są one udokumentowane. Dobór modelu obejmował:

- sposób tokenizacji - ze względu na znaczny rozmiar zbioru danych i konieczność bezpośredniego przejścia do reprezentacji rzadkiej, wybrana została natywna tokenizacja wykorzystywana przez obiekty z biblioteki Scikit-Learn,
- wybór metody zliczania występujących słów - proste zliczanie okazało się nieznacznie lepsze niż metoda TF-IDF,
- wybór metody filtrowania słów - odsiewanie stop-words negatywnie wpływało na model i nie zostało wykorzystane, najlepsze wyniki osiągnięto odrzucając słowa, które otrzymały niedostateczną liczbę zliczeń,
- wybór modelu - pomimo obiecujących wyników klasyfikatora Random Forest, ze względu na prostotę modelu i szybkość uczenia wybrana została regresja logistyczna,
- wybór relewantnych danych - końcowy model korzysta zarówno z tekstu recenzji, jak i jej podsumowania (ponieważ podsumowania są krótsze i bardziej informatywne, ekstrakcja danych z obu tych kolumn została zrealizowana z różnymi minimalnymi progami zliczeń), a także 2-gramów występujących w każdej z tych kolumn.

Bardziej szczegółowy opis dotyczący wszystkich decyzji optymalizacyjnych zawarty jest w załączonym notebooku.

3. Wyniki

Ostateczny model uzyskał trafność 67,9% na zbiorze testowym. Choć wynik nie jest doskonały, jest to istotna poprawa względem klasyfikatora większościowego (klasyfikator przewidujący zawsze ocenę 5 uzyskałby 50,6% trafności. Wykonywane na różnych etapach modele były oceniane również z pomocą innych metryk (metryka f1, a ze względu na uporządkowany charakter klas decyzyjnych również średni błąd absolutny i błąd średniokwadratowy). Ponieważ liczby przykładów dla różnych klas decyzyjnych są niezbilansowane, towarzysząca kolejnym modyfikacjom poprawa wyników modelu jest znacznie wyższa na innych metrykach niż trafność.

4. Zawartość projektu

W pliku “Sentiment analysis.ipynb” udokumentowany jest cały proces realizacji projektu. Dwa pliki .pickle posiadają dane konieczne, by uruchomić notebook “Test.ipynb”, który - po wymianie pliku “reviews_train.csv” na kompatybilny zbiór testowy powinien posłużyć do wygodnej oceny modelu.