

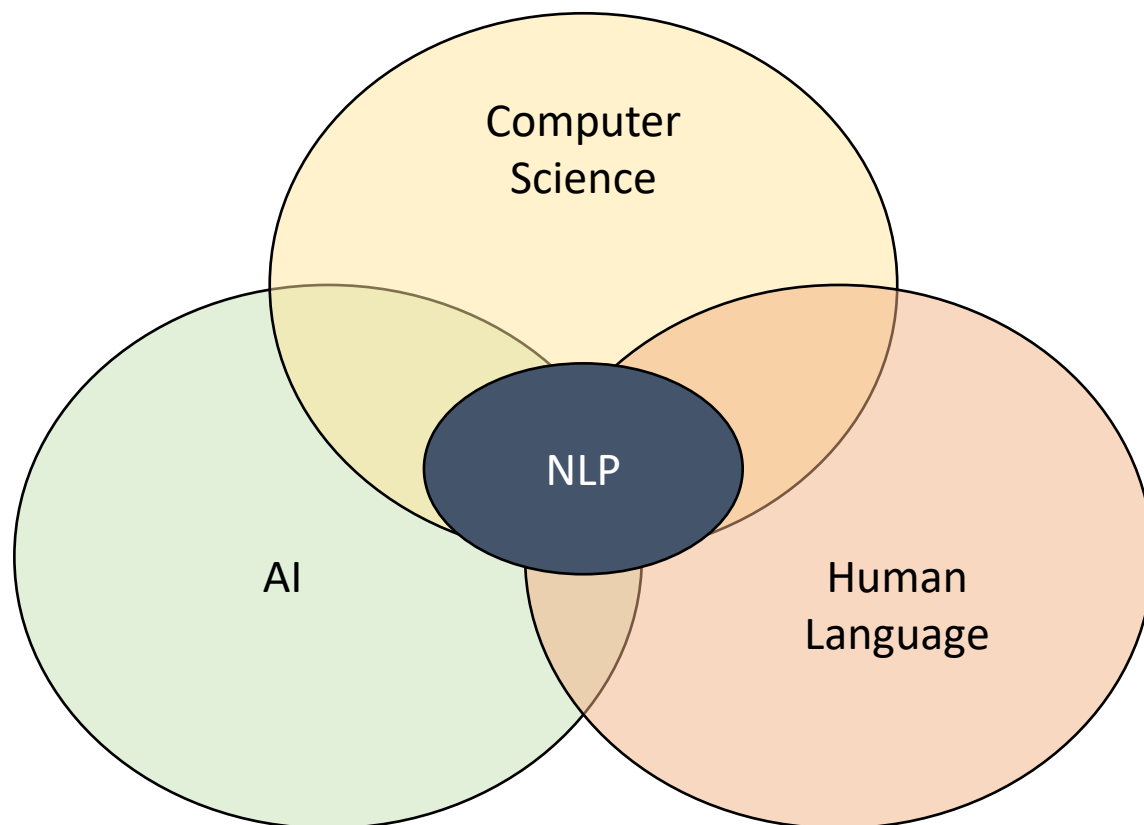
강원대학교  
AI 소프트웨어학과

---

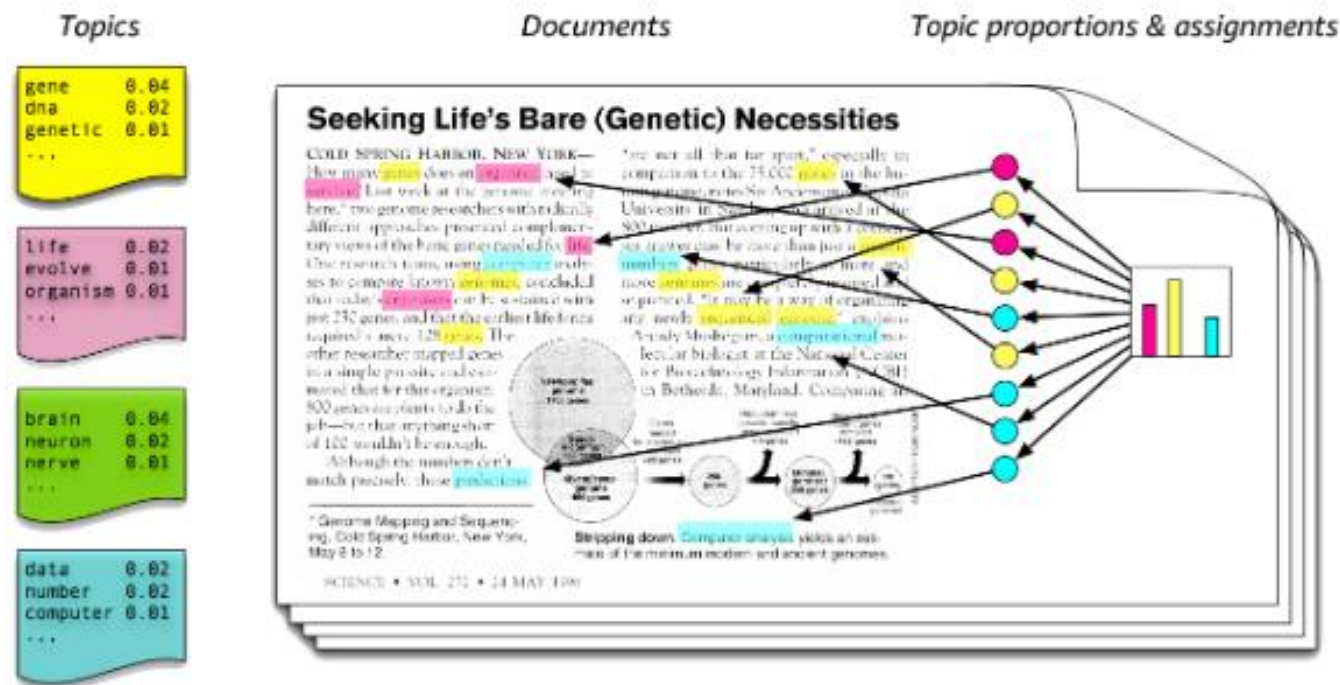
데이터 전처리  
- 텍스트 데이터 전처리 -

---

