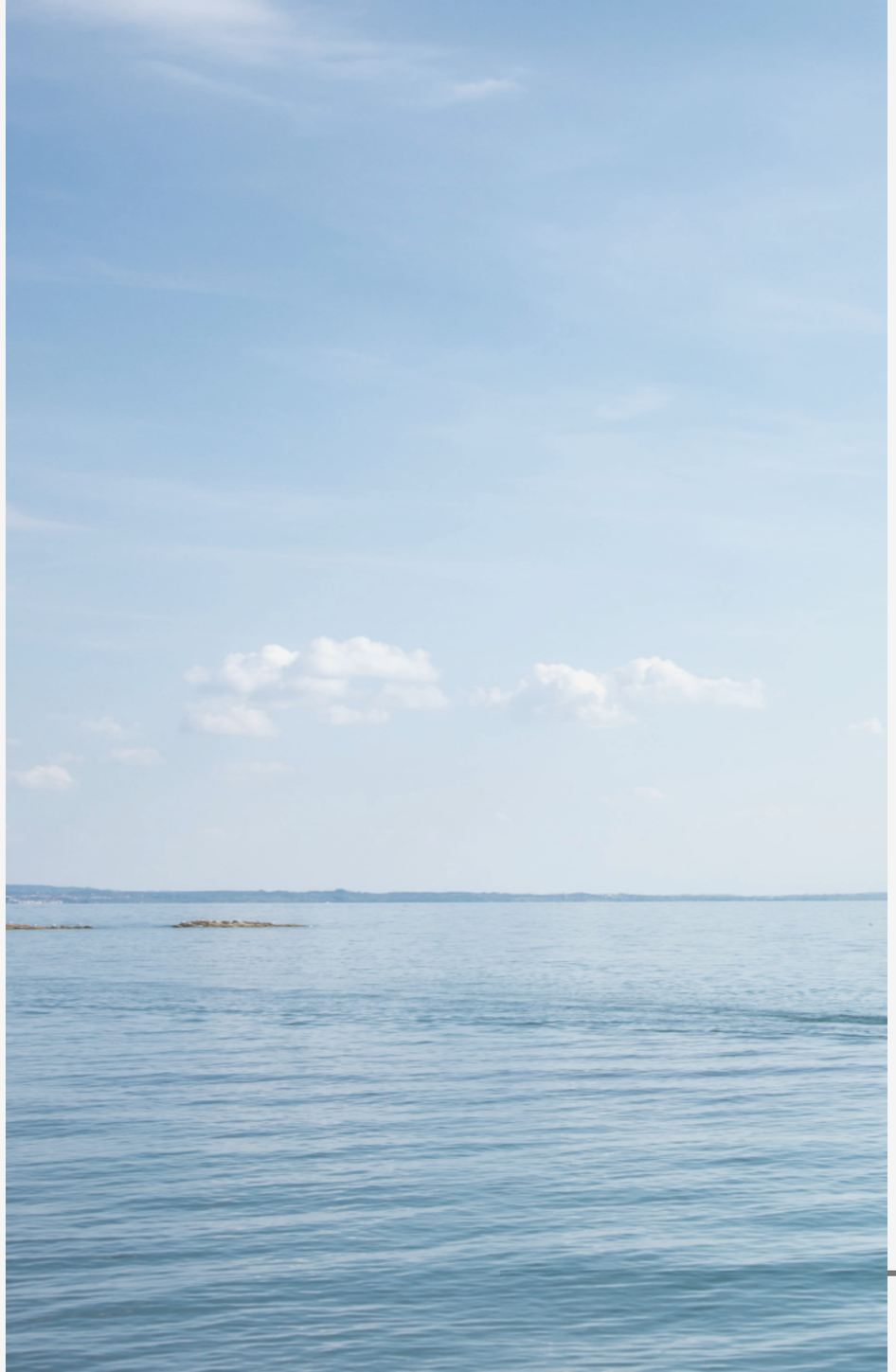
2021/1/16

2주차
Preprocessing and Tokenizing



목차

01 preprocessing

02 tokenizing

03 실습

01

Preprocessing

텍스트 전처리

아직 텍스트들이 분석하기에 충분하지 못함.
분석을 용이하게 하기 위해선 전처리 과정이 필수임.

