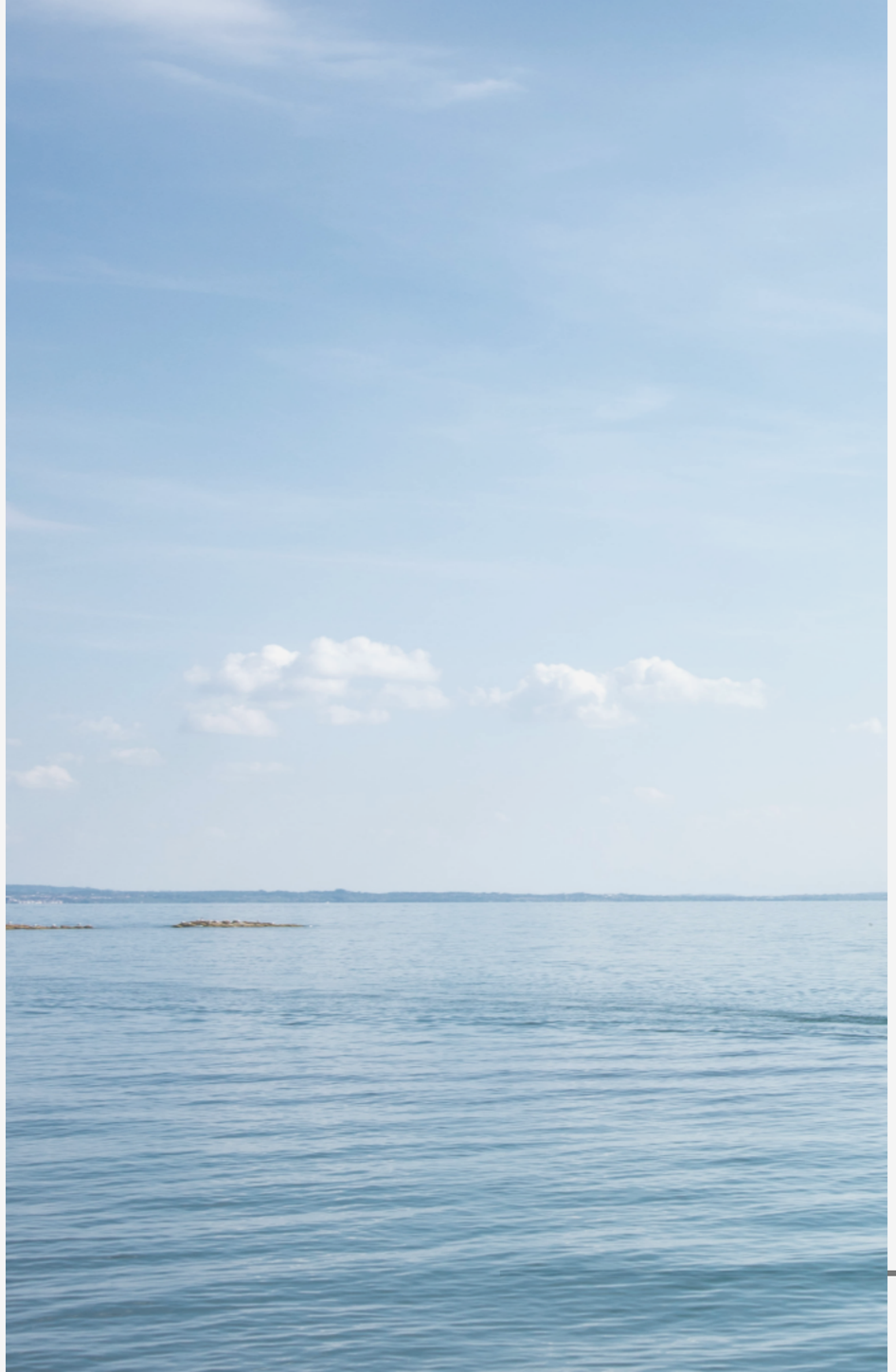
2021/1/7/

1주차
텍스트 마이닝이란? (OT)

