/*elice*/

파이썬 크롤링

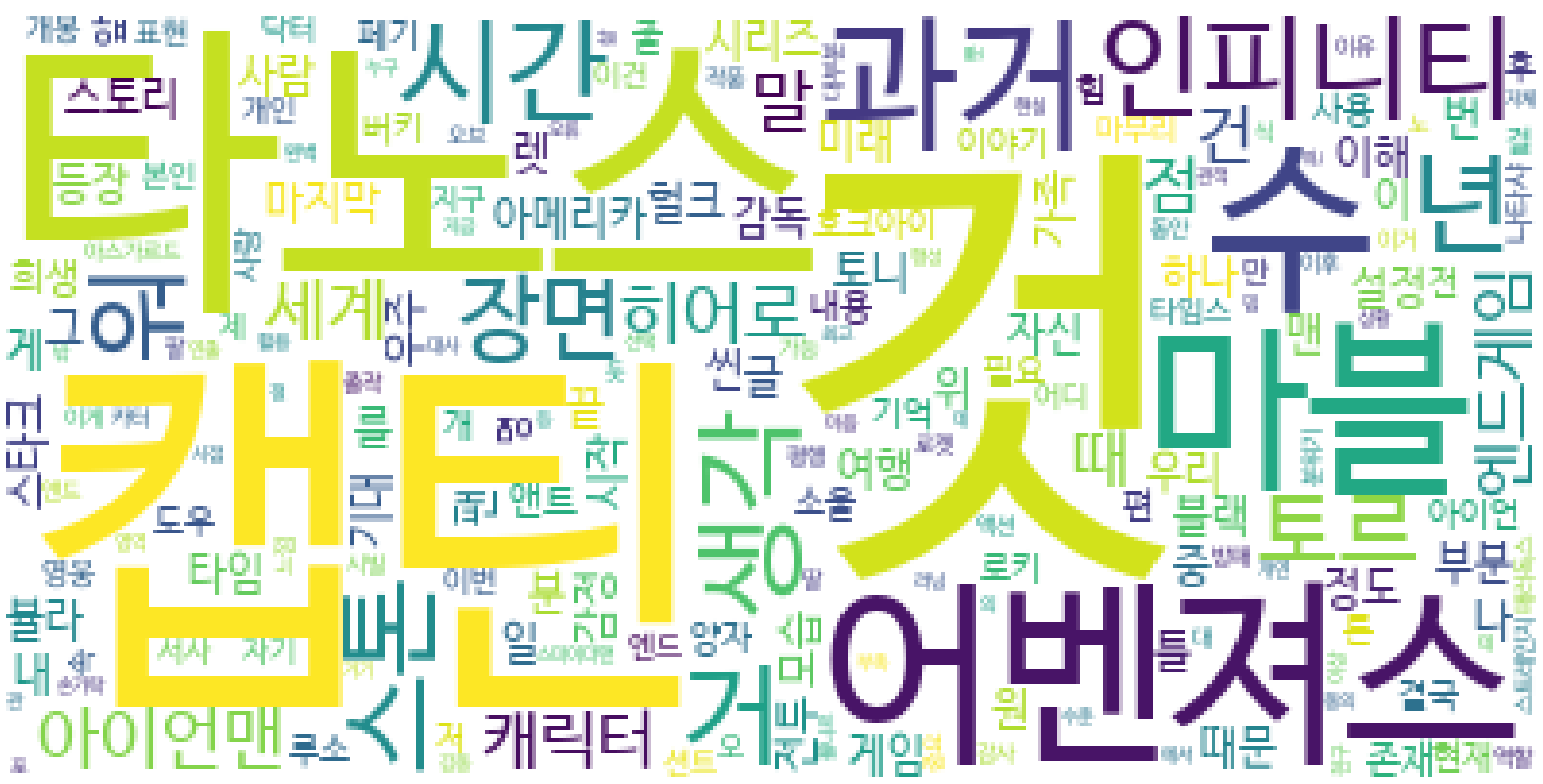
워드 클라우드 프로젝트

김경민 선생님

워드클라우드

워드클라우드란?

데이터에서 단어 빈도를 분석하여 시각화하는 기법



워드클라우드 준비

워드클라우드를 그리기 위해서 **텍스트 데이터**가 필요합니다.

네이버 **뉴스 기사**의 내용의 텍스트 데이터로

워드클라우드를 그려보도록 하겠습니다.

영어 문장 나누기

워드클라우드의 각 단어는 **빈도**에 따라 크기가 결정됩니다.

크기가 큰 단어일수록 **빈도**가 높습니다.

