#4

AI NEWSLETTER

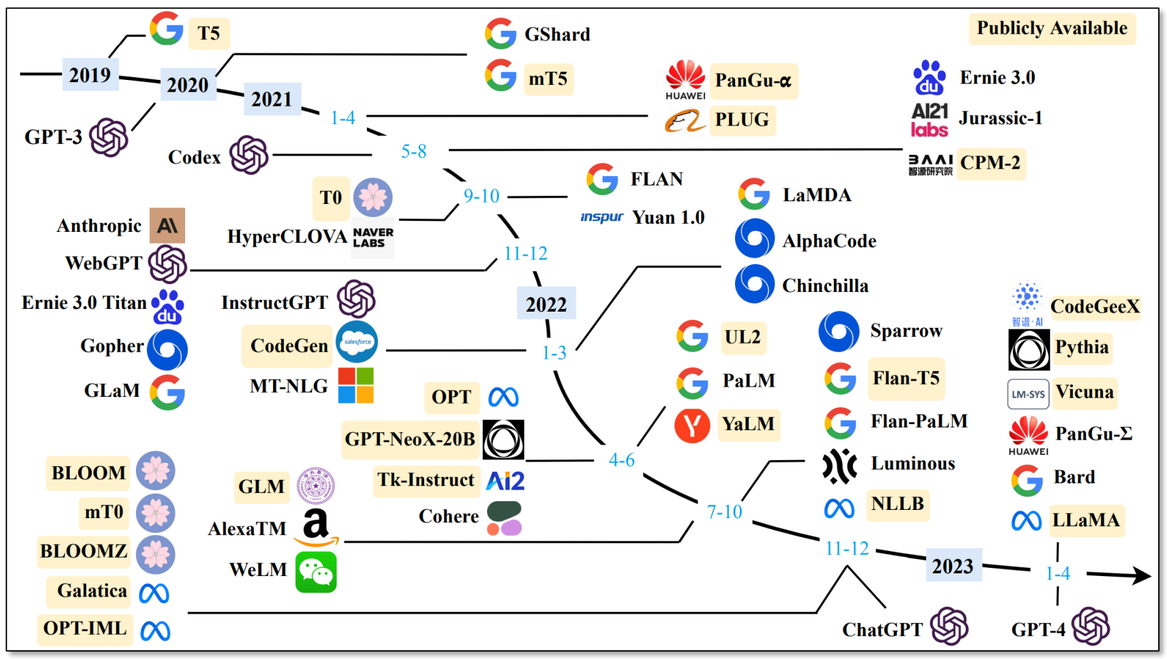
K-water
AI Research Lab.

08 JUN 2023

AI News

(논문) GPT와 아이들

OpenAI의 ChatGPT가 화제를 일으키며 수많은 언어모델들이 쏟아지고 있습니다. 최근 발표된 논문에서 정리한 현존하는 거대(1M 이상)언어모델을 발표시기에 따라 정리한 그림입니다.



여기에 특정 도메인에 특화되거나, 프롬프팅을 도와주는 각종 언어모델들이 있습니다.

(뉴스) DeepMind와 Google 합병

4월 21일 구글 모회사인 알파벳의 CEO 순다르 피차이는 구글 리서치 브레인 팀과 딥마인드를 '구글 딥마인드'로 합칠 것이라고 발표했습니다.

ChatGPT를 필두로 한 MS와 OpenAI사에 주도권을 빼앗기지 않으려는 움직임으로 여겨집니다.

