

Jin-Soo Kim
(jinsoo.kim@snu.ac.kr)

Systems Software & Architecture Lab.

Seoul National University

Jan. 6 – 17, 2020

Python for Data Analytics

Pandas Lab



Pandas lab

Lab Pandas-1. 영화 dataset

- movies.csv 는 영화 id 에 대한 영화의 이름과 장르 정보를

```
movieId,title,genres
1,Toy Story (1995),Adventure|Animation|Children|Comedy|Fantasy
2,Jumanji (1995),Adventure|Children|Fantasy
3,Grumpier Old Men (1995),Comedy|Romance
4,Waiting to Exhale (1995),Comedy|Drama|Romance
5,Father of the Bride Part II (1995),Comedy
6,Heat (1995),Action|Crime|Thriller
```

Lab Pandas-I. 영화 dataset

- ratings.csv 는 영화에 대한 사용자의 평점 데이터를 가지고 있다.
 - 이 두개 dataset을 활용해본다.

```
userId,movieId,rating,timestamp
1,1,4.0,964982703
1,3,4.0,964981247
1,6,4.0,964982224
1,47,5.0,964983815
1,50,5.0,964982931
1,70,3.0,964982400
1,101,5.0,964980868
1,110,4.0,964982176
1,151,5.0,964984041
```

Lab Pandas-1. 영화 dataset

- `def get_ratings_numpy(df)` 는 ratings.csv를 분석해서 영화 평점을 얻는 함수이다.
 - 함수에 대한 자세한 내용은 코드에
 - 단, 이 함수에서는 numpy만 이용해서 영화의 평점을 구한다.(pandas 이용 x)
 - output example

```
{id:rating}
1.0: 3.9209302325581397,
2.0: 3.4318181818181817,
3.0: 3.2596153846153846,
4.0: 2.357142857142857,
5.0: 3.0714285714285716,
```

Lab Pandas-I. 영화 dataset

- `def get_ratings_pandas(df)` 는 ratings.csv를 분석해서 영화 평점을 얻는 함수이다.
 - 함수에 대한 자세한 내용은 코드에
 - 단, 이 함수에서는 pandas를 이용해서 영화의 평점을 구한다.
 - output example

movied	
1	3.920930
2	3.431818
3	3.259615
4	2.357143
5	3.071429

Lab Pandas-1. 영화 dataset

- `def get_genres_columns(df)` 는 `movies.csv`를 사용해서, 영화의 장르들을 구하는 함수이다.
 - 함수에 대한 자세한 내용은 코드에
 - output example

```
['Romance', 'Comedy', 'Action', 'Film-Noir', 'Crime', 'Children', 'IMAX', 'Animation', ... ,  
'Mystery', 'Sci-Fi']
```

Lab Pandas-1. 영화 dataset

- `def make_df_to_onehot(df, genre_columns)`
 - 함수에 대한 자세한 내용은 코드에
 - 각 장르에 대한 column 값이 1이고 나머지는 0으로 encoding한다.
 - output example

movieId				title	Romance	Comedy	Action	Film-Noir	Crime	Children	IMAX	Animation	...	Western	Thriller	Musical	Adventure	Drama	Documentary		
War	Horror	Mystery	Sci-Fi																		
0	1			Toy Story (1995)	0		1	0	0	0	1	0	1 ...	0	0	0	1	0	0	0	0
1	2			Jumanji (1995)	0		0	0	0	0	1	0	0 ...	0	0	0	1	0	0	0	0
2	3			Grumpier Old Men (1995)		1	1	0	0	0	0	0	0 ...	0	0	0	0	0	0	0	0