영어 문장의 경우, 공백을 기준으로 나누어

각각의 단어를 얻을 수 있습니다.

영어 문장 나누기

영어로 이루어진 텍스트 데이터가 주어집니다.

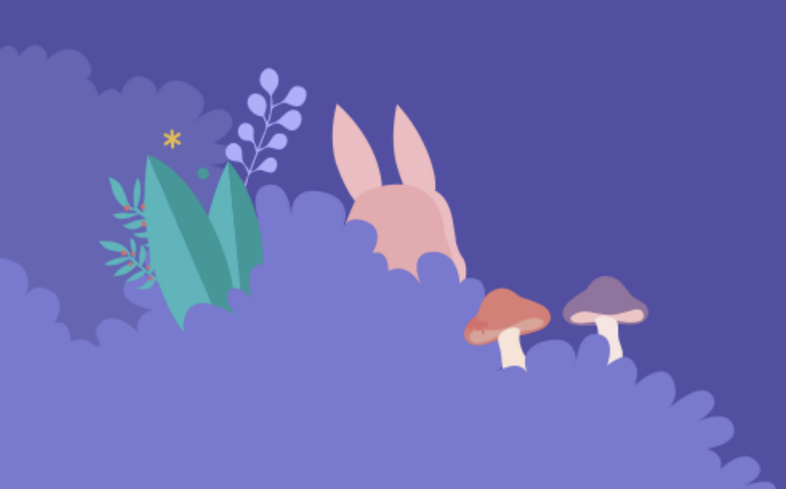
텍스트 데이터를 **공백**을 기준으로 나누어 빈도를 조사하고,

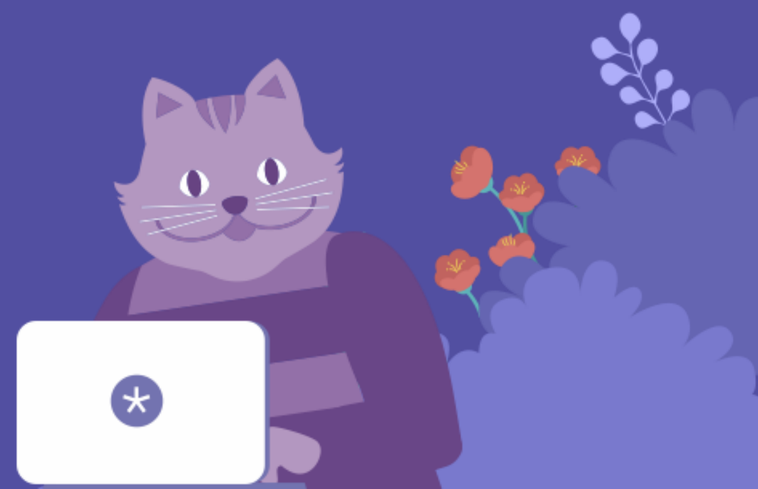
이를 바탕으로 워드클라우드를 그려보도록 하겠습니다.

/*elice*/

[실습1]