01 정규식 활용

import re
re.split(), re.findall(), re.sub()

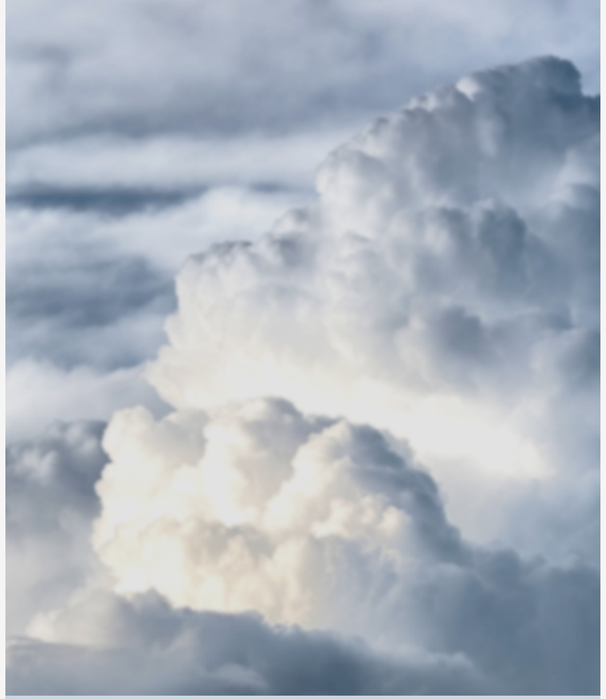
모듈 함수 종류	설명
re.compile()	정규표현식을 컴파일하는 함수
re.search()	문자열 전체에 대해서 정규표현식과 매치되는지를 검색
re.match()	문자열의 처음이 정규표현식과 매치되는지 검색
re.split()	정규 표현식을 기준으로 문자열을 분리하여 리스트로 리턴
re.findall()	문자열에서 정규 표현식과 매치되는 모든 경우의 문자열을 찾아서 리스트로 리턴
re.sub()	문자열에서 정규 표현식과 일치하는 부분에 대해서 다른 문자열로 대체

02 string 기본 메소드 활용

string.replace(), string.strip(), string.split(), string.lower()

메소드 종류	설명
str.replace(pattern,alternative)	해당 패턴과 매치되는 부분을 다른 문자열로 대체
str.lower()	문자열 전체에 대해서 소문자
str.upper()	문자열 전체에 대해서 대문자
str.strip()	문자열 공백 제거
str.split(sep='구분자')	문자열 구분자 기준으로 나누기
str.isalpha()	문자열이 영어로만 이루어졌으면 True 반환
str.isnumeric()	문자열이 숫자만 포함하고 있으면 True 반환

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.

간단히 말해서,

어휘/형태소 분석
형태소 - 뜻을 가진 가장 작은 말의 단위
형태소 분석 - 문장을 형태소 단위로 분리하고 품사를 태깅하는 과정
우리나라는 토큰을 형태소로 생각.

구분	내용																								
원문	손	흥	민	이	골	을	작	렬	하	며	토	트	넘	홋	스	퍼	의	승	리	를	이	끌	었	다	.
음절	손	흥	민	이	골	을	작	렬	하	며	토	트	넘	홋	스	퍼	의	승	리	를	이	끌	었	다	.
형태소	손흥민			이	골	을	작렬		하	며	토트넘			홋스퍼			의	승리		를	이끌		었	다	.
어절	손흥민이				골을		작렬하며				토트넘			홋스퍼의				승리를			이끌었다				.

