CHAPTER

01

빅데이터 프로그래밍의 개요

Introduction to Big Data Programming

컴퓨터소프트웨어공학과
김대영

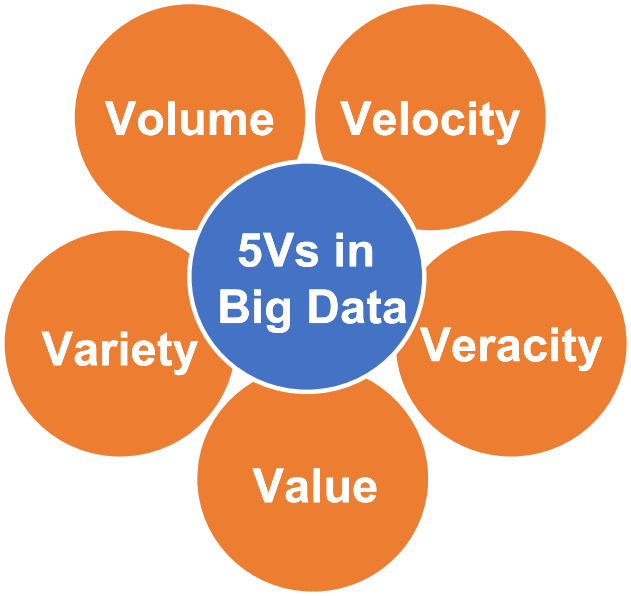
개 요

- 빅데이터란 무엇인가?
- 빅데이터 프로그래밍 소개
- 파이썬 개발 환경

빅데이터란

- **빅데이터**(Big Data)
  - 복잡하고 거대한 양의 데이터를 의미함
  - 전통적인 시스템으로부터 발생하는 데이터를 의미하지 않음
  - 사람, 기계, 자연 등에서 발생하는 다양한 종류의 데이터
  - 기술과 서비스의 발전을 통해 다양한 데이터 소스로부터 정형 및 비정형 데이터가 발생
  - 빅데이터의 분석을 통해서 서비스 사용자 및 시스템의 행동 패턴(Behavioral Pattern)과 의도(Intent)를 찾을 수 있음

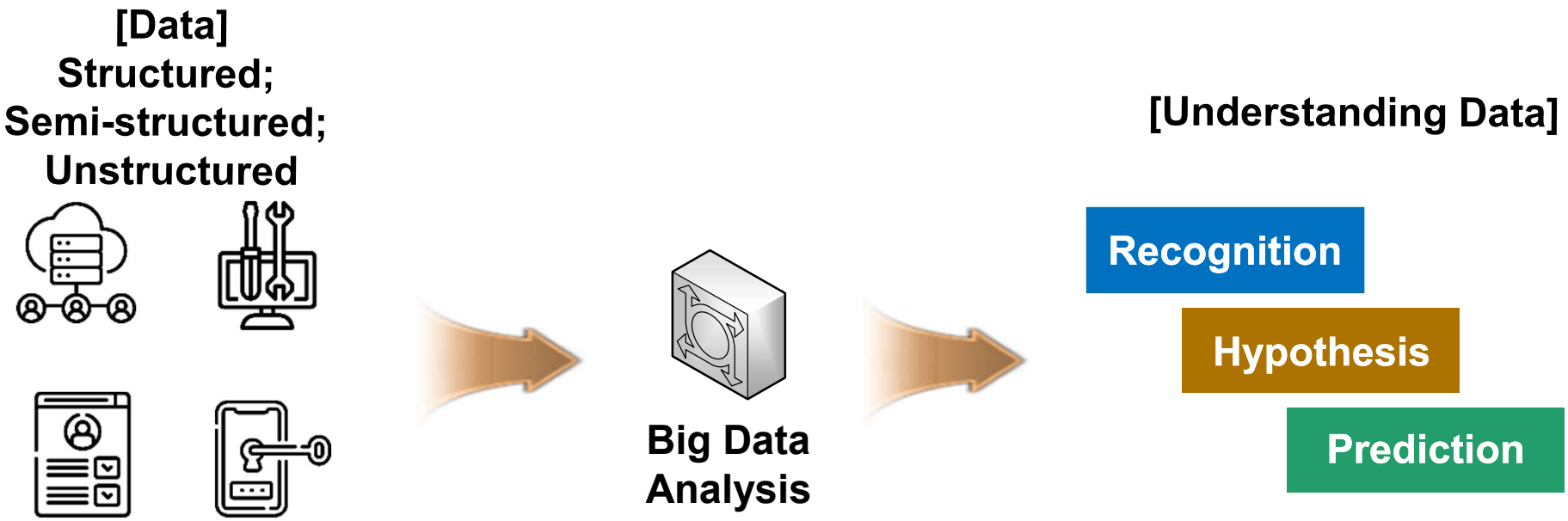
- **빅데이터의 특징**(Characteristics)
  - 빅데이터는 많은 양의 데이터로 발생 속도가 빠르고 다양한 형태의 데이터가 존재한다.
  - 핵심 개념 Key concepts (5Vs)
    - 데이터의 양(Volume)
    - 데이터의 발생 속도(Velocity)
    - 데이터의 종류(Variety)
    - 데이터의 정확성(Veracity)
    - 데이터가 가지는 의미(Value)