영어 문장 나누기

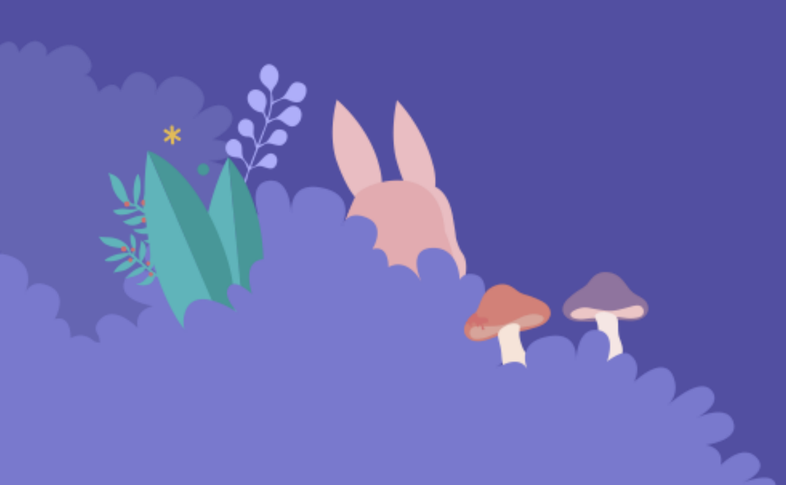


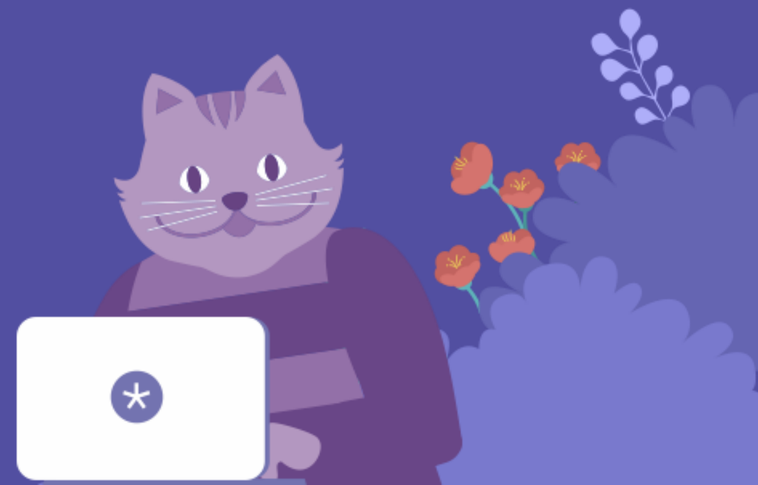


/*elice*/

[실습2]

워드클라우드 출력하기





네이버 뉴스 기사 워드클라우드

본격적으로 네이버 뉴스 기사의

워드클라우드를 그려보도록 하겠습니다.

이전 장의 실습에서 활용했던 코드로

네이버 뉴스 기사의 내용을 크롤링하세요.

네이버 뉴스 기사 워드클라우드

원하는 기사의 URL을 입력하시고,

워드클라우드를 출력해보신 후 출력된 모습을 관찰해보세요.

/*elice*/

[실습3]

네이버 뉴스 기사 내용 크롤링하기

/*elice*/

[실습4]

네이버 뉴스 기사 워드클라우드 출력하기

이전 실습의 문제점

이전 실습에서 그렸던 워드클라우드의 문제점은

단어에 **어미**와 **조사**가 붙어 분석이 왜곡되는 것입니다.

예를 들어 '대통령이' 와 '대통령은'은 둘 다

대통령이라는 공통된 키워드로 집계되어야 합니다.

형태소 추출

이를 추출하기 위해 한국어 단어에 붙는

어미와 조사를 제거하고, 단어의 어근만 집계되도록 하는

형태소 추출 과정이 필요합니다.

형태소 추출

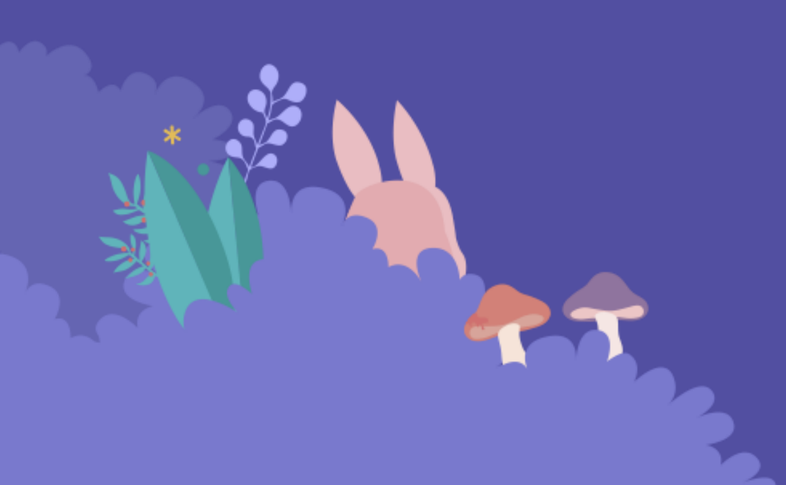
이 과정에서 한국어 자연어 처리 라이브러리인

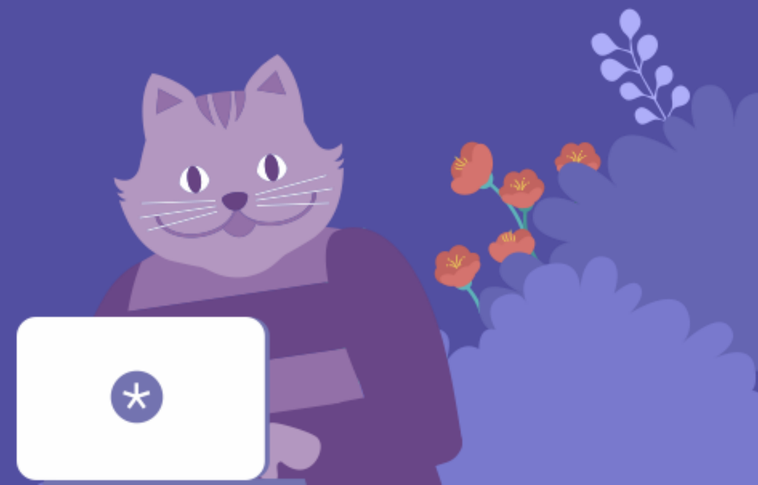
mecab을 사용합니다.

