Piotr Węsiora 16.12.2019

Numer indeksu - 253913

Komputerowe przetwarzanie tekstów - laboratorium 4

Klasyfikacja dokumentów tekstowych

Kod programu

```
import nltk
import random
from nltk.corpus import movie_reviews

TEST_REPETITIONS_TRAINING = 10
TEST_REPETITIONS_TEST = 10
TEST_REPETITIONS_WORDS = 10
documents = [(list(movie_reviews.words(fileid)), category)
             for category in movie_reviews.categories()
             for fileid in movie_reviews.fileids(category)]
all_words = nltk.FreqDist(w.lower() for w in movie_reviews.words())


def document_features(document, word_features):
    document_words = set(document)
    features = {}
    for word in word_features:
        features['contains({})'.format(word)] = (word in document_words)
    return features


def calculate_accuracy(featuresets, testing_set_size, test_set_size):
    train_set, test_set = featuresets[testing_set_size:], featuresets[:test_set_size]
    classifier = nltk.NaiveBayesClassifier.train(train_set)
    return nltk.classify.accuracy(classifier, test_set)


def naive_bayes_classifier_test(words_ammount, testing_set_size, test_set_size,
test_repetitions):
    result_list=[]
    word_features = list(all_words)[:words_ammount] #Najczęstsze słowa w korpusie
    for i in range (test_repetitions):
        random.shuffle(documents)
        featuresets = [(document_features(d, word_features), c) for (d, c) in documents]
        result_list.append(calculate_accuracy(featuresets, testing_set_size,
test_set_size))
    return result_list
```

Testy

```
big_training_result = naive_bayes_classifier_test(1000, 1000, 1000, TEST_REPETITIONS_TRAINING)
big_test_result = naive_bayes_classifier_test(1000, 1000, 1000, TEST_REPETITIONS_TEST)
big_word_count = naive_bayes_classifier_test(1000, 1000, 1000, TEST_REPETITIONS_WORDS)
medium_training_result = naive_bayes_classifier_test(1000, 100, 1000, TEST_REPETITIONS_TRAINING)
medium_test_result = naive_bayes_classifier_test(1000, 1000, 100, TEST_REPETITIONS_TEST)
medium_word_count = naive_bayes_classifier_test(100, 1000, 1000, TEST_REPETITIONS_WORDS)
small_training_result = naive_bayes_classifier_test(1000, 10, 1000, TEST_REPETITIONS_TRAINING)
small_test_result = naive_bayes_classifier_test(1000, 1000, 10, TEST_REPETITIONS_TEST)
small_word_count = naive_bayes_classifier_test(10, 1000, 1000, TEST_REPETITIONS_WORDS)
```

EKSPERYMENT 1 – BADANIE ZMIAN DOKŁADNOŚCI W ZALEŻNOŚCI OD WIELKOŚCI ZBIORU TRENINGOWEGO

Testy przeprowadzono na zbiorze testowym o rozmiarze 1000 i dla 1000 najczęstszych słów.

Wyniki

```
[0.763, 0.77, 0.715, 0.744, 0.729, 0.766, 0.757, 0.774, 0.777, 0.797]
Average accuracy using a big training set(1000 documents): 0.7592
[0.835, 0.824, 0.841, 0.834, 0.835, 0.839, 0.843, 0.809, 0.839, 0.831]
Average accuracy using a medium training set(100 documents): 0.833
[0.838, 0.842, 0.822, 0.835, 0.836, 0.841, 0.83, 0.832, 0.827, 0.823]
Average accuracy using a small training set(10 documents): 0.8326
```

Rys.1. Listy z wynikami kolejnych testów zmian dokładności wraz z zmianą rozmiaru zbioru treningowego i wartość średnia wyników.

Wnioski

Z przedstawionych wyników wynika, że w miarę wzrostu wielkości zbioru treningowego średnia dokładność maleje. Może to wynikać z nadmiernego dopasowanie (ang. overfitting) co wskazywałoby na większą jakość testów z większym zbiorem treningowym, bądź z błędu w logice programu.

EKSPERYMENT 2 – BADANIE ZMIAN DOKŁADNOŚCI W ZALEŻNOŚCI OD WIELKOŚCI ZBIORU TESTOWEGO

Testy przeprowadzono na zbiorze treningowym o rozmiarze 1000 i dla 1000 najczęstszych słów.

Wyniki

```
[0.75, 0.758, 0.706, 0.777, 0.793, 0.78, 0.765, 0.775, 0.775, 0.791]
Average accuracy with a big test set(1000 documents): 0.7670000000000001
[0.82, 0.74, 0.7, 0.73, 0.8, 0.75, 0.72, 0.8, 0.82, 0.79]
Average accuracy with a medium test set(100 documents)0.767
[0.8, 0.6, 0.9, 0.8, 0.7, 0.8, 0.7, 0.8, 0.7, 0.8]
Average accuracy with a  small test set(10 documents)0.76
```

Rys.2. Listy z wynikami kolejnych testów zmian dokładności wraz z zmianą rozmiaru zbioru testowego i wartość średnia wyników.

Wnioski

Odczytując wartości można stwierdzić, że wielkość zbioru testowego nie wpływa na dokładność klasyfikacji. Mały zbiór testowy może skutkować dużym zaokrągleniem dokładności klasyfikacji.

EKSPERYMENT 3 – BADANIE ZMIAN DOKŁADNOŚCI W ZALEŻNOŚCI OD PRZYJĘTEJ WIELKOŚCI ZBIORU NAJCZĘSTSZYCH SŁÓW

Testy przeprowadzono na zbiorze treningowym o rozmiarze 1000 i zbiorze testowym o rozmiarze 1000.

Wyniki

```
[0.786, 0.78, 0.77, 0.781, 0.809, 0.805, 0.784, 0.753, 0.776, 0.788]
Average accuracy using 1000 most common words: 0.7832000000000001
[0.627, 0.628, 0.637, 0.615, 0.606, 0.621, 0.623, 0.607, 0.645, 0.644]
Average accuracy using 100 most common words: 0.6253
[0.497, 0.491, 0.487, 0.501, 0.485, 0.501, 0.492, 0.479, 0.488, 0.499]
Average accuracy using 10 most common words: 0.4919999999999999
```

Rys.2. Listy z wynikami kolejnych testów zmian dokładności wraz z zmianą zbioru testowego i wartość średnia wyników.

Wnioski

Na podstawie otrzymanych wyników można stwierdzić, że wraz z zwiększającą się ilością najczęstszych słów wykorzystanych do klasyfikacji dokładność zwiększa się.