Московский государственный технический университет им. Н.Э. Баумана.

Факультет «Информатика и управление»

Кафедра ИУ5. Курс «Методы машинного обучения»

Отчет по лабораторной работе №4

«Создание рекомендательной модели»

Выполнил:
студент группы ИУ5-23Б
Белоусов Евгений
Подпись и дата:

Проверил:
преподаватель каф. ИУ5
Гапанюк Ю. Е.
Подпись и дата:

Москва, 2022 г.

Описание задания

1. Выбрать произвольный набор данных (датасет), предназначенный для построения рекомендательных моделей.
2. Опираясь на материалы лекции, сформировать рекомендации для одного пользователя (объекта) двумя произвольными способами.
3. Сравнить полученные рекомендации (если это возможно, то с применением метрик).

```
1 ! unzip /content/drive/MyDrive/Colab_data/MMO/wines-description.zip
```

```
Archive:  /content/drive/MyDrive/Colab_data/MMO/wines-description.zip
replace winemag-data-130k-v2.csv? [y]es, [n]o, [A]ll, [N]one, [r]ename: n
replace winemag-data-130k-v2.json? [y]es, [n]o, [A]ll, [N]one, [r]ename: n
replace winemag-data_first150k.csv? [y]es, [n]o, [A]ll, [N]one, [r]ename: n
```

```
1 import numpy as np
2 import pandas as pd
3 from sklearn.feature_extraction.text import TfidfVectorizer
4 from sklearn.neighbors import KNeighborsRegressor, KNeighborsClassifier
5 from sklearn.metrics.pairwise import cosine_similarity, manhattan_distances, euclidean_
```

```
1 data1 = pd.read_csv('winemag-data-130k-v2.csv')
2 data1.head()
```

	Unnamed: 0	country	description	designation	points	price	province	region_1	r
0	0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	
1	1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	
			Tart and						

```
1 data2 = pd.read_csv('winemag-data_first150k.csv')
2 data2.head()
```

	Unnamed: 0	country	description	designation	points	price	province	region_1	r
0	0	US	This tremendous 100% varietal wine hails from ...	Martha's Vineyard	96	235.0	California	Napa Valley	
			Ripe aromas of fig	Carodorum					

```
1 data = pd.concat([data1, data2], axis=0, ignore_index=True)
2 data.head()
```