텍스트분석/자연어처리



목차

01

텍스트 마이닝

텍스트 마이닝

비정형 또는 정형의 텍스트 데이터에 자연어 처리 기술과
문서 처리 기술을 적용하여 정보를 추출, 가공하는 목적으로 하는 기술

= 유용한 패턴이나 관계 찾아내는 프로세스

01 응용 분야

설명적 마이닝 & 예측적 마이닝

텍스트 마이닝 유형	활용분야	
	실무	연구
검색 (Information Retrieval)	스팸 필터링	사회동향 분석
분류 (Classification)	이슈 검출/트래킹	소셜미디어 분석
군집화 (Clustering)	정보검색	이슈 트래킹
웹마이닝 (Web Mining)	자살률 예측	온라인 행동 분석
정보추출 (Information Extraction)	주가 예측	연구분야 탐색
개념추출 (Concept Extraction)	소비자 인식 조사	질병관계 예측
~~자연어처리 (NLP)~~	경쟁사 분석	정책전략 수립

02 원리

언어학과 통계학적 기반에서 머신러닝을 통해 기계가 언어학적,
통계적 특징을 학습하는 형태로 발전하여 사용됨.

텍스트 마이닝이
마냥 쉽진 않은 이유

01 언어의 한계

맞춤법, 철자, 신조어, 단어 섞어쓰기, 줄여쓰기, 비문,
동음이의어, 문맥에 따른 다른 의미 등 총체적 난국.

02 데이터의 한계점

비정형 데이터인 만큼 전처리하기도 어렵고,
방대한 규모이며, 자연어처리에 대한 이해와 언어별 이해가 필요.

텍스트 마이닝이 마냥 쉽진 않은 이유

구분	내용
오탈자	“헝거게임 잼잇써요완전 대신 이전편꼭바여” “솔까 타노스 보석 하나도 못구했을때 다들 머했음? 3개 얻었을때도 그렇게 안쌔 뵈더만…”
동의어, 동음이의어	한혜진 : 1. 모델 한혜진 (달심), 2. 배우 한혜진 (기성용 부인), 3. 가수 한혜진 (트로트 가수) Close : 1. Opposite of open, 2. A preposition meaning not far IS : 1. Information System, 2. Islamic State, 3. International Standard
전처리	분석 데이터의 언어를 파악하고 언어의 특징 (교착어, 굴절어 등)에 맞는 전처리 작업 진행 댓글 단위로 분석할지, 문장 단위로 분석할지에 따라서 데이터 분리작업 진행
정보추출	해시 태그 (hash tag) 추출 : ‘#’ + (문자) 핸드폰 번호 추출 ‘010’ - (4자리 숫자) - (4자리 숫자)

07

특히 한글은, 더 힘들다

언어학적 특성으로 인해 전처리와 분석 과정이 까다로움.