- **Key Concepts** (5V)
  - 데이터의 양(Volume)
    - 로그 형태로 많은 양의 데이터가 수집되고, 이러한 데이터를 처리하는 능력이 빅데이터 분석의 주요 관심분야임
    - 빅데이터 분석을 수십 테라바이트 이상의 데이터가 활용될 수 있음
  - 데이터의 발생 속도(Velocity)
    - 데이터의 발생 속도는 데이터가 얼마나 빨리 증식되는지를 나타냄
    - 빠른 속도로 증가하는 데이터를 다루기 위해서는 데이터의 수집, 저장, 처리, 그리고 분석이 짧은 시간 내에 이루어져야 함

- **Key Concepts** (5V)
  - 데이터의 종류(Variety)
    - 가능한 여러 종류의 다양한 데이터가 다루어 짐
    - 비정형 데이터의 발생이 증가하고 있음
  - 데이터의 정확성(Veracity)
    - 3V의 다양한 데이터를 다룰 때, 모든 데이터가 정확한 것은 아님
    - 데이터의 정확성은 상황에 따라 달라질 수 있으며, 데이터의 정확성은 데이터의 분석에 영향을 미침
  - 데이터가 가지는 의미(Value)
    - 빅데이터에서 가장 중요한 부분이며, 잠재적 가치를 통해 비즈니스 로직 구현

- **빅데이터 분석**
  - 정형 및 비정형 데이터로부터 의미있는 가지를 추출하는 행위
  - 빅데이터에서 가치를 추출하는 방법은 데이터 분석을 통해서 패턴 인식, 가설 설정, 행위 예측을 수행하는 것임



- **프로그래밍 언어로써의 파이썬(Python)**
  - 빠르고 단순한 프로그래밍 언어
    - 스크립트 언어로써 쉽고 빠르게 데이터 분석에 활용할 수 있음
  - 다양한 라이브러리의 지원
    - 과학계산, 데이터 분석, 데이터 시각화를 위한 다양한 라이브러리를 지원함
  - 범용 프로그래밍 언어
    - 다양한 파이썬 라이브러리 지원을 통해 범용적 애플리케이션 개발에 활용됨
    - C, C++ 등의 코드와 통합이 가능하고, 기존 C 라이브러리를 공유할 수 있음
    - 인터프리터 언어이지만 하드웨어의 발전을 통해 연산시간을 대폭 줄임

• 데이터 분석을 위한 파이썬 라이브러리

- NumPy (Numerical Python)
  - 산술 데이터 처리를 위한 자료구조와 함수(알고리즘) 제공
  - 배열기반의 데이터 연산을 위한 자료구조와 함수 제공
  - 대부분의 과학 계산을 위한 함수(알고리즘) 제공

- Pandas
  - 구조화된 테이블 방식의 데이터를 효율적으로 다루기 위한 자료구조와 함수 (알고리즘) 제공
  - NumPy의 배열 연산과 테이블 방식의 관계형 데이터베이스의 데이터 처리를 위한 기능을 결합
  - 대표적 자료 구조: DataFrame (2차원 배열 객체), Series (1차원 배열 객체)

