Sztuczna inteligencja i inżynieria wiedzy

Laboratorium – lista 4

Jakub Jastrzębski, 254574

## Badane elementy

W ćwiczeniu zbadałem wpływ użytego modelu uczenia maszynowego i pewnych hiperparametrów (dla klasyfikatora SVM).

## Ocena rozwiązania

W ocenie danego rozwiązania badałem wynik funkcji f1, razem z precyzją danego modelu, do testów walidacyjnych wykorzystywana była 10-krotna walidacja krzyżowa. Jako łączny wynik brałem sumę: f1 \* 1.25 + precyzja.

## Wybrane modele

Modele Naive Bayess: GausianNB, MultinomialNB, BernoulliNB, ComplementNB

Model SVM z różnymi parametrami.

## Wstępne przetwarzanie danych

Dane wymagały oczyszczenia, więc wszystkie wybrakowane dane usunąłem, a jako końcowe klasy wziąłem 7 najpopularniejszych gatunków. Jeżeli jakaś książka nie miała żadnego gatunku z tych 7, usuwałem ten rekord. Ze wszystkich gatunków dla danej książki wybierałem ten najpopularniejszy. Do ekstrakcji cech z opisów użyłem ekstraktora TF-IDF.

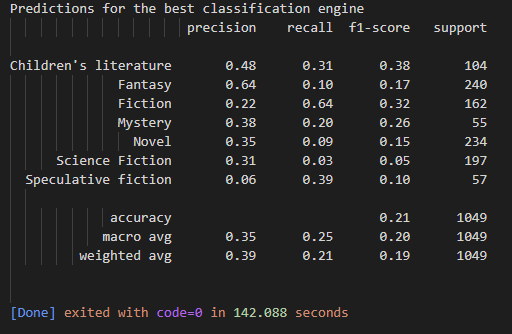
## Tabele z uzyskanymi wynikami

Porównanie wyników dla modeli Naive Bayess.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Lp.** | **GausianNB** | | **MultinomialNB** | | **BernoulliNB** | | **ComplementNB** | |
|  | Prec | F1 | Prec | F1 | Prec | F1 | Prec | F1 |
| 1 | 0.45 | 0.33 | 0.45 | 0.42 | 0.49 | 0.34 | 0.49 | 0.46 |
| 2 | 0.45 | 0.33 | 0.46 | 0.42 | 0.49 | 0.34 | 0.49 | 0.46 |
| 3 | 0.45 | 0.33 | 0.46 | 0.42 | 0.49 | 0.34 | 0.49 | 0.46 |
| 4 | 0.45 | 0.33 | 0.46 | 0.42 | 0.49 | 0.34 | 0.49 | 0.46 |
| 5 | 0.45 | 0.33 | 0.46 | 0.42 | 0.49 | 0.34 | 0.49 | 0.46 |
| 6 | 0.45 | 0.33 | 0.46 | 0.42 | 0.49 | 0.34 | 0.49 | 0.46 |
| Średnia | 0.45 | 0.33 | 0.46 | 0.42 | 0.49 | 0.34 | 0.49 | 0.46 |

Najlepszy model Naive Bayess: ComplementNB.

Wyniki testów dla najlepszego modelu Naive Bayess.



Porównanie wyników dla podejścia SVM.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Lp.** | **SVC-1** | | **SVC-2** | | **SVC-3** | |
|  | Prec | F1 | Prec | F1 | Prec | F1 |
| 1 | 0,48 | 0,43 | 0,49 | 0,35 | 0,25 | 0,33 |

SVC-1 -> kernel: rbf, C: 1.25, cache\_size: 325.0, shrinking: True, coef0: 0.0

SVC-2 -> kernel: poly, C: 1.0, cache\_size: 200.0, shrinking: False, coef0: 0.0

SVC-3 -> kernel: sigmoid, C: 1.0, cache\_size: 200.0, shrinking: True, coef0: 2.25

## Wnioski

* Czyszczenie danych jest niezwykle istotnym zadaniem w uczeniu maszynowym, gdyż to właśnie na tych danych model się uczy i jeżeli nie są one odpowiedniej jakości, nie będzie to dobry model
* Ze wszystkich podejść Naive Bayess w moim przypadku najlepsze wyniki dawał model ComplementNB, pomimo że zwykle najczęściej używanym modelem, dającym najlepsze wyniki jest MultinomialNB