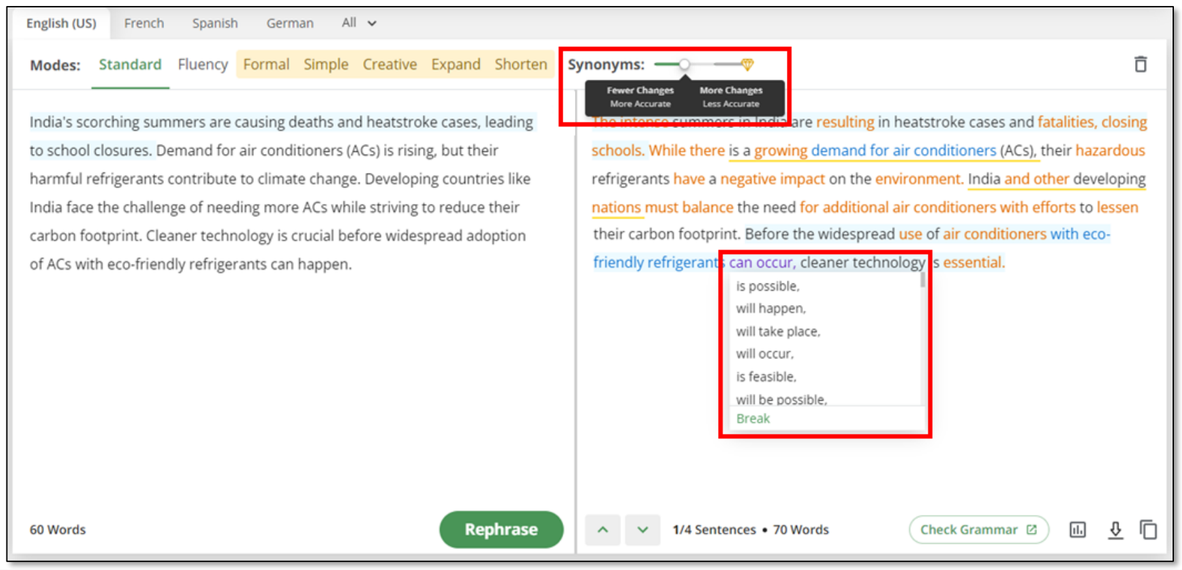
Google Brain

DeepMind

(어플리케이션) Quillbot

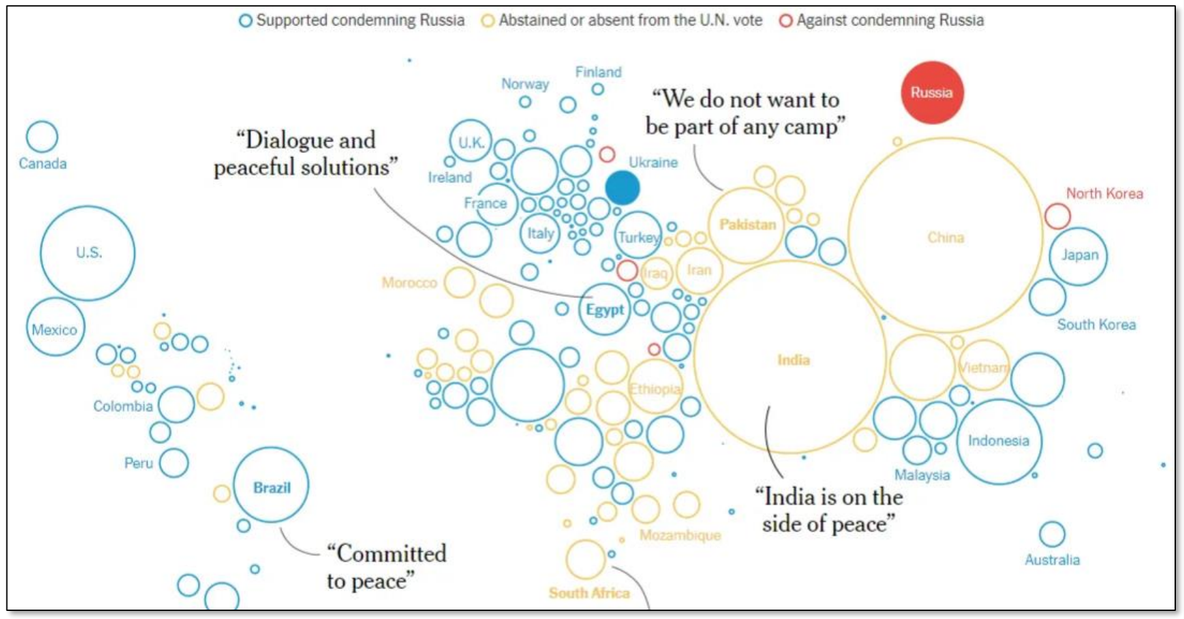
Quillbot은 인공지능을 활용한 작문 도구로, 문장 재구성, 문법확인, 표절검사, AI와 동시작업, 요약, 다양한 형태의 인용문 생성이 가능합니다.

