A background image showing a riot scene. In the foreground, there are police officers wearing helmets and holding large, transparent riot shields. Behind the shields, a crowd of people is visible, some holding flags, including a South African flag. A banner in the background reads "NO SHIRMS TO WHITE SUPREMACY". The scene is outdoors on a street with buildings in the background.

\*본 발표에는 혐오감을 유발할 수 있는 문구  
및 내용이 포함되어 있습니다.

---

머신 러닝을 통한 자연어 처리 기술의 기초

# 인공지능 기반 혐오 발언 분석에 관한 연구

-10729 류병우

---



# 연구 배경



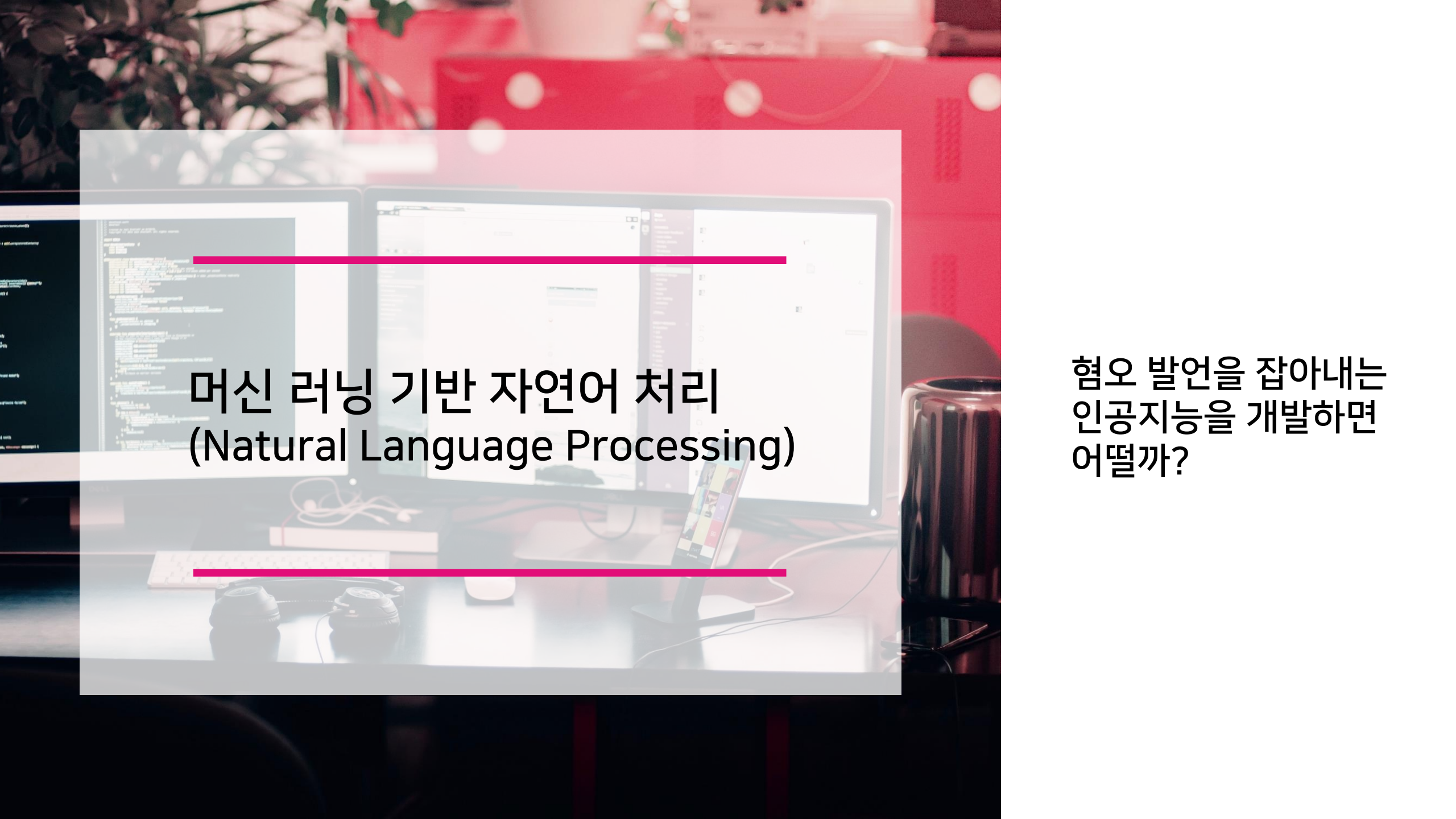
## 젠더 갈등에서 탄생한 “워마드”

