

<복제물에 대한 경고>

본 저작물은 **저작권법 제25조 수업목적 저작물 이용 보상금제도**에 의거, **한국복제전송저작권협회**와 약정을 체결하고 적법하게 이용하고 있습니다. 약정범위를 초과하는 사용은 저작권법에 저촉될 수 있으므로

저작물의 재 복제 및 수업 목적 외의 사용을 금지합니다.

2020. 03. 30.

건국대학교(서울)한국복제전송저작권협회

<전송에 대한 경고>

본 사이트에서 수업 자료로 이용되는 저작물은 **저작권법 제25조 수업목적 저작물 이용 보상금제도**에 의거,

한국복제전송저작권협회와 약정을 체결하고 적법하게 이용하고 있습니다.

약정범위를 초과하는 사용은 저작권법에 저촉될 수 있으므로

수업자료의 대중 공개·공유 및 수업 목적 외의 사용을 금지합니다.

2020. 03. 30.

건국대학교(서울)한국복제전송저작권협회

기계 학습 (Machine Learning)

컴퓨터공학부/인공지능학과(대학원)

김 학 수

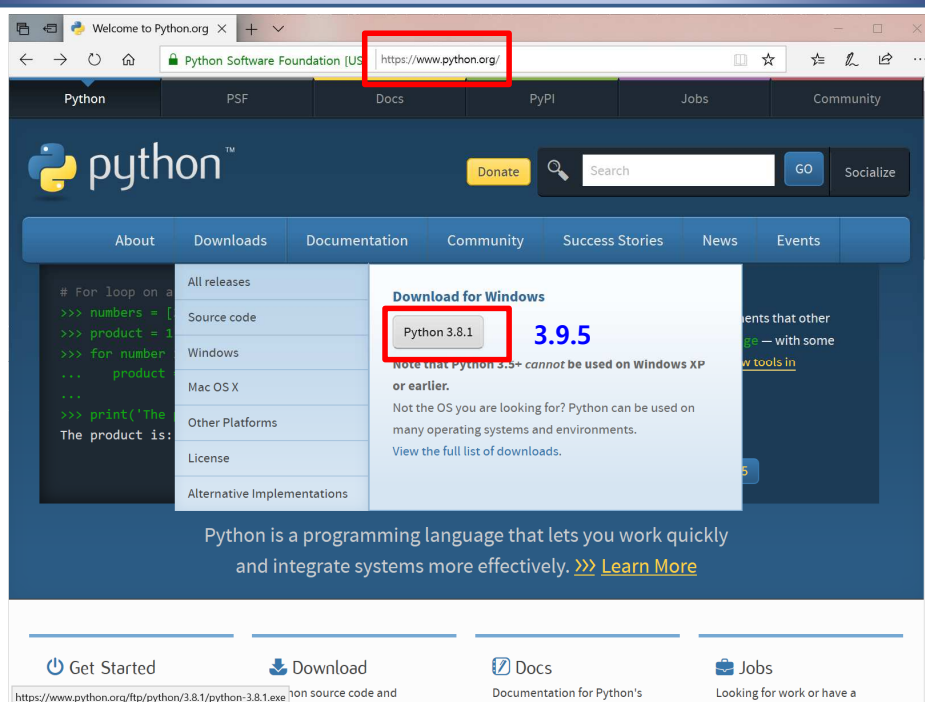
nlpdrkim@konkuk.ac.kr
<http://nlp.konkuk.ac.kr>

Prerequisite for Machine Learning

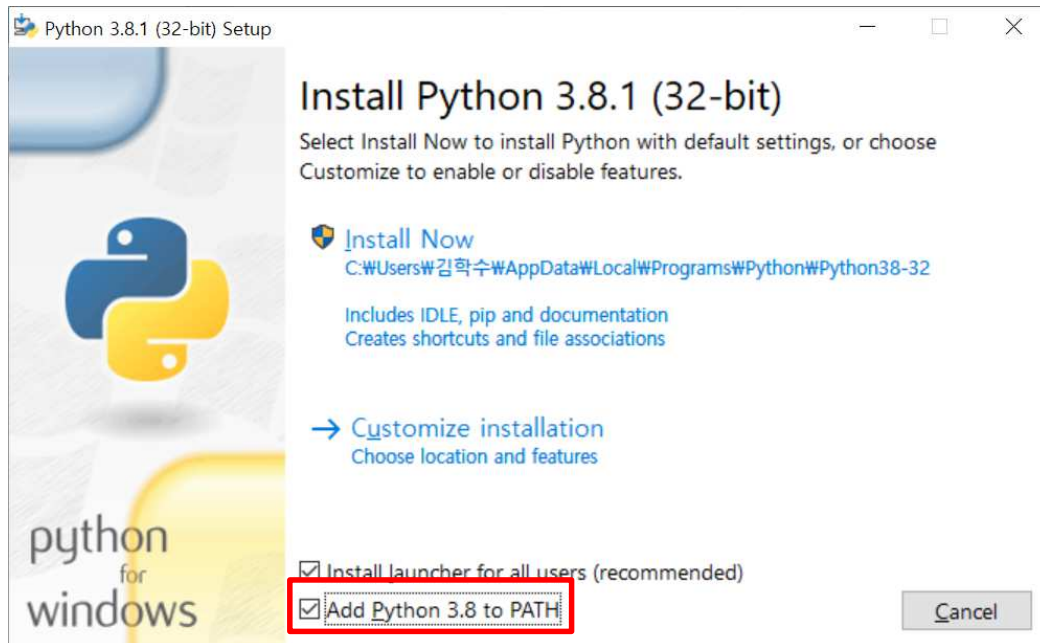
파이썬 언어 및 라이브러리(colab 포함) 설치가 되어 있는 분은
본 챕터를 건너뛰셔도 됩니다.



Python 설치

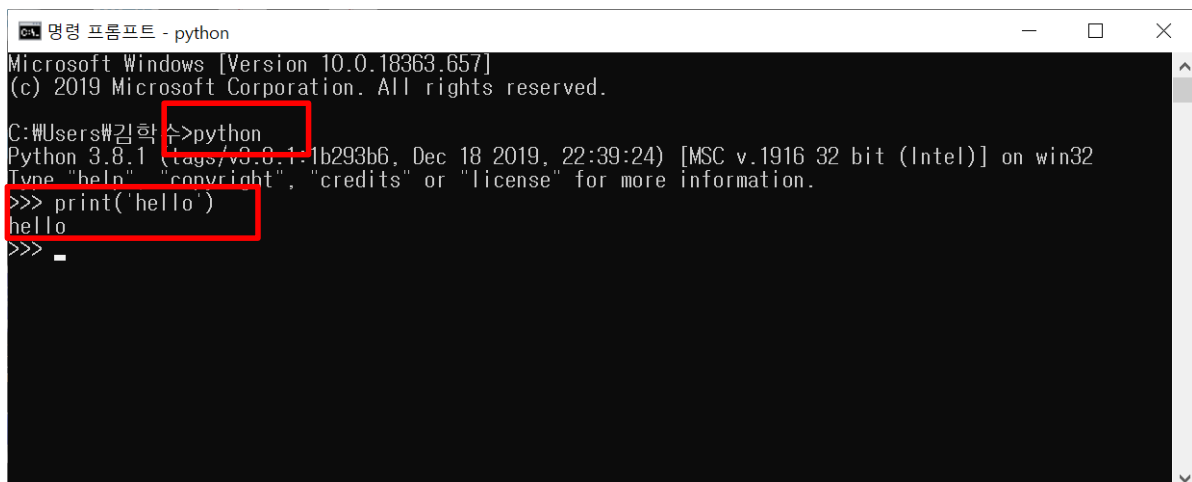


Python 설치



Edited by Harksoo Kim

Python 설치 확인



Edited by Harksoo Kim

PIP (Python Package Index)

- PIP: 파이썬으로 작성된 패키지 소프트웨어를 관리하는 패키지 관리 시스템

```
C:\Users\김수>python -m pip install --upgrade pip
Collecting pip
  Downloading https://files.pythonhosted.org/packages/54/0c/d01aa759fdc501a58f431eb594a17495f15b88da142ce14b5845662c13f3
/pip-20.0.2-py2.py3-none-any.whl (1.4MB)
    | 1.4MB 233kB/s
Installing collected packages: pip
  Found existing installation: pip 19.2.3
  Uninstalling pip-19.2.3:
    Successfully uninstalled pip-19.2.3
  Successfully installed pip-20.0.2
```



