Raport z klasyfikacji nagłówków

Maria Galanty - kandydatka na stanowisko Junior Data Scientist

Luty 2020

1 Wstep

Zadanie polegało na zbudowaniu modelu, który sklasyfikuje nagłówki jako sarkastyczne badź nie. W przedstawionym rozwiazaniu wykorzystałam Pythona.

2 Eksploracja danych

Dane zawierały 26709 rekordów w tym 11724 nagłówków satyrycznych i 14985 w drugiej klasie, nie było brakujacych zmiennych. Dwie klasy nie różnia sie znaczaco miedzy soba długościa tekstu ani ilościa zawartych znaków interpunkcyjnych.

3 Text Preprocessing

Przy przetwarzaniu jezyka naturalnego wystepuja dwa problemy zwiazane z formatem danych. Po pierwsze klasyfikatory potrzebuja zmiennych numerycznych, a nie tekstowych. Algorytmy te, wymagaja wektora liczbowego. Można to rozwiazać za pomoca podejścia *bag-of-words*, w którym każde unikalne słowo w tekście bedzie reprezentowane przez jedna liczbe. W zwiazku z czym pojawia sie drugi problem, który polega na ograniczeniu ilości wyrazów poddanych analizie. W tym celu najpierw zostały usuniete *stop words* sa to najbardziej popularne wyrazy w jezyku angielskim, które wiaża sie z gramatyka jezyka, należa tutaj takie wyrazy jak *a, the, you.* Nastepnie wykonany został *stemming*, który różne formy wyrazów sprowadza do formy podstawowej - korzenia wyrazu. Przykładowo *playing, play, plays* zostana zamienione na *play.* Kolejnym krokiem było usuniecie znaków interpunkcyjnych. Tak przygotowane dane zostały zmienione na wektory liczbowe. TF-IDF jest skrótem terminu *Term Frequency-Inverse Document Frequency* i jest algorytmem przekształcajacym tekst w reprezentacje liczbowa.

4 Klasyfikator

Ze wzgledu na zbalansowany zbiór danych, do ewaluacji modelu użyłam cross-walidacji i metryke *accuracy*. Przy cross-walidacji (umożliwiajacej sprawdzenie, czy model nie jest przeuczony, co jest szczególnie ważne przy tak dużym zbiorze danych) podzieliłam zbiór na 5 cześci, co pozwoliło trenować model na 4 z nich i testować na pozostałej. Zbudowałam nastepujace dwa klasyfikatory: Random Forest oraz Multinomail Naive Bayes. Drugi z nich uzyskał lepszy średni wynik klasyfikacji na danych testowych wynoszacy 0.78.