파이썬
추천시스템
surprise 패키지

Surprise – 파이썬 추천 패키지

- R은 recommenderlab , Spark 는 MLlib에서 쉽게 Recommendation을 수행할 수 있는 패키지를 가지고 있는 방면에 사이킷런에는 Recommendation을 쉽게 수행할 수 있는 package 를 가지고 있지 않습니다.
- Python 에서 recommendation을 쉽게 제공하는 대표적인 패키지로서 surprise 가 있습니다. Surprise는 Scikit learn의 API 와 유사하게 작성되어 있으며 이를 이용해 Recommendation Process 를 쉽게 적용할 수 있습니다.
- pip 또는 conda로 설치할 수 있으며, 윈도우 운영체제에 설치시에는 Visual studio build tools이 미리 설치되어 있어야 합니다.

Surprise 패키지를 이용한 추천 수행 프로세스

Surprise를 이용한 추천 구현 기본

1	필요한 라이브러리 로딩	from surprise import SVD, Dataset, accuracy from surprise.model_selection import train_test_split
2	필요한 데이터 세트를 로딩. 데이타는 Dataset 패키지를 이용 Csv 파일 및 Pandas Dataframe에서도 Loading 가능. 로딩한 데이터 세트를 학습용과 테스트용 데이터 세트로 분리.	data = Dataset.load_builtin('ml-100k') trainset, testset = train_test_split(data, test_size=.25)
3	행렬 분해를 수행할 알고리즘으로 SVD 생성하고 학습용 데이터로 학습.	algo = SVD() algo.fit(trainset)
4	테스트 데이터 세트에 대해서 prediction을 수행. 일반적인 scikit learn의 predict() 메소드는 surprise에서 test() 메소드 특정 사용자와 item에 대한 predict는 predict() 메소드 .	predictions = algo.test(testset)

```
Prediction(uid='120', iid='282', r_ui=4.0, est=3.4107789581968633, details={'was_impossible': False})
```

Surprise 주요 모듈 소개 - Dataset

- Surprise 는 무비렌즈 데이터 세트와 같이 userid, itemid , rating 컬럼들이 사용자(userid)를 기준으로 한 로우 레벨의 평점 데이터 로 구성된 데이터 세트만 입력 가능합니다.
- 입력받은 데이타의 첫번째 컬럼을 사용자 ID , 두번째 컬럼을 Item ID , 세번째 컬럼을 Rating으로 가정합니다. 네번째 부터는 Recommendation 알고리즘에 아예 사용하지 않습니다.
- 이렇게 로우 레벨로 입력 받은 사용자-아이템 데이터는 Dataset 객체로 로딩 후 사용자-아이템 평점 행렬로 변환 됩니다.

Movie Lens 사용자 평점 데이타

사용자 ID	아이템 ID	평점	Time Stamp
196	242	3	881250949
186	302	3	891717742
22	377	1	878887116
244	51	2	880606923
166	346	1	886397596

일반 데이터 파일 , 또는 Pandas Dataframe에서 로딩 가능합니다.
단 사용자 ID, 아이템 ID, 평점의 컬럼순은 반드시 지켜야 합니다.

Dataset 클래스의 주요 메소드

메소드 명	설명
Dataset.load_builtin(name='ml-100k')	무비렌즈 아카이브 FTP 서버에서 무비렌즈 데이터를 내려받습니다. ml-100k, ml-1M를 내려받을 수 있습니다. 일단 내려받은 데이터는 .surprise_data 디렉터리 밑에 저장되고, 해당 디렉터리에 데이터가 있으면 FTP에서 내려받지 않고 해당 데이터를 이용합니다. 입력 파라미터인 name으로 대상 데이터가 ml-100k인지 ml-1m인지를 입력합니다(name='ml-100k'). 디폴트는 ml-100k입니다
Dataset.load_from_file (file_path, reader)	OS 파일에서 데이터를 로딩할 때 사용합니다. 콤마, 탭 등으로 컬럼이 분리된 포맷의 OS 파일에서 데이터를 로딩합니다. 입력 파라미터로 OS 파일명, Reader로 파일의 포맷을 지정합니다
Dataset.load_from_df (df, reader)	판다스의 DataFrame에서 데이터를 로딩합니다. 파라미터로 DataFrame을 입력받으며 DataFrame 역시 반드시 3개의 컬럼인 사용자 아이디, 아이템 아이디, 평점 순으로 컬럼 순서가 정해져 있어야 합니다. 입력 파라미터로 DataFrame 객체, Reader로 파일의 포맷을 지정합니다.