구분	품사태그	설명
체언	**NNG**	**일반명사**
	NNP	**고유명사**
	NNB	의존명사
	NR	**수사**
	NP	대명사
용언	**VV**	**동사**
	VA	**형용사**
	VX	보조용언
	VCP	긍정지정사
	VCN	부정지정사
관형사	MM	관형사
부사	MAG	일반부사
	MAJ	접속부사
감탄사	IC	감탄사
조사	JKS	주격조사
	JKC	보격조사
	JKG	관형격조사
	JKO	목적격조사
	JKB	부사격조사
	JKV	호격조사
	JKQ	인용격조사
	JX	보조사
	JC	접속조사

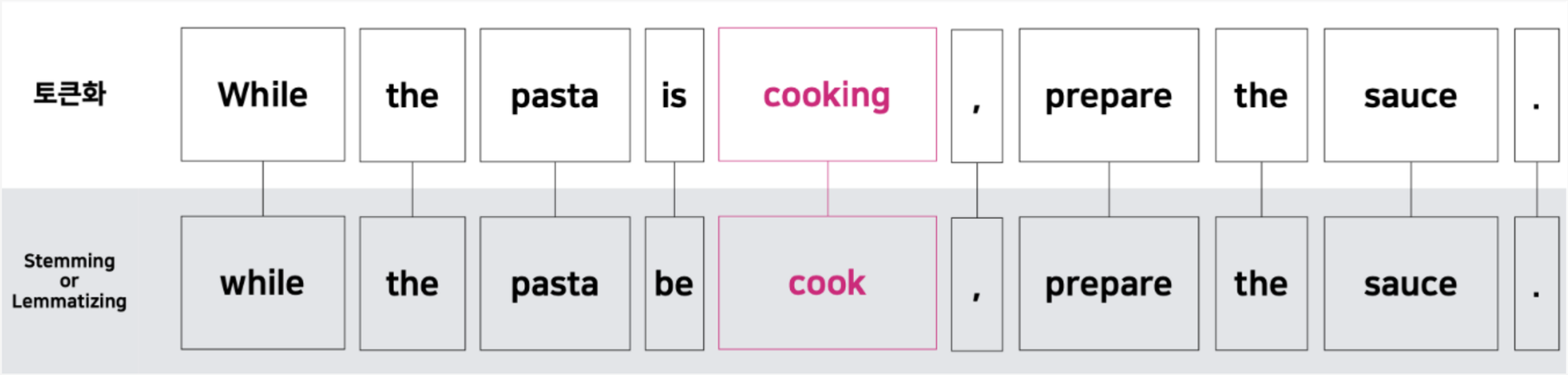
구분	품사태그	설명
선어말 어미	EP	선어말어미
어말 어미	EF	종결어미
	EC	연결어미
	ETM	명사형 전성어미
	ETD	관형형 전성어미
접두사	XPN	체언접두사
접미사	XSN	명사파생 접미사
	XSV	동사파생 접미사
	XSA	형용사 파생 접미사
어근	XR	어근
부호	SF	마침표, 물음표, 느낌표
	SP	쉼표, 가운뎃점, 콜론, 빗금
	SS	따옴표, 괄호, 줄표
	SE	줄임표
	SO	물결표, 숨길표, 빠짐표
	SW	기타 기호
미식별	NF	명사추정범주
	NV	용언추정범주
	NA	분석불능범주
외국어	SL	외국어
	SH	한자
	SN	숫자

영어의 형태소 분석

원형 복원: 어간 추출과 표제어 추출

어간 추출: 규칙 기반으로 단어의 변형된 형태 제거. (ing, -s, -es, -ed)