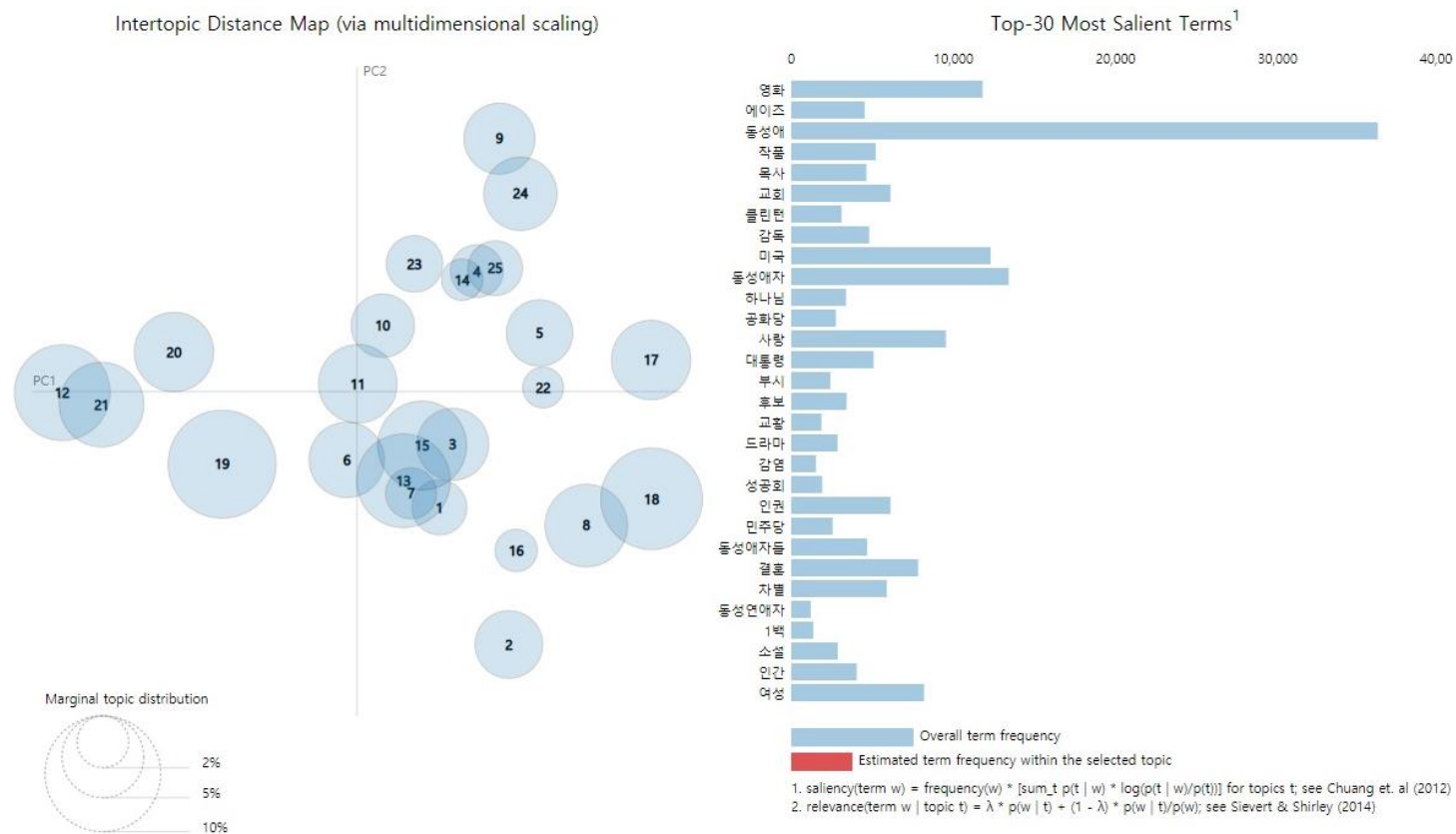
- NLP(Natural Language Processing)는 기계가 사람의 언어에 대해 처리하는 계산적 기술의 집합  
→ 의미분석, 감성분석, 음성인식, 번역 등이 존재