	Unnamed: 0	country	description	designation	points	price	province	region_1	r
0	0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	
1	1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	
			Tart and						

```
1 data.shape
```

```
(280901, 14)
```

```
1 description_data = data[data['description'].notnull()]
2 description_data.shape
```

```
(280901, 14)
```

```
1 titels = description_data['title'].values
2 titels[0:5]
```

```
array(['Nicosia 2013 Vulkà Bianco  (Etna)',
       'Quinta dos Avidagos 2011 Avidagos Red (Douro)',
       'Rainstorm 2013 Pinot Gris (Willamette Valley)',
       'St. Julian 2013 Reserve Late Harvest Riesling (Lake Michigan Shore)',
       "Sweet Cheeks 2012 Vintner's Reserve Wild Child Block Pinot Noir (Willamette \
      dtype=object)
```

```
1 descriptions = description_data['description'].values
2 descriptions[0:5]
```

```
array(["Aromas include tropical fruit, broom, brimstone and dried herb. The palate is
       "This is ripe and fruity, a wine that is smooth while still structured. Firm t
       'Tart and snappy, the flavors of lime flesh and rind dominate. Some green pine
       'Pineapple rind, lemon pith and orange blossom start off the aromas. The palat
```

```
       "Much like the regular bottling from 2012, this comes across as rather rough a
      dtype=object)
```

```
description_data.keys()
```

```
Index(['Unnamed: 0', 'country', 'description', 'designation', 'points',
       'price', 'province', 'region_1', 'region_2', 'taster_name',
       'taster_twitter_handle', 'title', 'variety', 'winery'],
      dtype='object')
```

```
wine_ids = description_data['Unnamed: 0'].values
wine_ids
```

```
array([     0,      1,      2, ..., 150927, 150928, 150929])
```

```
%%time
tfidf = TfidfVectorizer()
description_matrix = tfidf.fit_transform(descriptions)
description_matrix
```

```
CPU times: user 10.7 s, sys: 646 ms, total: 11.4 s
Wall time: 11.4 s
```

```
description_matrix
```

```
<280901x37137 sparse matrix of type '<class 'numpy.float64'>'
        with 9637987 stored elements in Compressed Sparse Row format>
```

```
class SimplerKnnRecomender:
  def __init__(self, X_matrix, X_ids, X_title, X_overview):
        """
        Входные параметры:
        X_matrix - обучающая выборка (матрица объект-признак)
        X_ids - массив идентификаторов объектов
        X_title - массив названий объектов
        X_overview - массив описаний объектов
        """
        #Сохраняем параметры в переменных объекта
        self._X_matrix = X_matrix
        self.df = pd.DataFrame(
            {'id': pd.Series(X_ids, dtype='int'),
            'title': pd.Series(X_title, dtype='str'),
            'overview': pd.Series(X_overview, dtype='str'),
            'dist': pd.Series([], dtype='float')})

  def recommend_for_single_object(self, K: int, \
                X_matrix_object, cos_flag = True, manh_flag = False):
        """
        Метод формирования рекомендаций для одного объекта.
        Входные параметры:
        K - количество рекомендуемых соседей
```

```
        X_matrix_object - строка матрицы объект-признак, соответствующая объекту
        cos_flag - флаг вычисления косинусного расстояния
        manh_flag - флаг вычисления манхэттэнского расстояния
        Возвращаемое значение: K найденных соседей
        """

        scale = 1000000
        # Вычисляем косинусную близость
        if cos_flag:
            dist = cosine_similarity(self._X_matrix, X_matrix_object)
            self.df['dist'] = dist * scale
            res = self.df.sort_values(by='dist', ascending=False)
            # Не учитываем рекомендации с единичным расстоянием,
            # так как это искомый объект
            res = res[res['dist'] < scale]

        else:
            if manh_flag:
                dist = manhattan_distances(self._X_matrix, X_matrix_object)
            else:
                dist = euclidean_distances(self._X_matrix, X_matrix_object)
            self.df['dist'] = dist * scale
            res = self.df.sort_values(by='dist', ascending=True)
            # Не учитываем рекомендации с единичным расстоянием,
            # так как это искомый объект
            res = res[res['dist'] > 0.0]

        # Оставляем K первых рекомендаций
        res = res.head(K)
        return res
```

Тест

```
test_id = 147
print(titels[test_id])
print(descriptions[test_id])
```

```
Vincent Vineyards 2012 Family Reserve Cabernet Sauvignon (Santa Ynez Valley)
Black cherry, black plum and black currant are integrated well into the fresh dill ar
```

```
test_matrix = description_matrix[test_id]
test_matrix
```

```
<1x37137 sparse matrix of type '<class 'numpy.float64'>'
        with 43 stored elements in Compressed Sparse Row format>
```

```
skr1 = SimplerKnnRecomender(description_matrix, wine_ids, titels, descriptions)
```

```
rec = skr1.recommend_for_single_object(15, test_matrix)
rec
```

	id	title	overview	dist
113991	113991	Rancho Sisquoc 2013 Cabernet Sauvignon (Santa ...	Smashed black rocks mesh with dried dill, oreg...	342602.135622
135479	5508	NaN	Ripe and jammy, this is a rich wine from a fin...	322750.946229
57464	57464	Bailli de Bourg 2010 Côtes de Bourg	Ripe and jammy, this is a rich wine from a fin...	322750.946229
217265	87294	NaN	With good concentration as well as stalky tann...	322274.198901
80952	80952	Clos des Cordeliers 2006 Prestige (Saumur-Cha...	With good concentration as well as stalky tann...	322274.198901
110633	110633	Clos des Cordeliers 2006 Prestige (Saumur-Cha...	With good concentration as well as stalky tann...	322274.198901
31770	31770	Sutcliffe 2014 Cabernet Franc (Colorado)	The nose is full of baking spices, cinnamon an...	322237.645911
87026	87026	Sutcliffe 2014 Cabernet Franc (Colorado)	The nose is full of baking spices, cinnamon an...	322237.645911
110381	110381	Santos & Seixo 2013 Reserva Red (Douro)	This well-structured, ripe wine has spice, as ...	321790.762687
20683	20683	Georges Vigouroux 2010	This is a fruity wine with juicy	314741.908173

```
1 rec2 = skr1.recommend_for_single_object(15, test_matrix, cos_flag=False)
2 rec2
```

	id	title	overview	dist
		Rancho Sisquoc 2013 Cabernet	Smashed black rocks mash	

fastText

			Ripe and jammy, this is a rich	

```
1 ! pip install fasttext
```

```
Requirement already satisfied: fasttext in /usr/local/lib/python3.7/dist-packages (0
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (from
Requirement already satisfied: pybind11>=2.2 in /usr/local/lib/python3.7/dist-package
Requirement already satisfied: setuptools>=0.7.0 in /usr/local/lib/python3.7/dist-pac
```

80952	80952			1.164239e+06

```
1 import fasttext
```

217265	87294	NaN		1.164239e+06

```
1 !gunzip /content/drive/MyDrive/Colab_data/MMO/cc.en.300.bin.gz
```

87026	87026		spices, cinnamon an	1.164270e+06

```
1 ft = fasttext.load_model('/content/drive/MyDrive/Colab_data/MMO/cc.en.300.bin')
```

```
Warning : `load_model` does not return WordVectorModel or SupervisedModel any more, b
```

		Red (Douro)	has spice, as ...	

```
1 description_matrix_ft = []
2 for description in descriptions:
3   description_matrix_ft.append(ft[description])
4 description_matrix_ft
```

```
[array([-3.68458708e-03, -1.17978789e-02, -8.43475666e-03,  2.02767048e-02,
        -1.95700433e-02, -1.75447129e-02, -1.31429371e-03,  3.46399471e-03,
        -5.78520959e-03, -7.98403285e-03,  1.21923732e-02, -2.03261530e-04,
         1.26048932e-02, -5.42361801e-03, -1.15254428e-02, -4.75562401e-02,
         1.63089018e-02, -1.21812418e-03,  1.01590184e-02,  2.36188732e-02,
        -1.31515041e-02,  1.09748114e-02,  2.32175062e-03, -4.30774502e-03,
         3.61965187e-02,  3.49890105e-02, -5.73193096e-03,  7.27625331e-03,
         2.61073164e-03,  3.57478410e-02, -2.18761247e-03, -2.82946363e-04,
         5.08598657e-03,  5.19180391e-03,  2.01852974e-02, -9.76802781e-04,
         9.22914071e-04,  1.79132894e-02,  1.86295598e-03,  2.37991307e-02,
         2.13875365e-03, -1.32100610e-03, -1.08428882e-03,  1.53334846e-03,
        -3.40479892e-03,  2.42982507e-02, -2.02872208e-03,  5.58797503e-03,
        -5.06552286e-04,  2.83353467e-04,  1.12648718e-02, -1.81287657e-02,
        -9.15385201e-04, -8.54227040e-03,  6.04627701e-03,  6.92596799e-03,
         1.56389512e-02, -2.47626449e-03, -2.13058908e-02, -2.16606678e-03,
         4.69146203e-03,  1.30530866e-03,  3.25641921e-03,  2.19029877e-02,
         3.02401371e-03, -5.16555598e-03,  2.34230068e-02,  5.59921470e-03,
        -1.42499143e-02, -5.72375208e-03,  2.09960807e-02, -5.48990024e-03,
        -6.91474695e-03,  1.02537936e-02, -8.68808292e-03, -2.85980641e-03,
         5.97267738e-03,  4.96197026e-03, -1.09779434e-02,  1.89949460e-02,
        -1.17272721e-03,  5.19637659e-04, -2.97866366e-03,  6.98823947e-03,
        -1.41893895e-02,  2.80324998e-03, -2.09303517e-02, -2.44304836e-02,
         1.25627099e-02,  4.22734814e-03, -4.98524518e-04, -8.05275142e-03,
        -2.42222175e-02,  3.97287309e-03,  1.38222445e-02, -6.86501298e-05,
         3.49551961e-02, -9.34907049e-03, -3.55202798e-03, -5.54964226e-03,
         1.57819549e-03,  4.09523910e-03, -1.44580742e-02,  1.22106923e-02,
         3.17316060e-03, -5.79977175e-03,  1.35909785e-02,  9.54005402e-03,
```

```
            -1.41152423e-02, -4.06228099e-03, -2.52920669e-02,  9.03851259e-03,
            -5.35519095e-03, -1.52418781e-02, -4.97860601e-04,  1.79471411e-02,
            -5.24512492e-03,  1.03180623e-02, -2.07145422e-04,  1.16569968e-02,
            -5.22898510e-03,  1.84096349e-03, -4.61022044e-03,  1.17521146e-02,
             2.38551293e-03,  2.26342715e-02,  1.10772997e-02,  1.64676603e-04,
             1.66861825e-02, -1.06739663e-02,  7.66480062e-03, -5.34854736e-03,
            -3.13132710e-04, -6.22675661e-03,  2.91860886e-02,  9.30599496e-03,
            -4.20451816e-03, -6.53517060e-03, -1.64806265e-02,  8.38649645e-03,
             1.18816895e-02, -1.55502027e-02,  1.60941272e-04,  6.93022739e-03,
            -2.63898131e-02, -2.07829708e-03, -2.91024037e-02, -4.26418334e-03,
             3.80725623e-03, -4.17405012e-04,  1.22516891e-02,  3.97527678e-04,
            -2.10073590e-03, -5.82508370e-03, -3.47377150e-04,  1.24703804e-02,
             8.99744686e-03, -5.60176186e-03, -6.49045361e-03, -1.03788627e-02,
            -6.54661679e-04, -4.53409180e-03, -2.16112696e-02,  1.33514041e-02,
            -3.89213092e-05, -7.79301487e-03, -4.51588584e-03, -6.65698899e-03,
             1.05322227e-02,  1.32866642e-02,  1.50387799e-02,  8.29816610e-03,
            -1.00667253e-02,  1.02605494e-02,  2.09176238e-03, -8.76046717e-03,
            -1.93075370e-03, -2.75472971e-03,  4.27537132e-03, -1.38256010e-02,
            -1.13010537e-04,  2.30136048e-02,  3.74199380e-03, -8.63977335e-03,
            -8.38049036e-03, -1.83874648e-02, -1.14306090e-02,  4.12548985e-03,
            -7.90136866e-03, -1.04491180e-03,  1.18658869e-02, -2.10305043e-02,
            -7.70762842e-03,  3.04868817e-03,  3.07237240e-03, -1.01640029e-03,
             7.79729197e-03,  5.03537385e-03,  1.00220405e-02,  1.99493691e-02,
            -1.40973125e-02,  1.06779300e-02, -6.75673690e-03, -4.64632502e-03,
            -1.72339589e-03, -5.08606667e-03,  2.37186458e-02,  7.27834413e-03,
            -1.39557375e-02, -1.82862757e-04, -1.85826924e-02, -1.39134815e-02,
             8.68175365e-03, -1.32100808e-03,  3.99251282e-03,  1.72778778e-03,
            -8.41213297e-03,  7.97637273e-03, -3.25597619e-04, -3.68739013e-03,
             3.71845695e-03,  3.34221194e-03,  2.30807811e-03,  1.53618371e-02,
             4.06395225e-03, -4.14198311e-03,  7.95530714e-03,  1.40668452e-02,
            -1.56306941e-03,  1.46461055e-02,  2.17132270e-02, -3.49055720e-03,
```

```
1 len(description_matrix_ft)
```

```
    280901
```

```
1 skr2 = SimplerKnnRecomender(description_matrix_ft, wine_ids, titels, descriptions)
```

```
1 test_matrix_ft = description_matrix_ft[test_id]
```

```
1 test_matrix_ft = np.array([test_matrix_ft])
```

```
1 rec3 = skr2.recommend_for_single_object(15, test_matrix_ft)
2 rec3
```

	id	title	overview	dist
105074	105074	San Simeon 2012 Estate Reserve Petite Sirah (P...	Elegant aromas of blackcurrant and black cassi...	930990.695953
259339	129368	NaN	Quite well-oaked, which is the house style for...	930847.227573
124696	124696	Margerum 2013 Über Syrah (Santa Barbara County)	Heaps of crushed black pepper meet with fresh ...	930402.696133
113991	113991	Rancho Sisquoc 2013 Cabernet Sauvignon (Santa ...	Smashed black rocks mesh with dried dill, oreg...	930130.124092
129065	129065	Oso Libre 2011 Querida Cabernet Sauvignon (Pas...	There is a density and unity of dark fruit on ...	929983.019829
246852	116881	NaN	From the onset this wine struts its stuff. The...	929539.918900
259752	129781	NaN	From the onset this wine struts its stuff. The...	929539.918900
83810	83810	Longoria 2014 Sanford & Benedict Vineyard Pino...	A strong black-pepper character on the nose of...	929375.410080
193719	63748	NaN	Shows a lifted fruit character on the nose—may...	929302.692413
277479	147508	NaN	Shows a lifted fruit character on the nose—may...	929302.692413