UNISTUDY

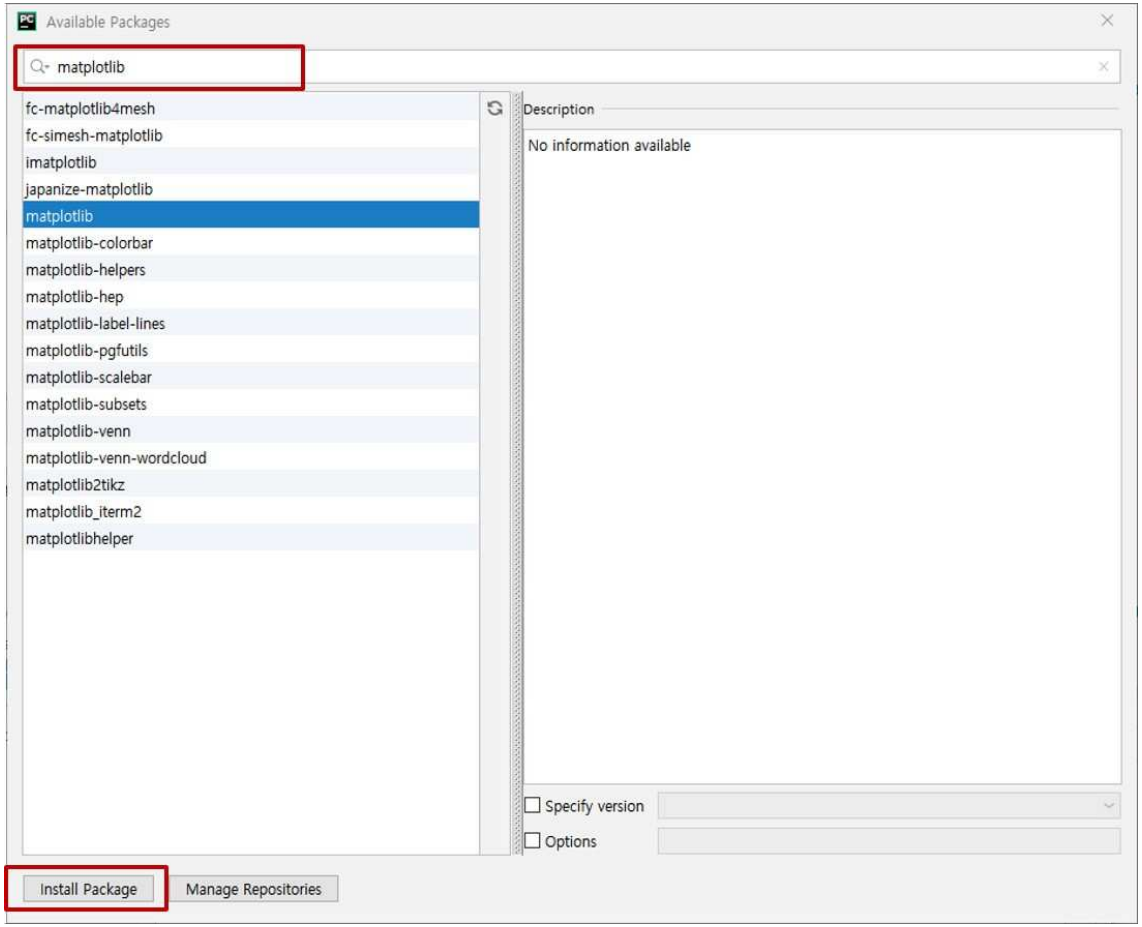
파이썬

36강. 자료 시각화

1. 자료 시각화

- 자료 분석에서 그래프를 이용한 시각화는 다양한 형태로 이용되어진다. 자료 분석의 도입부에서는 전 체적인 자료의 구조를 분석하거나 분석 방향을 제시하고, 중반부에서는 잘못된 처리 결과를 확인하며, 후반부에서는 분석 결과를 도식화하여 의사결정에 반영하기 위해서 자료를 시각화한다. 이처럼 자료 시각화는 자료 분석의 전 과정에서 유용하게 사용된다. 파이참에서 그래프를 그리기 위해서는 matplotlib 패키지를 설치해야한다.