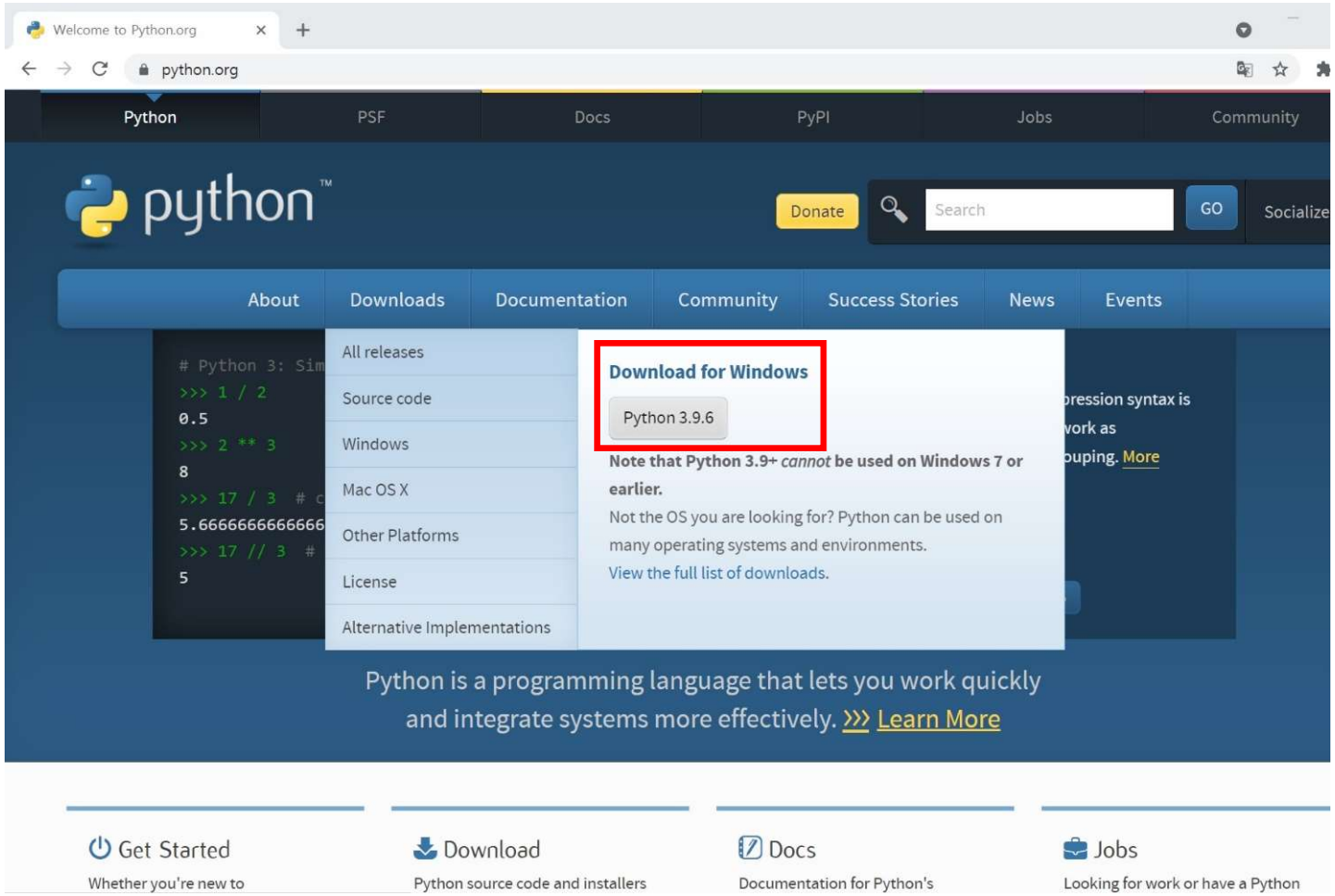
• 데이터 분석을 위한 파이썬 라이브러리

- Matplotlib
  - 데이터 시각화를 위한 파이썬 라이브러리

- SciPy
  - 미분방정식의 계산처럼 과학 계산을 위한 컴퓨팅 함수 지원
  - NumPy와 SciPy를 함께 사용하면 대부분의 과학계산 기능을 수행할 수 있음

- Scikit-learn
  - 범용 머신러닝 도구로써 다음의 하위 기능을 포함함
    - 분류(k-NN, LR, Random Forest, SVM), 회귀(Lasso, ridge regression), 클러스터링(k-means clustering, spectral clustering), 차원축소(PCA, feature selection, etc.), 모델 선택(grid search, cross-validation, metrics), 데이터 전처리(normalization, feature extraction)

