Python 데이터 분석

2016년도 2학기 2학년 방과후학교

빅데이터 분석

‘빅데이터’는 기존의 방법으로 처리하기 힘든 초대용량 데이터로, 보통은 수백 테라(tera) 바이트에서 수 페타(peta) 바이트 이상의 대용량을 의미합니다. 빅 데이터의 정의는 매우 다양하지만, 통상 3V로 설명됩니다. 처리 대상 데이터가 매우 큰 크기(Volume), 매우 빠른 속도(Velocity) 그리고 매우 다양한 유형(Variety)의 특징을 가졌을 때, 우리는 이를 ‘빅데이터’라 얘기 합니다.

특히, 빅 데이터의 대부분은 텍스트와 이미지 등의 비정형성을 가지고 있습니다. ‘빅 데이터’는 단순히 그 규모성뿐 아니라, 데이터가 매우 빠르게 변화, 전파되고, 파편화 되어 가기 때문에 그 전체를 이해하기 힘들며, 수 많은 노이즈 속에서 중요한 패턴을 발견하기란 점점 더 어려워지고 있습니다.

빅데이터 분석 기술 (1/2)

자연언어처리(NLP)
자연언어처리는 글로 된 인간 언어를 컴퓨터를 통해 처리하기 위한 기술입니다. 형태소 분석, 구문 분석, 개체명 인식 등의 기술을 포함합니다.

정보 검색(IR)
빅데이터 처리를 위해서는 정보 검색이 필수입니다. 대규모 데이터를 색인하고, 이 중에서 주제와 관련된 데이터를 빠르게 찾아 분석에 적용합니다. 기존의 검색과 다른 점은 기존 검색은 인간을 위한 정보 검색이라면, 빅 데이터 분석에서의 정보 검색은 컴퓨터가 검색 시스템을 사용하는 수요자라는 것입니다.

정보 수집(Crawling)
기존의 웹 검색을 위한 수집기보다 매우 발전된 정보 수집 기술이 필요합니다. 예를 들어, 트위터와 같은 소셜 미디어의 실시간 수집은 스트림 데이터에 대한 처리를 필요로 하며, AJAX, 자바스크립트 처리가 가능한 포커스드 크롤링(focused crawling) 기술이 적용 되어야 합니다.

기계 학습(Machine Learning)
빅 데이터 분석에서의 영웅 중 하나는 바로 기계학습입니다. 기계 학습은 충분한 학습 데이터로부터 모델을 생성하고, 해당 모델을 통해 대용량 데이터를 자동 분석, 귀납 추론하는 시스템을 의미합니다. 통상 SVMrhk 같은 통계 이론에 기반하며, 자동 분류, 자동 군집, 베이지안 네트워크 기반 추론 등 강력한 데이터 분석 기능을 제공합니다.

빅데이터 분석 기술 (2/2)

텍스트 마이닝(Text mining)
대규모 텍스트 말뭉치로부터 의미 있는 정보를 추출, 분석하는 기술입니다. 기계 학습 기반의 통계적 방법과 규칙 기반의 방법이 있으며, 최근에는 이들이 하이브리드 형태로 결합되어 사용됩니다. 기존의 분류, 군집 기능 외에 감성(평판) 분석과 같은 기능 구현에 텍스트 마이닝은 필수적입니다.

클라우드 컴퓨팅과 NoSQL
초 대용량 데이터의 저장과 관리, 운영을 위해서는 클라우드 컴퓨팅 기술이 기본이며, 특히 Hadoop, HBase, Cassandra, MongoDB와 같은 NoSQL 기술이 적절히 활용되어야 합니다.

시맨틱(Semantic) 기술
심층 분석을 위해서는 데이터에 대한 의미적 분석이 매우 중요합니다. 시맨틱 기술은 시맨틱 메타데이터 자동 추출, 시맨틱 네트워크 생성, 지식 베이스 구축, 온톨로지의 활용, 논리 및 통계적 추론 등을 포함합니다. 시맨틱 기술은 비정형 데이터와 정형 데이터를 의미적으로 연결하고, 분석하기 위한 핵심이며, 왓슨 컴퓨터, 애플의 시리, 울프람 알파 등이 이런 사실을 증명하고 있습니다

통계 기술
빅 데이터의 통계적 의미를 찾고, 그 패턴을 분석하기 위해서 강력한 통계 기능을 필요로 합니다. 통계 패키지인 R은 이런 의미에서 매우 활용도가 높습니다. Hadoop 상에서 R을 사용함으로써 과거에 생각하기 힘든 규모의 데이터에 대한 통계 처리가 가능하게 되었습니다.