최근 국내 여혐 및 남혐 문제를 기반으로 힘을 얻고 있는 혐오 사이트, 워마드

혐오 발언을 기초로 낙태 태아 훼손 사진 업로드, '성체 모독' 논란 등 사회적 갈등 확산

> 온라인 **혐오 발언**을 통해 퍼지는 갈등





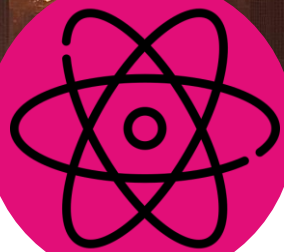
# 머신 러닝 기반 자연어 처리 (Natural Language Processing)

혐오 발언을 잡아내는  
인공지능을 개발하면  
어떨까?

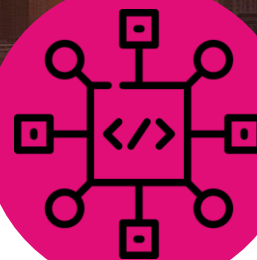
# 목차



1.  
머신 러닝 이론 기초



2.  
데이터 수집 과정



3.  
자연어 처리 기법을  
통한 인공지능 개발



4.  
자연어 처리 응용



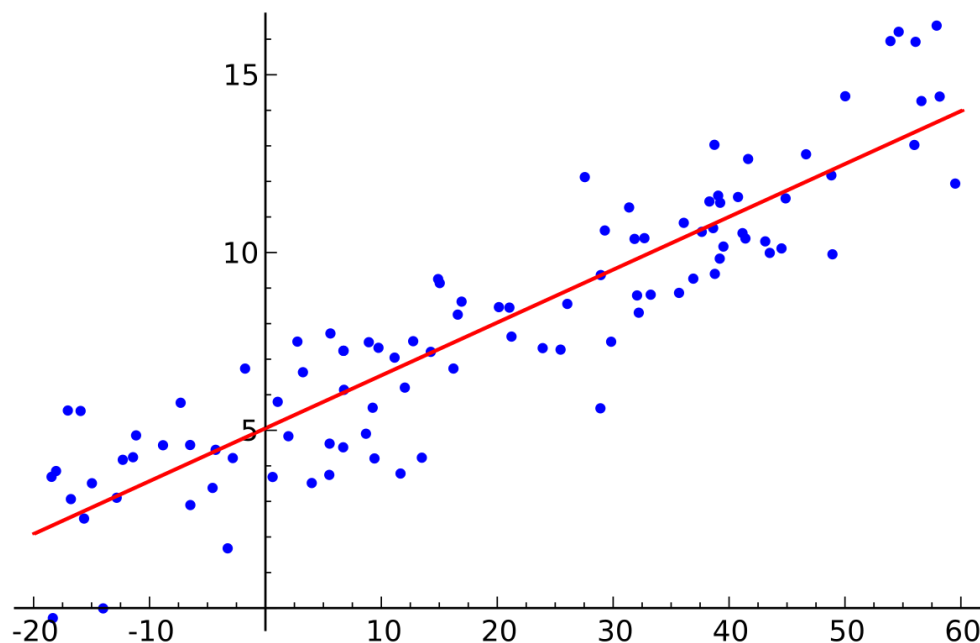
# 1. 머신 러닝 이론 기초

· 머신 러닝의 정의: 머신 러닝이란, 대량의 정보를 통해 기계를 학습시켜 기존에 알지 못했던 사실을 예측하는 알고리즘을 개발하는 분야이다.

