МГТУ им. Н. Э. Баумана

Факультет: Информатика, искусственный интеллект и системы управления

Кафедра: Системы обработки информации и управления

Дисциплина: Методы машинного обучения

Рубежный контроль №2 "Методы обработки текстов"

Выполнил: Солохов И. Р. ИУ5-23М

In [1]:
```
import numpy as np
import pandas as pd
```

In [21]:
```
data = pd.read_csv('titles.csv')
data.head()
```

Out[21]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Docu
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	In TV D
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	In TV
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	D
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	In Rc Sh

In [22]:
```
data.keys()
```

Out[22]:
```
Index(['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added',
       'release_year', 'rating', 'duration', 'listed_in', 'description'],
      dtype='object')
```

In [23]:
```
import sklearn
from sklearn.svm import LinearSVC
from sklearn.naive_bayes import MultinomialNB
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
from sklearn.model_selection import cross_val_score
```

In [24]:
```
tfidfv = TfidfVectorizer()
tfidf_features = tfidfv.fit_transform(data['description'])
tfidf_features
```

Out[24]:
```
<8807x19159 sparse matrix of type '<class 'numpy.float64'>'
        with 189832 stored elements in Compressed Sparse Row format>
```

In [25]:
```
countv = CountVectorizer()
countv_features = countv.fit_transform(data['description'])
countv_features
```

Out[25]:
```
<8807x19159 sparse matrix of type '<class 'numpy.int64'>'
        with 189832 stored elements in Compressed Sparse Row format>
```

In [26]:
```
y = data['type'].values
```

In [27]:
```
cross_val_score(LinearSVC(), tfidf_features, y, scoring='accuracy', cv=3).mean()
```

Out[27]:
```
0.742477142507895
```

In [28]:
```
cross_val_score(LinearSVC(), countv_features, y, scoring='accuracy', cv=3).mean()
```

Out[28]:
```
0.7092074418950095
```

In [29]:
```
cross_val_score(MultinomialNB(), tfidf_features, y, scoring='accuracy', cv=3).mean()
```

Out[29]:
```
0.7023956075242114
```

In [30]:
```
cross_val_score(MultinomialNB(), countv_features, y, scoring='accuracy', cv=3).mean()
```

Out[30]:
```
0.7318038657748028
```

In [31]:
```
print('Наилучшее значение при LinearSVC и tfidf:', cross_val_score(LinearSVC(), tfidf_feat
```

```
Наилучшее значение при LinearSVC и tfidf: 0.742477142507895
```