Akademia Górniczo-Hutnicza
im. Stanisława Staszica w Krakowie

AGH University of Krakow

Analiza sentymentu recenzji filmowych za pomocą naiwnej klasyfikacji bayesowskiej

Łukasz Obrzut Mateusz Mglej

AGH

Część I

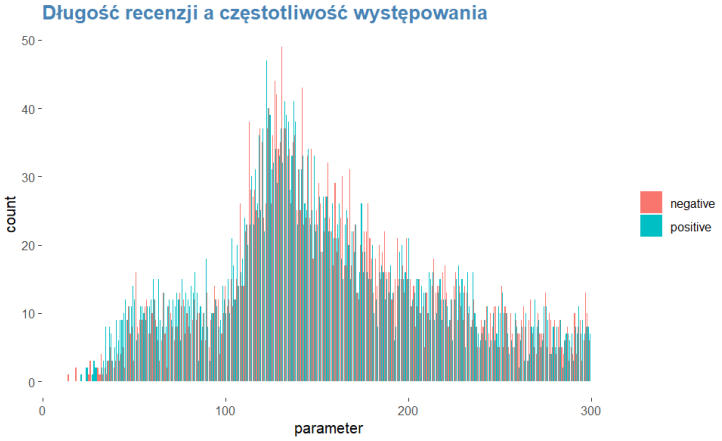
Zbieranie danych

Baza danych oraz przedstawienie problemu

Naszym celem będzie rozwiązanie problemu klasyfikacji recenzji filmowych. Do projektu wykorzystamy zbiór danych składający się z 50 tysięcy recenzji filmowych z serwisu IMDB zarówno pozytywnych i negatywnych. Do budowy modelu wykorzystamy podzbiór, który będzie składał się z 10 000 recenzji (zostanie to uzasadnione). Na podstawie ilości recenzji, częstotliwości występowania słów oraz odpowiedniego parametru wygładzenia Laplace'a zastosujemy stosowną technikę uczenia maszynowego do tego typu zadań.
Baza danych dostępna jest na stronie: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews.

Rozkład długości recenzji

Rozkład długości recenzji jest podobny zarówno dla recenzji pozytywnych i negatywnych.



Przykłady recenzji

Dane zawierają pewne mankamenty (duplikaty, znaki HTML) oraz wzorce (np. słowa kluczowe dla danej opinii), które będziemy mieć na uwadze w naszej analizie.

Przykładowe recenzje

- pozytywna: *"Clearly an hilarious movie.< br/ >< br/ >It angers me to see the poor ratings given to this piece of comic genius< br/ >< br/ >Please look at this for what it is, a funny, ridiculous enjoyable film. Laugh for christ sake!< br/ >< br/ >"*
- negatywna: *"This is the weakest of the series, not much of a plot and a rather odd-looking Wallace. But it's still pretty good, considering. A sign of greater things to come!< br/ >< br/ >6/10"*

AGH

Część II

Eksploracja i przygotowanie danych

www.agh.edu.pl

Czyszczenie recenzji

Czyszczenie danych w naszej analizie będzie opierało się kolejno na:

1. usunięciu duplikatów recenzji (418 przypadków),
2. konwersji na małe litery,
3. usunięciu liczb oraz stop-words (słów popularnych, czyli o małym znaczeniu w problemie klasyfikacji),
4. usunięciu znaków HTML,
5. usunięciu znaków interpunkcyjnych,
6. redukcji słów do ich rdzenia.

Porównanie oryginalnej recenzji do oczyszczonej

AGH

Oryginalna recenzja

"We brought this film as a joke for a friend, and could of been our worst joke to play. The film is barely watchable, and the acting is dire. The worst child actor ever used and Hasslehoff giving a substandard performance. The plot is disgraceful and at points we was so bored we was wondering what the hell was going on. It tries to be gruesome in places but is just laughable.$< br/ >< br/ >$Just terrible"