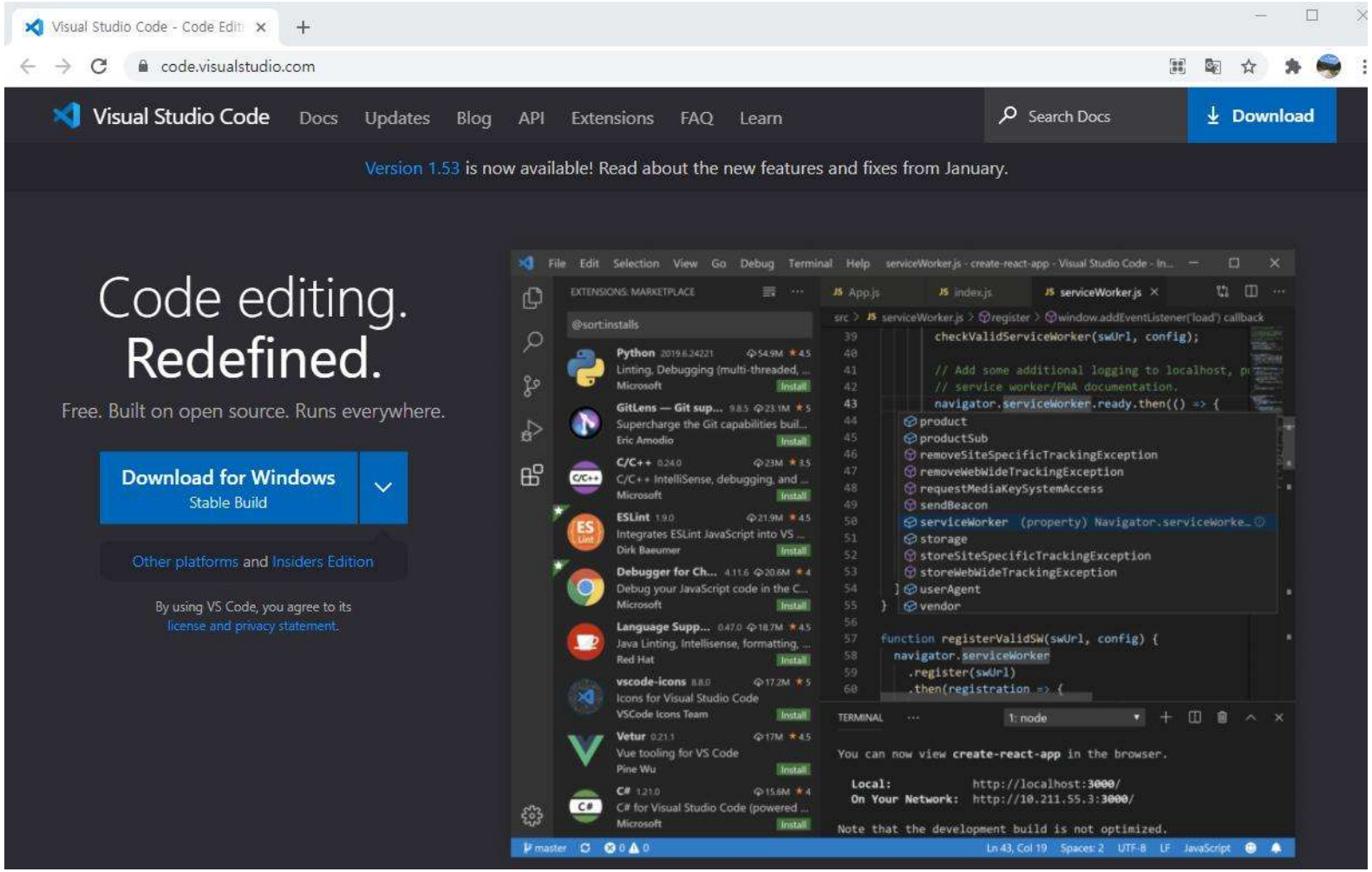
- **파이썬 설치**
  - http://python.org 에 접속하여 사용 운영체제에 맞는 파이썬 설치



• 통합 개발환경 VSCODE 설치

- VSCODE (Visual Studio Code)
  - VSCODE는 Microsoft의 무료 소스 코드 편집기
  - 윈도우, macOS, 리눅스 지원, 디버깅 및 Git 제어 등을 지원
  - 다양한 프로그래밍 언어와 연동 가능한 통합 개발환경 도구 이며, 각 언어와 함께 사용할 다양한 기능을 제공함
  - 소스 코드 편집에 최적화된 인터페이스 제공