Surprise 주요 모듈 소개 - Reader

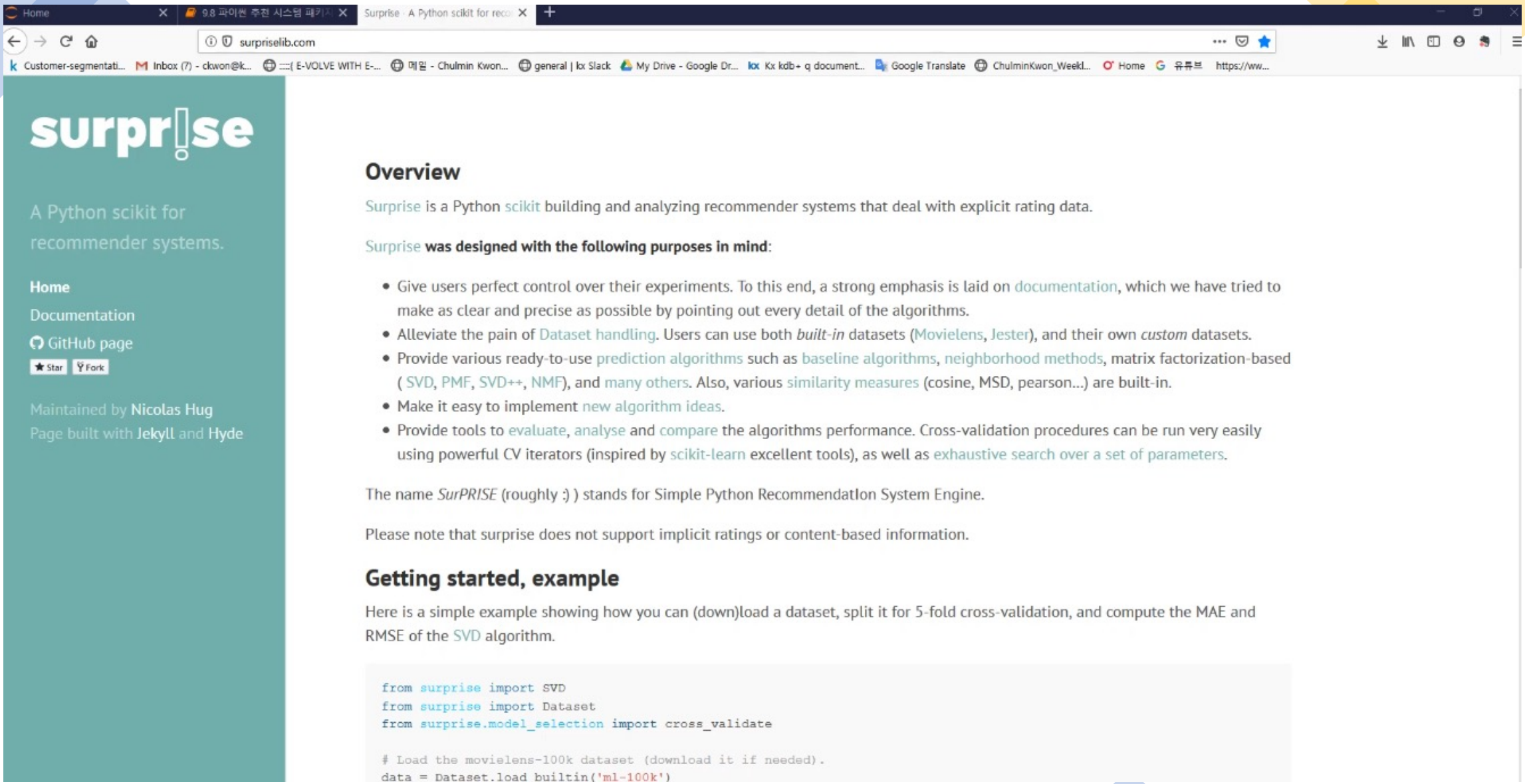
- Raw 데이터 소스에서 Dataset로 로딩 규칙을 지정하기 위해 사용됩니다.
- Surprise 데이터 세트는 기본적으로 무비렌즈 데이터와 같은 로우 레벨의 사용자-아이템 평점 데이터 형식을 따르므로 무비렌즈 데이터 형식이 아닌 경우 이를 변환하여 Dataset로 로딩해야 합니다.

```
from surprise import Reader
reader = Reader(line_format='user item rating timestamp', sep=', ', rating_scale=(0.5, 5))
data=Dataset.load_from_file('./ml-latest-small/ratings_noh.csv', reader=reader)
```

- line_format (string): 컬럼을 순서대로 나열합니다. 입력된 문자열을 공백으로 분리해 컬럼으로 인식합니다.
- sep (char): 컬럼을 분리하는 분리자이며, 디폴트는 '₩t'입니다. 판다스 DataFrame에서 입력받을 경우에는 기재할 필요가 없습니다.
- rating_scale (tuple, optional): 평점 값의 최소 ~ 최대 평점을 설정합니다. 디폴트는 (1, 5)이지만 ratings.csv 파일의 경우는 최소 평점이 0.5, 최대 평점이 5이므로 (0.5, 5)로 설정했습니다.