Edited by Harksoo Kim

Numpy, Scipy, Matplot 설치

- Numpy: 행렬이나 다차원 배열을 쉽게 처리 할 수 있도록 지원하는 라이브러리
- Scipy: 과학 컴퓨팅과 기술 컴퓨팅에 사용되는 라이브러리
- Matplot: 유사한 그래프 표시를 가능케 하는 라이브러리

```
명령 프롬프트
(c) 2019 Microsoft Corporation. All rights reserved.
C:\Users\김수>python -m pip install numpy scipy matplotlib
Collecting numpy
  Downloading https://files.pythonhosted.org/packages/0e/c3/be53614c4e3490778050e1df48fd463837297d5
dd402dae3b500f2050eba/numpy-1.18.1-cp38-cp38-win32.whl (10.8MB)
    | 10.8MB 1.1MB/s
Collecting scipy
  Downloading https://files.pythonhosted.org/packages/db/9e/465a416eb04114e3722b17b0f4fa5235bab8a76
961de51db0e5850183fb1/scipy-1.4.1-cp38-cp38-win32.whl (27.9MB)
    | 27.9MB 3.3MB/s
Collecting matplotlib
  Downloading https://files.pythonhosted.org/packages/86/e4/1ef1cb7f2c52345a7e3c8efd3de7ec943818c00
11c839b4880c0ba0bb7b1/matplotlib-3.1.3-cp38-cp38-win32.whl (8.9MB)
    | 8.9MB 6.8MB/s
Collecting kiwisolver<=1.0.1 (from matplotlib)
  Downloading https://files.pythonhosted.org/packages/b9/b1/118f3d5dee660bbe4548f06dcd0e1a10e454583
26c3d0efad7dbbf28be24/kiwisolver-1.1.0-cp38-cp38-win32.whl (43kB)
    | 51kB ...
```



Edited by Harksoo Kim

Numpy, Scipy, Matplot 설치

```
C:\Users\김학수>python
Python 3.8.1 (tags/v3.8.1:1b293b6, Dec 18 2019, 22:39:24) [MSC v.1916 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license()" for more information
>>> import numpy
>>> import scipy
>>> import matplotlib
>>>
```

설치확인!



Edited by Harksoo Kim

Jupyter Notebook 설치

- Jupyter Notebook: 웹 브라우저에서 파이썬 코드를 작성하고 실행해 볼 수 있는 개발도구

```
명령 프롬프트
C:\Users\김학수>python -m pip install jupyter
Collecting jupyter
  Downloading jupyter-1.0.0-py2.py3-none-any.whl (2.7 kB)
Collecting nbconvert
  Downloading nbconvert-5.6.1-py2.py3-none-any.whl (455 kB)
  | 455 kB 234 kB/s
Collecting ipywidgets
  Downloading ipywidgets-7.5.1-py2.py3-none-any.whl (121 kB)
  | 121 kB 6.4 MB/s
Collecting qtconsole
  Downloading qtconsole-4.6.0-py2.py3-none-any.whl (121 kB)
  | 121 kB 6.8 MB/s
Collecting ipykernel
  Downloading ipykernel-5.1.4-py3-none-any.whl (116 kB)
  | 116 kB 6.4 MB/s
Collecting jupyter-console
  Downloading jupyter_console-6.1.0-py2.py3-none-any.whl (21 kB)
```



Edited by Harksoo Kim

Jupyter 실행

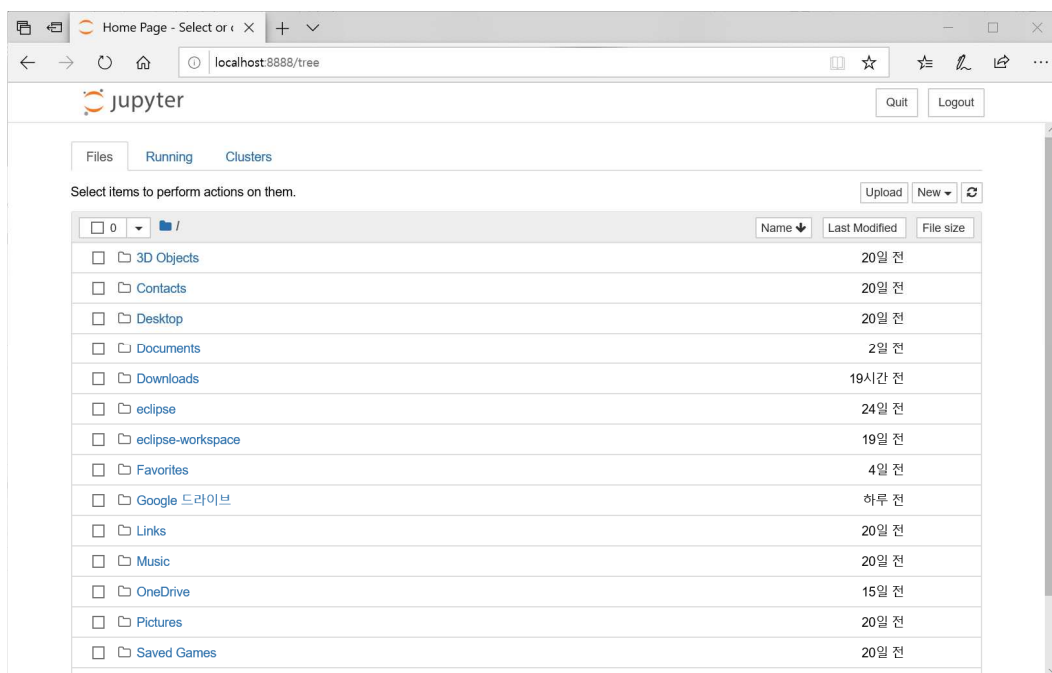
```
명령 프롬프트 - jupyter notebook
C:\Users\김학수>
C:\Users\김학수>jupyter notebook
[I 10:24:19.840 NotebookApp] Writing notebook server cookie secret to C:\Users\김학수\AppData\Roaming\jupyter\runtime\notebook_cookie_secret
[W 10:24:20.300 NotebookApp] Terminals not available (error was No module named 'winpty'.cywinpty)
[I 10:24:20.300 NotebookApp] Serving notebooks from local directory: C:\Users\김학수
[I 10:24:20.300 NotebookApp] The Jupyter Notebook is running at:
[I 10:24:20.300 NotebookApp] http://localhost:8888/?token=a3bf7025eebf7610c5973563a16494be8c39ce8082947215
[I 10:24:20.300 NotebookApp] or http://127.0.0.1:8888/?token=a3bf7025eebf7610c5973563a16494be8c39ce8082947215
[I 10:24:20.300 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 10:24:20.331 NotebookApp]

To access the notebook, open this file in a browser:
file:///C:/Users/%EA%B9%80%ED%95%99%EC%88%98/AppData/Roaming/jupyter/runtime/notebook-14056-open.html
0 copy and paste one of these URLs:
http://localhost:8888/?token=a3bf7025eebf7610c5973563a16494be8c39ce8082947215
or http://127.0.0.1:8888/?token=a3bf7025eebf7610c5973563a16494be8c39ce8082947215
[I 10:25:28.675 NotebookApp] 302 GET /?token=a3bf7025eebf7610c5973563a16494be8c39ce8082947215 (::1) 0.00ms
```