- 토픽모델링(Topic Modeling) : 단어, 말뭉치(corpus)로 부터 숨겨진 의미를 찾고 키워드별로 주제를 묶어 주는 모델로 문서에 대한 확률 분포를 가정해 분류해주는 방법



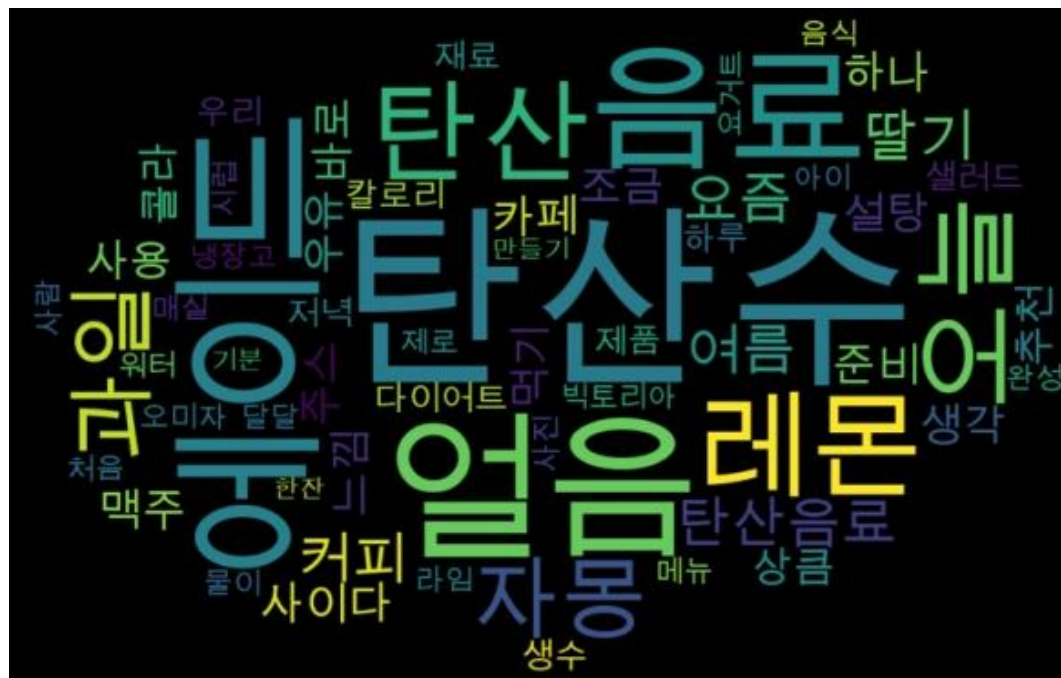
- 토픽모델링(Topic Modeling) : 단어, 말뭉치(corpus)로 부터 숨겨진 의미를 찾고 키워드별로 주제를 묶어 주는 모델로 문서에 대한 확률 분포를 가정해 분류해주는 방법



- 워드 클라우드(Word Cloud)는 텍스트를 분석해 사람들의 관심사, 키워드, 개념 등을 파악할 수 있도록 빈도수를 단순히 카운트하여 시각화 시킨 방법



- **활용 사례 : 사람들의 댓글 및 의견들을 통해 새로운 가치 및 의미를 찾는 것에 활용**  
→ **음식의 새로운 조합, 사람들의 흥미요소, 전혀 연관이 없는 새로운 가치**

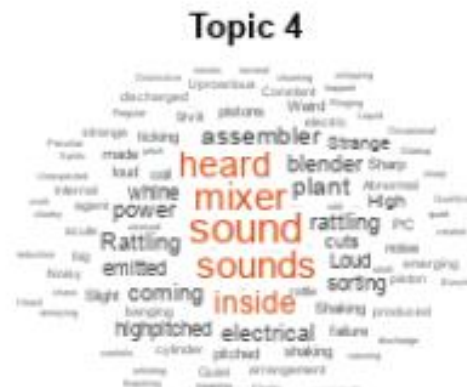




- 활용 사례 : 사람들의 댓글 및 의견들을 통해 새로운 가치 및 의미를 찾는 것에 활용  
→ 음식의 새로운 조합, 사람들의 흥미요소, 전혀 연관이 없는 새로운 가치



- **활용 사례 : 음악의 장르 파악, 논문의 키워드를 파악, 글쓴이의 성향을 파악**





- 감성 분석(Sentiment Analysis)이란 텍스트에 들어있는 의견이나 감성, 평가, 태도 등의 주관적인 정보를 컴퓨터를 통해 분석하는 과정

‘백신 접종’ 관련 SNS 키워드 감성 분석

(SNS Data: 유튜브 외 4)