/*elice*/

[실습5]

형태소 추출하기

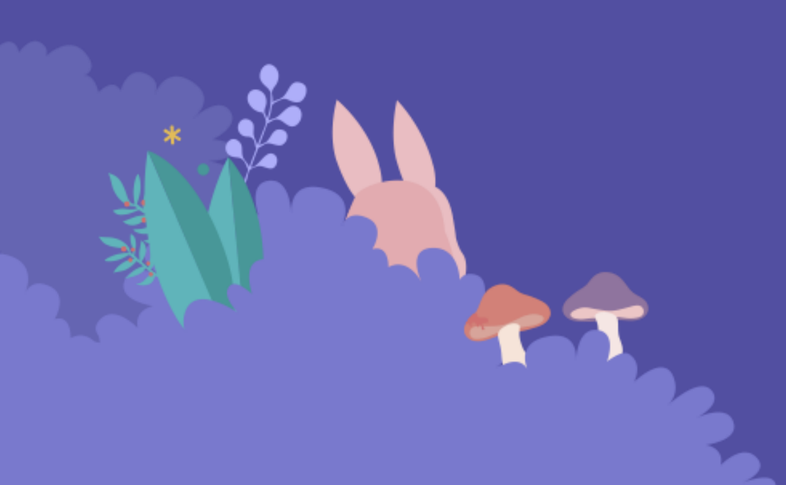


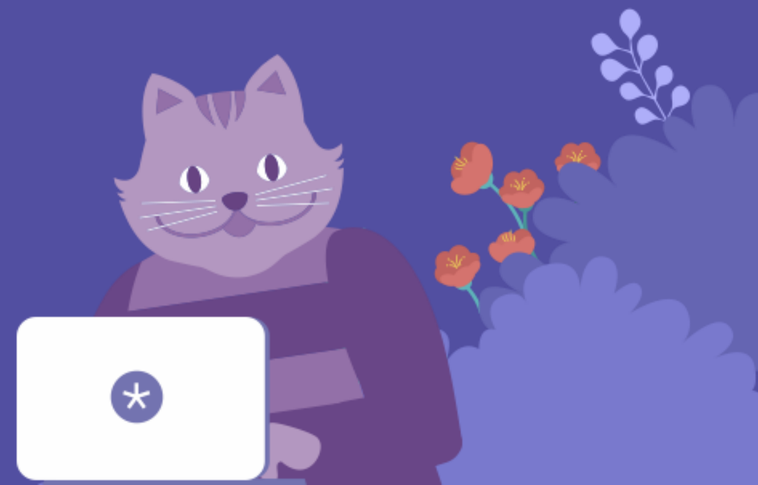


/*elice*/

[실습6]

형태소를 추출한 워드클라우드 출력하기





여러 개의 기사 내용 크롤링하기

하나의 기사만으로는 단어의 빈도수를 파악하기 어려울 수 있습니다.