1. 자료 시각화


Available Packages
matplotlib
fc-matplotlib4mesh
fc-simesh-matplotlib
imatplotlib
japanize-matplotlib
matplotlib
matplotlib-colorbar
matplotlib-helpers
matplotlib-hep
matplotlib-label-lines
matplotlib-pgfutils
matplotlib-scalebar
matplotlib-subsets
matplotlib-venn
matplotlib-venn-wordcloud
matplotlib2tikz
matplotlib_iterm2
matplotlibhelper
Description
No information available
Specify version
Options
Install Package
Manage Repositories

2. 단어 출현빈도수 시각화

- news 자료를 대상으로 출현빈도수가 Top5에 해당하는 단어를 선 그래프와 막대 그래프로 시각화하는 방법에 대해서 알아본다.

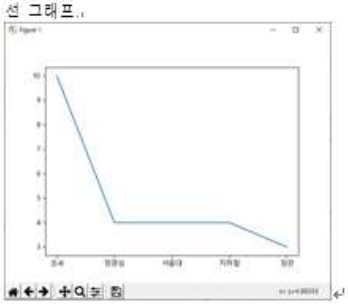
2. 단어 출현빈도수 시각화

chapter09.lecture.step06_word_count.py

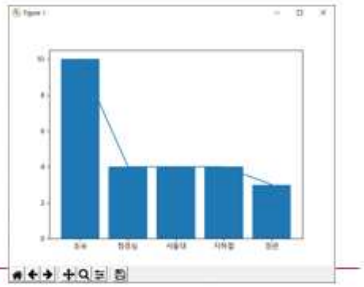
```
# (1) 단어와 출현 빈도수 만들기
words = [] # 단어
counts = [] # 출현빈도수

for word, count in top5_word :
    words.append(word)
    counts.append(count)

print(words)
print(counts)

# (2) pyplot 모듈 import
import matplotlib.pyplot as plt

# (3) 차트에서 한글 지원
from matplotlib import font_manager, rc
font_name = font_manager.FontProperties( fname="c:/Windows/Fonts/malgun.ttf").get_name()
rc('font', family=font_name)

# (4) 선 그래프
print('선 그래프')
plt.plot(words, counts)  # 그리기
plt.show()  # 보이기

# (5) 막대 그래프
print('막대 그래프')
plt.bar(words, counts) # 그리기
plt.show() # 보이기
```

Python Console

```
['조국', '정경심', '서울대', '지하철', '장관']
[10, 4, 4, 4, 3]
선 그래프
```



```
막대 그래프
```



2. 단어 출현빈도수 시각화

- 단어 빈도수 시각화 예
- # (1) 단어와 출현 빈도수 벡터 만들기
- 차트를 그리기 위해서 단어와 단어의 출현 빈도수를 각각 벡터(vector)로 만든다. 벡터란 반복 가능한 배열 형태의 자료구조를 의미한다.
- # (2) pyplot 모듈 import
- 차트를 그릴 수 있는 pyplot 모듈을 import한다.
- ※ matplotlib 패키지가 설치되어 있어야 한다.
- # (3) 차트에서 한글 지원
- 파이썬의 차트에서는 기본적으로 한글 지원을 하지 않는다. 따라서 단어가 한글이 경우 차트 내에서 한글을 나타내기 위해서는 별도로 한글 폰트를 사용할 수 있도록 관련 내용을 import 해야 한다. 예문에서는 c:/windows/Fonts 폴더에 있는 맑은 고딕체를 이용하여 한글이 깨지지 않도록 폰트를 지정하고 있다.

2. 단어 출현빈도수 시각화

- # (4) 선 그래프
- 단어와 단어의 출현 빈도수를 이용하여 선 그래프를 그린다. pyplot의 plot() 함수는 선 스타일 (line style)을 기본으로 제공하므로 x축에 단어 벡터를 y축에는 단어의 출현 빈도수를 이용하여 선 그래프가 그려진다.
- # (5) 막대 그래프
- 단어와 단어의 출현 빈도수를 이용하여 막대 그래프를 그린다. pyplot의 bar() 함수는 x축에 단어 벡터를 y축에는 단어의 출현 빈도수를 이용하여 막대 그래프를 그린다.

3. 기본 차트 그리기

- matplotlib 패키지에서 제공하는 함수를 이용하여 기본 차트를 시각화 하는 방법에 대해서 알아본다.