긍정

1. 백신	65,009 건
2. 현황	42,366 건
3. 마스크	42,332 건



중립

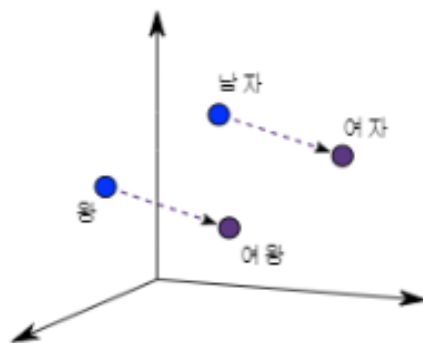
1. 서울	19,348 건
2. 공무원	14,003 건
3. 노인	13,092 건



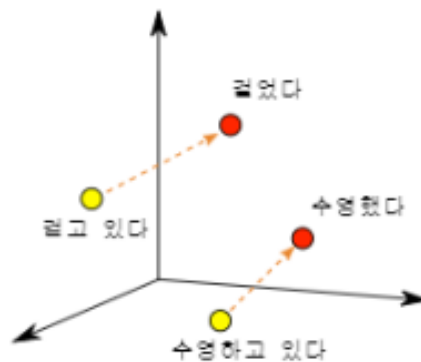
부정

1. 코로나	106,344 건
2. 확진자	42,340 건
3. 바이러스	34,509 건

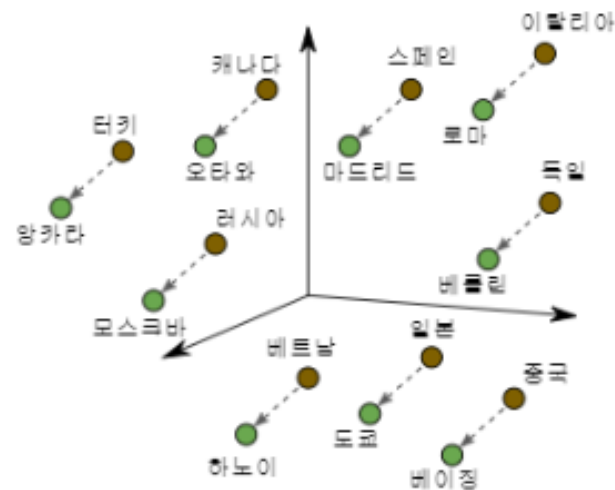
- NLP를 하기위해 텍스트를 컴퓨터가 이해할 수 있도록 숫자로 바꾸는 작업이 필요함
- 사람의 경우는 문맥을 통해 문장 및 의미를 구별하는 것이 가능함 → 임베딩(Embedding)
- 즉 자연어를 수치화 한 것으로 벡터로 표현하는 것을 말하고 임베딩은 그 과정까지 모두를 포함하는 의미



남자-여자

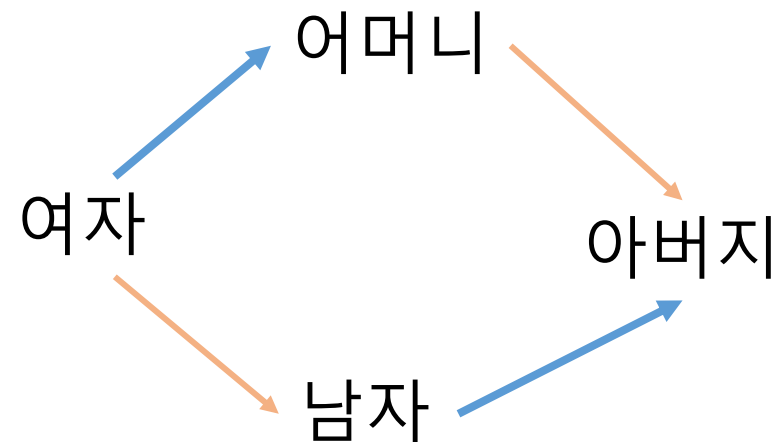
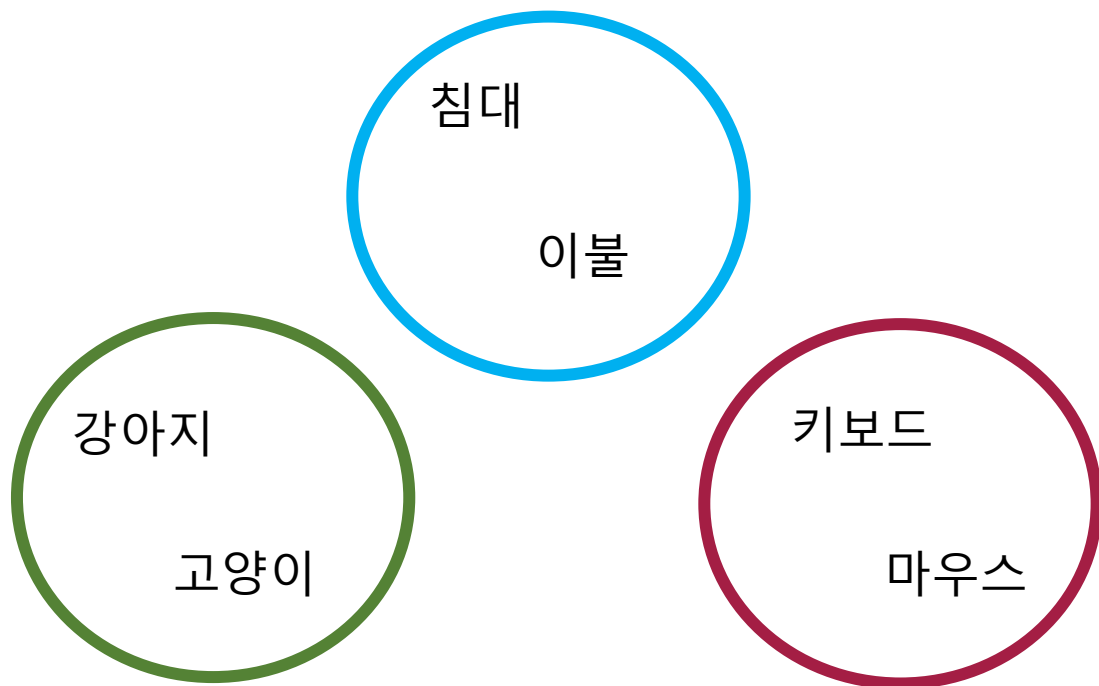