시각화(Visualization)
분석된 결과의 통찰력 있는 이해를 위해 데이터 시각화는 점점 더 중요해지고 있습니다. 그래프와 같이 기본적인 시각화로부터 네트워크 표현, 지도와의 매시업 등 그 표현 유형과 분석가와의 상호작용 기술이 빠르게 발전하고 있습니다.

데이터 분석을 위한 도구

IPython

- 대화형컴퓨팅으로 분석프로그래밍에 적합함
- 운영체제의 쉘파일 시스템과 통합되어있음
- 웹기반의 대화형 노트북지원으로 수식, 표, 그림 등을 표현가능
- 가볍고 빠른 병렬컴퓨팅 엔진이용
- 코딩과 문서화, 테스트까지 한 화면에OK

데이터 분석을 위한 도구

Python + IPython + 과학 계산용 각종 Package와 그것들의 Dependencies를 묶어 배포판으로 제공

- Anaconda (https://www.continuum.io/downloads)
- Canopy (https://store.enthought.com/downloads)

For new users who want to install a full Python environment for scientific computing and data science, we suggest installing the Anaconda or Canopy Python distributions, which provide Python, IPython and all of its dependences as well as a complete set of open source packages for scientific computing and data science.

1. Download and install Continuum's Anaconda or the free edition of Enthought's Canopy.
2. Update IPython to the current version using the Terminal:

Anaconda:

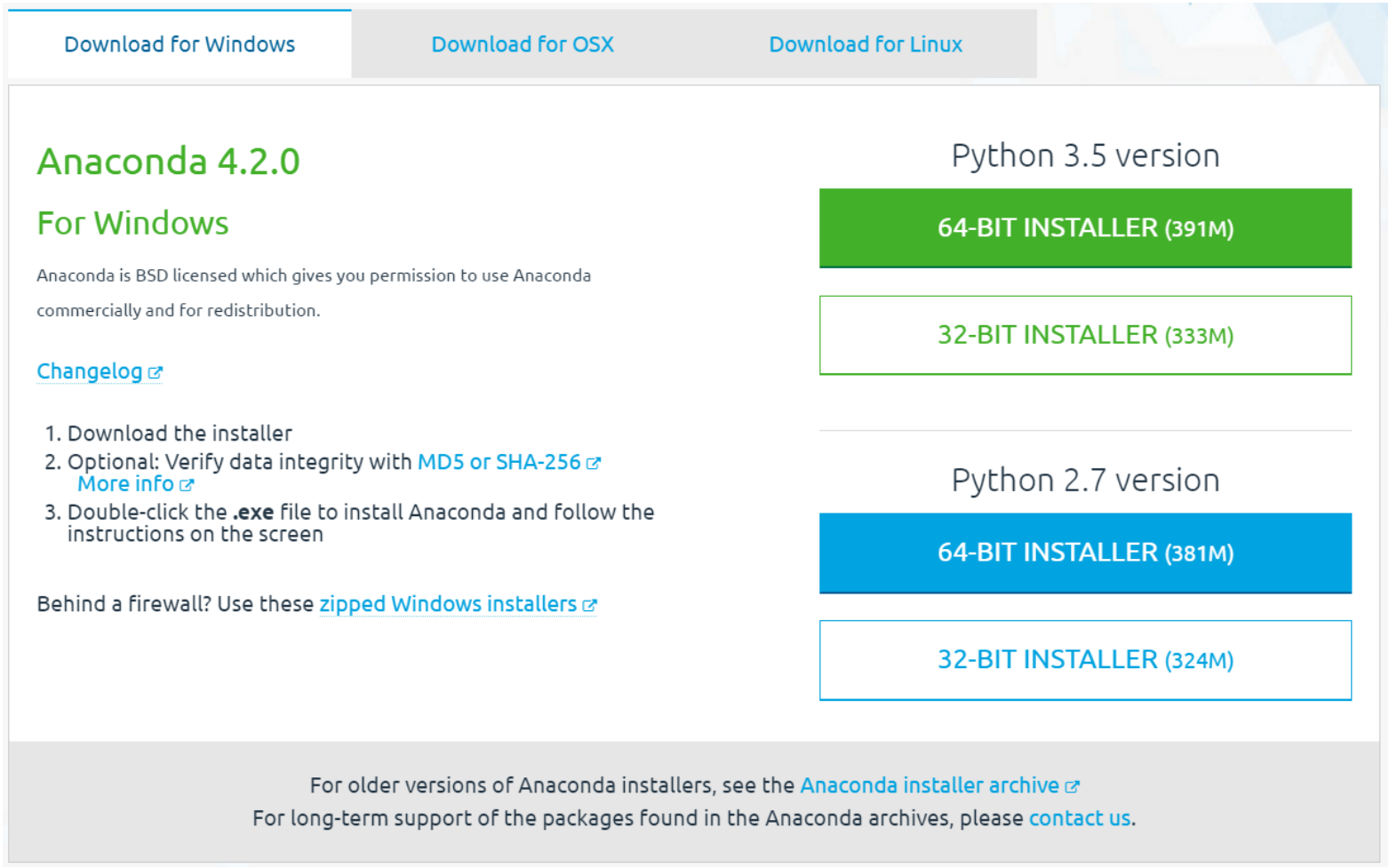
```
conda update conda
conda update ipython
```