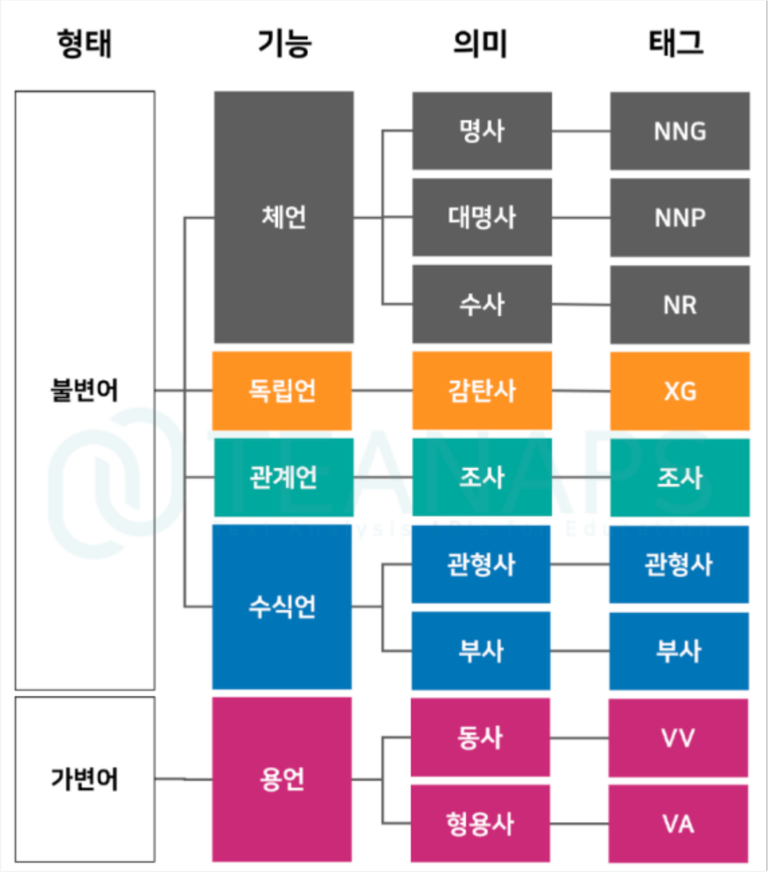
표제어 추출: 단어의 형변환 사전을 기반으로 품사에 맞는 단어의 원형으로 복구



하지만 한글은,,,,

한글의 5언 9품사

한글은 단어를 기능, 의미, 형태 세 가지 기준으로 나눔.



한국어의 언어학적 특징

교착어: 어근에 접사가 결합되어 각 단어의 기능을 나타냄
굴절어: 단어 자체의 형태 변화로 문법성 나타내줌
고립어: 단어의 형태변화 없이 문법적 관계는 어순에 의해 정해지는것