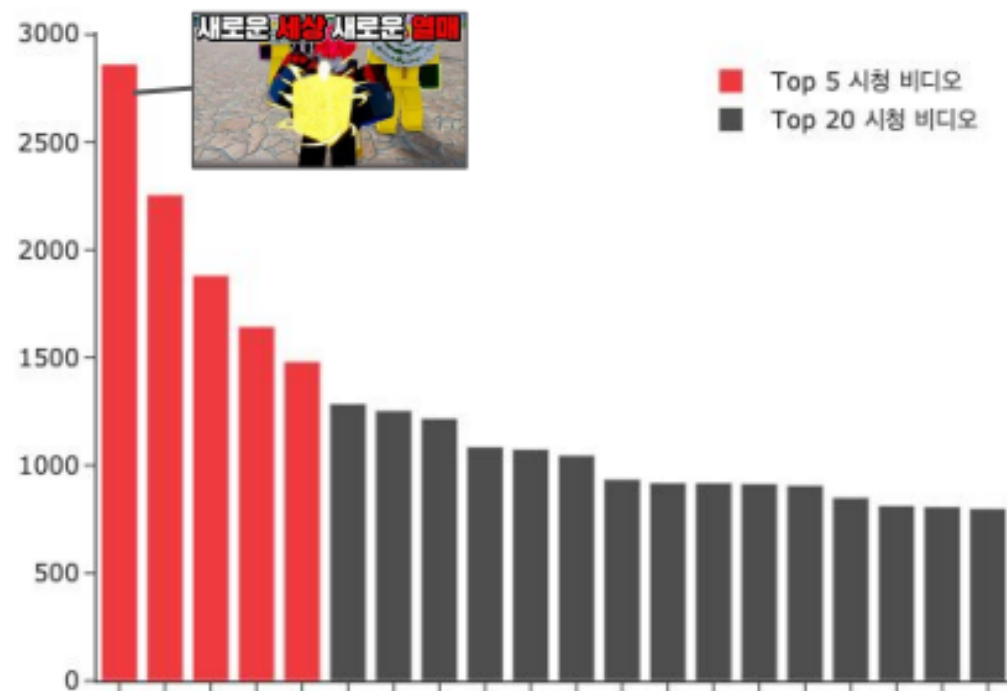
동사 시제



국가-수도

- 임베딩(Embedding) : 전체 단어들 간의 관계에 맞춰 해당 단어의 특성을 갖는 벡터로 바꿔주므로 단어 간의 의미를 파악해 문법적 관계를 알 수 있음





\* 그래프에 첨부된 링크를 열어 자세히 확인할 수 있습니다.

## 채널의 시청시간이 가장 높은 영상은 새로운 열매 콘텐츠입니다.

가장 높은 시청시간을 갖고있는 영상의 경우 새로운 아이템 혹은 how to를  
다른 정보성 영상으로 로블록스 게임에 적용하고자 하는 니즈를 갖고  
있습니다.

