데이터분석 with 파이썬

텍스트 처리 및 시각화 - WordCloud

천양하

목차

- 텍스트 데이터란?
- wikipedia 텍스트 데이터
- WordCloud 텍스트 데이터 시각화

텍스트 데이터란 무엇인가

- 텍스트는 인간이 오랫동안 정보를 효율적으로 교환하는 데에 가장 중요한 수단으로 사용
- 텍스트 데이터는 구조화된 문서(HTML, XML, CSV, JSON 파일)와 구조화되지 않은 문서(자연어로 된 텍스트)로 나눌 수 있다.
- 일반적으로 원천 데이터는 가공된 형태가 아니기 때문에 우리는 이들 데이터를 수정하여서 완전한 데이터로 만들어야 한다.

정보를 한 눈에 보여주는 워드 클라우드(word cloud)

- 위의 그림은 wordart.com에서 제공하는 워드 클라우드word cloud라는 시각화 기술의 예시이다.
- 워드 클라우드word cloud는 각 단어의 크기가 빈도 또는 중요성을 나타내는 텍스트 데이터 시각화 기술
- 텍스트 데이터에서 핵심 정보를 시각적으로 제공할 때 유용
- 워드 클라우드는 소셜 네트워크, 웹 사이트의 데이터를 분석하는데 널리 사용된다.

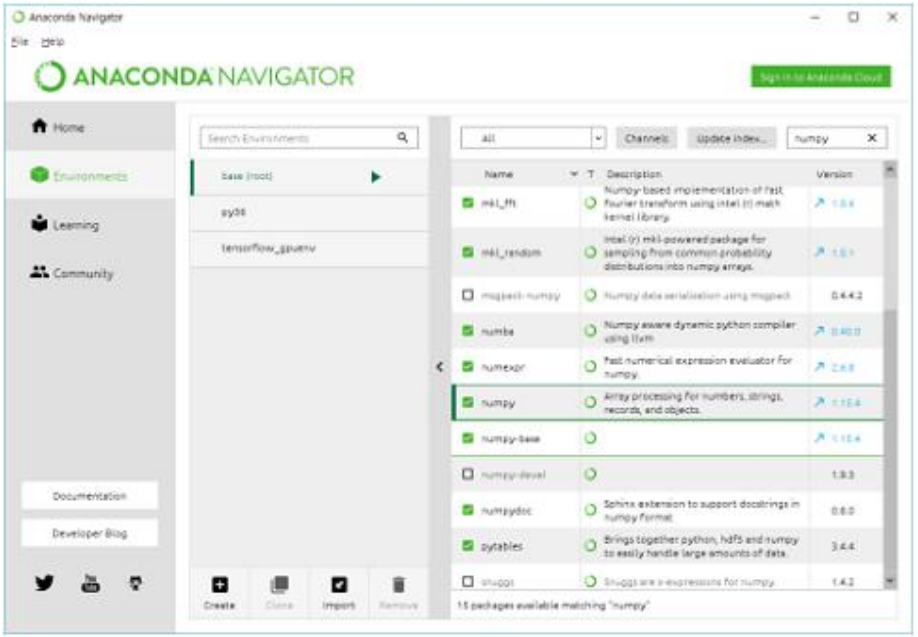
wikipedia 라이브러리 개요

- 워드 클라우드를 생성하기 위해서는 우선 원천 데이터의 역할을 수행할 텍스트를 준비해야 한다.
- 텍스트를 준비하기 위한 방법으로 위키백과의 내용을 가져오는 wikipedia 모듈 필요 - wikipedia 모듈을 먼저 설치

- **아나콘다 내비게이터**
  - 박스에 체크가 되지 않았다면 체크하여 설치

- **Google Colab 환경**
  - wikipedia 라이브러리가 없으면 아래 코딩으로 설치하면 된다.

```
pip install wikipedia
```



wikipedia 텍스트 데이터 가져오기

- wikipedia.page(title)이라고 하여 title을 제목으로 하는 위키백과 페이지를 얻을 수 있다. 이 페이지의 텍스트 데이터를 얻고 싶으면 해당 페이지의 content을 사용하면 된다.