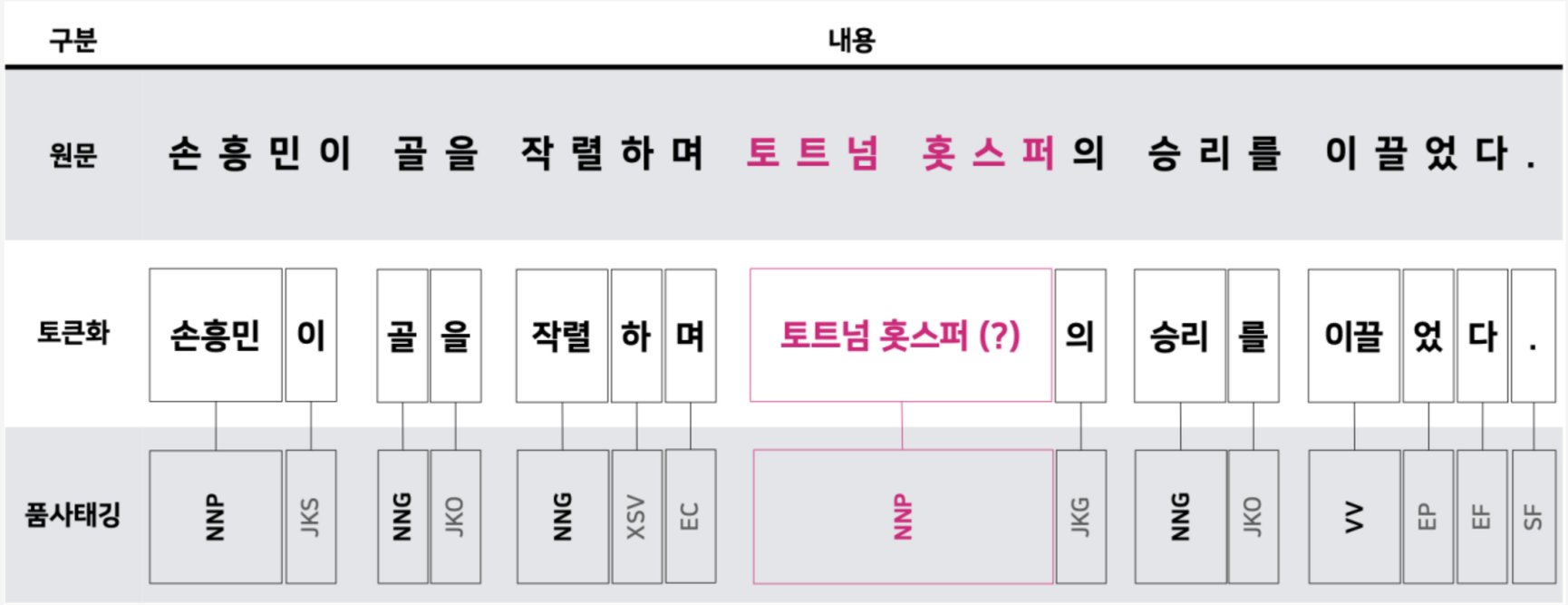
어근	피동	높힘	과거	추측	전달	어미	파생된 단어
일어나						+다	일어나다
일어나	+지					+다	일어나지다
일어나	+지	+시				+다	일어나지시다
일어나	+지	+시	+었			+다	일어나지셨다
일어나			+았			+다	일어났다
일어나				+겠		+다	일어나겠다
일어나					+더라		일어나더라
일어나		+지	+었			+다	일어나졌다
일어나		+지	+었	+겠		+다	일어나졌겠다
일어나	+지	+었	+겠		+더라		일어나졌겠더라
일어나			+았	+겠		+다	일어났겠다
일어나	+지	+시	+았	+겠	+더라		일어나지셨겠더라

한국어의 언어학적 특징_교착어

어근에 따라 단어가 무한정 생성되기 때문에 비슷한 의미의 단어가 계속 생기나,
형태소 분석 시 수많은 경우의 수를 다르게 처리하게 되어 추가적인 토큰화가 필요.
단, 어순은 영향 많이 받지 않음.

한글의 형태소 분석

문장을 형태소 단위로 구분하고, 품사를 구별하여 태깅,
용언의 변형으로 탈락한 형태소 복원.



한글 형태소 분석기

01 Mecab

검색에서 쓸만한 오픈소서 한국어 형태소 분석기를 목적으로 개발,
사용자 사전 등록기능 제공해 다양한 도메인에서 생성된 단어들 인식 가능.

02 Okt

트위터에서 개발한 한국어 형태ㅐ소 분석기.
유명한 인물이나 신조어 잘 인식하는 편이지만 형태소 태깅 면에서 품질이 낮음.

03 kkma

서울대에서 개발, 가능한 모든 형태소 후보와 조합을 찾아 가장 적합한 것을 골라냄.
=> 계산이 길어 아주 느림.

