Николай Шиваров

Откриване на негативни мнения за хотели в TripAdvisor и Booking.com със Scikit-learn

Мненията за TripAdvisor взех от:

<https://www.kaggle.com/datasets/andrewmvd/trip-advisor-hotel-reviews>

Мненията за Booking.com взех от:

<https://data.world/opensnippets/bookingcom-reviews-dataset>

Данните за TripAdvisor бяха във формат, който изглеждаше удобен за работа по задачата, същото не може да бъде казано за .csv файла за Booking. Там имаше доста излишни колони. Затова първото нещо, което направих бе да променя формата на данните, с които щях да работя. За целта, създадох файла “dataTransformer.py”, в който взимам двете изтеглени csv-та и създадох 4 нови csv-та. В началото премахнах всички излишни колони в booking, след това махнах всички редове със съдържание "There are no comments available for this review", в случая това имитира null-value, тези редове определено щяха да пречат на моделите. Накрая преименувах и разместих колоните за да могат да съвпадат в двата файла. След това запазих двата Dataframe-а в това им състояние в нови csv файлове. В TripAdvisor оценките са само цели числа между 1 и 5, а в Booking може да имаме и дробни числа като най-ниската оценка е 1, а най високата 10. В Rating колоната на Booking, където имаше рейтинг над 7 го смених със стойността “positive”, в противен случай слагах стойността “negative”. Направих същото и за TripAdvisor като условието за “positive” беше рейтинг повече от 3.

Реших да ползвам два вида модели- логистична регресия и линейна регресия, като за всеки вид създадох по 3 модела- общ за двата файла и с всеки файл по отделно. За всеки от моделите тестовото множество беше с големина 20% и превръщах текста в TF-IDF характеристики.

За да тествам моделите, които ползваха логистична регресия проверявах процента точност и създавах графика с confusion-matrix. За модела с всички данни точността беше 89.2% и матрицата беше такава:

.A blue squares with numbers

Description automatically generated

Очвидно най-честата грешка е, че негативни мнения се смятат за позитивни.

Моделът, който работеше само с данни за Booking даде точност 87.4% и матрица:

A graph of a number of blue squares

Description automatically generated with medium confidence

Тук по-рядко се бъркат позитивни мнения за негативни, но негативните мнения доста по-често се бъркат за позитивни.

Моделът, който работеше само с данни за TripAdvisor даде точност 90.6% и матрица:

A blue squares with numbers

Description automatically generated

Този модел най-добре се справя с негативните мнения.

За да тествам другия вид модели използвам residual plot и mean square error. Тъй като някои мнения дават стойности под 1 и над 10 накрая смятам и mean square error отново като преди това намалям на 10 стойностите над 10 и увеличавам на 1 стойностите над 1. В модела, който има всички данни умножих рейтингите на TripAdvisor по 2.

mean square error даде: 1.799

residual plot:

A graph with blue and red lines

Description automatically generated

mean square error след граници даде: 1.757

**Моделът, който ползва само booking:**

mean square error даде: 1.421

residual plot:

A graph of blue and red lines

Description automatically generated

mean square error след граници даде: 1.408

**Моделът, който ползва само TripAdvisor:**

mean square error даде: 0.5186, но ако я сравняваме с предните модели трябва да я умножим по 4 и тогава ще стане над 2

residual plot:

A graph of blue lines and red lines

Description automatically generated

mean square error след граници даде: 0.4922

Накрая може да си направим извода, че линейната регресия се спрява по-добре с данните за booking, а логистичната с тези за TripAdvisor