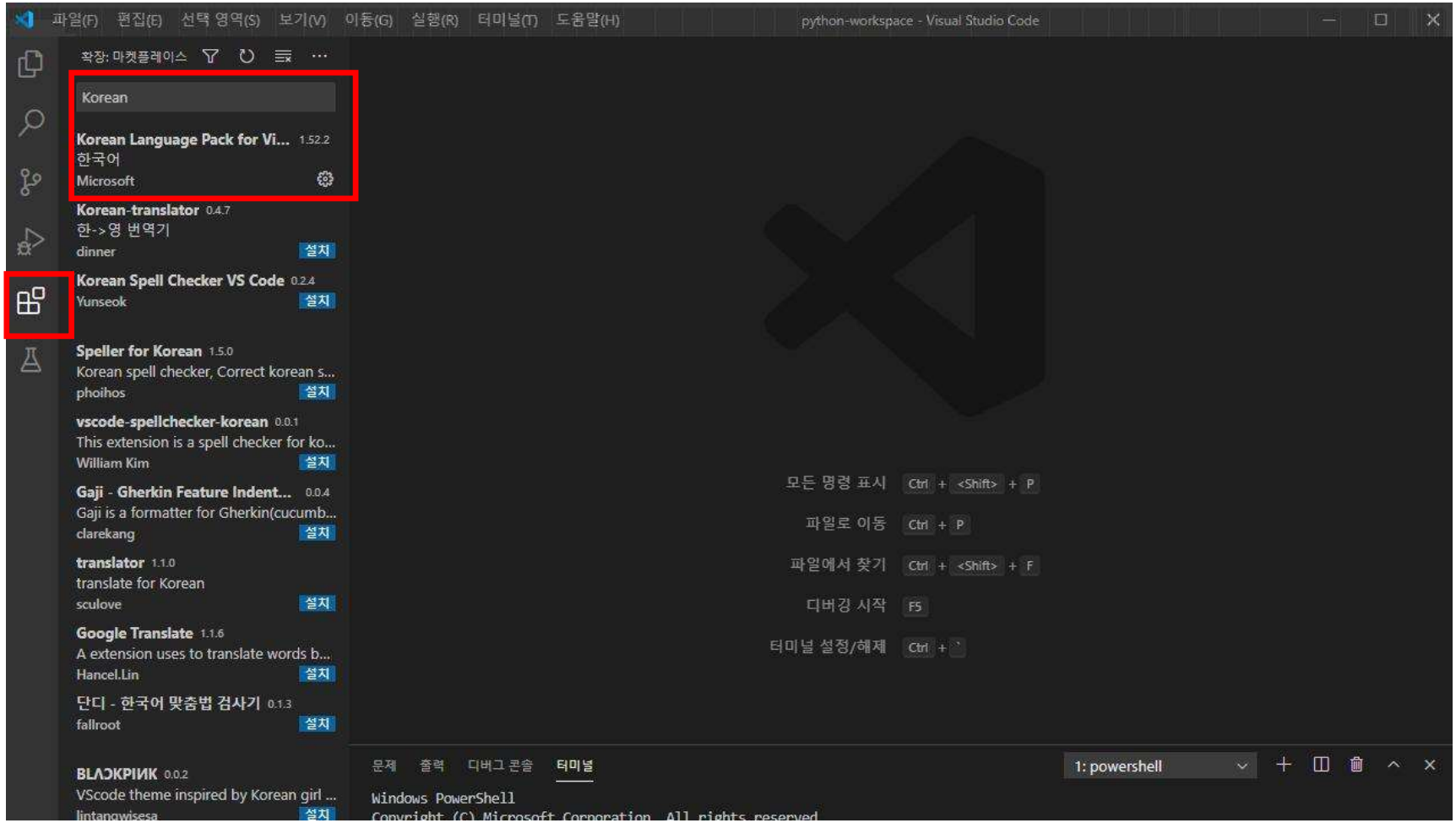
- **통합 개발환경 VSCODE 설치**
  - https://code.visualstudio.com 에서 VSCODE 다운로드 후 설치