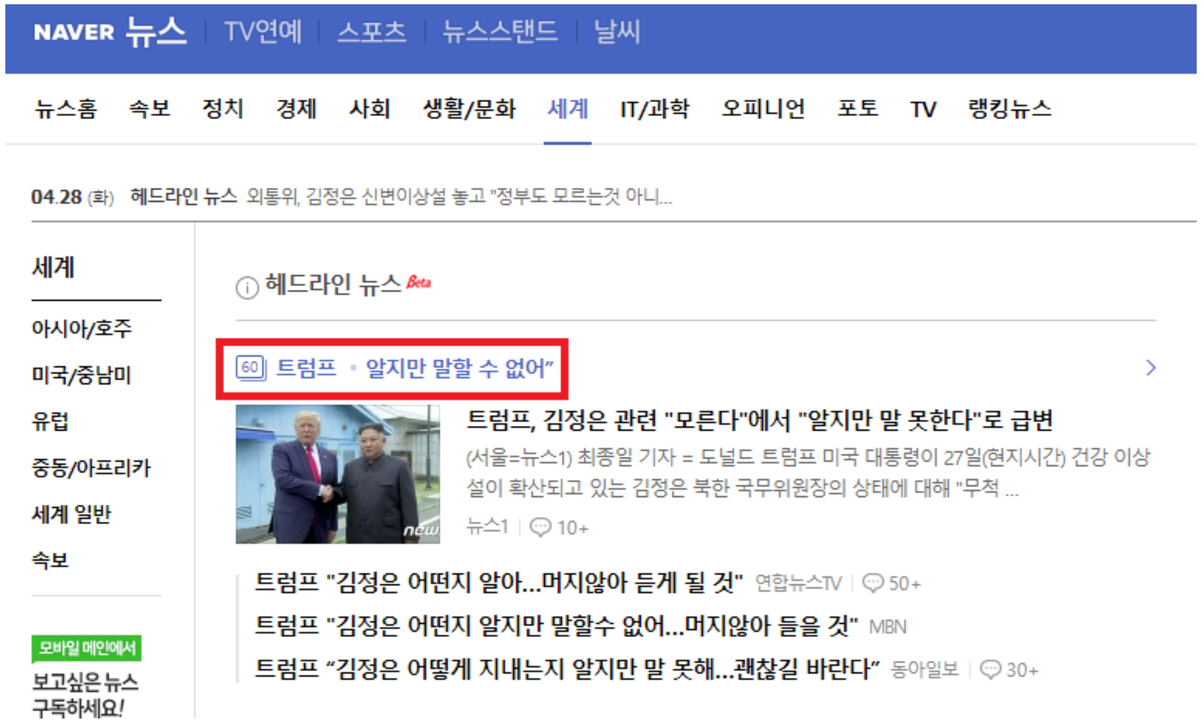
기사의 분량, 기자의 성향 등 여러 요인이 반영되기 때문입니다.

여러 개의 기사 내용 크롤링하기

따라서 공통된 주제에 대한 **여러 기사**의 텍스트 데이터를 같이 분석하면

효과적인 워드클라우드를 출력할 수 있습니다.

여러 개의 기사 내용 크롤링하기



네이버 뉴스 페이지는 관련된 주제의

여러 기사를 묶어서 보여주고 있습니다.

여러 개의 기사 내용 크롤링하기

각각의 분야(정치, 경제, 사회, 생활, 세계, 과학)에 대해

페이지 최상단에 보이는 **주제에 해당하는 기사**들의

텍스트 데이터로 워드클라우드를 출력해봅시다.

/*elice*/

[실습7]

여러 개의 기사 내용 크롤링하기

/*elice*/

[실습8]