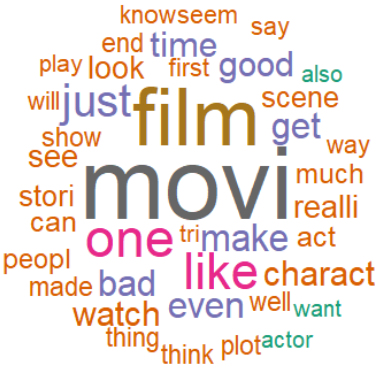
Recenzja oczyszczona

"brought film joke friend worst joke play film bare watchabl act dire worst child actor ever use hasslehoff give substandard perform plot disgrac point bore wonder hell go tri gruesom place just laughabl just terribl"

Chmura słów

Z wyczyszczonych danych możemy graficznie przedstawić za pomocą chmury wyrazów najczęściej występujące słowa w naszych recenzjach.

Rysunek: Pozytywne recenzje



Rysunek: Negatywne recenzje

Przygotowanie danych do budowy klasyfikatora

Po oczyszczeniu danych możemy przejść do kolejnych operacji:

- tokenizacji tekstu, czyli podziału tekstu na wyrazy,
- utworzenia macierzy DTM, która w swoich komórkach zawiera liczbę wystąpień danego słowa w danej recenzji,
- podzielenia naszych danych losowo na zbiór treningowy oraz testowy w stosunku $3:1$,
- wyrzucenia potencjalnie nieprzydatnych słów, czyli takich które wystąpiły w mniej niż 10 recenzjach ze zbioru treningowego (parametr $\text{Freq} = 10$),
- wprowadzenia zmiennych indykatorowych, które będą mówiły, czy dane słowo wystąpiło w danej recenzji, czy nie.

AGH

Część III

Budowa modelu

Uzasadnienie wyboru zastosowanych narzędzi analizy danych

Narzędziem danych odpowiednim dla naszej analizy może okazać się naiwny klasyfikator bayesowski, ponieważ:

- na podstawie dużej liczby cech musimy wyekstrahować nieznaczny wpływ,
- mała złożoność (istotne, ponieważ macierz DTM jest dużych rozmiarów),
- znany z dobrego sprawowania się z zadaniami klasyfikacji tekstu.

Specyfikacja zastosowanego modelu analitycznego

W naszym problemie klasyfikacji, prawdopodobieństwo Bayesowskie będzie tożsame przykładowo z problemem

$$P(\textbf{negatywna}|\textbf{bad}) = \frac{P(\textbf{bad}|\textbf{negatywna})P(\textbf{negatywna})}{P(\textbf{bad})},$$

gdzie:

- **negatywna** - dana recenzja jest negatywna,
- **bad** - dana recenzja zawiera słowo "bad".

Uogólniając, nasz model może estymować prawdopodobieństwa należenia do klasy recenzji (pozytywna bądź negatywna) pod warunkiem wystąpienia konkretnych słów $S_1, \ldots, S_n$ za pomocą wzoru:

$$P(\text{recenzja}_i|S_1, \ldots, S_n) = \frac{1}{Z}P(\text{recenzja}_i)\prod_{i=1}^{n} P(S_i|\text{recenzja}_i),$$

gdzie Z jest odpowiednim czynnikiem skalującym.

AGH

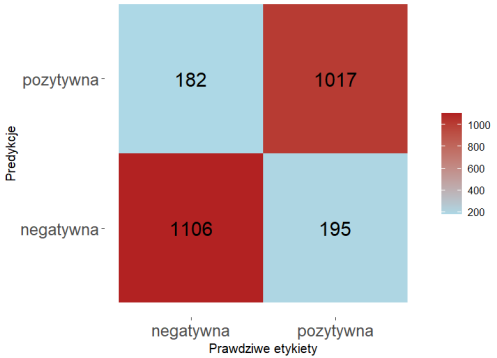
Część IV

Ocena modelu

www.agh.edu.pl

Wyniki

Po wykonaniu treningu klasyfikatora oraz sprawdzeniu jakości modelu na zbiorze testowym otrzymujemy skuteczność **84.9%**.
Poniżej przedstawiamy wyniki w postaci tabeli krzyżowej:



Rysunek: Rezultaty predykcji 2500 recenzji

Ocena modelu

Rezultaty dla zbioru testowego zawierającego 2500 recenzji:

- 2123 recenzji przewidzianych poprawnie,
- 377 błędów.