용언의 변형

모르네, 모르다, 모른다, 몰랐더니,
몰랐다, 몰랐니,,,,,

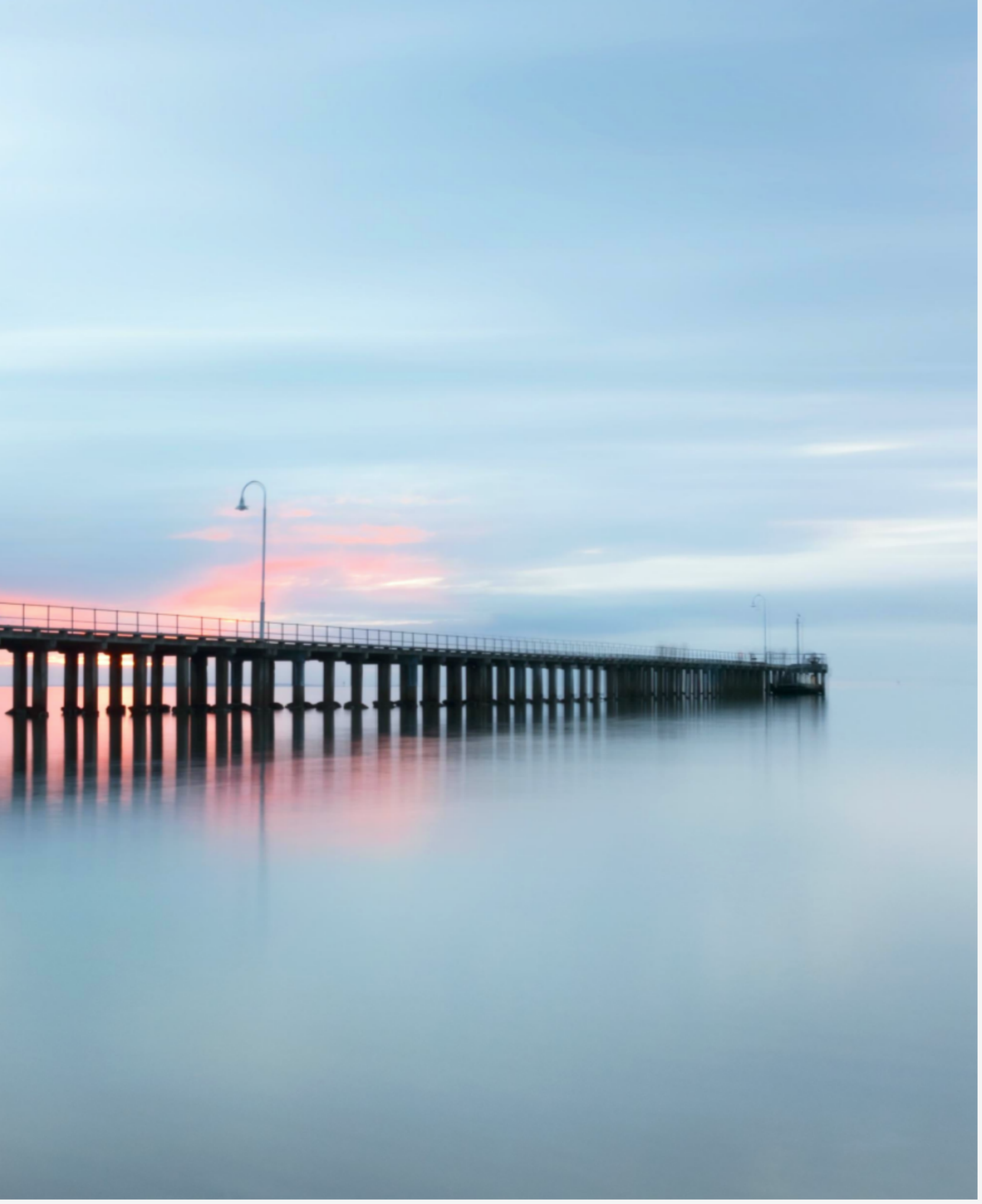
같은 말도 다른 글자로...

=연세 연세대 연잡대 연대
욘세이 욘세 연잡대 연머,,,,

신조어

형태소 분석기가 아직 부족.

그래서 아직도 번역기가 매생이 전복죽을 Every life is ruined. 라 번역하는 겁니다



자연어처리

**간단한 텍스트 시각화나
네트워크 분석이 아닌
인공지능 활용.**

자연어 처리는 음성 인식, 내용 요약, 번역, 사용자의 감성 분석,
텍스트 분류 작업(스팸 메일 분류, 뉴스 기사 카테고리 분류),
질의 응답 시스템, 챗봇과 같은 곳에서 사용되는 분야입니다.

02

nltk, koNLPy, tensorflow 등 설치

필요한 패키지 설치

nltk 설치

파이썬이 설치되었다는 전제 하

01 pip install nltk

```
(base) odageon-ui-MacBookAir:~ dagunoh$ pip install nltk
Requirement already satisfied: nltk in ./opt/anaconda3/lib/python3.8/site-packages (3.5)
Requirement already satisfied: click in ./opt/anaconda3/lib/python3.8/site-packages (from n
ltk) (7.1.2)
Requirement already satisfied: joblib in ./opt/anaconda3/lib/python3.8/site-packages (from
nltk) (0.16.0)
Requirement already satisfied: regex in ./opt/anaconda3/lib/python3.8/site-packages (from n
ltk) (2020.6.8)
Requirement already satisfied: tqdm in ./opt/anaconda3/lib/python3.8/site-packages (from nl
tk) (4.47.0)
```

