دانشگاه اصفهان
دانشکده مهندسی کامپیوتر

مبانی داده‌کاوی

عنوان

پروژه‌ی پایانی

مشخصات

فرینام همتی‌زاده، ۹۶۳۶۱۳۱۰۳

خرداد ۱۴۰۰

نحوه‌ی اجرا

باید کد naïve_bayes.py برای سؤال اول و کد knn.py برای سؤال ۲ ران شود که هر دو قبل از ساختن مدل کد preprocess.py را برای تمیز کردن متن مجموعه‌ی داده‌های آموزش و آزمون فراخوانی می‌کنند.

توضیحات

لازم به ذکر است که حذف کلمات ایست دقت را افزایش می‌داد و تمام نتایج زیر با حذف آن‌ها به دست آمده است و ترتیب برچسب‌ها در نمودارهای گرافیکی ماتریس درهم‌ریختگی به‌صورت زیر است:

```
categories = ['اجتماعی', 'ادیان', 'اقتصادی', 'سیاسی', 'فناوري', 'مسائل راهبردي ايران', 'ورزشی']
categories_number = ['1', '2', '3', '4', '5', '6', '7']
```

شکل ۱

و بر اساس سؤال در مدل بیز کلمات مهم برای هر دو مجموعه‌ی آموزش و آزمون حساب شد که در کد قابل مشاهده است.

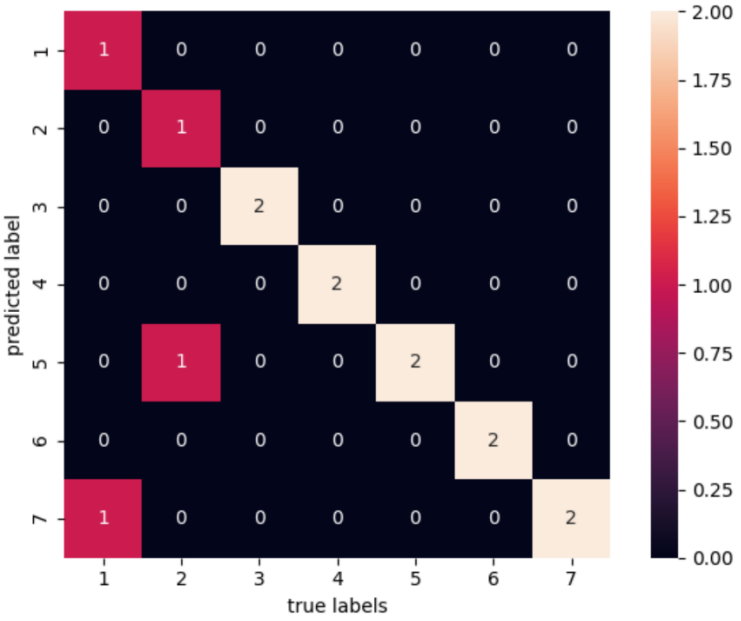
سؤال ۱

در سؤال اول از CountVectorizer و MultinomialNB استفاده شد که نتایج آن به‌صورت زیر است:

```
Naive Bayes Confusion Matrix:
 [[1 0 0 0 0 0 1]
 [0 1 0 0 1 0 0]
 [0 0 2 0 0 0 0]
 [0 0 0 2 0 0 0]
 [0 0 0 0 2 0 0]
 [0 0 0 0 0 2 0]
 [0 0 0 0 0 0 2]]

Naive Bayes Classification Report:
                    precision    recall  f1-score   support

           2         0.67      0.50      1.00       اجتماعی
             2         0.67      0.50      1.00       ادیان
           2         1.00      1.00      1.00       اقتصادی
             2         1.00      1.00      1.00       سیاسی
            2         0.80      1.00      0.67       فناوري
2         1.00      1.00      1.00       مسائل راهبردي ايران
             2         0.80      1.00      0.67       ورزشی

          accuracy                           0.86        14
         macro avg       0.90      0.86      0.85        14
      weighted avg       0.90      0.86      0.85        14


Naive Bayes accuracy:
 0.8571428571428571
```

شکل ۲

شکل گرافیکی ماتریس درهم‌ریختگی برای مدل بیز به‌صورت زیر است:



شکل ۳

بعد برای بهبود عملکرد در سؤال اول از TFIDF استفاده شد که نتایج به‌صورت زیر است:

```
Naive Bayes Confusion Matrix:
 [[1 0 0 0 0 0 1]
 [0 2 0 0 0 0 0]
 [0 0 2 0 0 0 0]
 [0 0 0 2 0 0 0]
 [0 0 0 0 2 0 0]
 [0 0 0 0 0 2 0]
 [0 0 0 0 0 0 2]]

Naive Bayes Classification Report:
                    precision    recall  f1-score   support

           2         0.67      0.50      1.00       اجتماعی
             2         1.00      1.00      1.00       ادیان
           2         1.00      1.00      1.00       اقتصادی
             2         1.00      1.00      1.00       سیاسی
            2         1.00      1.00      1.00       فناوري
2         1.00      1.00      1.00       مسائل راهبردي ايران
             2         0.80      1.00      0.67       ورزشی

          accuracy                           0.93         14
         macro avg       0.95      0.93      0.92         14
      weighted avg       0.95      0.93      0.92         14


Naive Bayes accuracy:
 0.9285714285714286
```

شکل ٤

سؤال ۲

در سؤال دوم از KNeighborsClassifier، CountVectorizer و TfidfTransformer استفاده شد که نتایج آن برای Kهای مختلف به‌صورت زیر است: (با افزایش K دقت کم شده است)

(از روش فاصله‌ی اقلیدسی استفاده شده است)

```
With k = 1
KNN Confusion Matrix:
 [[1 0 0 0 0 0 1]
 [0 2 0 0 0 0 0]
 [0 0 2 0 0 0 0]
 [0 0 0 2 0 0 0]
 [0 0 0 1 1 0 0]
 [0 0 0 0 0 2 0]
 [0 0 0 0 0 0 2]]

KNN accuracy:
 0.8571428571428571
```

```
With k = 5
KNN Confusion Matrix:
 [[1 0 0 0 0 0 1]
 [0 2 0 0 0 0 0]
 [1 0 1 0 0 0 0]
 [0 0 0 2 0 0 0]
 [1 0 0 0 1 0 0]
 [0 0 1 0 0 1 0]
 [0 0 0 0 0 0 2]]

KNN accuracy:
 0.7142857142857143
```

```
With k = 15
KNN Confusion Matrix:
 [[1 0 1 0 0 0 0]
 [0 2 0 0 0 0 0]
 [0 0 2 0 0 0 0]
 [0 0 0 2 0 0 0]
 [0 0 0 0 2 0 0]
 [0 0 2 0 0 0 0]
 [0 0 0 0 0 0 2]]

KNN accuracy:
 0.7857142857142857
```

شکل‌های گرافیکی ماتریس درهم‌ریختگی برای مدل‌های KNN به‌صورت زیر است: