



به نام خدا
دانشگاه تهران



دانشکده مهندسی برق و کامپیوتر

درس شبکه‌های عصبی و یادگیری عمیق

تمرین اول

نام و نام خانوادگی	حسام اسداله زاده – مسعود طهماسبی
شماره دانشجویی	810198429 – 810198346
تاریخ ارسال گزارش	۱۴۰۱.۰۸.۱۹

پاسخ ۳ – Restricted Boltzmann Machine

۳-۱. سیستم توصیه‌گر

(A) پنج مورد نخست و آخر از فیلم‌ها:

```
[4] 1 movies_df
```

	movieId	title	genres
0	1	Toy Story (1995)	Adventure Animation Children Comedy Fantasy
1	2	Jumanji (1995)	Adventure Children Fantasy
2	3	Grumpier Old Men (1995)	Comedy Romance
3	4	Waiting to Exhale (1995)	Comedy Drama Romance
4	5	Father of the Bride Part II (1995)	Comedy
...
9737	193581	Black Butler: Book of the Atlantic (2017)	Action Animation Comedy Fantasy
9738	193583	No Game No Life: Zero (2017)	Animation Comedy Fantasy
9739	193585	Flint (2017)	Drama
9740	193587	Bungo Stray Dogs: Dead Apple (2018)	Action Animation
9741	193609	Andrew Dice Clay: Dice Rules (1991)	Comedy

9742 rows x 3 columns

پنج مورد نخست و آخر از امتیازها:

```
[5] 1 ratings_df
```

	userId	movieId	rating	timestamp
0	1	1	4.0	964982703
1	1	3	4.0	964981247
2	1	6	4.0	964982224
3	1	47	5.0	964983815
4	1	50	5.0	964982931
...
100831	610	166534	4.0	1493848402
100832	610	168248	5.0	1493850091
100833	610	168250	5.0	1494273047
100834	610	168252	5.0	1493846352
100835	610	170875	3.0	1493846415

100836 rows x 4 columns

ابعاد هر دو دیتاست:

```
[8] 1 print("Shape of movies dataset:", movies_df.shape)
2 print("Shape of ratings dataset:", ratings_df.shape)

Shape of movies dataset: (9742, 3)
Shape of ratings dataset: (100836, 4)
```

در دیتاست فیلم‌ها، ۹۷۴۲ فیلم وجود دارد که هر کدام دارای ۳ ویژگی movieId, title, genres می‌باشند. در دیتاست امتیازها نیز ۱۰۰۸۳۶ امتیاز مربوط به ۶۱۰ کاربر وجود دارد.

اضافه کردن ستون List index:

```
[10] 1 movies_df['List index'] = movies_df.index
      2 movies_df
```

	movieId	title	genres	List index
0	1	Toy Story (1995)	Adventure Animation Children Comedy Fantasy	0
1	2	Jumanji (1995)	Adventure Children Fantasy	1
2	3	Grumpier Old Men (1995)	Comedy Romance	2
3	4	Waiting to Exhale (1995)	Comedy Drama Romance	3
4	5	Father of the Bride Part II (1995)	Comedy	4
...
9737	193581	Black Butler: Book of the Atlantic (2017)	Action Animation Comedy Fantasy	9737
9738	193583	No Game No Life: Zero (2017)	Animation Comedy Fantasy	9738
9739	193585	Flint (2017)	Drama	9739
9740	193587	Bungo Stray Dogs: Dead Apple (2018)	Action Animation	9740
9741	193609	Andrew Dice Clay: Dice Rules (1991)	Comedy	9741

9742 rows x 4 columns

(B) ادغام کردن دو دیتاست:

```
[11] 1 joint_df = ratings_df.merge(movies_df, left_on='movieId', right_on='movieId')
      2 joint_df
```

	userId	movieId	rating	timestamp	title	genres	List index
0	1	1	0.777778	964982703	Toy Story (1995)	Adventure Animation Children Comedy Fantasy	0
1	5	1	0.777778	847434962	Toy Story (1995)	Adventure Animation Children Comedy Fantasy	0
2	7	1	0.888889	1106635946	Toy Story (1995)	Adventure Animation Children Comedy Fantasy	0
3	15	1	0.444444	1510577970	Toy Story (1995)	Adventure Animation Children Comedy Fantasy	0
4	17	1	0.888889	1305696483	Toy Story (1995)	Adventure Animation Children Comedy Fantasy	0
...
100831	610	160341	0.444444	1479545749	Bloodmoon (1997)	Action Thriller	9325
100832	610	160527	0.888889	1479544998	Sympathy for the Underdog (1971)	Action Crime Drama	9330
100833	610	160836	0.555556	1493844794	Hazard (2005)	Action Drama Thriller	9342
100834	610	163937	0.666667	1493848789	Blair Witch (2016)	Horror Thriller	9389
100835	610	163981	0.666667	1493850155	31 (2016)	Horror	9390

100836 rows x 7 columns

(C) ستون timestamp حاوی اطلاعات زمانی مربوط به امتیازات کاربران است که واضحاً ارتباط یا correlation با خروجی سیستم پیشنهادگر ما ندارد پس قابل حذف است. ستون‌های title و genres حاوی اطلاعات متنی (string) مربوط به فیلم‌ها هستند که می‌توان به جای آن از movieId به عنوان ویژگی عددی استفاده کرد. البته از ویژگی movieId نیز نمی‌توان استفاده کرد چراکه تعداد فیلم‌ها و در نتیجه تعداد نورون‌های visible layer، ۹۷۴۲ تا هستند ولی بازه‌ی movieId از ۱ تا ۱۹۳۶۰۹ است در نتیجه به جای این ویژگی از ستون List index که در قسمت الف ایجاد شد استفاده می‌کنیم و ستون movieId را نیز حذف می‌کنیم.

```
1 joint_df.drop(['timestamp', 'title', 'genres', 'movieId'], axis=1, inplace=True)
2 joint_df.head()
```

	userId	rating	List index
0	1	0.777778	0
1	5	0.777778	0
2	7	0.888889	0
3	15	0.444444	0
4	17	0.888889	0

(D)

```
1 df = joint_df.groupby(by="userId")
2 df.get_group(1)
```

	userId	rating	List index
0	1	0.777778	0
215	1	0.777778	2
267	1	0.777778	5
369	1	1.000000	43
572	1	1.000000	46
...
16056	1	0.777778	2802
16075	1	1.000000	2836
16208	1	0.777778	2847
16243	1	0.777778	2991
16250	1	1.000000	3673

232 rows x 3 columns

(E)

```
[9] 1 rating = ratings_df['rating']
2 train_X = (rating-rating.min())/(rating.max()-rating.min())
3 ratings_df['rating'] = train_X
4 ratings_df.head()
```

	userId	movieId	rating	timestamp
0	1	1	0.777778	964982703
1	1	3	0.777778	964981247
2	1	6	0.777778	964982224
3	1	47	1.000000	964983815
4	1	50	1.000000	964982931

(H) در مرحله‌ی ترین کردن مدل باید به این نکته توجه کنیم که نباید اطلاعات کاربر ۷۵ را به مدل بدهیم چراکه این کار موجب data leakage می‌شود. نتایج زیر، فیلم‌های پیشنهادی توسط مدل را نشان می‌دهد.

```
[41] 1 movies_df['title'][ind]
```

418	Jurassic Park (1993)
2078	Sixth Sense, The (1999)
2226	Fight Club (1999)
257	Pulp Fiction (1994)
97	Braveheart (1995)
911	Star Wars: Episode VI - Return of the Jedi (1983)
900	Raiders of the Lost Ark (Indiana Jones and the...)
507	Terminator 2: Judgment Day (1991)
2145	American Beauty (1999)
659	Godfather, The (1972)
1503	Saving Private Ryan (1998)
510	Silence of the Lambs, The (1991)
898	Star Wars: Episode V - The Empire Strikes Back...
224	Star Wars: Episode IV - A New Hope (1977)
1939	Matrix, The (1999)

Name: title, dtype: object

نکته‌ی جالب توجه در فیلم‌های پیشنهادی این است که در میان این فیلم‌ها، برخی فیلم‌ها توسط کاربر ۷۵ نمره‌دهی شده‌اند. به عنوان مثال فیلم Star Wars: Episode IV - A New Hope توسط این کاربر با امتیاز ۵ رتبه‌بندی شده است؛ این یعنی مدل به خوبی توانسته است تشخیص دهد که این کاربر فیلم ذکرشده را می‌پسندد. همچنین مدل به خوبی تشخیص داده است که این کاربر به مجموعه فیلم‌های Star Wars علاقه‌مند است و قسمت‌های دیگر این مجموعه را نیز پیشنهاد داده است؛ به طوریکه کاربر ۷۵، ۴ قسمت از این مجموعه فیلم را تماشا کرده است و مدل، قسمت ششم این مجموعه را نیز پیشنهاد داده است.