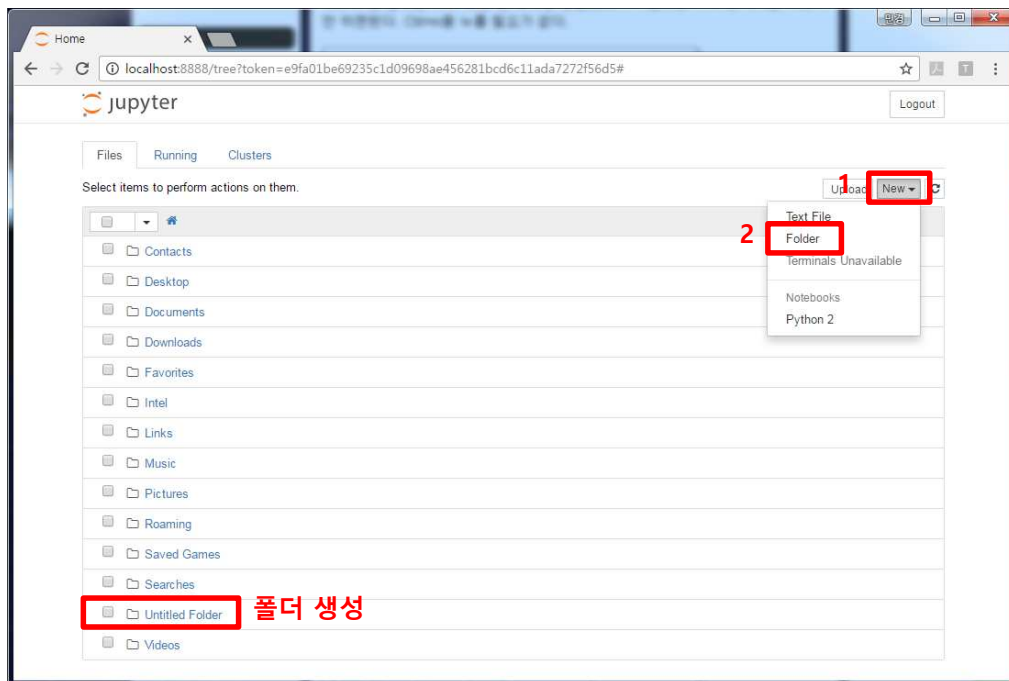
Edited by Harksoo Kim

Jupyter 실행 화면



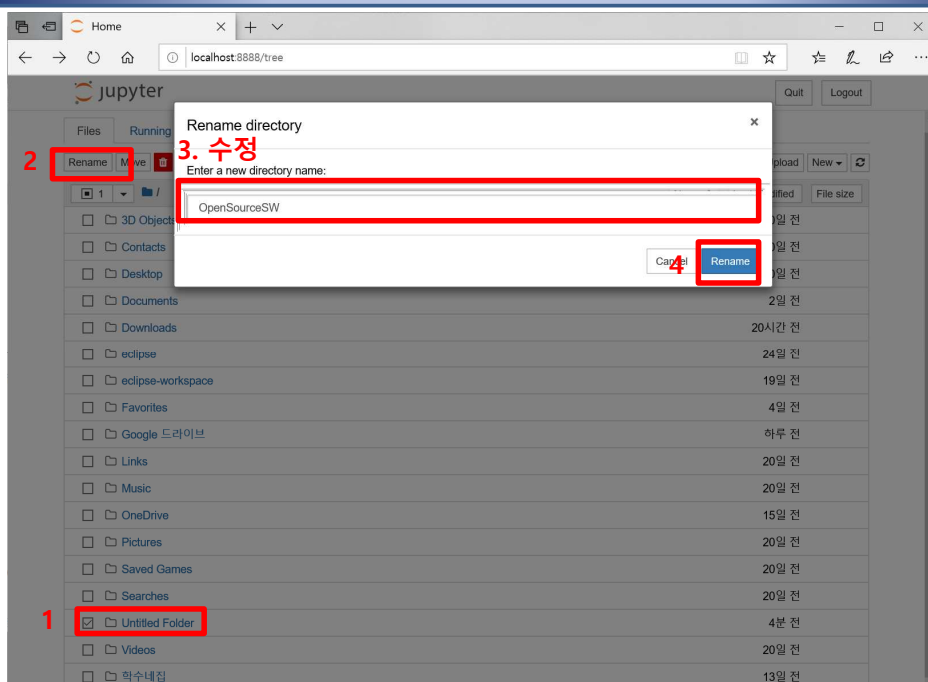
Edited by Harksoo Kim

Jupyter 폴더 생성



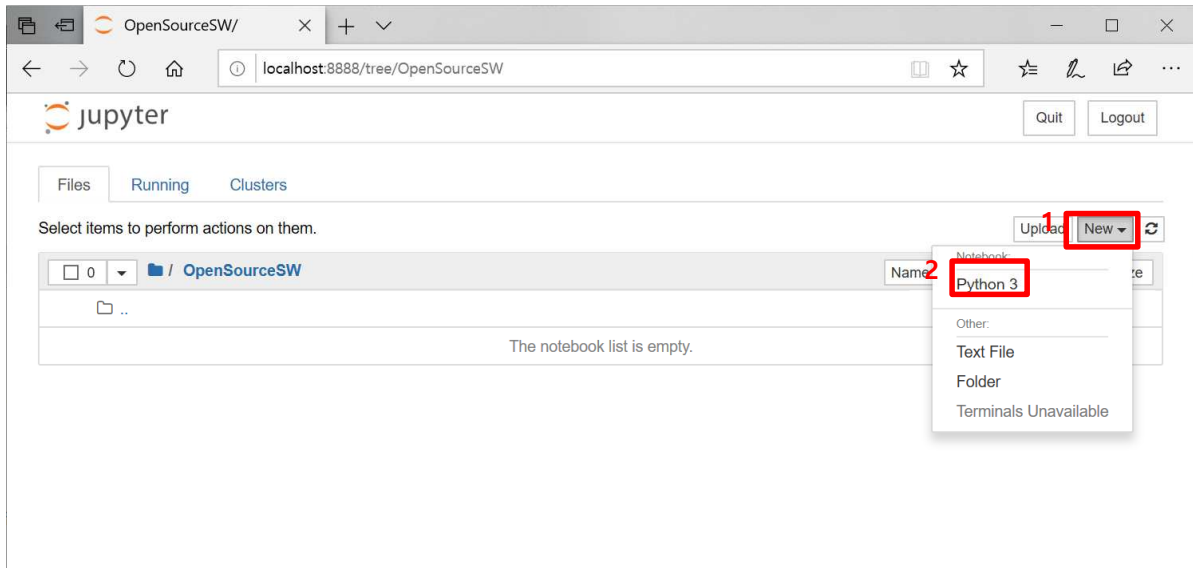
Edited by Harksoo Kim

Jupyter 폴더 이름 수정



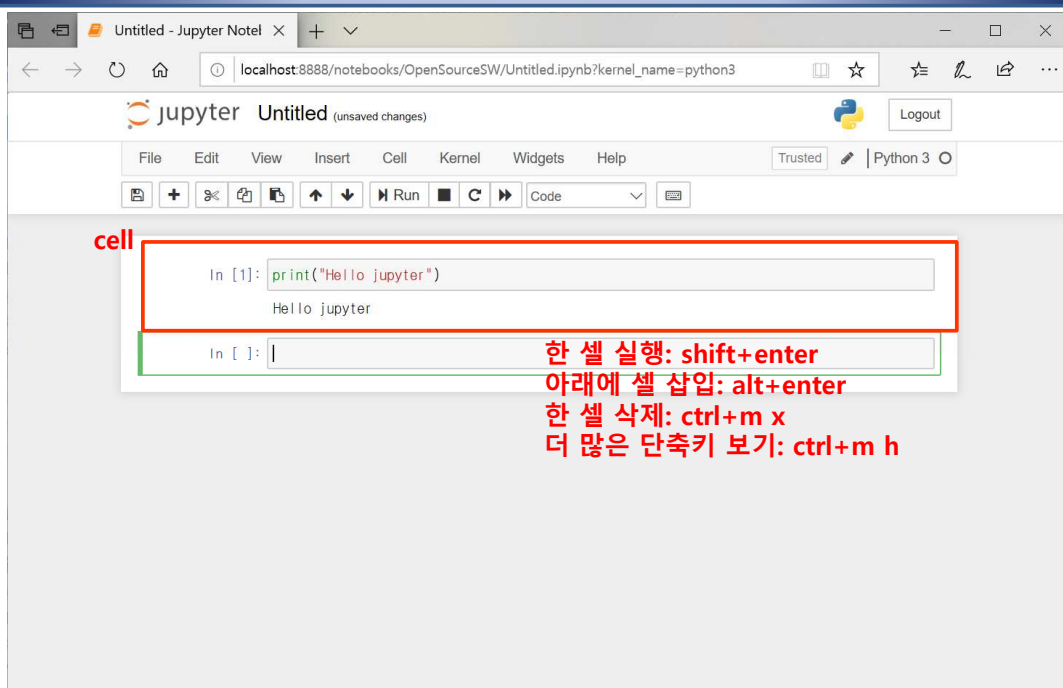
Edited by Harksoo Kim

Jupyter로 python 실행



Edited by Harksoo Kim

Jupyter로 python 코딩하기



Edited by Harksoo Kim

Scikit-learn

- scikit-learn: 파이썬으로 작성된 데이터 분석을 위한 범용 오픈 소스 라이브러리

```
명령 프롬프트
Microsoft Windows [Version 10.0.18363.657]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\Users\김학수>python -m pip install scikit-learn
Collecting scikit-learn
  Downloading scikit_learn-0.22.1-cp38-cp38-win32.whl (5.6 MB)
    |#####| 5.6 MB 6.8 MB/s
Requirement already satisfied: numpy>=1.11.0 in c:\Users\김학수\AppData\Local\Programs\Python\Python38-32\lib\site-packages (from scikit-learn) (1.18.1)
Requirement already satisfied: scipy>=0.17.0 in c:\Users\김학수\AppData\Local\Programs\Python\Python38-32\lib\site-packages (from scikit-learn) (1.4.1)
Collecting joblib>=0.11
  Downloading joblib-0.14.1-py2.py3-none-any.whl (294 kB)
    |#####| 294 kB 89 kB/s
Installing collected packages: joblib, scikit-learn
Successfully installed joblib-0.14.1 scikit-learn-0.22.1

C:\Users\김학수>python
Python 3.8.1 (tags/v3.8.1:1b293b6, Dec 18 2019, 22:39:24) [MSC v.1916 32 bit (Intel)] on
win32
Type "help()", "copyright()", "credits()" or "license()" for more information.
>>> import sklearn
>>>
```