• 통합 개발환경 VSCODE 설치

- 한국어 패키지 설치

Search: Korean
Install: Korean Language Package

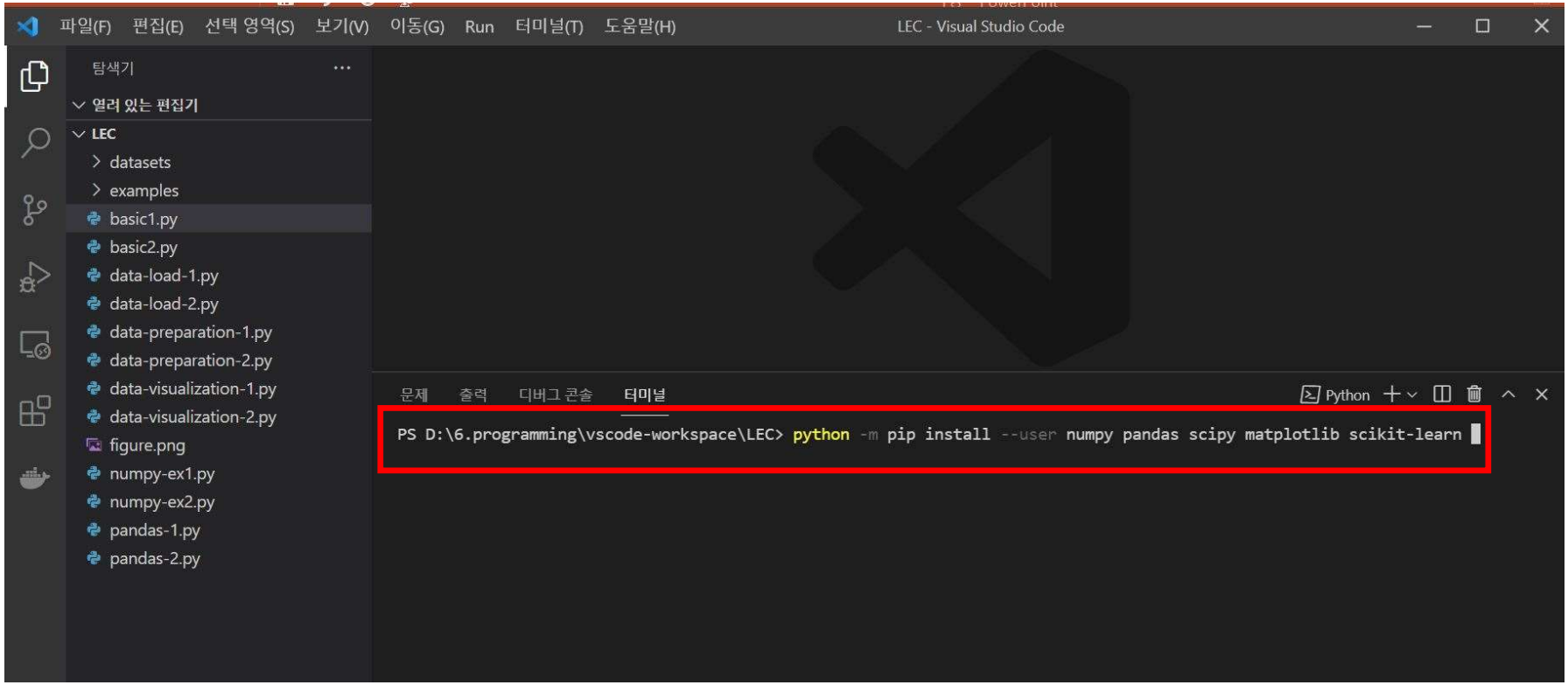