해당 채널의 총 시청 시간(hr) : 59,735.22

상위 5개 비디오의 시청 시간(hr) : 10,126.0

상위 5개 비디오의 시청 시간 점유율 : 16.95%

# 01

블록스피스

검색 횟수 : 38,048

전체 검색 대비: 24.0%

02

비밀

검색 횟수 : 22,774

전체 검색 대비: 14.36%

03

공공 대안 프로그램

검색 횟수 : 8,116

전체 검색 대비: 5.12%

04

로봇복스

검색 횟수 : 7,022

전체 검색 대비: 4.43%

06

## 탕탕특공대 챕터1

검색 횟수 : 6,070

전체 검색 대비: 3.83%

07

탕탕특공대 챕터5

검색 횟수 : 5,388

전체 검색 대비: 3.4%

08

## 꼬임

검색 횟수 : 5,336

전체 검색 대비: 3.37%

09

별피 코드

검색 횟수 : 3,425

전체 검색 대비: 2.16%





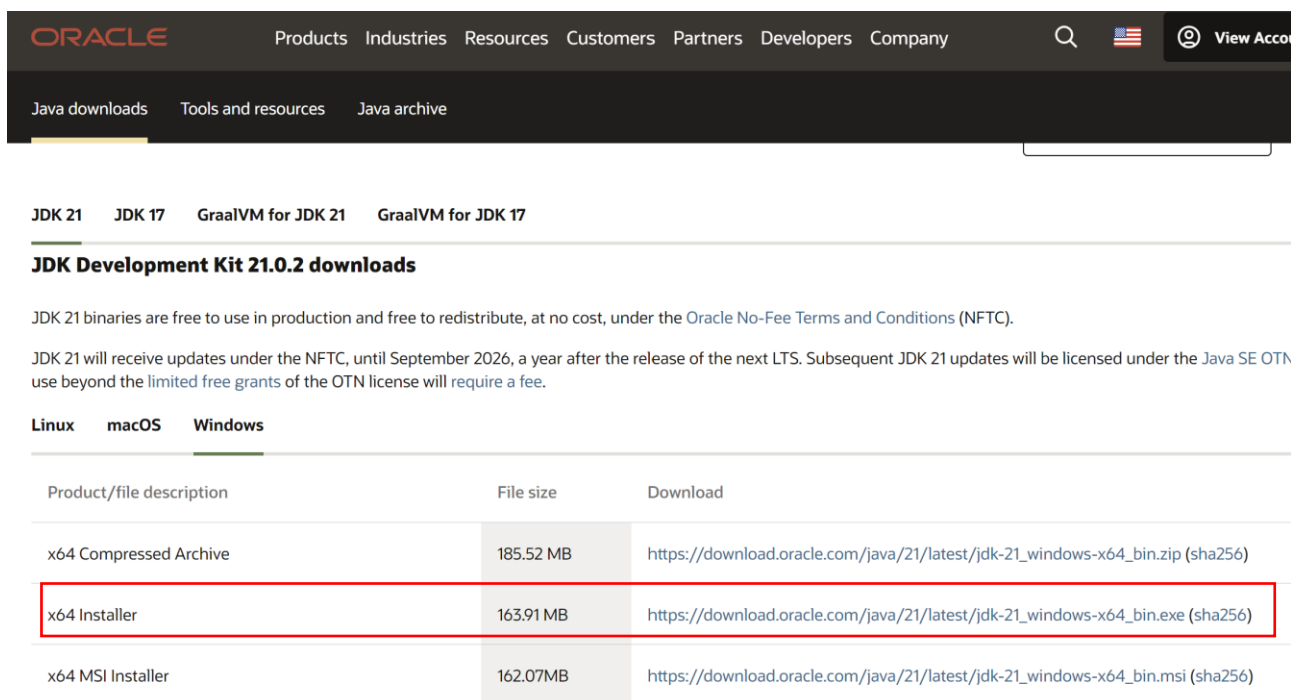
다시 도약하는 대한민국  
함께 잘사는 국민의 나라

신직업	1인 미디어 특화 데이터 분석가			
정의	소셜 미디어 데이터를 수집 및 분석해 마케팅 전략을 도출하고 콘텐츠 가치에 따른 비즈니스 정책 수립			
필요 역량	<ul style="list-style-type: none"> <li>데이터 수집, 분석, 시각화</li> <li>비즈니스 매니지먼트, 전략</li> </ul>			
교육	기간	단기 <input type="checkbox"/>	중기 <input checked="" type="checkbox"/>	장기 <input type="checkbox"/>
	난이도	하 <input type="checkbox"/>	중 <input checked="" type="checkbox"/>	상 <input type="checkbox"/>
현황	<ul style="list-style-type: none"> <li>인플루언서 마케팅, 기업의 소셜 미디어 마케팅이 활성화되며 소셜 미디어 내 데이터 분석을 통한 판매, 마케팅 전략의 중요성 증대</li> </ul>			
향후 전망	<ul style="list-style-type: none"> <li>소셜 미디어의 데이터 수집, 분석과 이를 활용한 비즈니스 전략을 마련할 수 있는 분석가에 대한 수요가 높아질 것으로 예상(전문가 양**)</li> </ul>			

!pip install konlpy

Open Korean Text란 것이 JAVA환경으로 구성되어 있기 때문

<https://www.oracle.com/java/technologies/downloads/#jdk21-windows>



ORACLE Products Industries Resources Customers Partners Developers Company

Java downloads Tools and resources Java archive

JDK 21 JDK 17 GraalVM for JDK 21 GraalVM for JDK 17

**JDK Development Kit 21.0.2 downloads**

JDK 21 binaries are free to use in production and free to redistribute, at no cost, under the [Oracle No-Fee Terms and Conditions \(NFTC\)](#).

JDK 21 will receive updates under the NFTC, until September 2026, a year after the release of the next LTS. Subsequent JDK 21 updates will be licensed under the Java SE OTN use beyond the [limited free grants](#) of the OTN license will [require a fee](#).