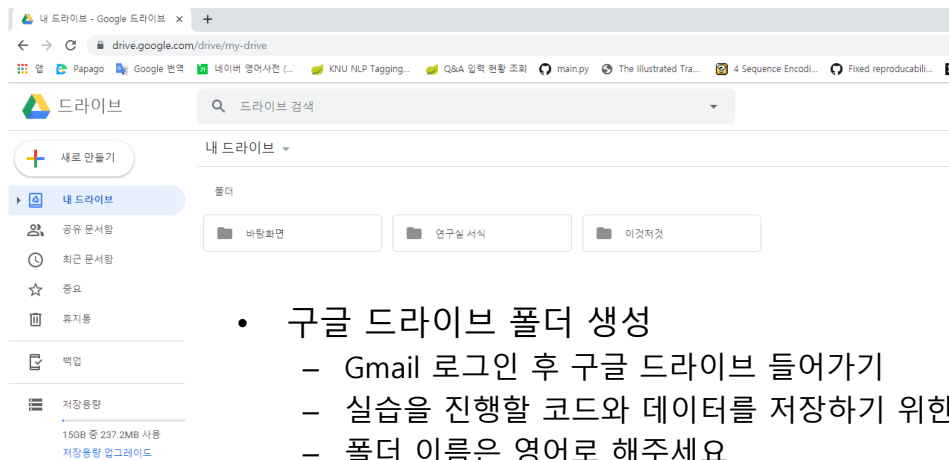
설치 확인



Edited by Harksoo Kim

Google Colab

- Google Colab: AI 개발자들을 위해 구글에서 제공하는 무료 클라우드 서비스

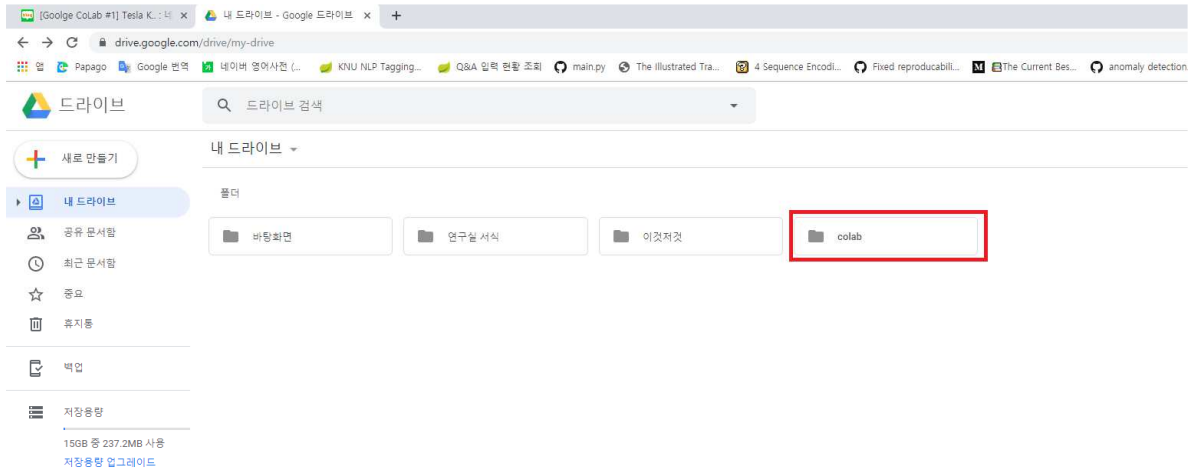


- 구글 드라이브 폴더 생성
 - Gmail 로그인 후 구글 드라이브 들어가기
 - 실습을 진행할 코드와 데이터를 저장하기 위한 폴더 생성
 - 폴더 이름은 영어로 해주세요.



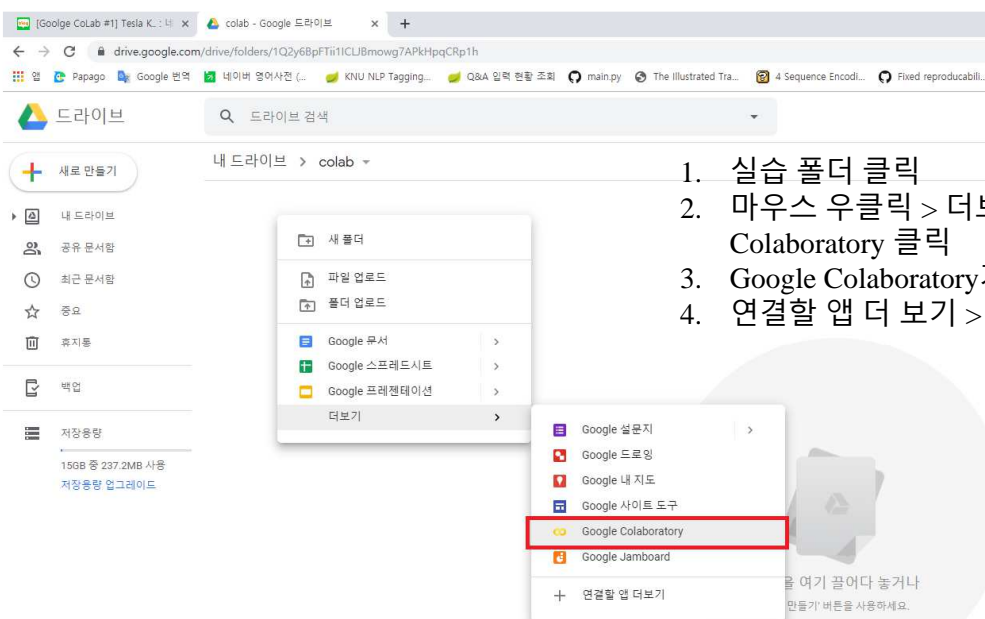
Edited by Harksoo Kim

Google Colab 설치



Edited by Harksoo Kim

Google Colab 설치

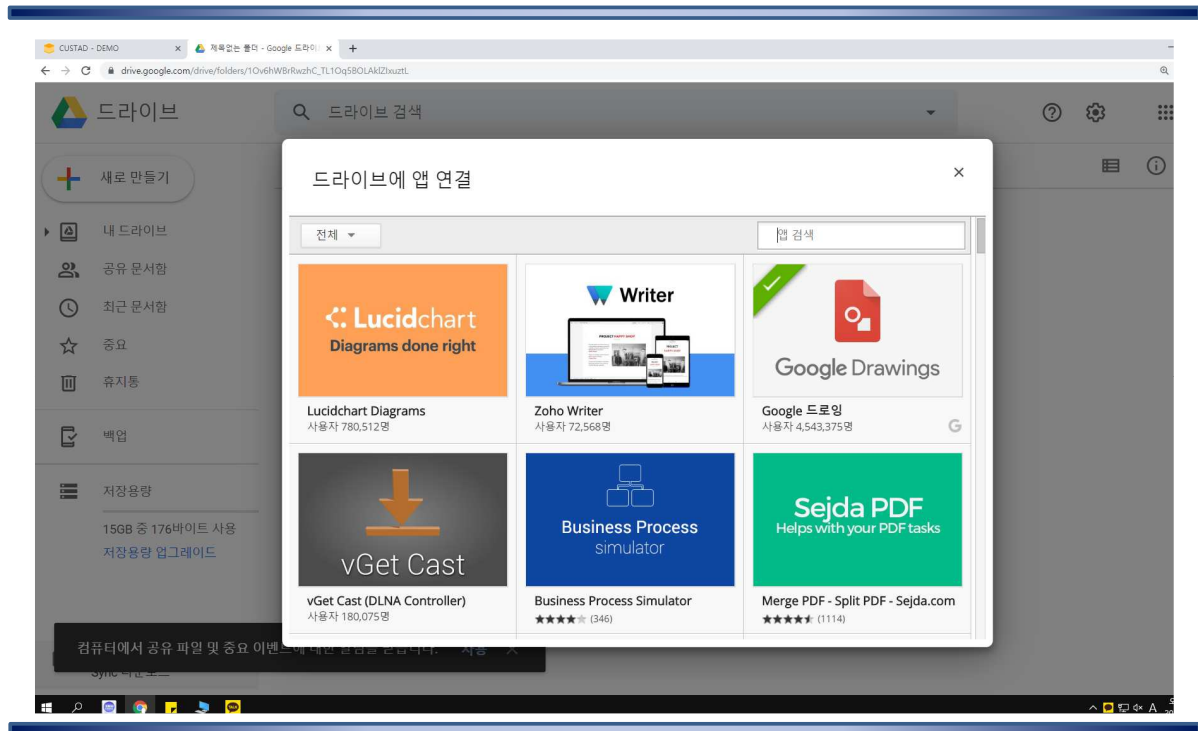


1. 실습 폴더 클릭
2. 마우스 우클릭 > 더보기 > Google Colaboratory 클릭
3. Google Colaboratory가 없는 경우
4. 연결할 앱 더 보기 > colab 검색 > 연결