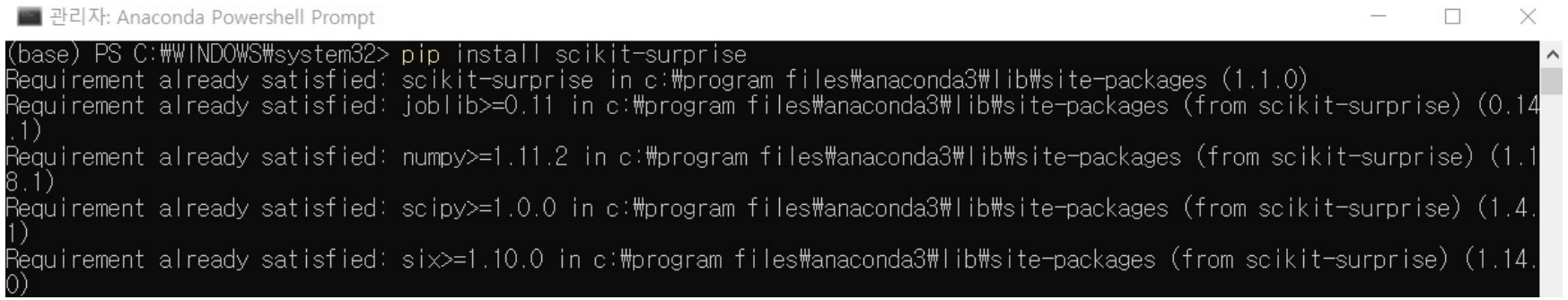
무비렌즈 협업필터링 추천시스템 구현 실습 (surprise 패키지 기본 모델)

surpriselib.com



서프라이즈 패키지 설치

pip install scikit-surprise



Surprise 추천 알고리즘 클래스

클래스명	설명
SVD	행렬 분해를 통한 잠재 요인 협업 필터링을 위한 SVD 알고리즘.
KNNBasic	최근접 이웃 협업 필터링을 위한 KNN 알고리즘.
BaselineOnly	사용자 Bias와 아이템 Bias를 감안한 SGD 베이스라인 알고리즘.

지원 알고리즘은 surprise 사이트 문서에서 참조할 수 있습니다
(http://surprise.readthedocs.io/en/stable/prediction_algorithms_package.html)

사용자의 성향을 반영한 Baseline rating

$$\hat{r}_{ui} = b_{ui} = \mu + b_u + b_i$$

사용자 u의 아이템 i에 대한 예측 평점은
전체 사용자의 평균 영화 평점 + 사용자 편향점수 + 아이템 편향 점수

모든 사용자의 평균 영화 평점 : 3.5

난 진정한 영화 매니아.
영화 평가는 언제나 깐깐하게

사용자 A 평균 평점
3.0

어벤저스 3편 평균 평점
4.2

사용자 A 의 어벤저스 3편 베이스 라인 평점= 3.5 – 0.5 + 0.7= 3.7

모든 사용자의 평균 영화 평점		사용자 편향 점수		아이템 편향 점수
3.5	+	3.0 -3.5 = - 0.5	+	4.2 -3.5 = 0.7
		특정 사용자 평균 평점 – 전체 사용자 평균 평점		특정 영화 평균 평점 – 전체 사용자 평균 평점

Baseline rating을 반형한 행렬 분해의 비용 최소화 함수

$$min(\sum_{r_{ui} \in R_{train}} (r_{ui} - \hat{r}_{ui})^2 + \lambda \left(b_i^2 + b_u^2 + ||q_i||^2 + ||p_u||^2 \right))$$

b_i 는 아이템 편향 점수

b_u는 사용자 편향 점수

SVD의 튜닝 파라미터

파라미터명	내용
n_factors	잠재 요인 K의 개수. 디폴트는 100. 커질수록 정확도가 높아질 수 있으나 과적합 문제가 발생할 수 있습니다.
n_epochs	SGD(Stochastic Gradient Descent) 수행 시 반복 횟수. 디폴트는 20.
biased (bool)	베이스라인 사용자 편향 적용 여부이며, 디폴트는 True입니다.

교차 검증과 하이퍼 파라미터 튜닝

Surprise는 교차 검증과 하이퍼 파라미터 튜닝을 위해 사이킷런과 유사한 cross_validate( )와 GridSearchCV 클래스를 제공합니다

추천 Summary

추천 시스템의 중요성

추천 시스템의 유형

- 콘텐츠 기반 필터링
- 협업 필터링
  - **최근접 이웃 기반(Nearest Neighbor)**
    - 사용자 기반 (User-user CF)
    - 아이템 기반 (Item-item CF)
  - **잠재 요인 기반(Latent Factor)**
    - 행렬 분해 기반(Matrix Factorization)

Surprise – 파이썬 추천 패키지