• 통합 개발환경 VSCODE 설치

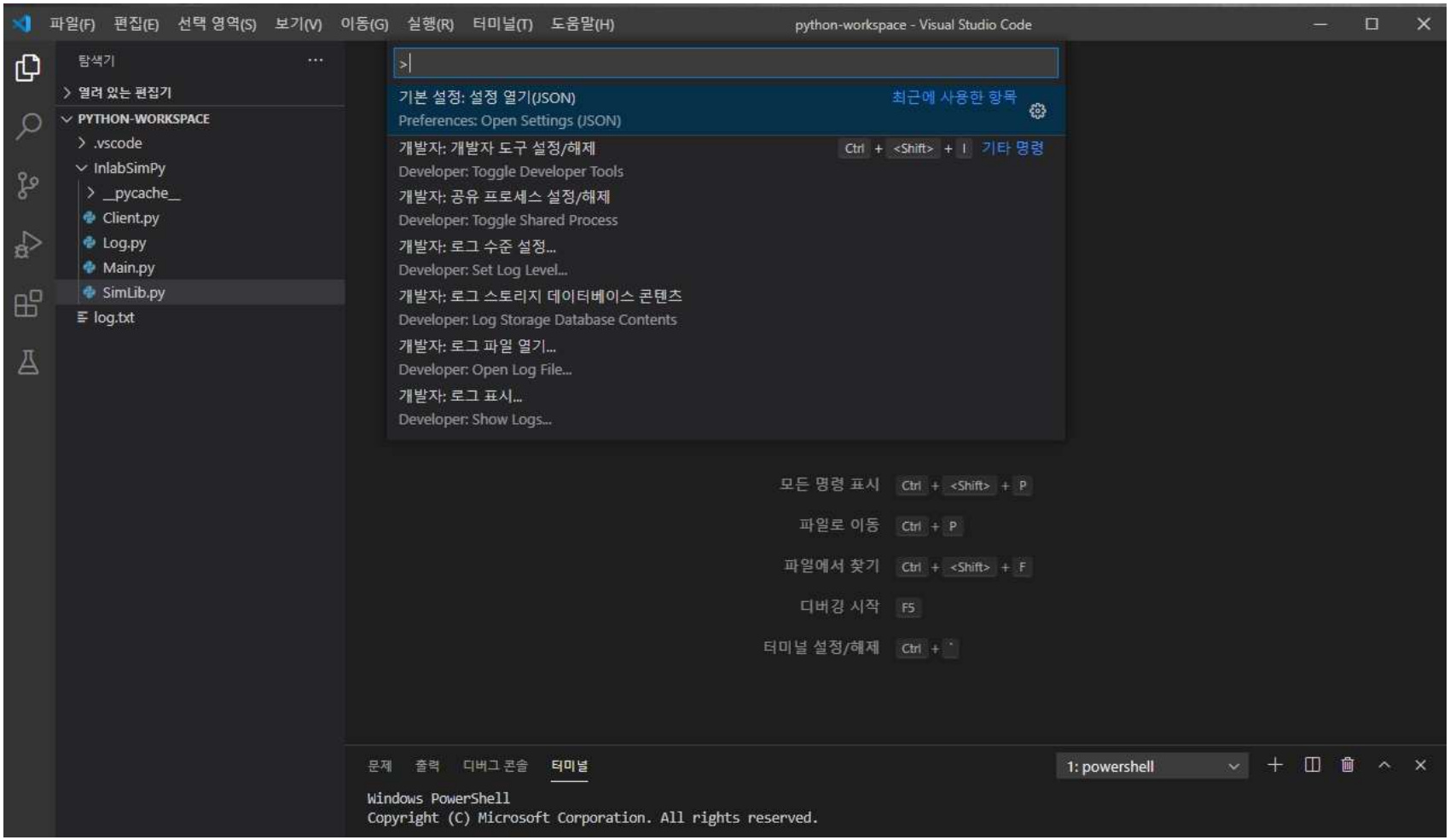
- Python Extension 설치 *Search*: Python
  Install: Python