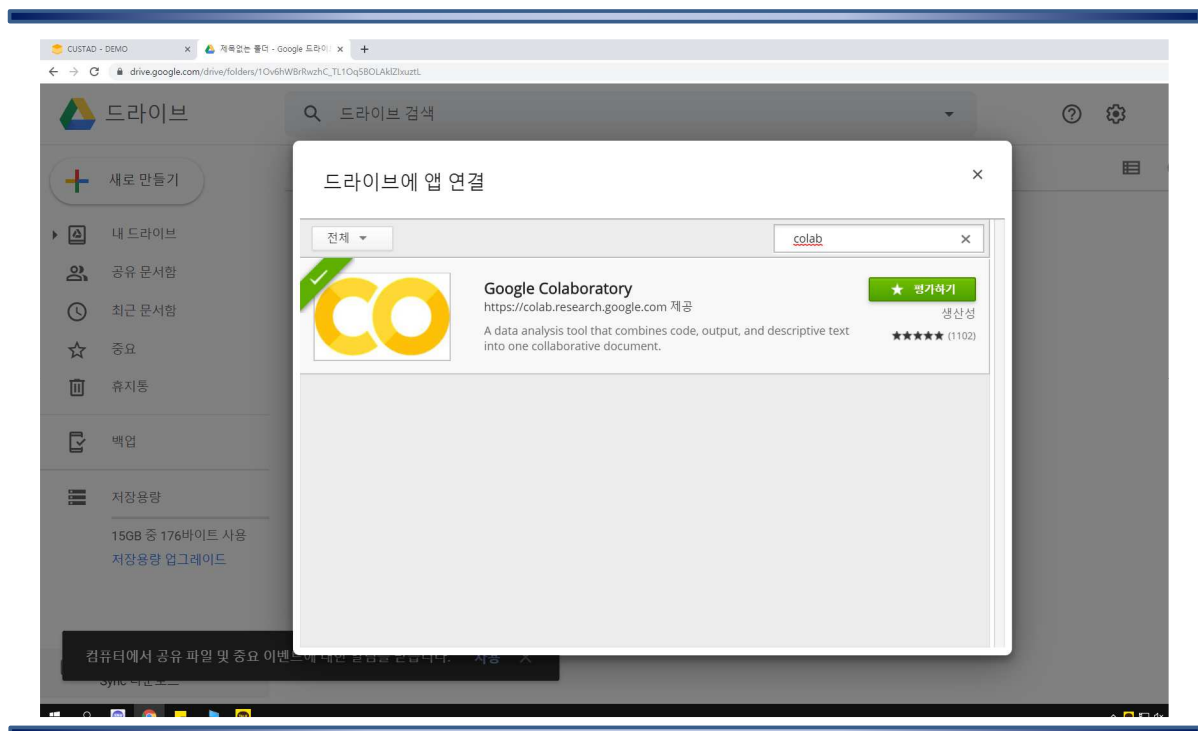
Edited by Harksoo Kim

Google Colab 설치



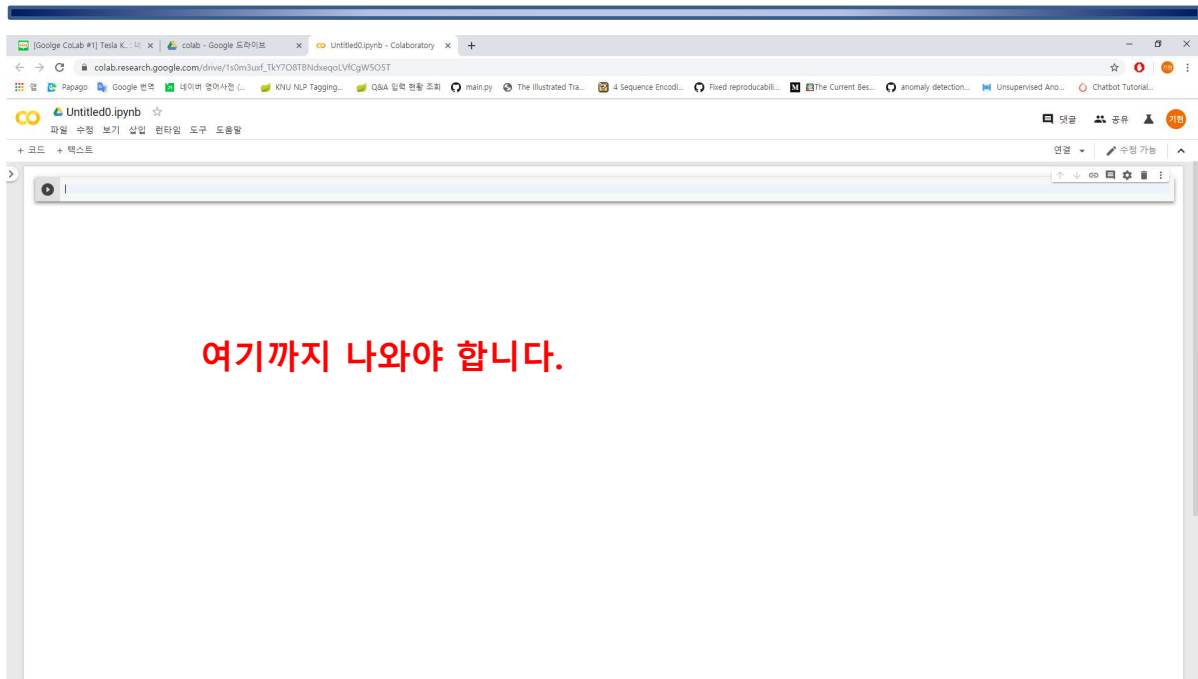
Edited by Harksoo Kim

Google Colab 설치



Edited by Harksoo Kim

Google Colab 설치



Edited by Harksoo Kim

구글 드라이브 연동

- 개인의 구글 드라이브와 코랩 클라우드를 연동
 - 코랩 클라우드에 바로 파일을 업로드하면 일정 시간 후에 해당 파일이 삭제됨
 - 따라서 구글 드라이브에 파일을 업로드 하고 이를 코랩 클라우드와 연동하여 사용



Edited by Harksoo Kim

구글 드라이브 연동

```
from google.colab import drive
drive.mount('/gdrive', force_remount=True)
```

위의 코드 입력 후 **shift+enter** 을 눌러 코드 실행

```
from google.colab import drive
drive.mount('/gdrive', force_remount=True)
```

... Go to this URL in a browser: https://accounts.google.com/o/oauth2/auth?client_id=947318989803

Enter your authorization code:

링크 선택 > 자신의 계정 선택 > 액세스 허용 > 인증 코드 복사하여 입력



Edited by Harksoo Kim

구글 드라이브 연동

```
from google.colab import drive
drive.mount('/gdrive', force_remount=True)
```

Go to this URL in a browser: https://accounts.google.com/o/oauth2/auth?client_id=947318989803-6bn6qk8qdc