4. 차트 자료 만들기

- 다음 예문은 차트를 시각화하기 위해서 필요한 패키지를 import하고 차트를 만드는데 필요한 자료를 생성하는 과정이다.

chapter09.lecture.step07_matplot.py

```
# (1) 패키지 import
import matplotlib.pyplot as plt  # 차트 생성
import random # 차트 자료 생성
# 음수 부호 지원
import matplotlib
matplotlib.rcParams['axes.unicode_minus'] = False

# (2) 차트 자료 생성
print(random.randint(a=1, b=5))  # 1~5 난수 정수
print(random.random()) # 0~1 난수 실수
print(random.normalvariate(mu=0, sigma=1))
# 평균=0, 표준편차=1를 갖는 표준정규분포 난수
```

Python Console

```
3
0.47927395326232514
-0.4161786445881056
```

4. 차트 자료 만들기

- 해설 패키지 import와 차트 자료 만들기 예
- # (1) 패키지 import
- plt 별칭을 이용하여 기본 차트를 그릴 수 있고, random 모듈을 이용하여 차트를 만드는데 필요한 자료를 만든다. 또한 차트 내에서 음수 부호가 깨지지 않도록 음수 부호를 지원하는 패키지를 import 한다.
- # (2) 차트 자료 생성
- 모든 차트는 숫자 자료를 대상으로 만들 수 있기 때문에 난수를 생성할 수 있는 관련 함수를 이용한다. randomint(a, b) 함수는 a~b 사이의 임의의 난수 정수를 만든다. random() 함수는 0과 1사이의 임의의 난수 실수를 만든다. normalvariate() 함수는 평균과 표준편차를 인수로 하여 정규분포를 따르는 난수를 만든다. 이때 평균=0, 표준편차=1인 경우 표준정규분포를 따르는 난수가 만들어진다.

5. 기본 차트 그리기

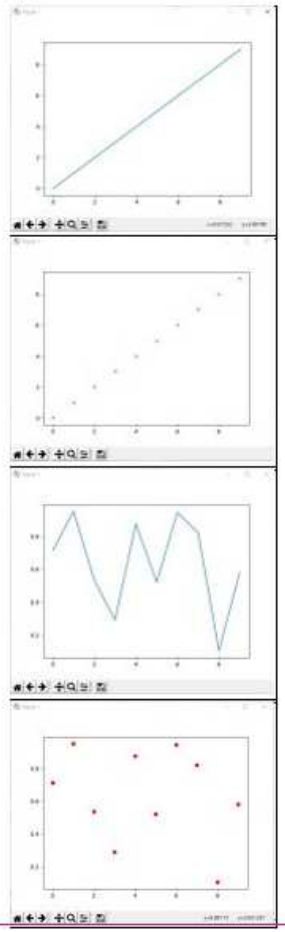
- 다음 예문은 생성된 자료를 토대로 pyplot 모듈에서 제공하는 plot() 함수를 이용하여 차트를 그리는 과정이다.

5. 기본 차트 그리기

chapter09.lecture.step07_matplot.py

```
# (1) plot()함수 도움말
help(plt.plot)
'''
plot(x, y)  # plot x and y using default li
ne style and color
plot(x, y, 'bo')  # plot x and y using blue
circle markers
plot(y)       # plot y using x as index arr
ay 0..N-1
plot(y, 'r+')     # ditto, but with red plus
ses
'''

# (2) 기본 차트 그리기

# (2-1) 1개 data
data = range(10)  # range(n) 동일 - 0~9
plt.plot(data) # plot(y), x : index
plt.plot(data, 'r+')  # plot(y, 'r+')



# (2-2) 2개 data
data2 = [random.random() for i in range(1
0)] # 난수 실수
plt.plot(data, data2) # line

plt.plot(data, data2, 'ro')  # point
```

Python Console



5. 기본 차트 그리기

- 해설 기본 차트 그리기 예

- # (1) plot()함수 도움말
- pyplot 모듈의 plot()함수에 대한 도움말을 이용하여 함수의 파라미터에 대한 의미를 알 수 있다. 다음은 plot()함수의 4가지 파라미터 사용에 대한 도움말의 내용이다.
- plot(x, y) : 2개 인수를 사용할 경우 x축과 y축의 자료로 사용되고, 기본 선 그래프와 색상으로 차트가 그려진다.
- plot(x, y, 'bo') : 파랑색의 타원 마커 효과를 추가하여 차트가 그려진다.
- plot(y) : 1개의 인수를 사용할 경우 y축의 자료로 사용되고, x축은 자료의 색인으로 차트가 그려진다. plot(y, 'r+') : y축의 값이 빨강색의 플러스 기호로 차트가 그려진다.