02 jupyter notebook에서 나중에 사용할 때,

```
import nltk
nltk.download()
```

koNLPy 설치

JDK 설치되었다는 전제 하

01 pip install konlpy

```
(base) odageon-ui-MacBookAir:~ dagunoh$ pip install koNLPy
Requirement already satisfied: koNLPy in ./opt/anaconda3/lib/python3.8/site-packages (0.5.
)
Requirement already satisfied: JPype1>=0.7.0 in ./opt/anaconda3/lib/python3.8/site-package
 (from koNLPy) (1.1.2)
Requirement already satisfied: numpy>=1.6 in ./opt/anaconda3/lib/python3.8/site-packages (
rom koNLPy) (1.18.5)
Requirement already satisfied: tweepy>=3.7.0 in ./opt/anaconda3/lib/python3.8/site-package
 (from koNLPy) (3.9.0)
Requirement already satisfied: lxml>=4.1.0 in ./opt/anaconda3/lib/python3.8/site-packages
from koNLPy) (4.5.2)
```

02 JDK 오류 뜨면....

1) JDK 설치

우선 JDK를 1.7 버전 이상으로 설치해야 합니다.

설치 주소 : https://www.oracle.com/technetwork/java/javase/downloads/index.html

설치한 후에는 JDK가 설치된 경로를 찾아야 합니다.
예를 들어 저자의 경우에는 jdk가 아래의 경로에 설치되어 있습니다.

C:₩Program Files₩Java₩jdk-11.0.1

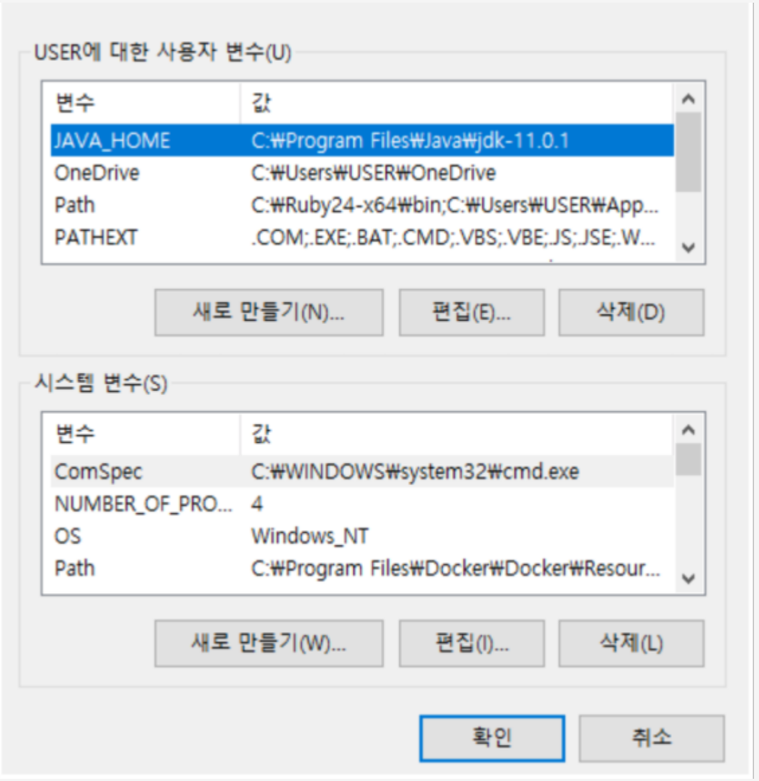
11.0.1과 같이 버전에 대한 숫자는 어떤 버전을 설치했느냐에 따라 다를 수 있습니다.

2) JDK 환경 변수

설치 경로를 찾았다면 해당 경로를 복사합니다.
해당 경로를 윈도우 환경 변수에 추가해야하기 때문입니다.

윈도우 10기준)
제어판 > 시스템 및 보안 > 시스템 > 고급 시스템 설정 > 고급 > 환경 변수
새로 만들기(N)...를 누르고 JAVA_HOME이라는 환경 변수를 만듭니다.

