Lab - Build a matrix

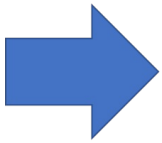
Introduction

PDF 파일 다운로드

Machin Learning의 두 번째 랩은 Pandas와 Numpy를 활용하여 Rating Matrix 또는 Frequent Matrix를 만드는 것입니다. 추천 시스템 개발 등 머신러닝을 하다 보면 누가, 어떤 물건(또는 서비스)를 얼마나 이용하고 평가 하였는 가를 Matrix 형태로 변형하여 분석하는 일이 많은데 이를 위한 전처리 과정이 필요합니다. 흔한 예제로 생각해보면 영화를 본 사용자들이 각 영화를 평가한 별점 점수를 Matrix 형태로 표현하는 것이 있습니다. 일반적으로 데이터 베이스는 저장 공간의 효율성을 위해 Matrix 형태로 저장하는 것이 아니라 Event과 발생한 정보를 Row 단위로 저장합니다. 이렇게 DB에 쌓인 정보를 Matrix로 변환하는 게 이번 랩의 목표입니다. 실제 데이터의 변환은 아래 그림과 같습니다.

Movie	User B rating
M1	3
M2	7
M3	8
M4	0
M5	0

DB에 쌓인 영화평가 정보



User	M1	M2	M3	M4	M5
A	9	4		8	
B	3	7	8		
C		8	7		4
D		5		8	

User-Movie-Rating으로 변환된 Matrix

backend.ai 설치

숙제를 제출하기 앞서, 레블업의 backend.ai를 여러분의 파이썬에 설치하셔야 합니다. 설치하는 과정은 매우 쉽습니다. 아래처럼 터미널 또는 cmd 창에서 입력을 하시면 됩니다.

```
pip install backend.ai-client
```

숙제 파일(lab_bulid_matrix.zip) 다운로드

먼저 해야 할 일은 숙제 파일을 다운로드 받는 것 입니다. 아래링크를 다운로드 하거나 Chrome 또는 익스플로러와 같은 웹 브라우저 주소창에 아래 주소를 입력합니다.

- 링크 lab_numpy.zip
- https://s3.ap-northeast-2.amazonaws.com/teamlab-gachon/mooc_pic/2_lab_bulid_matrix.zip

  또는 Mac OS에서는 아래 명령을 쓰셔도 됩니다.

```
wget https://s3.ap-northeast-2.amazonaws.com/teamlab-gachon/mooc_pic/2_lab_buli
```

  다운로드 된 `lab_bulid_matrix.zip` 파일을 작업 폴더로 이동한 후 압축해제 후 작업하시길 바랍니다.
  압축해제 하면 폴더가 `linux_mac` 과 `windows` 로 나눠져 있습니다. 자신의 OS에 맞는 폴더로 이동해서 코드를 수정해 주시기 바랍니다.

bulid_matrix.py 코드 구조

본 Lab은 Pandas의 기본적인 동작과 Numpy를 결합하여 일반적으로 쌓여있는 데이터를 Matrix 형태로 변경합니다. 변환되는 Matrix 형태는 두가지이며, 본 Lab도 두 가지 모두를 지원하는 것을 목표로 합니다.

get_rating_matrix

첫 번째 함수는 Rating Matix을 만드는 것 입니다. Rating Matrix는 영화, 책 처럼 사용자가 제품에 대한 평가를 Matrix 형태로 표현한 것입니다. 저희는 `movie_rating.csv` 라는 파일을 활용하여 rating matrix를 구성한다. `movie_rating.csv` 는 아래처럼 구성된다.

source	target	rating
Mick LaSalle	Superman Returns	3.0
Mick LaSalle	The Night Listener	3.0
Claudia Puig	Snakes on a Plane	3.5
Claudia Puig	Just My Luck	3.0
Claudia Puig	The Night Listener	4.5

Lisa Rose	Lady in the Water	2.5
Lisa Rose	Snakes on a Plane	3.5

본 랩에서 다루는 모든 csv파일의 column은 source, target으로 구성되며, source는 row의 index 정보가, target는 column의 기준 정보가 된다. rating 정보는 `get_rating_matrix` 함수에서만 사용되며, 사용자가 영화에 대한 평가를 정보를 담고 있다.