- **빅데이터 관련 라이브러리 설치**
  - 파이썬 버전 확인: *python3 --version*
  - PIP를 사용하여 빅데이터 분석 관련 라이브러리 설치:

    python3 –m pip install –user numpy pandas scipy matplotlib scikit-learn

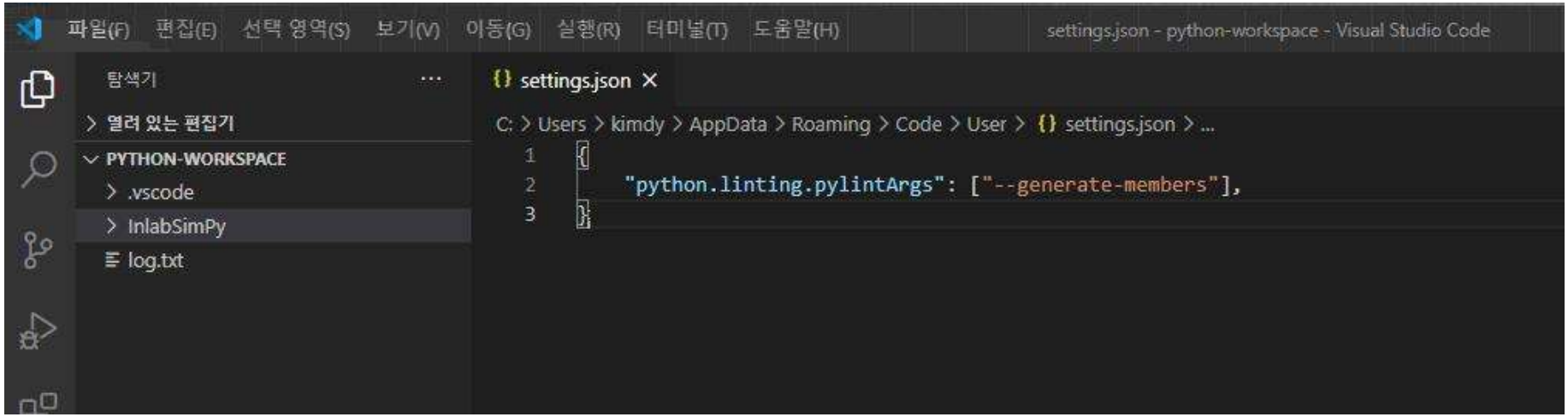


- **파이썬 NumPy 설정 (VSCODE에서 NumPy 사용하기)**
  - Ctrl + Shift + P in VSCODE



- **파이썬 NumPy 설정 (VSCODE에서 NumPy 사용하기)**

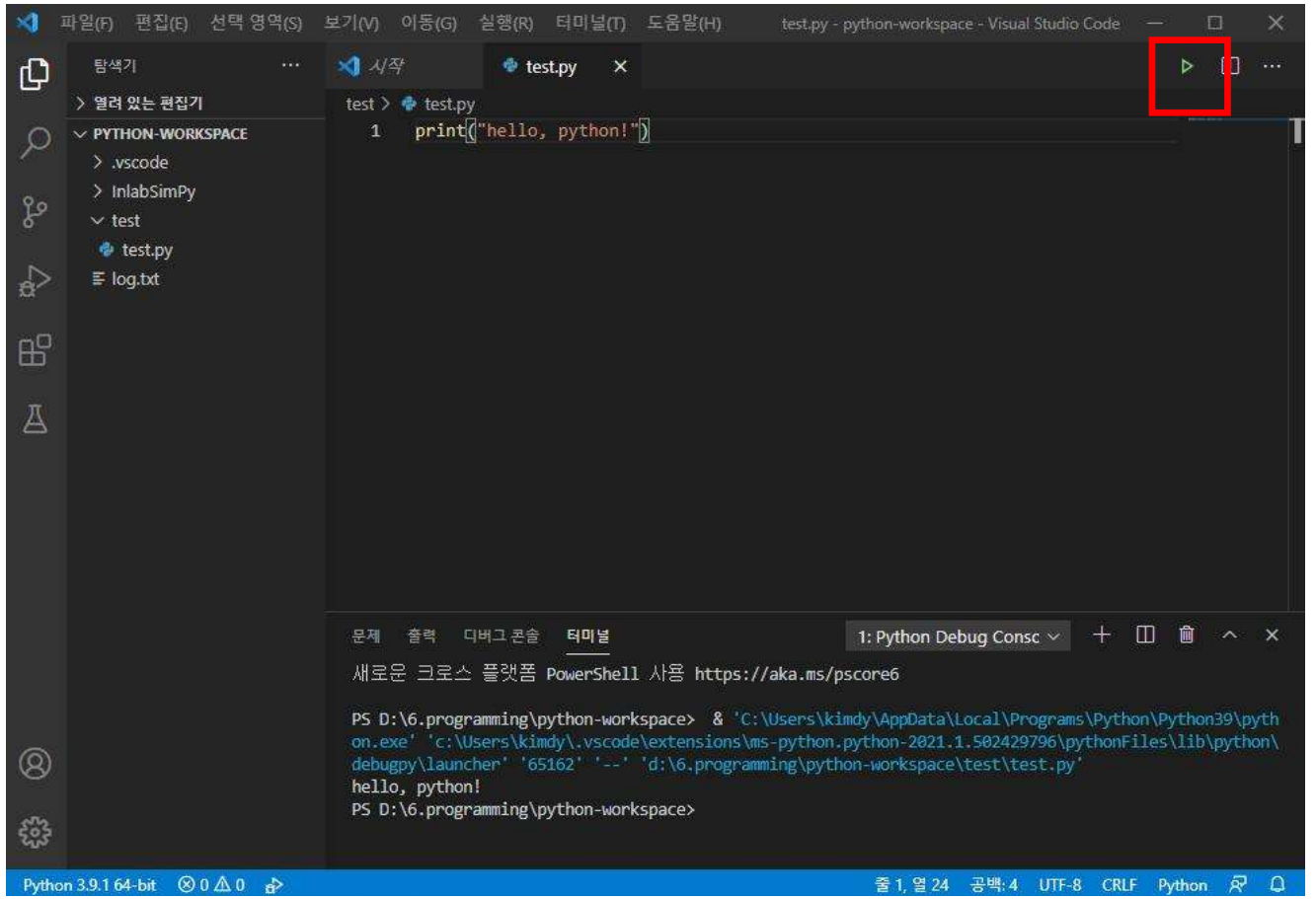
```
{
    "python.linting.pylintArgs": ["--generate-members"],
}
```



연습

- VSCODE에서 파이썬을 사용하여 다음의 문장 출력
  - "hello, python!"
    - Write *print("hello, python!")*
    - Run (Ctrl+5)

Ctrl+5



요약

- 빅데이터
  - 다양한 정형 및 비정형 데이터 분석을 통해 의미 있는 정보 추출
  - 의미있는 정보를 통해 서비스 및 시스템 운용에 활용
- 빅데이터 프로그래밍 도구: Python
  - 빠르고 단순하며 다양한 라이브러리를 지원하여 빅데이터 및 머신 러닝 범용 애플리케이션 개발에 활용
- 빅데이터 프로그래밍 환경 구축
  - Python 및 Python 라이브러리 설치
  - VSCODE 통합 개발 환경과 연동