환경 변수의 값은 앞서 찾았던 jdk 설치 경로입니다.



3) JPype 설치

이제 JAVA와 Python을 연결해주는 역할을 하는 JPype를 설치해야 합니다.

설치 주소 : https://www.lfd.uci.edu/~gohlke/pythonlibs/#jpype

해당 링크에서 적절한 버전을 설치해야 하는데 cp27은 파이썬 2.7,
cp36은 파이썬 3.6을 의미합니다.
우리는 파이썬 3.6을 사용하고 있으므로 cp36이라고 적힌 JPype를 설치해야 합니다.

또 사용하는 윈도우 O/S가 32비트인지, 64비트인지에 따라서 설치 JPype가 다른데,
윈도우 32비트를 사용하고 있다면 win32를,
윈도우 64비트를 사용하고 있다면 win_amd64를 설치해야 합니다.
예를 들어 파이썬 3.6, 윈도우 64비트를 사용 중이라면
JPype1-0.6.3-cp36-cp36m-win_amd64.whl를 다운로드합니다.

프롬프트에서 해당 파일의 경로로 이동하여 아래 커맨드를 통해 설치합니다.

> pip install JPype1-0.6.3-cp36-cp36m-win_amd64.whl
이제 JPype의 설치가 완료되었다면, KoNLPy를 사용할 준비가 되었습니다

tensorflow 설치

머신러닝 라이브러리 오픈소스

01 pip install tensorflow

02 jupyter notebook에서 나중에 사용할 때,

```
> ipython
...
In [1]: import tensorflow
as tf
In [2]: tf.__version__
Out[2]: '2.0.0'
```