본 랩의 목적은 위 테이블과 같이 구성된 정보를 Matrix 형태로 바꾸는 거다. Matrix 형태로 바꾸는 규칙은 다음과 같다.

- source는 row, target은 column의 기준이 된다.
- source와 target의 정렬된 값을 활용하여 index를 설정한다. 즉 위 Table에서는 `Claudia Puig`과 row의 0번째 index로 설정된다.
- rating의 정보는 Matrix에서 각 Element 값에 할당된다.
- 생성되는 Matrix Ndarray로 나타난다.
- dict, collection 모듈 등 파이썬의 Built-in Module은 사용할 수 있으나, for 문은 사용할 수 없다.

생성하는 함수의 Template은 아래와 같으며, 입력값은 처리하는 csv 파일의 이름만 넣을 수 있다.

```
def get_rating_matrix(filename):
    pass
```

실제한 구현한 예제와 결과물은 아래와 같다.

```
>>> import numpy as np
>>> import build_matrix as test_code
>>> test_code.get_rating_matrix("movie_rating.csv")
array([[ 3. ,  0. ,  3.5,  0. ,  4.5,  0. ],
       [ 0. ,  3. ,  3.5,  0. ,  3. ,  3.5],
       [ 0. ,  3. ,  4. ,  5. ,  3. ,  3.5],
       [ 3. ,  2.5,  3.5,  3.5,  3. ,  2.5],
       [ 2. ,  3. ,  4. ,  3. ,  3. ,  0. ],
       [ 0. ,  0. ,  4.5,  4. ,  0. ,  0. ]], dtype=float32)
```

get_frequent_matrix

두 번째 함수는 얼마나 빈번하게 제품을 구매했는지를 표현하는 Frequent Matrix를 만드는 것 입니다. Frequent Matrix는 사용자가 특정 제품을 구매한 횟수를 기록하는 Matrix이다. 저희가 제공하는 csv파일은 `1000i.csv` 라는 파일로 아래처럼 구성되어 있습니다.

source	target
source,target	
3	7
4	15
2	49
5	44
1	1
2	19
4	22
4	34
4	40
5	31
4	17
5	16
2	43
5	20
3	48

본 함수에서는 기존 함수와 달리 Rating column이 없습니다. 대시신 source와 target의 조합이 한 개 이상으로 중복될 수 있고, 이것이 Frequent로 처리해야 합니다. 즉 Rating이 명시적으로 있는게 아니라 데이터를 통해 Frequent를 찾아내는 것이 목적입니다. Matrix 형태로 바꾸는 규칙은 다음과 같습니다.

- source는 row, target은 column의 기준이 된다.
- source와 target의 정렬된 값을 활용하여 index를 설정한다. 즉 위 Table에서는 `1` 은 row의 0번째 index로 설정된다.
- Source와 Target이 출현한 정보는 Frequent로 Matrix에서 각 Element 값에 할당되어야 한다.
- 생성되는 Matrix Ndarray로 나타내며, dtype은 np.float32
- dict, collection 모듈 등 파이썬의 Built-in Module은 사용할 수 있으나, for 문은 사용할 수 없다.