Ex) 선형회귀모델  $f(x) = a_0x_0 + a_1x_1 + \dots + a_nx_n + b$

· 응용: 인구증가예측, 분야별 기사 분류 등

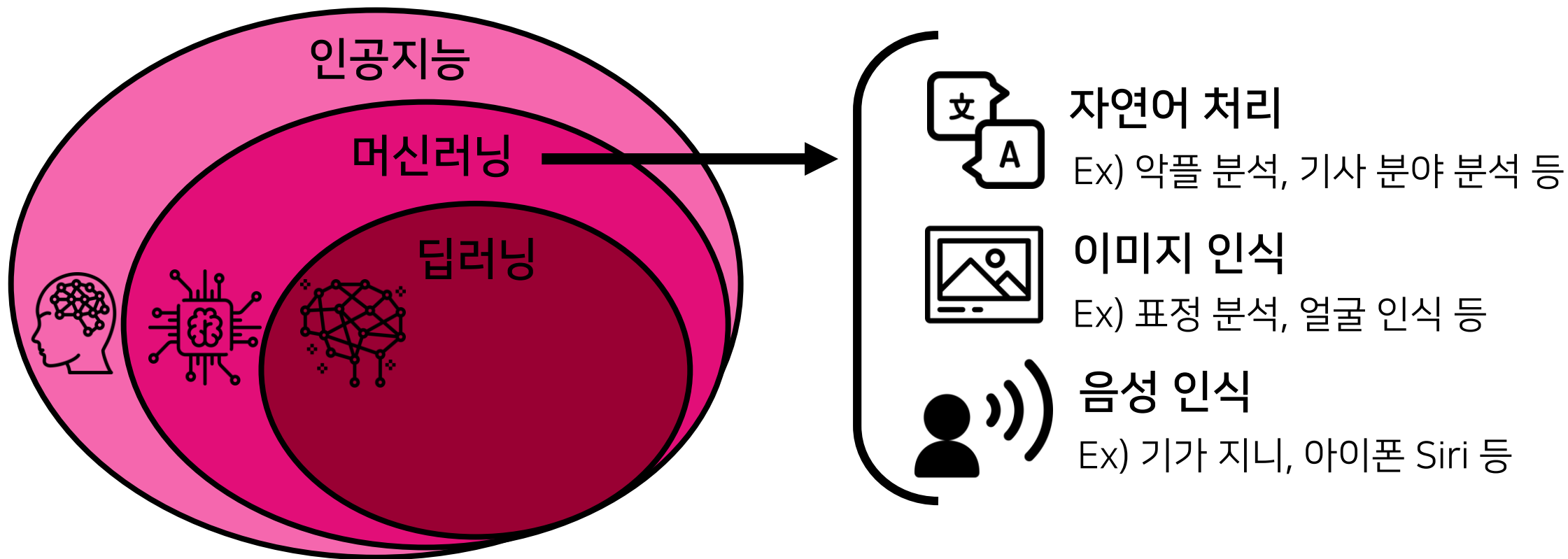
· 비유: 학생이 교과서 내용을 공부한 뒤 학습 내용을 바탕으로 시험 문제 정답을 고르는 것과 유사하다.