영어, 프랑스어, 중국어, 일본어 등 23개 언어의 의역을 지원하며, 동의어·유의어 추천, 모드 선택(Standard, Fluency, Formal 등)도 가능합니다.



(시각화) 우크라이나-러시아 전쟁

최근 141개국이 러시아의 무조건적인 철수를 촉구하는 유엔 결의안을 지지하였으며, 뉴욕타임스는 각국의 반응을 버블차트로 시각화하였습니다.



Hands-On AI Project

가독성을 위해 일부 세부적인 라인은 생략되어 있습니다. 전체 코드는 QR코드 링크를 참고해주세요.

로우(low)코드 머신러닝 (pycaret)

Pycaret은 기계학습을 수행할 수 있는 대표적인 로우코드(low-code) python 패키지입니다.

```
!pip install pycaret
```

간단한 회귀(regression)분석을 위해 미국 건강보험 데이터셋을 준비해봅시다.

- 입력값: 나이, 성별, BMI, 자녀수, 흡연, 거주지역
- 목표값: 청구된 의료비(charges)

```
from pycaret.datasets import get_data
data = get_data('insurance')
```

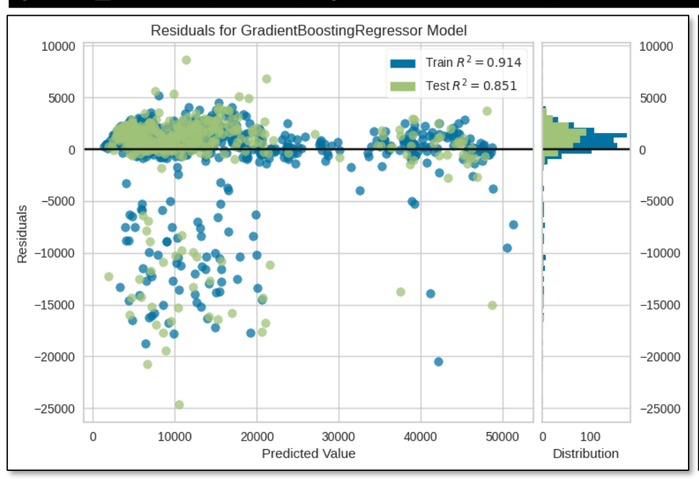
데이터가 준비되면 아래 두줄로 분석이 끝납니다.

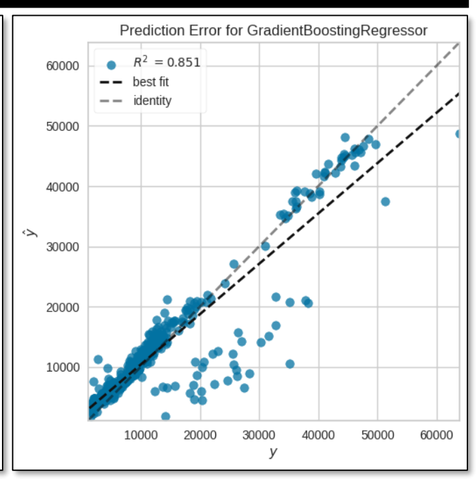
```
data_setup = setup(data, target='charges', session_id=101)
best = compare_models()
```

Models	MAE	…	RMSE	R2
Gradient Boosting Regressor	2538.4	…	4600.2	0.853
Random Forest Regressor	2598.1	…	4751.1	0.844
Light Gradient Boosting Regressor	2809.7	…	4843.1	0.837
Extra Trees Regressor	2565.0	…	4987.5	0.829
AdaBoost Regressor	3956.0	…	5171.0	0.813
Extreme Gradient Boosting	2918.0	…	5193.5	0.813
…	…	…	…	…