Biorąc pod uwagę prostotę algorytmu oraz złożoność danych (czasem sentyment wynika z kontekstu całej wypowiedzi, a nie pojedynczych słów) otrzymany wynik **84.9%** wydaje się bardzo poprawny.
Co dalej?

- próba modyfikacji parametrów naszego modelu,
- zastosowanie innego modelu do klasyfikacji: drzewo decyzyjne.

Część V

Dopracowanie modelu

Parametr Freq

W oryginalnym modelu parametr Freq wynosi 10, czyli model pomija te słowa które nie wystąpiły w przynajmniej 10 recenzjach ze zbioru treningowego.

Parametr Freq	Liczba słów w zbiorze treningowym	Skuteczność
10	7171	84.9 %
5	11110	85.4 %
15	5533	85.0 %
20	4516	85.1 %
25	3939	84.9 %
30	3472	84.8 %

Znaczącej poprawy nie ma. Najlepsze rezultaty dla wartości parametru 5 oraz 20. Ze względów praktycznych (szybkość obliczeń) dalsze poprawki będą uwzględniały parametr $Freq = 20$.

Parametr Laplace

Wygładzanie Laplace'a pomaga uniknąć sytuacji, gdy model źle klasyfikuje daną recenzję, ponieważ dane słowo kluczowe akurat nie wystąpiło w recenzjach ze zbioru uczącego. Domyślnie model nie korzysta z tego parametru, a użycie go może pomóc go dopracować.

Parametr Laplace	Skuteczność
0	85.1%
0.5	85.1%
1	85.2%
5	85.1%
10	84.9%

Brak istotnej poprawy, najlepszy wynik dla parametru 1.

Liczba użytych danych

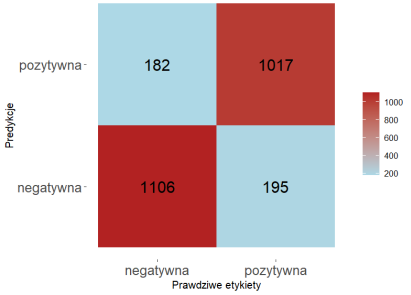
Pierwszy model korzysta z 10000 recenzji. Odpowiednio dla tej wartości parametr Freq wynosi 20. Czy zwiększenie liczby danych poprawi model?

Liczba danych	**Parametr Freq**	**Skuteczność**
10 000	20	85.1 %
20 000	40	85.7 %
49 582	100	84.2 %

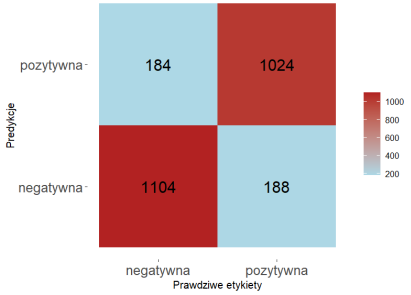
Korzystanie z większej ilości danych nie poprawia znacząco modelu. Lepszy wynik przy 20000 wynika raczej z lepszej jakości danych. Można założyć, że używanie 10000 recenzji jest wystarczające.

Podsumowanie modyfikacji parametrów

- Pierwszy model: 10 000 danych, Freq: 10, Laplace: 0 $\Rightarrow$ 84.9%.
- Dopracowany model: 10 000 danych, Freq: 20, Laplace: 1 $\Rightarrow$ 85.2%.



Rysunek: Pierwszy model



Rysunek: Dopracowany model

Drzewa decyzyjne

Model drzewa decyzyjnego z użyciem algorytmu C5.0

Skuteczność	Parametry modelu
78.9%	bez wzmacniania
81.6%	trials: 8
81.7%	trials: 10
82.2%	trials: 13
82.3%	trials: 14
82.1%	trials: 15

Skuteczność pierwszego modelu wynosi 78.9%. Próba dopracowania za pomocą metody AdaBoost (iteracyjne budowanie wielu drzew, parametr trials) daje lepsze wyniki, ale nie przewyższają wyników uzyskanych za pomocą klasyfikacji bayesowskiej.