# 1. 머신 러닝 이론 기초



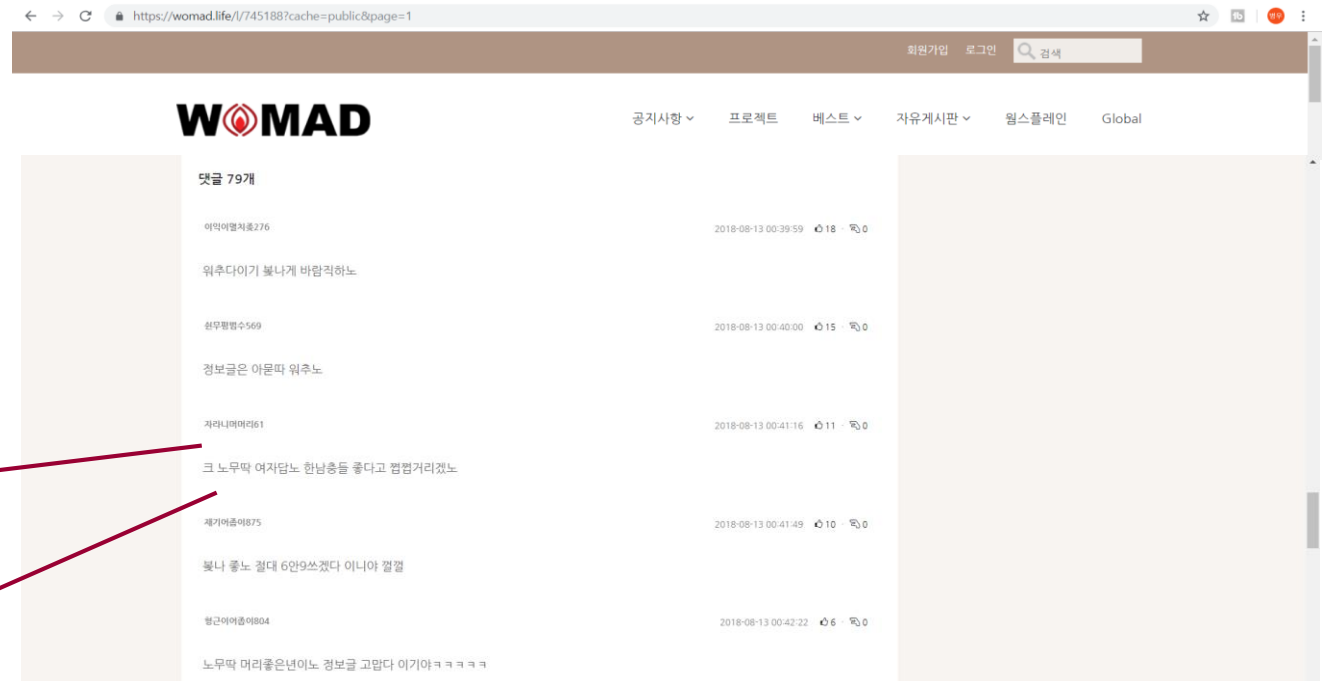


## 2. 데이터 수집 과정

- 워마드 웹사이트에서 인기 포스트의 댓글, 학습 데이터로서 크롤링

- 웹크롤링: 웹사이트에서 원하는 정보를 자동으로 가져오는 것. 파이썬(프로그래밍 언어의 일종)에서는 beautifulsoup 와 selenium 라이브러리를 활용하여 이를 실행한다.

“크 노무딱 여자답노 한남충들 좋다고 찹찹거리겠노”





## 2. 데이터 수집 과정

댓글	
0	땡문에 묻은 소추 나노 말말 달라라\n개인적으로는 큰 깨달음 얻었노 나는 야망이 없...
1	멋진 글이노
2	땡문이노 내일 일어나자마자 필사하노
3	말말 안 쓰노?
4	말말 달라라 이기 그거 빼면 땡문이노
5	권력은 복종하지 않기위해 얻어야하는 것이다 ㅇㄱㅇ이노. 난 누구의 지배를 받는것 ...
6	말머리 달라라 소추준다
7	추하고 역겨운 길을 걸어, 무감각함과 평온의 세계로 가는 것이야말로 권력이다.\n누...
8	땡문이노 야망보지 더 힘주겠노
9	땡문이노. 내가 워마드에 들어가본 것이 인생의 최대 터닝포인트라고 느끼는 이유 중 ...
10	땡문이노 일기에 붙여놓고 보겠노
11	땡문이노 유입들은 닥눈삼 자세로 이 글 정독하고 다른 땡문들도 정독해라 이기야
12	구구절절 땡문이노 자트릭스에서 추하지 않은 방법으로 권력을 획득할 수 있는 방법은 ...
13	지금 아주 자트릭스 심한 집단 (과)에서 스트레스 붓나 받고 살았는데 개돼지들보다 ...