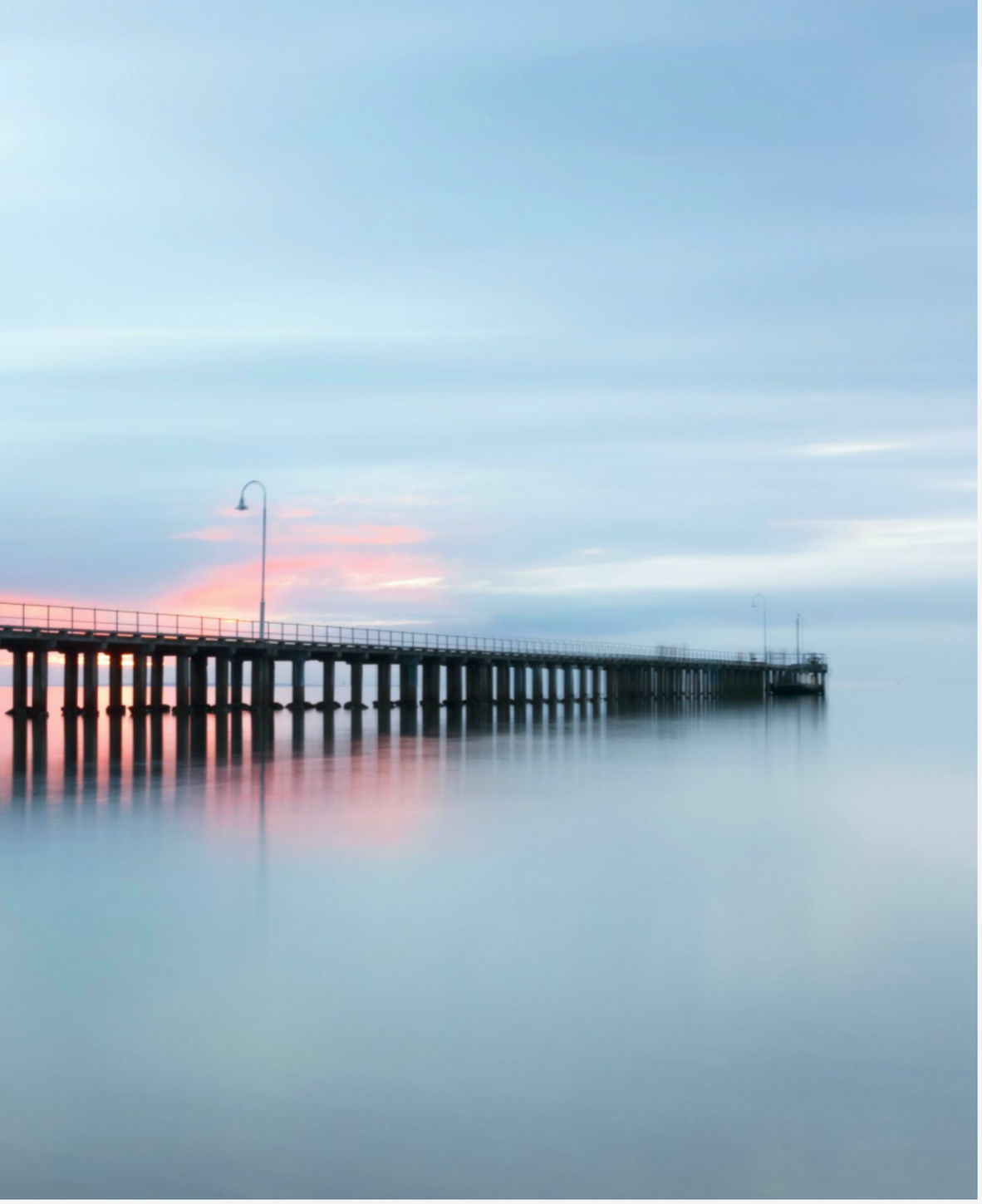
문서 표현 방법 중_stopwords

분석 결과에 출현하더라도 큰 의미 없거나 방해하는 것들.

이,그,저,이거,저거,
$,%,@ 등

조사 같은 기능적인 역할

nltk.stopwords 이용



실습파일로

가봅시다!