Enthought Canopy:

```
enpkg ipython
```

Anaconda download & install

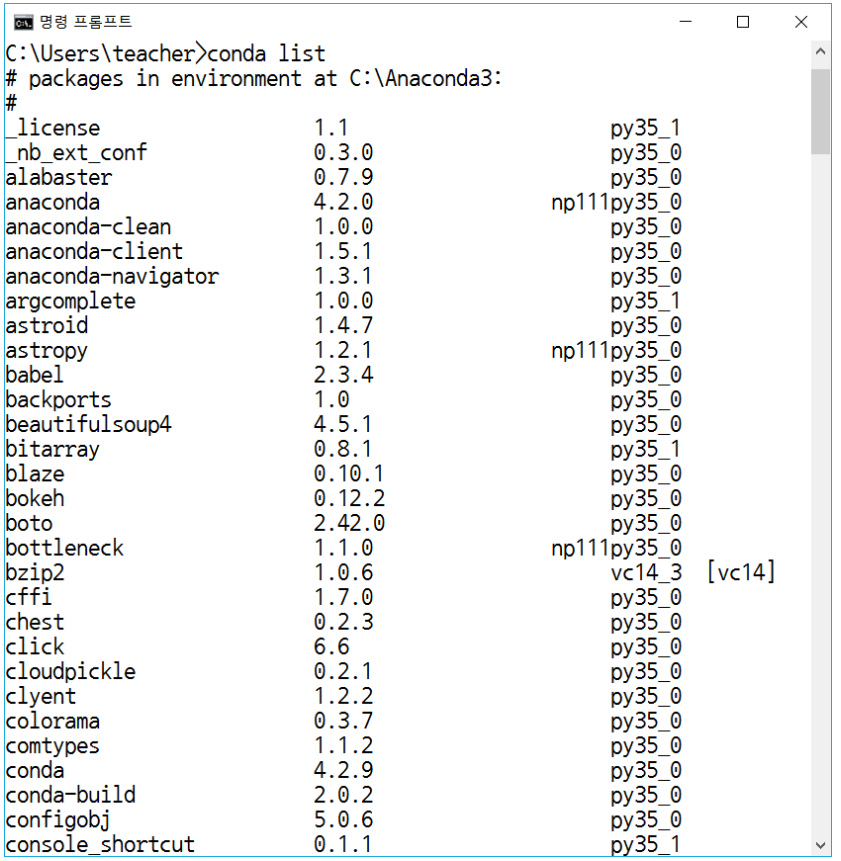
https://www.continuum.io/downloads에서 OS에 맞는 파일 download 하여 디폴트로 설치한다.