- 워마드 웹사이트 대부분의 댓글을 혐오 발언이라 가정

- 특정한 단어("한남충", "~이기", "~노")가 반복되고, 극단적인 혐오 내용이 자주 출현하기에 혐오 발언 데이터로서 적합하다 판정

- 약 3만 개의 댓글을 크롤링하여 데이터셋(그림과 같은 표)으로 저장




$$TF(\text{단어의 빈도수}) * IDF(\text{단어의 희소성})$$

$$= \frac{n(\text{단어 출현 빈도})}{N(\text{문서 내 총 단어 수})} * \log_{10} \frac{D(\text{문서 내 문장 수})}{d(\text{단어를 포함한 문장 수})}$$



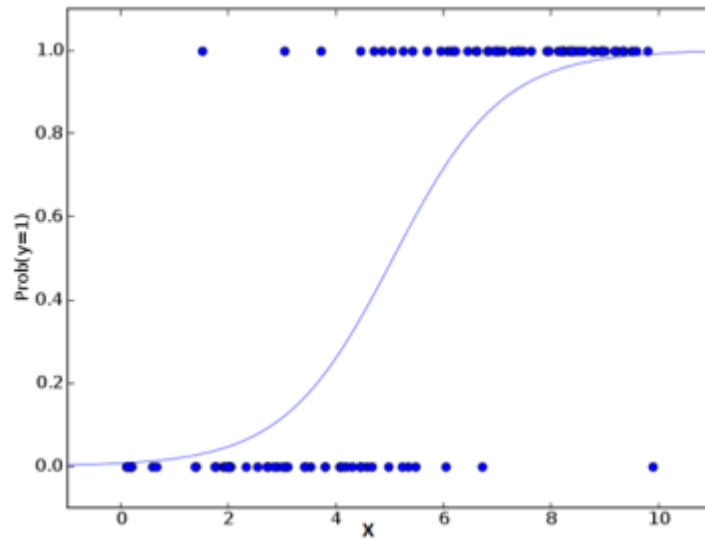

## 3. 자연어 처리 기법을 통한 인공지능 개발

### 2. 로지스틱 회귀(logistic regression)

· 문장이 혐오발언일 경우엔 1, 정상발언일 경우엔 0으로 나타내어, 혐오발언(1)이 될 확률  $f(x)$ 가 0.5보다 큰가 작은가에 따라 혐오발언 또는 정상발언으로 분류한다.

· 앞서 각 특징어에 부여한 TF-IDF 가중치의 값에 따라 로지스틱 함수를 제작한다.