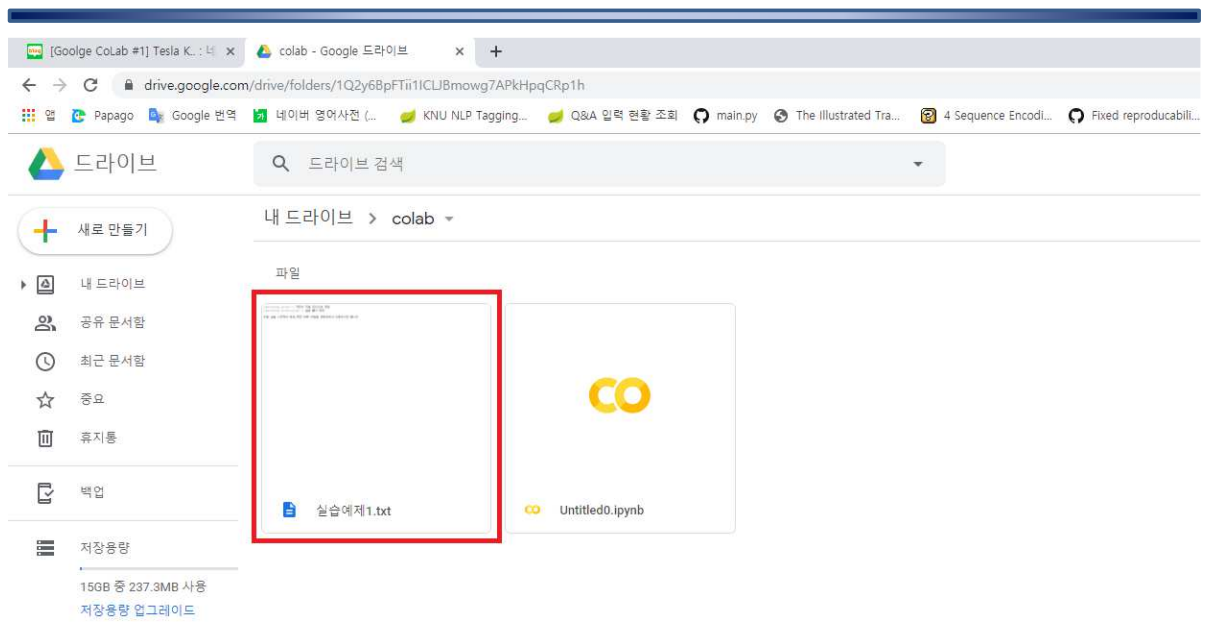
Enter your authorization code:
.....
Mounted at /gdrive

“Mounted at /gdrive” 라는 메시지가 뜨면 연동 완료



Edited by Harksoo Kim

파일 업로드

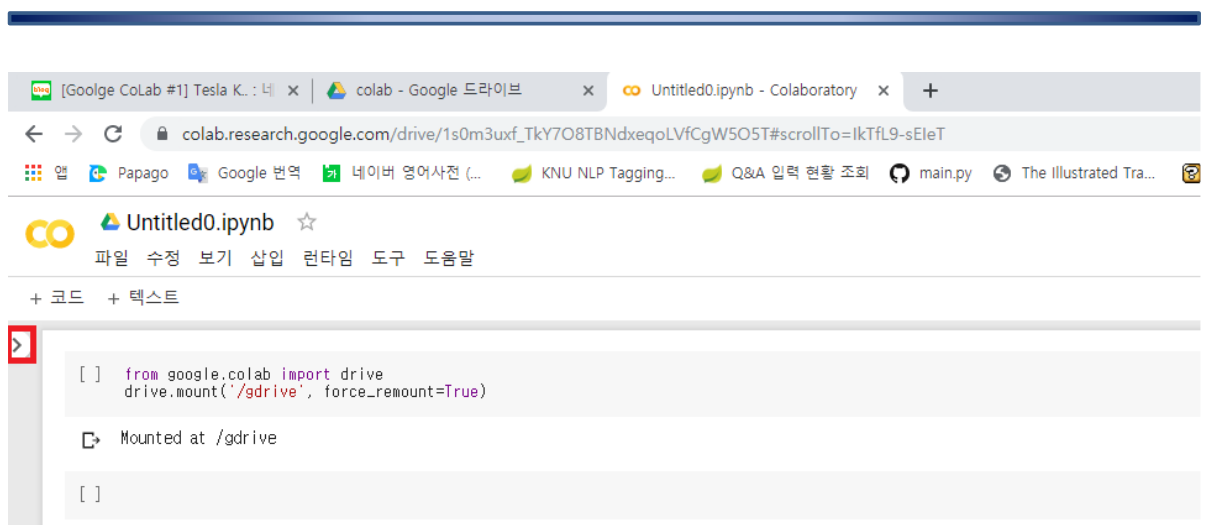


실습 폴더에 "실습예제1.txt" 파일 업로드



Edited by Harksoo Kim

파일 업로드

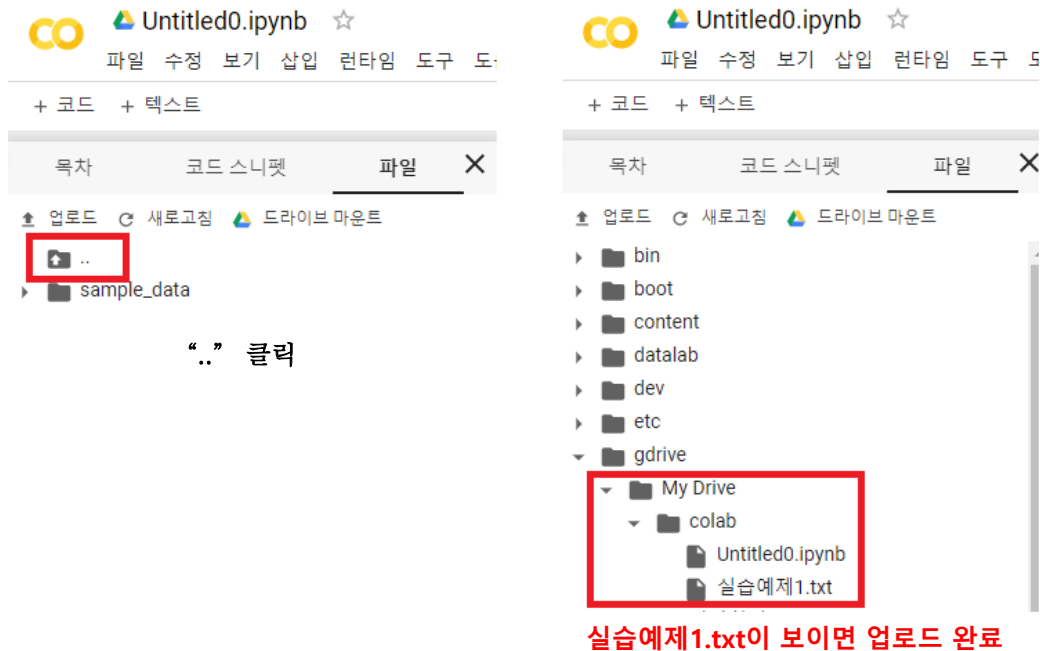


왼쪽 상단 화살표 클릭



Edited by Harksoo Kim

파일 업로드



“..” 클릭

실습예제1.txt이 보이면 업로드 완료



Edited by Harksoo Kim

업로드 한 파일 출력

```
[9] from google.colab import drive
    drive.mount('/gdrive', force_remount=True)
```

Mounted at /gdrive

```
file_path = "/gdrive/My Drive/colab/실습예제1.txt"
with open(file_path, "r", encoding="utf8") as inFile:
    lines = inFile.readlines()
    for line in lines:
        print(line)
```