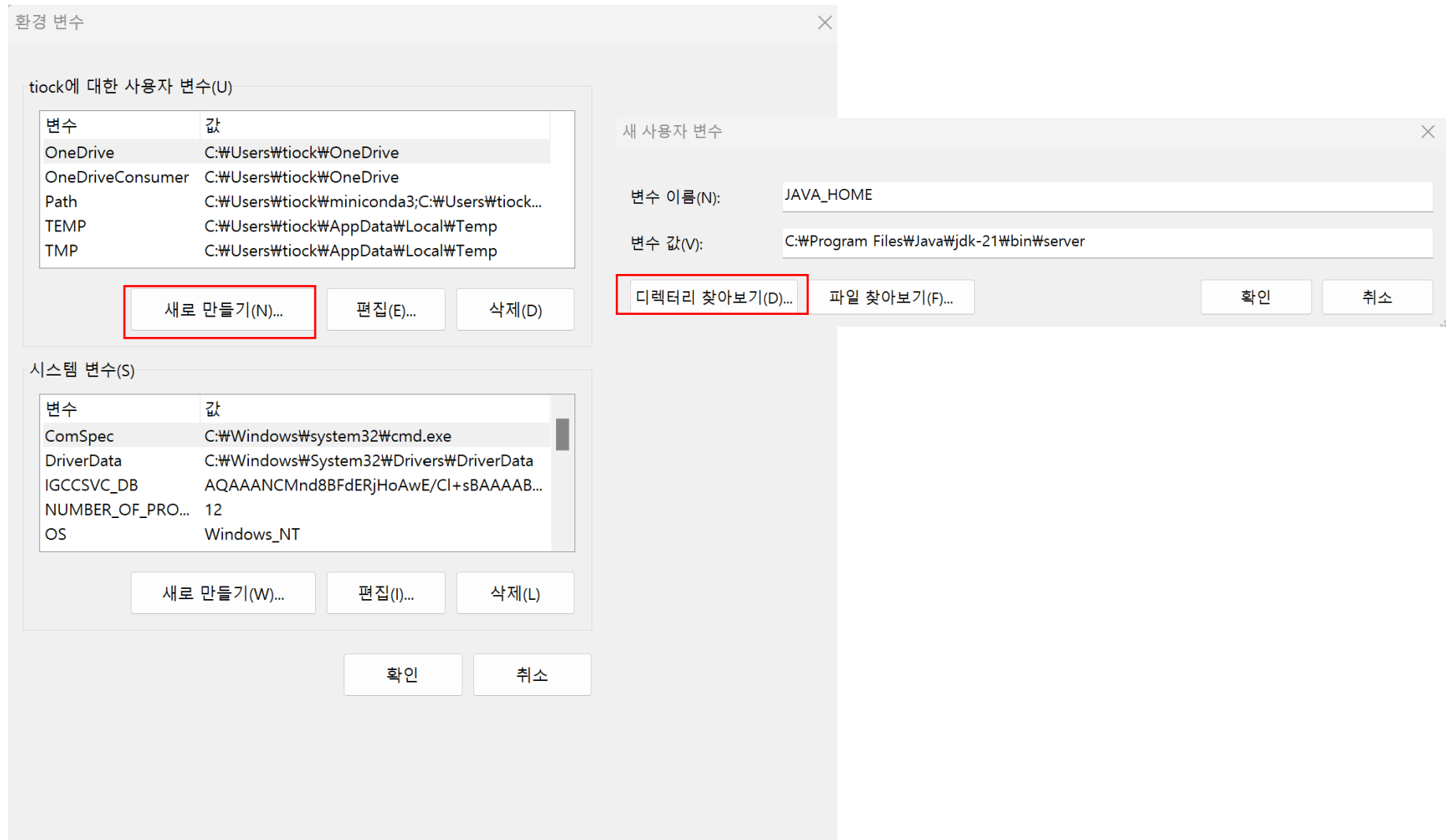
Linux macOS **Windows**

Product/file description	File size	Download
x64 Compressed Archive	185.52 MB	<a href="https://download.oracle.com/java/21/latest/jdk-21_windows-x64_bin.zip">https://download.oracle.com/java/21/latest/jdk-21_windows-x64_bin.zip</a> (sha256)
x64 Installer	163.91 MB	<a href="https://download.oracle.com/java/21/latest/jdk-21_windows-x64_bin.exe">https://download.oracle.com/java/21/latest/jdk-21_windows-x64_bin.exe</a> (sha256)
x64 MSI Installer	162.07MB	<a href="https://download.oracle.com/java/21/latest/jdk-21_windows-x64_bin.msi">https://download.oracle.com/java/21/latest/jdk-21_windows-x64_bin.msi</a> (sha256)