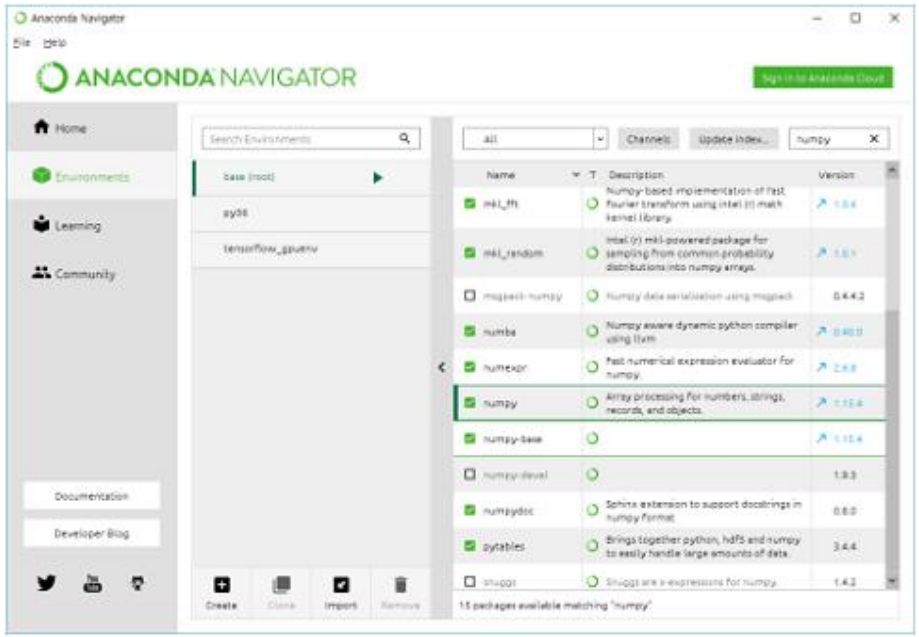
```
import wikipedia

# 위키백과 사전의 컨텐츠 제목을 명시해 준다
wiki = wikipedia.page('Artificial intelligence')
# 이 페이지의 텍스트 컨텐츠를 추출하도록 한다
text = wiki.content
```

- 텍스트 데이터가 준비되면, 이 데이터를 이용하여 워드 클라우드 이미지를 생성한다

WordCloud 설치 : WordCloud 설정

❖ **WordCloud는 일반적으로 많이 사용하는 모듈이기 때문에 기본으로 설치되어 있음**



- **아나콘다 내비게이터**
  - 박스에 체크가 되지 않았다면 체크하여 설치

- **Google Colab 환경**
  - WordCloud 라이브러리가 없으면 설치하면 된다.

명령문	pip list # 설치된 라이브러리들을 보여줌
명령문	pip install wordcloud #라이브러리가 없다면 설치실행 from wordcloud import WordCloud #WordCloud 라이브러리 불러오기

WordCloud 문법 : WordCloud 설정

- 가로와 세로 크기를 클래스의 생성자에 넘겨 주어 이미지의 크기를 정하고, generate() 함수를 불러 워드 클라우드를 만들 재료가 될 텍스트 데이터를 인자로 넘겨준다
- 인자들 중에서 width는 워드 클라우드 이미지의 너비이고 height는 높이를 픽셀단위로 표현한 것이다.

```
from wordcloud import WordCloud

# 워드 클라우드를 생성하기 위해 위의 코드를 삽입할 것
wordcloud = WordCloud(width = 2000, height = 1500).generate(text)
```

텍스트 데이터 시각화 : matplotlib

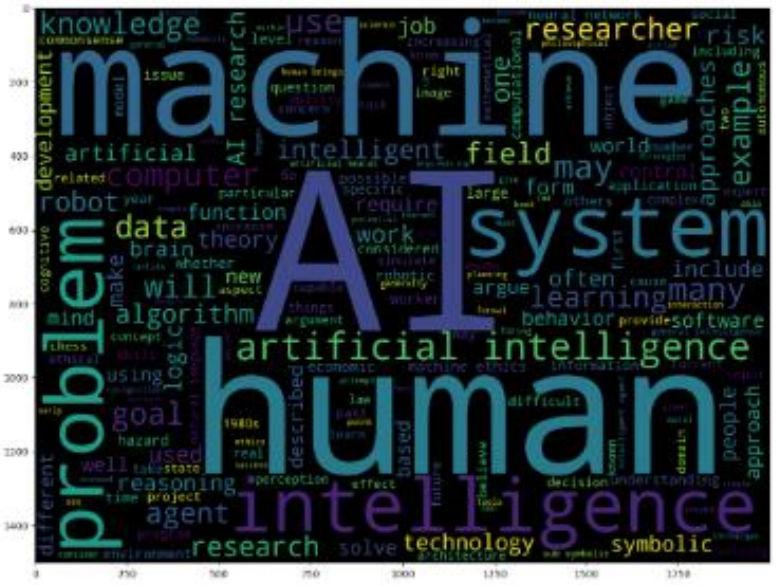
- 워드 클라우드 이미지를 화면에 그리기
- 파이썬의 데이터 시각화와 차트를 그려주는 matplot 라이브러리 사용
- matplotlib의 이미지 그리기 함수인 imshow()를 이용하면 쉽게 그릴 수 있다.

```
import matplotlib.pyplot as plt

plt.figure(figsize=(40, 30))
# 화면에 이미지를 그려준다
plt.imshow(wordcloud)
plt.show()
```

텍스트 데이터 시각화 : 중지어 설정

- 그림과 같이 워드 클라우드가 잘 그려졌다.
- 많은 텍스트 데이터에는 자주 쓰이지만, 특별히 중요한 의미를 갖지 않는 단어들이 있다.
- 이러한 단어를 자연어 처리에서는 중지어stop word라고 한다.