Anaconda Install 확인

1. Install된 packages 확인

> conda list

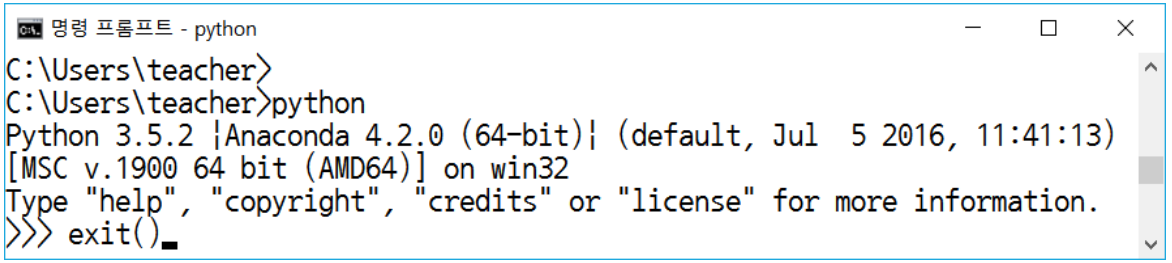


2. Python shell에서 information 확인

> python

Anaconda Information 정보 출력

>>>



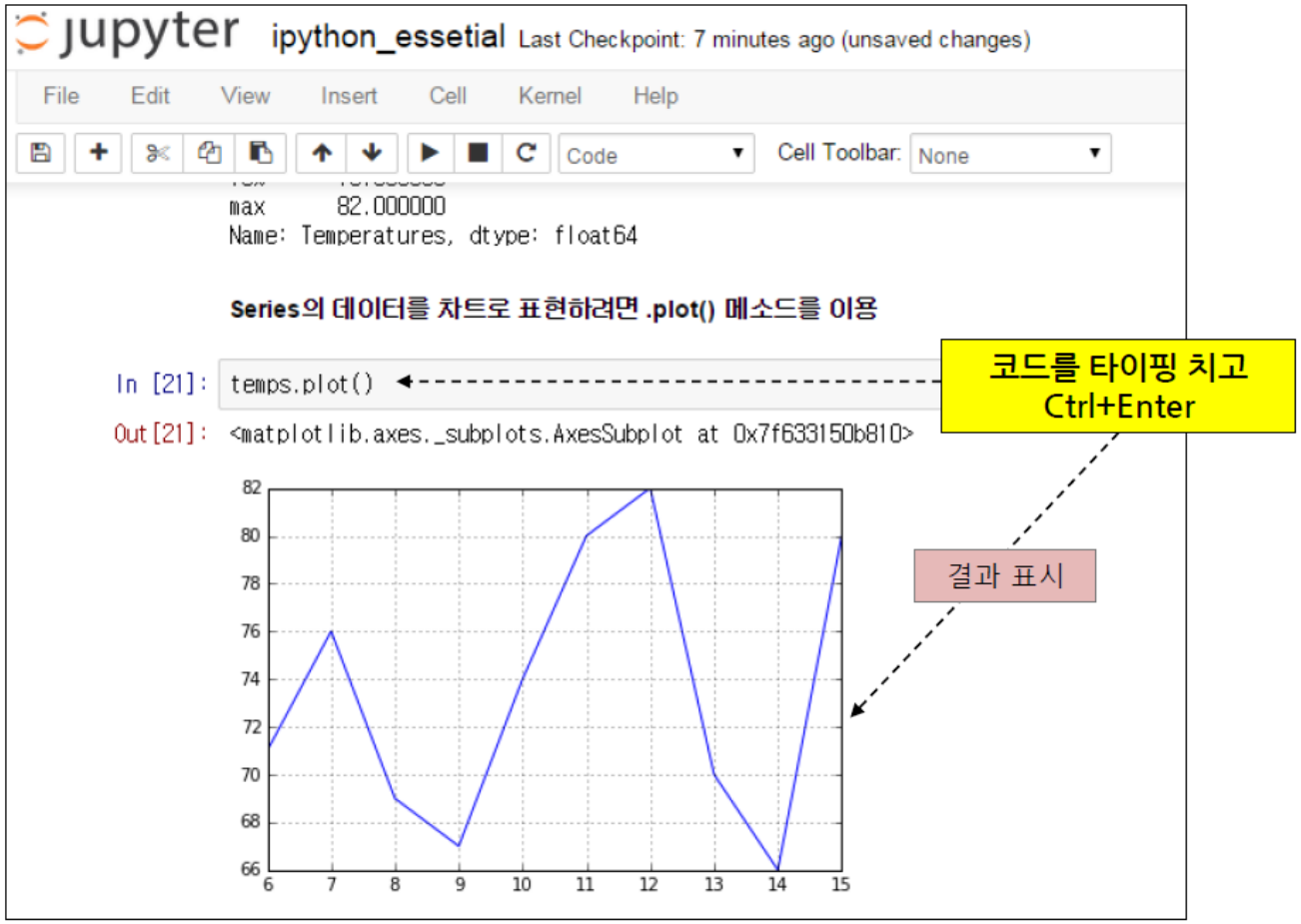
IPython

명령 프롬프트

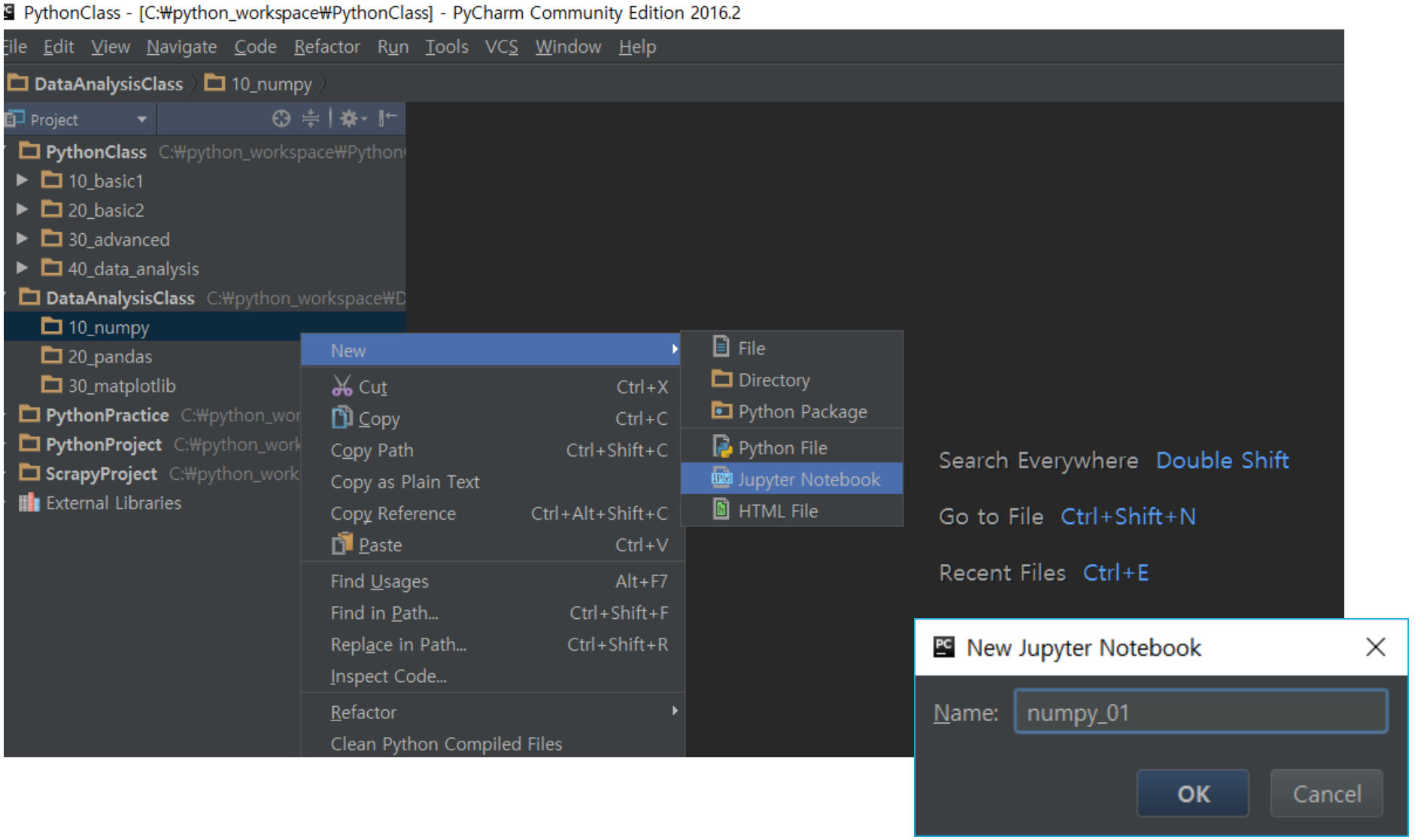
```
C:\Users\teacher>ipython
Python 3.5.2 |Anaconda 4.2.0 (64-bit)| (default, Jul  5 2016, 11:41:13) [MSC v.1900 64 bit
(AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 5.1.0 -- An enhanced Interactive Python.
?         -> Introduction and overview of IPython's features.
%quickref -> Quick reference.
help      -> Python's own help system.
object?   -> Details about 'object', use 'object??' for extra details.

In [1]: import numpy as np

In [2]: arr = np.array([[1,2,3],[4,5,6]])

In [3]: arr
Out[3]:
array([[1, 2, 3],
       [4, 5, 6]])

In [4]: arr + 10
Out[4]:
array([[11, 12, 13],
       [14, 15, 16]])

In [5]: arr * arr
Out[5]:
array([[ 1,  4,  9],
       [16, 25, 36]])

In [6]: exit()

C:\Users\teacher>_
```

IPython Notebook (Jupyter)

그래서 나온 것이 IPython Notebook!! 웹기반의 IPython 개발 도구로 자체 웹서버에 의해 동작되며 개발 시 편리한 기능들을 제공하고 있음



Jupyter Notebook 파일 생성



Jupyter Notebook 파일 실행