Gensim 설치

word2vec 등

01 pip install gensim

02 jupyter notebook에서 나중에 사용할 때,

import gensim

Scikit-learn 설치

머신러닝 라이브러리

01 **pip install scikit-learn**

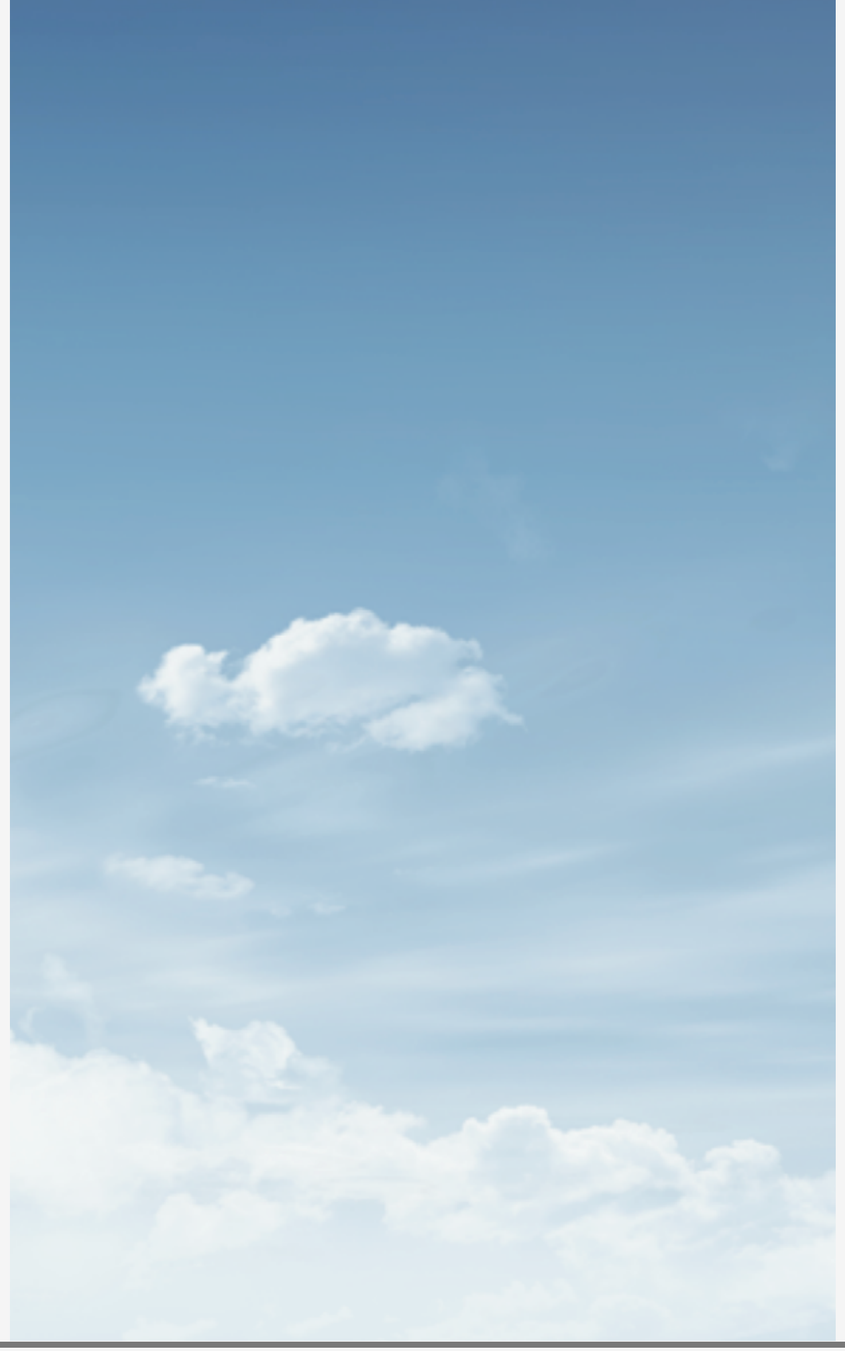
02 jupyter notebook에서 나중에 사용할 때,

import sklearn

03

웹크롤링

언제나 예쁘게 csv 파일로 존재하는 것이 아님...!
우리가 긁어서 써와야 할 수도 있음...!



실습해봅시다

모두 실습 파일을 켜주세요

○ BeatifulSoup

Beautiful Soup 객체 형성하여 웹크롤링 하게 해줌.

○ IDMB review crawling/
코로나 확진자 수 crawling

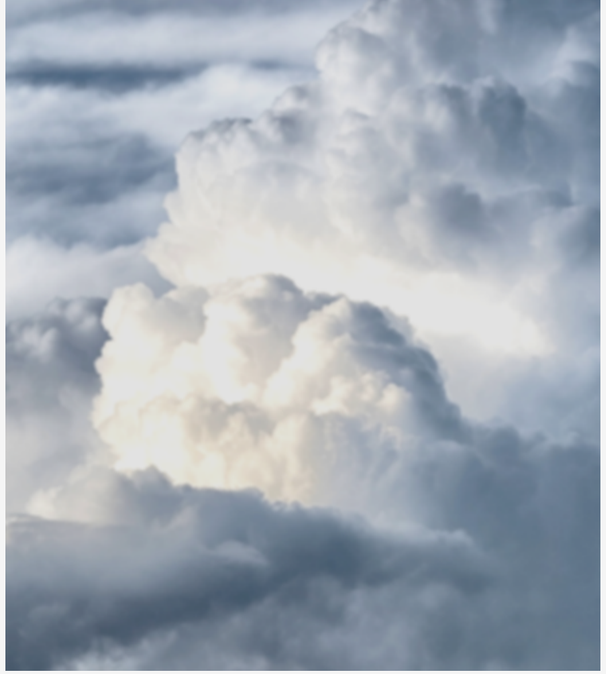
○ page 가 여러개인 경우/ 파일에 저장

04

텍스트 전처리 맛보기

이렇게 받아온 텍스트를 전처리해야 좋은 결과를
얻을 수 있음.

4-1. 토큰화
nltk tokenizer 이용



Corpus?
말뭉치 또는 코퍼스(Corpus)는
자연언어 연구를 위해 특정한 목적을 가지고
언어의 표본을 추출한 집합

Token?
토큰(Token)이란 문법적으로 더 이상 나눌 수 없는 언어요소
텍스트 토큰화(Text Tokenization)란 말뭉치로부터
토큰을 분리하는 작업을 뜻함.

발표를 들어주셔서
감사합니다 :)