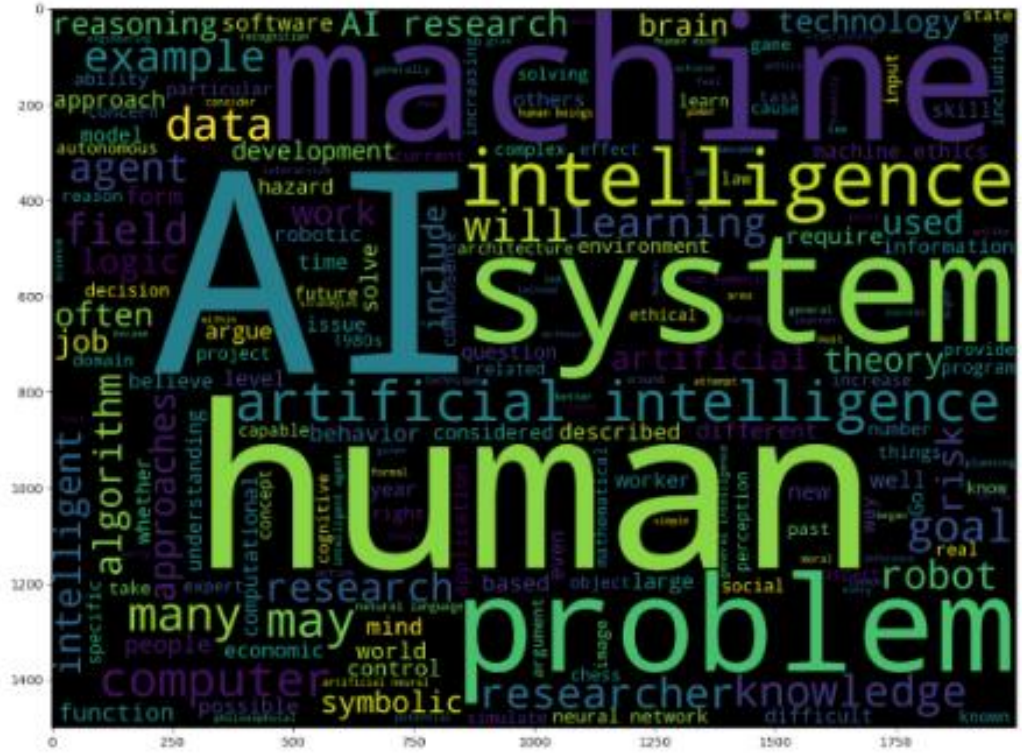
텍스트 데이터 시각화 : 중지어 설정

- 워드 클라우드를 만들 때 이러한 중지어를 제외하는 방법
- 중지어 리스트는 wordcloud 모듈의 STOPWORDS에 정리되어 있다.
- STOPWORDS는 집합set 데이터로 정의되어 있으며, 우리가 중지어를 추가하고 싶으면 다음과 같이 새로운 집합과 합집합을 구해서 사용하면 된다.

```
from wordcloud import WordCloud, STOPWORDS
# 중지어가 제외된 워드 클라우드를 만들자
s_words = STOPWORDS.union( {'one', 'using', 'first', 'two', 'make', 'use'} )
wordcloud = WordCloud(width = 2000, height = 1500,
                      stopwords = s_words).generate(text)
```

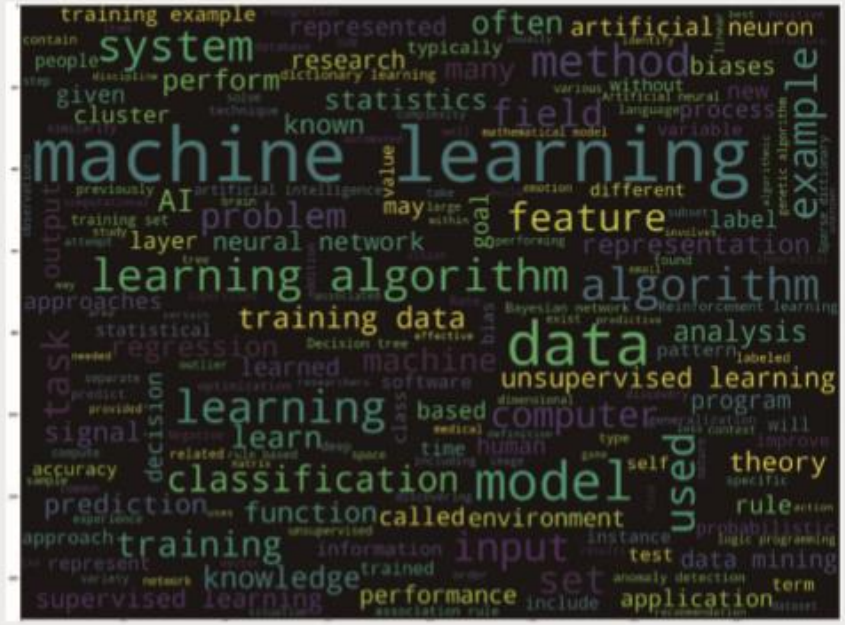
텍스트 데이터 시각화 : 중지어 설정

- 다음과 같이 이러한 중지어가 없어진 워드 클라우드를 얻을 수 있을 것이다.



확인학습

- 다음과 같이 Python과 Machine Learning을 이용하여 워드 클라우드를 각각 표현해보자.



THANK YOU FOR YOUR ATTENTION