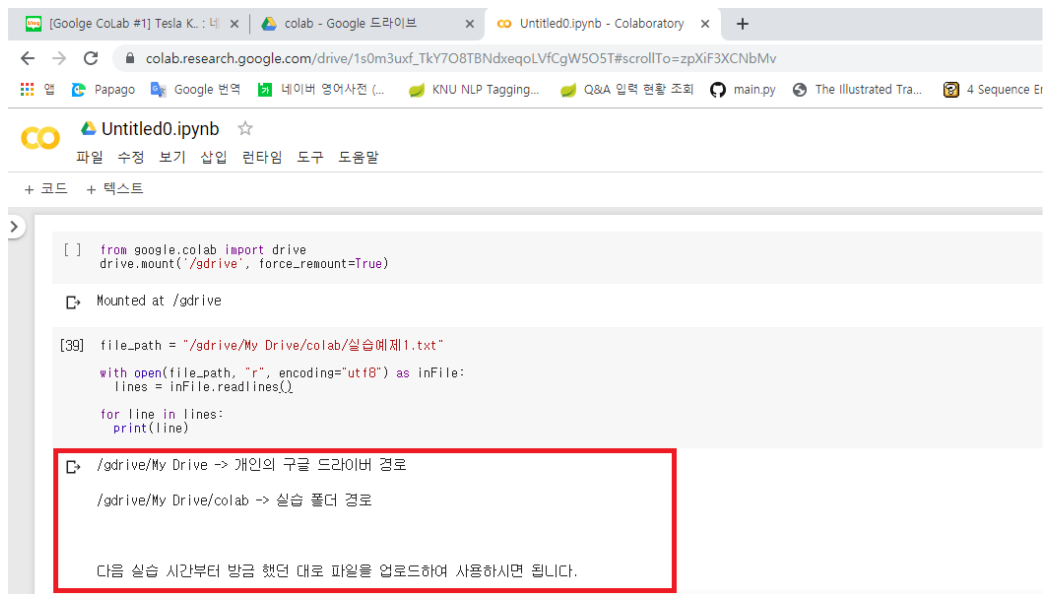
본인이 생성한 실습 폴더 이름 입력

위의 코드 입력 후 코드 실행



Edited by Harksoo Kim

업로드한 파일 출력



The screenshot shows a Google Colab notebook interface. The browser tabs at the top include 'Google CoLab #1 Tesla K...', 'colab - Google 드라이브', and 'Untitled0.ipynb - Colaboratory'. The address bar shows the URL 'colab.research.google.com/drive/1s0m3uxf_TkY7O8TBNdxqoLVfCgW5O5T#scrollTo=zxXf3XCnbMv'. The notebook title is 'Untitled0.ipynb'. Below the title, there are tabs for '파일' (File), '수정' (Edit), '보기' (View), '삽입' (Insert), '런타임' (Runtime), '도구' (Tools), and '도움말' (Help). The main area shows a code cell with the following Python code:

```
[ ] from google.colab import drive
drive.mount('/gdrive', force_remount=True)

[39] file_path = "/gdrive/My Drive/colab/실습예제1.txt"
with open(file_path, "r", encoding="utf8") as inFile:
    lines = inFile.readlines()
    for line in lines:
        print(line)
```

The output of the code cell is shown below the code:

```
Mounted at /gdrive

/gdrive/My Drive -> 개인의 구글 드라이브 경로
/gdrive/My Drive/colab -> 실습 폴더 경로

다음 실습 시간부터 방금 했던 대로 파일을 업로드하여 사용하시면 됩니다.
```

위 내용이 출력되면 성공!



Edited by Harksoo Kim



What is Machine Learning?

기계 학습

- Mitchell의 정의

- "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E .
- 어떤 컴퓨터 프로그램이 T 라는 작업을 수행한다. 이 프로그램의 성능을 P 라는 척도로 평가했을 때 경험 E 를 통해 성능이 개선된다면 이 프로그램은 학습을 한다고 말할 수 있다.



Edited by Harksoo Kim

기계 학습 문제

- 분류 (classification)

- 미리 정의된 범주에 입력 데이터를 할당하는 문제



- 군집화 (clustering)

- 미리 정의된 규칙에 따라 데이터를 그룹화하는 문제

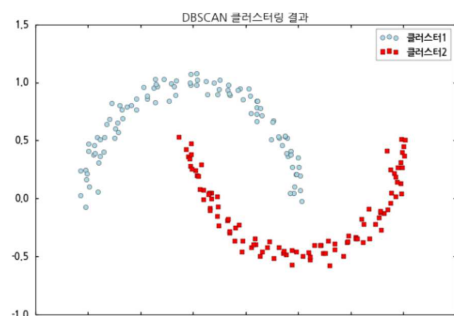


그림 출처: 블로그 참스터디-GodGo



Edited by Harksoo Kim

기계 학습 방법론

- 지도 학습 (supervised learning)
 - 정답이 부착된 데이터를 바탕으로 학습을 진행
 - 장단점
 - 비지도 학습에 비해 높은 성능을 보임
 - 데이터 구축에 많은 시간과 노력을 필요로 함
 - 데이터 구축 예
 - 기사 분류: 정치/경제/스포츠/연예 레이블
 - 영화평 예측: 댓글과 긍정, 부정 레이블 (또는 댓글의 점수)
 - 기계 독해(질의응답): 질문과 문서 내 정답 위치
 - 기계 번역: 한국어와 영어 대역 문장 쌍
 - 챗봇: 질문과 응답 문장 쌍
-



Edited by Harksoo Kim

기계 학습 방법론

- 비지도 학습 (unsupervised learning)
 - 정의된 척도(measure)에 따라 학습을 진행
 - 장단점
 - 데이터 구축이 쉬움 (정답 부착 불필요)
 - 지도 학습에 비해 낮은 성능을 보임
 - 척도의 예
 - 유클리디언 거리
 - 코사인 유사도
 - 실행 예
 - 유사한 문장/문서들을 그룹화
 - 비슷한 의미의 단어들을 그룹화
-



Edited by Harksoo Kim

기계 학습 방법론

- 반지도 학습 (semi-supervised learning)
 - 소량의 정답 부착 데이터를 바탕으로 모델(학습된 결과를 담고 있는 파일)을 만들고 대량의 정답 미부착 데이터를 활용하여 성능을 개선하는 방향으로 학습을 진행
 - 장단점
 - 데이터 구축이 지도 학습 보다 쉬움
 - 지도 학습에 비해 낮은 성능 (but, 근접한 성능을 보임)
 - 대표적인 학습법: 능동 학습(active learning)
 - (1) 정답 부착 데이터로 모델 학습
 - (2) 학습된 모델을 이용하여 정답 미부착 데이터에 정답 자동 부착
 - (3) 일정 수준 이하의 자동 부착 정답을 수정하여 데이터에 추가
 - (4) 수렴할 때까지 (1)~(3)을 반복
-



Edited by Harksoo Kim

데이터 구성

- 기계학습 데이터
 - 훈련 데이터(80%) + 개발 데이터(10%) + 평가 데이터(10%)
 - 훈련 데이터 (training data, train set)
 - 모델을 만들기 위해 사용되는 데이터
 - 개발 데이터 (developing data, dev set)
 - 학습이 잘되고 있는지 평가하는데 사용되는 데이터
 - 훈련 데이터에 속하지 않는 데이터에서도 잘 작동하는지 테스트
 - 평가 데이터 (test data, test set)
 - 최종 학습된 모델을 평가하기 위해 사용되는 데이터
-



Edited by Harksoo Kim

지도 학습 모델 개발 절차

- 데이터 수집
- 데이터 변환
 - 자질(특징) 추출 (feature extraction)
 - 정답 부착
- 모델 학습
- 모델 평가



Edited by Harksoo Kim

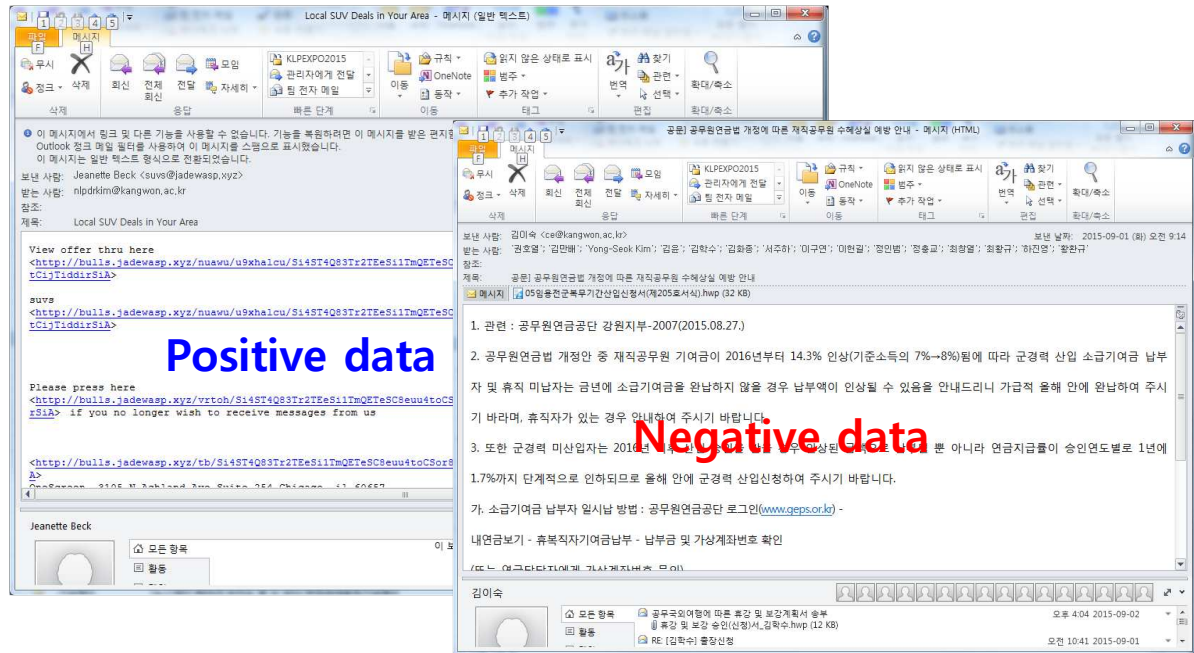
데이터 수집

- 데이터 수집 시 고려사항
 - 기계학습 대상이 되는 문제를 명확히 선정
 - 문제 해결을 위한 데이터를 가능한 많이 수집
 - 문제 대상 데이터(positive data)와 비대상 데이터(negative data)를 균형 있게 수집
- 데이터 수집 예
 - 문제
 - 스팸 메일 검출
 - 데이터
 - 스팸 메일 (positive data)
 - 비스팸 메일 (negative data)