여러 개의 기사 내용으로 워드클라우드 출력하기

더 많은 기사 내용 크롤링하기

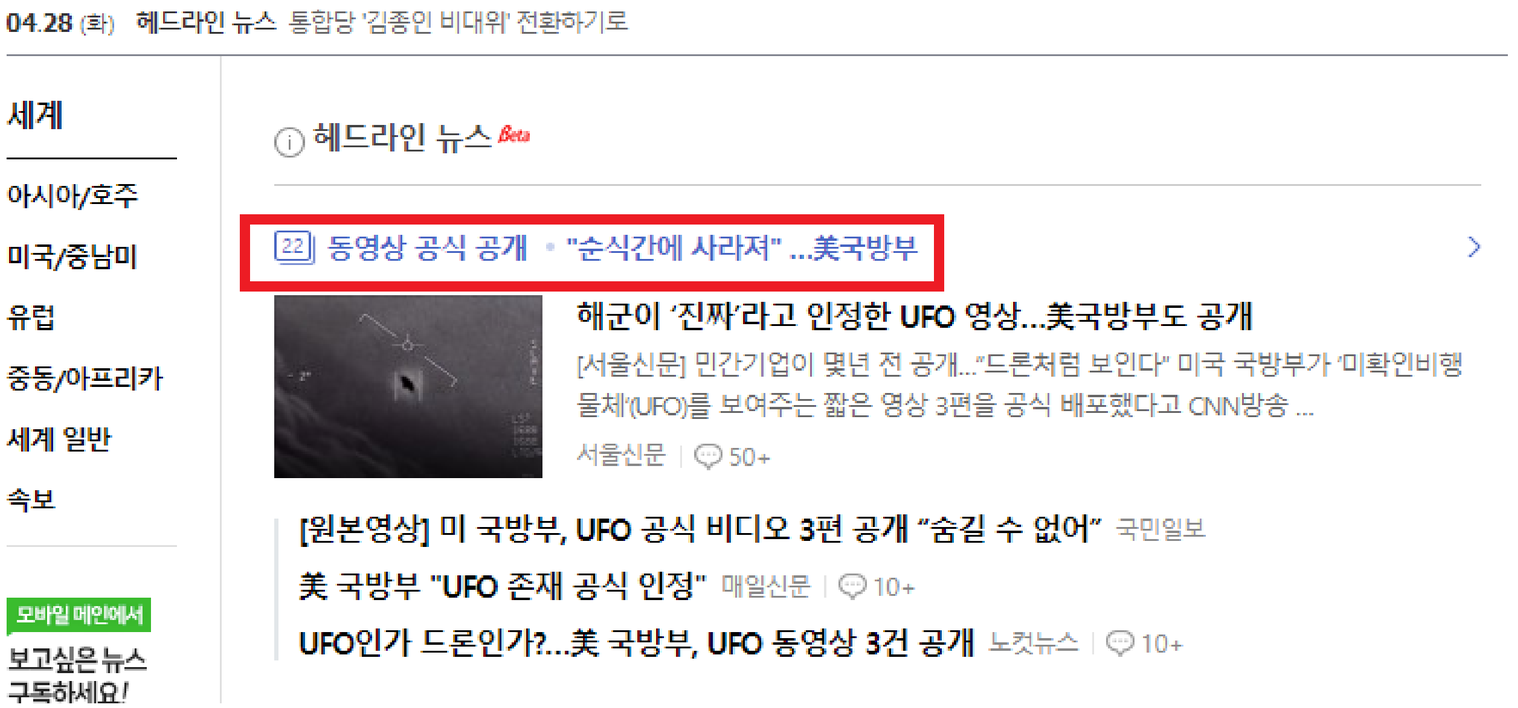
이전 실습으로 각 주제마다 **3~4개 기사**의

텍스트 데이터를 크롤링 할 수 있게 되었습니다.

이 상태에서, **더 많은 기사**의 내용을 크롤링하면

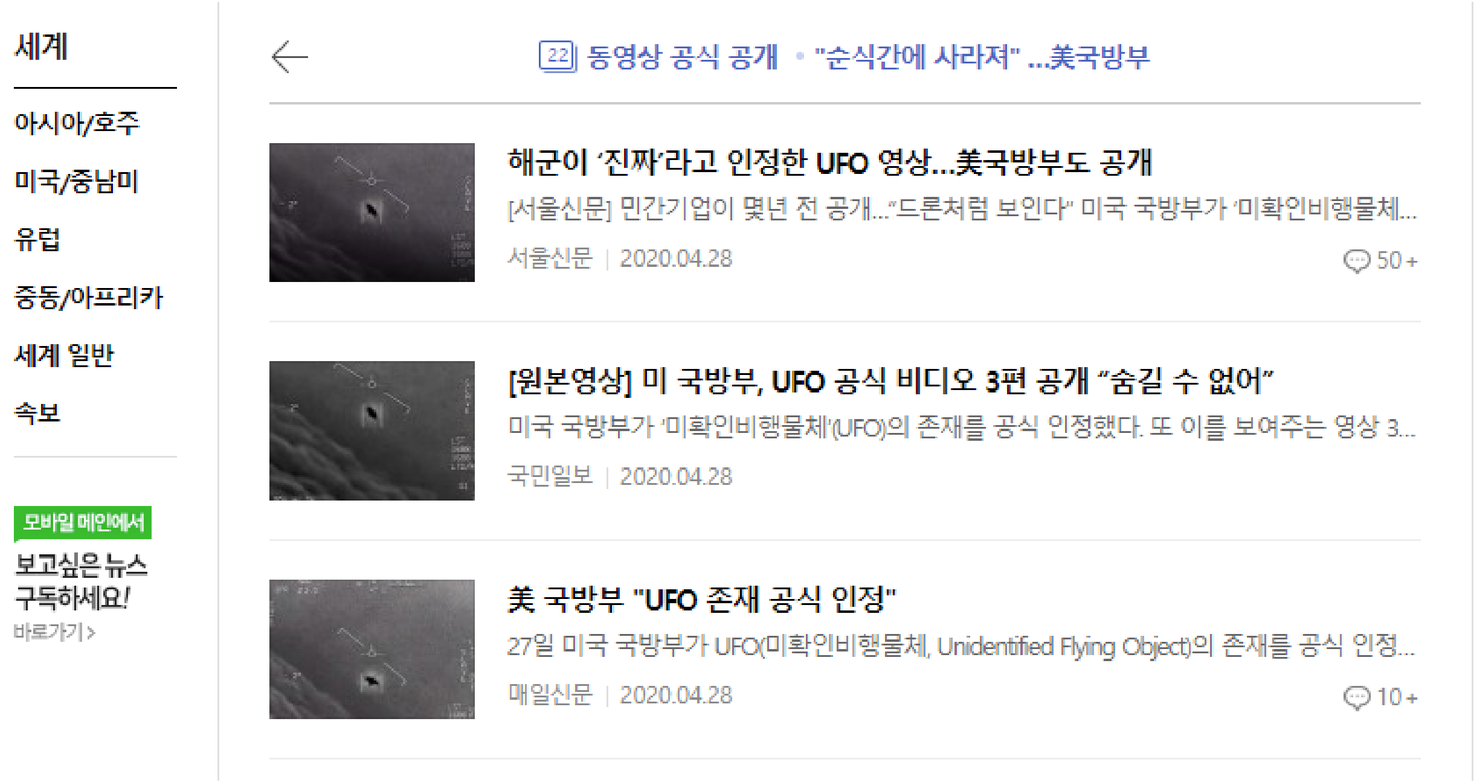
텍스트 데이터를 풍부하게 만들 수 있습니다.