생성하는 함수의 Template은 아래와 같으며, 입력값은 처리하는 csv 파일의 이름만 넣을 수 있다.

```
def get_frequent_matrix(filename):
    pass
```

실제한 구현한 예제와 결과물은 아래와 같다.

```
>>> import numpy as np
>>> import build_matrix as test_code
>>> test_code.get_frequent_matrix("1000i.csv")
array([[ 19.,  17.,  14.,  11.,  17.,  25.,   7.,  22.,   5.,  18.,  10.,
         13.,  13.,   8.,  20.,  10.,   9.,  10.,  16.,  15.,   9.,  11.,
         17.,  15.,  14.,   8.,   6.,  12.,  18.,  12.,   6.,  18.,   9.,
         24.,   7.,  19.,  14.,   6.,   4.,  12.,  15.,  14.,  20.,   9.,
         12.,  16.,  11.,   9.,  11.,  12.],
       [ 20.,  16.,  10.,  15.,  17.,  18.,  10.,  13.,   5.,  19.,   8.,
         14.,  14.,   9.,  15.,  14.,  13.,   8.,  12.,   9.,   5.,  10.,
         28.,  18.,   7.,   8.,   6.,  19.,  14.,  13.,  11.,  12.,  18.,
         15.,   7.,  11.,  17.,   9.,   5.,   5.,  13.,  12.,  15.,   9.,
         13.,  16.,  16.,  10.,  16.,   9.],
       [ 12.,  16.,  13.,  19.,  23.,  19.,   5.,  14.,   5.,  18.,   7.,
          6.,  14.,   8.,  20.,  17.,  14.,  11.,  16.,  12.,   7.,   9.,
         23.,  12.,  12.,   8.,   7.,  23.,  26.,  10.,   9.,  20.,  16.,
         11.,   4.,  19.,  12.,  12.,   5.,  10.,  10.,  14.,  10.,  17.,
         15.,  16.,  11.,  17.,   9.,  11.],
       [ 14.,  14.,  19.,  11.,  11.,  18.,   7.,  16.,   7.,  17.,   6.,
         19.,  18.,  12.,  13.,  13.,  14.,   9.,  21.,  16.,   6.,   6.,
         19.,  14.,  19.,   5.,  12.,  14.,  18.,  11.,  11.,  21.,  15.,
         10.,  11.,  14.,  17.,  21.,   6.,  14.,   9.,  16.,  18.,  12.,
         16.,  16.,  26.,  16.,  12.,  20.],
       [ 13.,   7.,   8.,  15.,  13.,  16.,   3.,  19.,  11.,  12.,   7.,
         10.,  13.,  14.,  16.,  14.,  23.,   9.,  13.,  10.,  11.,   3.,
         11.,  14.,   9.,   6.,  11.,  16.,  18.,  11.,   5.,  14.,  10.,
         16.,  10.,   5.,  14.,  11.,   3.,   9.,  11.,  10.,  16.,   8.,
         13.,  20.,  14.,  18.,  21.,   3.]], dtype=float32)
```

숙제 template 파일 제출하기 (윈도우의 경우)

1. `windows` + `r` 를 누르고 cmd 입력 후 확인을 클릭합니다.
2. 작업을 수행한 폴더로 이동 합니다.
3. 밑에 명령어를 cmd창에 입력합니다.

```
install.bat
submit.bat [YOUR_HASH_KEY]
```

숙제 template 파일 제출하기 (Mac or Linux)

1. 터미널을 구동합니다.
2. 작업을 수행한 디렉토리로로 이동 합니다.
3. 밑에 bash창을 입력합니다.

```
bash install.sh
bash submit.sh [YOUR_HASH_KEY]
```

> backend.ai 서비스의 업데이트에 의해 실행전 반드시 `bash install.sh` 또는 `install.bat` 수행을 바랍니다.

Next Work

고생하셨습니다. Numpy와 Pandas를 함께 해야함 성공할 수 있는 랩입니다. 아직 Matrix와 Vector 데이터를 핸들링하는 방법이 익숙하지 않았다면 상당히 어렵게 푸셨을 것 같습니다. 그럼에도 불구하고, 우리는 계속 전진해야 합니다. Code가 당신과 함께 하길…

> **Human knowledge belongs to the world** - from movie 'Password' -