시스템 환경 변수 편집 → 환경변수 → 새로 만들기  
→ 변수 이름 : JAVA\_HOME, 변수 값 : 디렉토리  
로 찾아 입력



시스템 환경 변수 편집  
→ 환경변수 → 새로  
만들기 → 변수 이름 :  
JAVA\_HOME, 변수  
값 : 디렉토리로 찾아  
입력



python =3.10.9 버전 가상환경 설치

pip install -r requirements.txt

python -m ipykernel install --user --name=커널이름 --display-name '원하는이름'



```
from konlpy.tag import Okt
import re
```

```
# Okt 형태소 분석기 객체 생성
okt = Okt()
```

```
# 문장 텍스트
```

```
text = "나는 AI소프트웨어학과에서 김창균 교수님의 강의로 데이터전처리에서 자연어 처리를 배우고 있습니다. 이제 한글 텍스트 마이닝을 시작해보겠습니다!"
```

```
# 1. 텍스트 전처리
```

```
clean_text = re.sub(r'[^가-힣Ws]', '', text)
```

→ 한글과 공백을 제외한 모든 문자 제거, `Ws`: 공백 문자 `^`: 부정, 한글의 모든 음절:가-힣

```
# 2. 단어 토큰화
```

```
words = okt.morphs(clean_text) → 문장을 형태소로 분해하여 단어 토큰화# 3. 형태소 분석
```

```
pos_tags = okt.pos(clean_text) → 형태소에 품사 태깅
```

```
from konlpy.tag import Okt  
import re
```

```
print(("전처리된 텍스트:{}".format(clean_text))  
print(("단어 토큰화 결과:{}".format(words))  
print(("형태소 분석 결과:{}".format(pos_tags))
```

## 명사(Noun), 조사(Josa), 접두사(Suffix), 형용사(Adjective), 동사(Verb)

전처리된 텍스트:나는 소프트웨어학과에서 김창균 교수님의 강의로 데이터전처리에서 자연어 처리를 배우고 있습니다 이제 한글 텍스트 마이닝을 시작해보겠습니다

단어 토큰화 결과:['나', '는', '소프트웨어', '학과', '에서', '김창균', '교수', '님', '의', '강의', '로', '데이터', '전', '처리', '에서', '자연어', '처리', '를', '배우고', '있습니다', '이제', '한글', '텍스트', '마', '이닝', '을', '시작', '해보겠습니다']

형태소 분석 결과:[('나', 'Noun'), ('는', 'Josa'), ('소프트웨어', 'Noun'), ('학과', 'Noun'), ('에서', 'Josa'), ('김창균', 'Noun'), ('교수', 'Noun'), ('님', 'Suffix'), ('의', 'Josa'), ('강의', 'Noun'), ('로', 'Josa'), ('데이터', 'Noun'), ('전', 'Modifier'), ('처리', 'Noun'), ('에서', 'Josa'), ('자연어', 'Noun'), ('처리', 'Noun'), ('를', 'Josa'), ('배우고', 'Verb'), ('있습니다', 'Adjective'), ('이제', 'Noun'), ('한글', 'Noun'), ('텍스트', 'Noun'), ('마', 'Noun'), ('이닝', 'Noun'), ('을', 'Josa'), ('시작', 'Noun'), ('해보겠습니다', 'Verb')]

## Stopword

- 불용어는 언어에서 일반적으로 사용되는 단어로, 분석에 덜 의미 있는 정보라 필터링 되는 경우가 많음
- 불용어를 제거하면 텍스트에서 더 중요한 단어에 집중하는 데 도움이 됨(input\_sample.csv)

이 문장에는 단어1과 단어2가 존재합니다. 하지만 단어2는 매우 불필요한 단어 입니다.



이 문장 과 존재

stopwords - Windows 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

|불 필요

매우

단어

## 감성분석

- 감정 분석은 텍스트에 표현된 감정적 어조나 태도를 결정하는 데 사용되는 자연어 처리(NLP) 기술
- 사용된 단어와 문맥에 따라 텍스트를 긍정적, 부정적, 중립과 같은 범주로 분류(output\_sample.csv)

"이 영화는 정말 재미와 흥미를 나에게 주었고, 스케일이 엄청나다."

\*positive\_words - Windows 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

재미+1

흥미+1

엄청나다+2

" 이 영화는 주인공의 연기가 노잼이어서 싫다. 하지만 흥미는 있었다. "

\*negative\_words - Windows 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

나쁘다+1

노잼+1

싫다+2

## 워드클라우드

- 각 단어의 크기가 텍스트에서의 빈도나 중요성을 시각적으로 표현함 (output\_sample.csv)



# 최소 단어 길이  
min\_word\_length = 2

# 빈도 설정  
min\_word\_frequency = 2




## 동의어처리

- 동의어는 다른 단어와 동일하거나 거의 동일한 의미를 갖는 단어
- 텍스트를 풍부하게 하고 반복을 방지하여 언어의 다양성을 제공하는 데 사용(text\_dict.csv)

가라케가 갈라카이 가지말라 케가 안갈라카이 가라가라 카드라



 \*word\_replacement\_rules - Windows 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

가라고 해서 : 가라케가, 가라캐가

가려고 하니 : 갈라카이

해서 : 케가, 캐가

계속가라 : 가라가라

하더라 : 카드라, 카더라