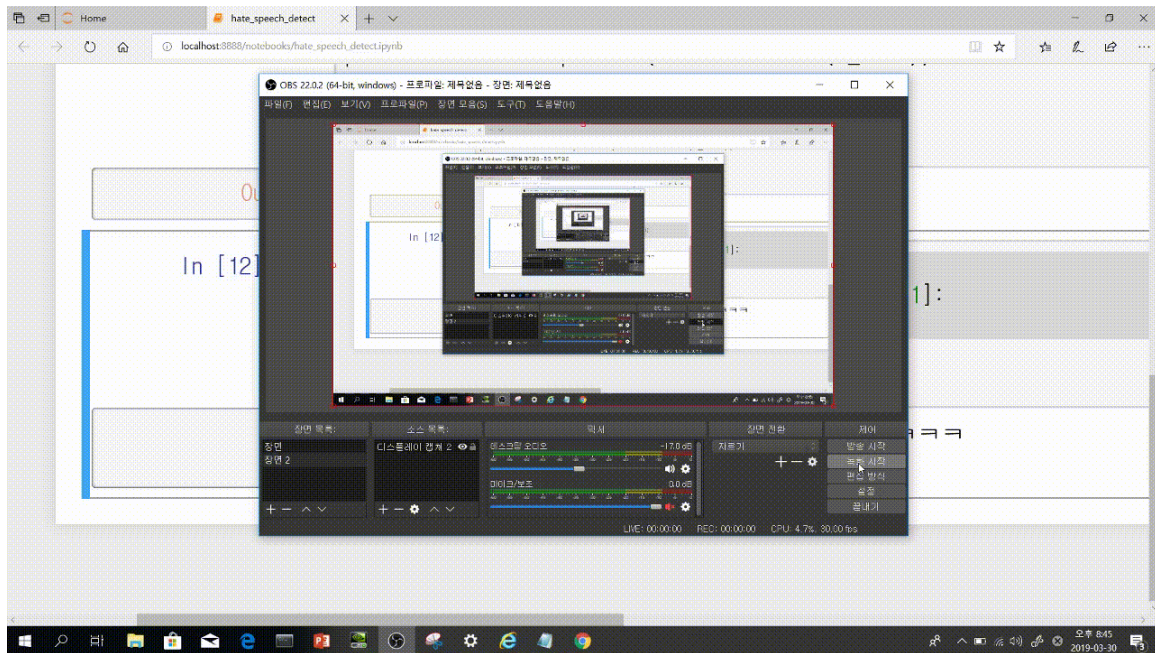
$$f(x) = \frac{1}{1 + \exp[-(b + a_1x_1 + \cdots a_nx_n)]}$$





# 3. 자연어 처리 기법을 통한 인공지능 개발

## 3. 결과물



```
#TfidfVectorizer Logistic Regression
import pandas as pd
from sklearn.utils import shuffle
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import roc_auc_score

list = []
normal_data = pd.read_csv("korean_sentence.csv", encoding="ANSI")
for i in range(len(normal_data["문장"])):
    normal_data["문장"][i] = normal_data["문장"][i].replace("##t", "")
for i in range(len(normal_data["문장"])):
    list.append(i)
for i in range(len(list)):
    list[i] = 0
list
normal_data["혐오 여부"] = list
womad_data = pd.read_csv("womad_comment_train_data.csv")
womad_data["문장"] = womad_data["댓글"]
list = []
for i in range(len(womad_data["문장"])):
    list.append(i)
for i in range(len(list)):
    list[i] = 1
womad_data["혐오 여부"] = list
womad_data = womad_data.drop(["'댓글'", "Unnamed: 0", "level_0", "Unnamed: 0.1", "index", "Unnamed: 0.1.1"], axis=1)
train_data = shuffle(pd.concat([womad_data.sample(n=1000), normal_data])).reset_index(drop=True)
hate_data = shuffle(pd.concat([womad_data, normal_data])).reset_index(drop=True)

X_train, X_test, y_train, y_test = train_test_split(train_data["문장"], train_data["혐오 여부"], random_state=0)
vect = TfidfVectorizer().fit(X_train)
X_train_vectorized = vect.transform(X_train)

model = LogisticRegression()
model.fit(X_train_vectorized, y_train)
```





## 4. 자연어 처리 응용

- 미래에 **딥러닝**을 통해 단어의 의미 분석과 관한 연구 또한 진행할 예정
- 자연어 처리** 기술이 혐오 발언 감지, 분야에 따른 문헌 분류, 가짜 뉴스 감지 등으로 발전할 가능성
- 인공지능**을 어떻게 개발하고 사용할지 결정하는 것은 결국 인간, 우리의 선택에 따라 미래가 좌우된다.