5. 기본 차트 그리기

- # (2-1) 1개 data
- range(10)에 의해서 생성된 0~9사이의 10개 정수를 y축의 자료로 사용하여 plot() 함수에 의해서 기본 선 그래프가 그려진다. 또한 plot(data, 'r+') 함수에 의해서 빨강색의 플러스 기호로 차트를 그린다.
- # (2-2) 2개 data
- data는 x축에 data2는 y축의 자료로 사용하여 기본 선 그래프가 그린다. 이때 data2는 0과 1사이의 임의의 난수 실수를 저장한 변수이다. 또한 빨강색의 타원 마커 효과('ro')를 추가하여 차트가 그린다.

6. 산점도와 히스토그램 그리기

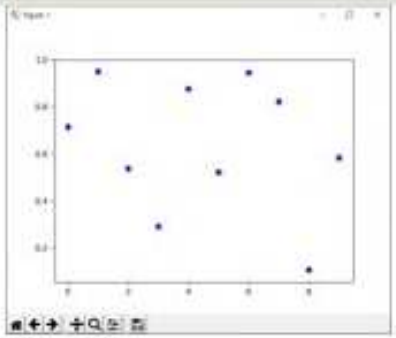
- pyplot 모듈에서 제공하는 scatter()와 hist() 함수를 이용하여 산점도와 히스토그 램을 그리는 과정이다.

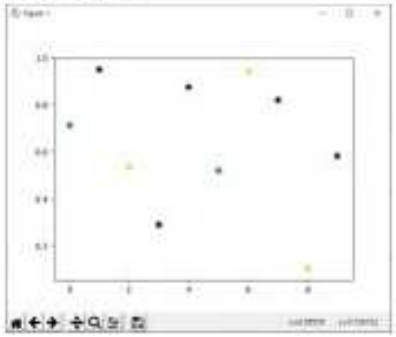
6. 산점도와 히스토그램 그리기

chapter09.lecture.step07_matplot.py

```
# (1) 산점도 그리기

# (1-1) 단색 산점도
plt.scatter(x=data, y=data2, c='b', marker='o')

# (1-2) 여러 가지 색 산점도
cdata = [random.randint(a=1, b=3) for i in range(10)]
cdata  # [3, 2, 2, 1, 3, 2, 3, 1, 3, 3]
plt.scatter(x=data, y=data2, c=cdata, marker='o')



# (2) 히스토그램
data3 = [random.normalvariate(mu=0, sigma=1) for i in range(1000)]
```

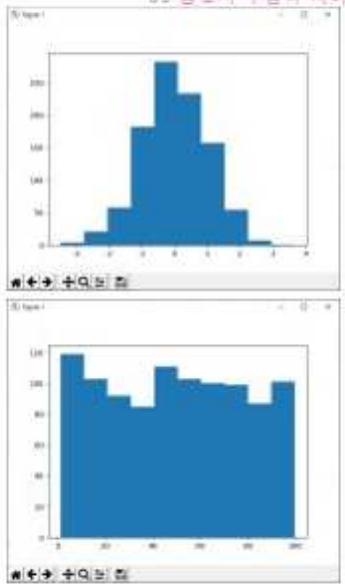
Python Console





6. 산점도와 히스토그램 그리기


```
plt.hist(data3) # 정규분포

data4 = [random.uniform(a=1, b=100) for i
in range(1000)]
plt.hist(data4) # 균등분포
```



6. 산점도와 히스토그램 그리기

- 해설 산점도와 히스토그램 그리기 예
- # (1-1) 단색 산점도
- 산점도는 x축과 y축의 교차점을 시각화한 차트를 말한다. scatter(x=data, y=data2, c='b', marker='o') 함수에 의해서 단색 산점도가 그려진다. c는 산점도 색상, marker는 산점도를 나타내 는 마커 효과를 지정하는 속성이다.
- # (1-2) 여러 가지 색 산점도
- 1~3 사이의 난수 정수를 이용하여 범주(category)를 생성하고 각 범주별로 산점도의 색상을 적용하 는 예문이다.
- # (2) 히스토그램
- 평균 0, 표준편차 1을 갖는 표준정규 분포를 따르는 난수 1,000개를 갖는 data3을 이용하여 히스토그 램을 그리면 평균을 중심으로 좌우 대칭성을 갖는 차트가 그려진다.
- 또한 uniform() 함수를 이용하여 1~100 사이의 난수 1,000개를 갖는 data4를 이용하여 히스토그램 을 그리면 균등분포의 차트가 그려진다.

UNISTUDY

THANK YOU