Edited by Harksoo Kim

데이터 수집



Edited by Harksoo Kim

데이터 변환

• 자질 추출

- 주어진 문제를 해결하는 데 중요한 단서가 될 수 있는 정보를 선별하여 기계가 읽을 수 있는 형태로 변환하는 작업

Features (자질)



비정형
텍스트
데이터

	보낸 사람	제목 특수문자	제목 의심단어	내용 이미지	내용 의심단어
mail 1	Fred	8	3	true	20
mail 2	Jane	0	2	false	1
mail 3	Billy	0	0	false	3
mail 4	Fred	10	2	true	20
mail 5	Jessica	23	4	false	8
mail 6	John	0	0	false	4

Examples

Feature value (자질 값)



Edited by Harksoo Kim

데이터 변환

- 정답 부착(labeling, tagging, annotation)
 - 지도 학습을 위해서 자질 추출 데이터에 정답을 부착하는 작업
 - 매우 중요하며 시간과 노력이 많이 드는 작업

	보낸 사람	제목 특수문자	제목 의심단어	내용 이미지	내용 의심단어	정답
mail 1	Fred	8	3	true	20	O
mail 2	Jane	0	2	false	1	X
mail 3	Billy	0	0	false	3	X
mail 4	Fred	10	2	true	20	O
mail 5	Jessica	23	4	false	8	O
mail 6	John	0	0	false	4	X



Edited by Harksoo Kim

데이터 변환

- 대표적인 기계학습 데이터 형식
 - ARFF (Attribute-Relation File Format): WEKA라는 기계학습 툴킷(toolkit)에서 사용하는 데이터 포맷

```

@relation credit
@attribute checking_balance {1-200DM,<0DM,>200DM,unknown}
@attribute months_loan_duration numeric
@attribute credit_history {poor,perfect,good,critical,verygood}
@attribute phone {yes,no}
@attribute default {yes,no}

@data
<0DM,6,critical,yes,no
unknown,10,poor,no,yes
.....
    
```

문제: credit (자질 명칭)

numeric (자질 값 형식)

{yes,no} (정답 형식)

데이터

요즘에는 헤더(relation, attribute) 없이 데이터만으로 구성하는 경우가 빈번함!



Edited by Harksoo Kim

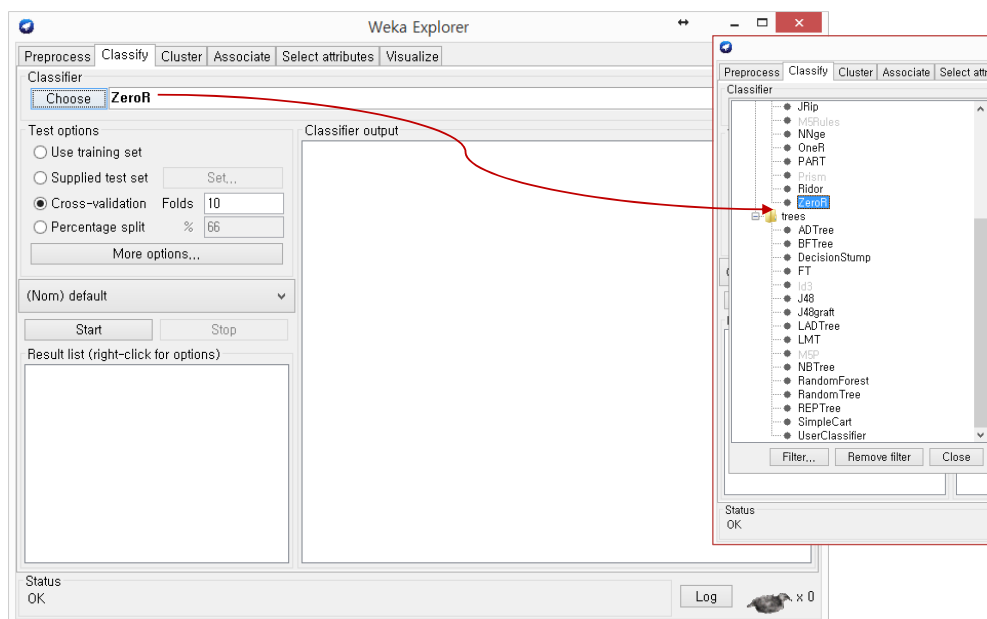
모델 학습

- 주어진 문제에 적합한 모델 선택
 - Decision Tree
 - Support Vector Machine
 - Conditional Random Fields
 - Artificial Neural Network
- 다양한 툴킷을 이용하여 모델 구성 및 학습
 - GUI 기반 기계학습: WEKA
 - 전통적인 기계학습: Scikit-learn
 - 딥러닝(deep learning): Tensorflow, Pytorch



Edited by Harksoo Kim

모델 학습



WEKA(Waikato Environment for Knowledge Analysis)



Edited by Harksoo Kim

모델 평가

- 평가 종류

- Closed test

- 학습 데이터를 이용한 평가
 - 학습이 되고 있는지 단순 확인하는 과정

- Development test

- 학습에 참여하지 않은 데이터에서 잘 작동하는지, 올바른 방향으로 학습되고 있는지를 확인하는 과정

- Open test

- 학습에 참여하지 않은 데이터를 이용한 평가
 - 실제 환경에서 어느 정도 성능이 나오는지 확인하는 과정



Edited by Harksoo Kim

모델 평가

- 10배 교차검증 (10-fold cross validation)

- 데이터를 10등분
 - 9개로 학습하고 나머지 1개로 평가
 - 상기 평가를 10회 시행
 - 10회 평가의 평균 값을 모델의 성능으로 사용

- 평가 척도 (evaluation measure)

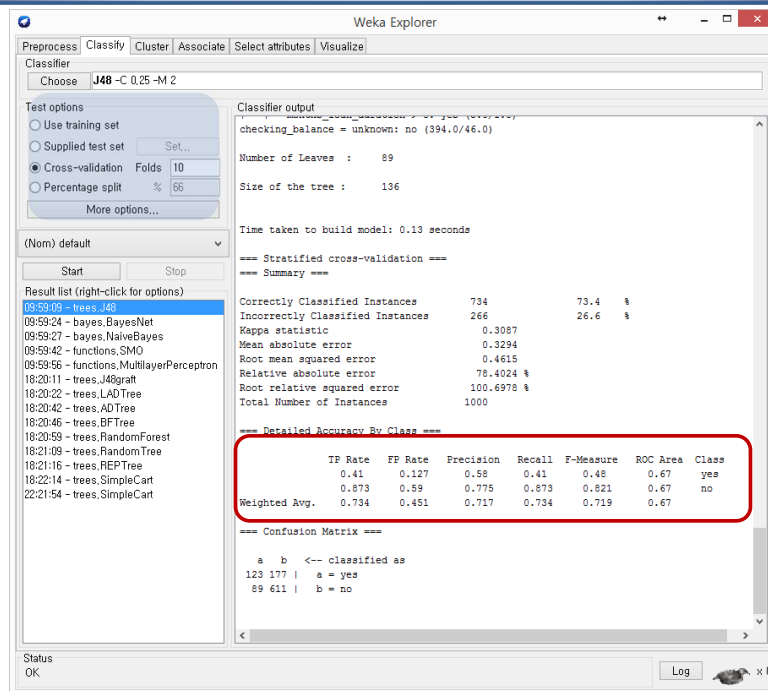
- 정밀도 (accuracy): $(TP+TN)/(TP+FP+FN+TN)$
 - 범주(레이블)에 상관없이 모델의 출력들 중에 맞은 것의 비율
 - 정확률 (precision): $TP/(TP+FP)$
 - 범주 별(레이블 별) 모델의 출력 중에 맞은 것의 비율
 - 재현율 (recall): $TP/(TP+FN)$
 - 범주 별(레이블 별) 정답 중에 모델이 맞춘 것의 비율
 - F1-score: 정확률과 재현율의 조화평균
 - $(2 * precision * recall) / (precision + recall)$

	정답 (1)	정답 (0)
예측 (1)	TP	FP
예측 (0)	FN	TN



Edited by Harksoo Kim

모델 평가



Edited by Harksoo Kim

질의응답

Q&A

Homepage: <http://nlp.konkuk.ac.kr>
E-mail: nlpdrkim@konkuk.ac.kr



Edited by Harksoo Kim