더 많은 기사 내용 크롤링하기



기사 페이지에서 더 많은 기사를 확인할 수 있습니다.

더 많은 기사 내용 크롤링하기



세부 페이지에서 더 많은 기사에 각각 접근하실 수 있습니다.

/*elice*/

[실습9]

더 많은 기사 내용 크롤링하기

/*elice*/

[실습10]

더 많은 기사로 워드클라우드 출력하기

CREDIT

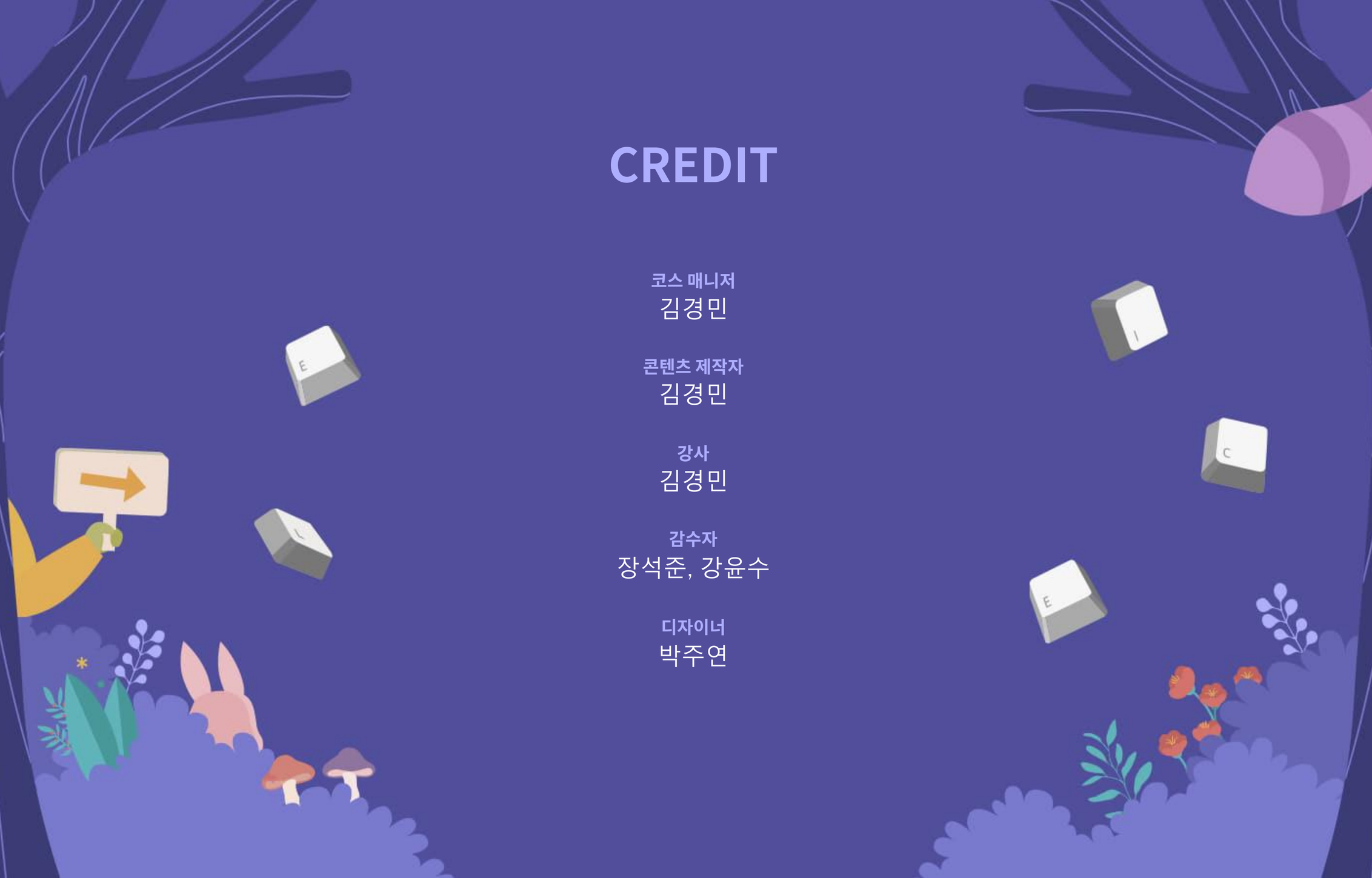
코스 매니저
김경민

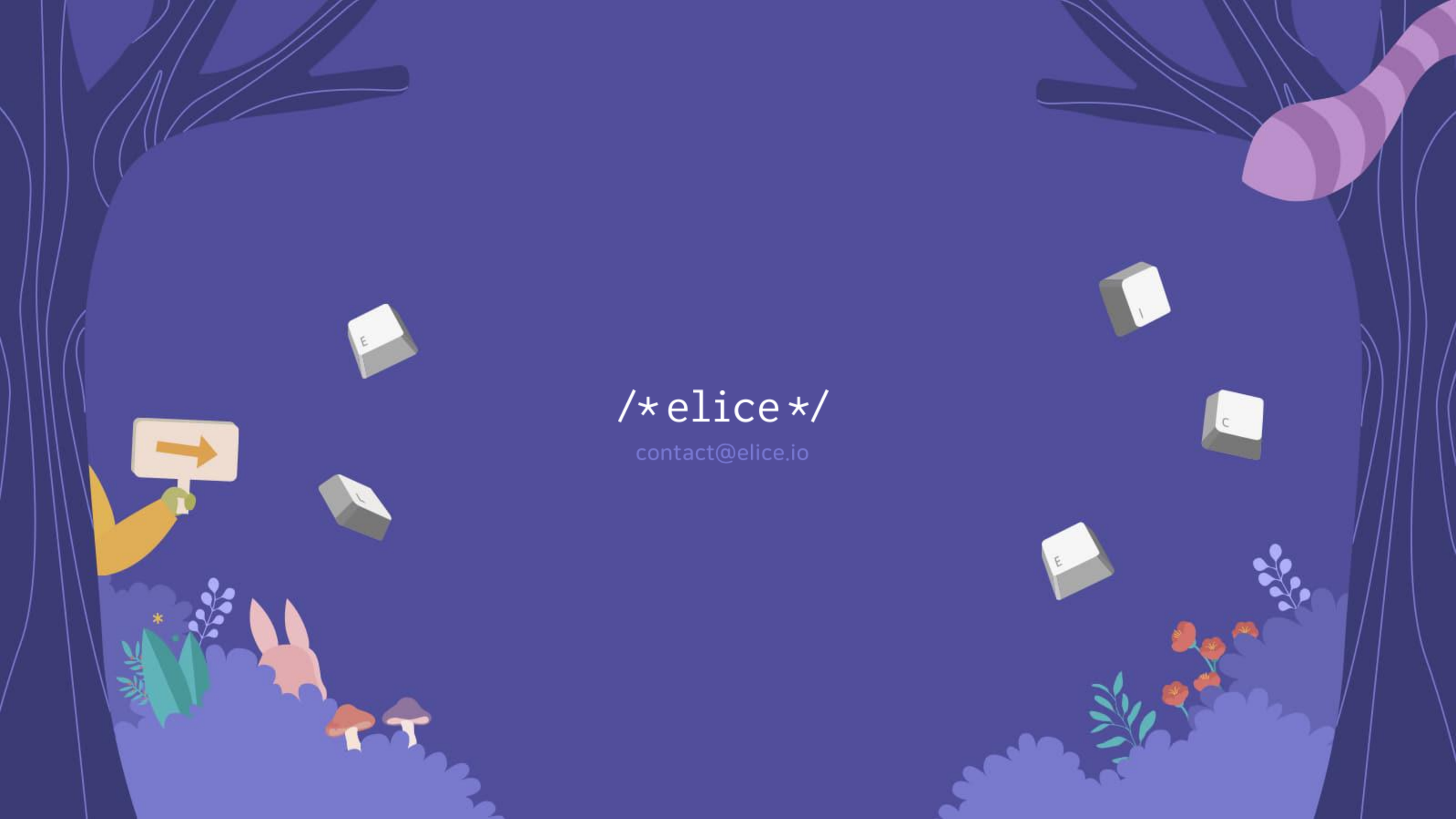
콘텐츠 제작자
김경민

강사
김경민

감수자
장석준, 강윤수

디자이너
박주연




/*elice*/
contact@elice.io