MAE, RMSE, R^2와 같은 성능지표를 통해 여러 기계학습 모델의 성능을 비교해줍니다. Gradient Boosting 모델이 가장 좋은 것으로 나타났네요.

```
plot_model(best, plot='residuals')
plot_model(best, plot='error')
```





유튜브 요약기

간단한 정보도 유튜브로만 검색될 때가 많습니다. 하지만 영상보다는 텍스트가 편할 때도 있습니다.

유튜브 영상의 자막 정보를 가져오는 함수입니다.

```
def get_transcript(url, lang='ko'):
    video_id = parse_qs(urlparse(url).query)['v'][0]
    formatter = TextFormatter()
    transcript = YouTubeTranscriptApi.get_transcript(
                        video_id, languages=[lang])
    text = formatter.format_transcript(transcript)
    return video_id, text
```

텍스트를 요약하는 함수입니다.

```
model_name = 'eenzeenee/t5-small-korean-summarization'
model = AutoModelForSeq2SeqLM.from_pretrained(model_name)
tokenizer = AutoTokenizer.from_pretrained(model_name)
def summarize_youtube(text):
    input = tokenizer(['summarize: ' + text], max_length=4096)
    output = model.generate(**input, max_length=512)
    decoded_output = tokenizer.batch_decode(output)[0]
    result = nltk.tokenize(decoded_output.strip())[0]
    return result
```

K-water 유튜브 채널의 최근 영상 url 입력

```
url = 'https://www.youtube.com/watch?v=AHa9Ls19O2I'
video_id, text = get_transcript(url)
result = summarize_youtube(text)
```

자동으로 생성된 자막을 받아오기 때문에 조금씩 오타가 보이기는 하지만 나름 잘 된 것 같습니다.

```
# text (원문)
[음악] 안녕하십니까 강원지역 협력단 지역협력부 하선혜 대리입니다 저희 강원지역 협력단은 현대화 사업소 특성상 거리가 떨어진 곳들에 많이 위치하다 … (중략) … 너무 뜻깊은 하루였던 것 같습니다 잘 마실게요 오늘 보람이 섰습니다 커피 잘 먹었습니다 아이 신나라 중요한 것은 강원지역 협력단 파이팅 [음악]
```

```
# result (요약결과)
강원지역 협력단은 강원도 18개 시군의 거버넌스 역할과 함께 국책사업을 수행하고 있는 조직으로 강원도 18개 지자체 중 10개의 지자체와 위스탁 협약을 맺어 사업을 수행하고 있으며 사업 목표 이수율 85 달성이라는 미션을 수행하기 위해 전 직원이 힘을 다하고 있다.
```

TIPs

set

리스트에 저장되어 있는 변수들의 중복이 있는지 궁금할 때가 있습니다. set 함수는 중복된 원소들을 제거하기 때문에 이를 활용할 수 있습니다.

```
def all_unique(lst):
    return len(lst) == len(set(lst))

x = [1, 1, 2, 2, 3, 2, 3, 4, 5, 6]
y = [1, 2, 3, 4, 5]

all_unique(x) # False
all_unique(y) # True
```

pandas.groupby()

엑셀에서 간단하게 할 수 있는 필터나 피벗테이블 같은 작업들도 pandas에서는 groupby()를 활용해 수행할 수 있습니다.

```
# 기본적인 사용방법
dataframe.groupby(분리할 열).agg({분석할 열: 통계량}, …)

# 분리할 column을 지정하고 column을 기준으로 통계량 산출
dataframe.groupby('column').sum() # 합계
dataframe.groupby('column').var() # 분산
dataframe.groupby('column').count() # 데이터 수

# 피벗테이블처럼 여러 열에 대해 계층적으로 분리
dataframe.groupby(['col1', 'col2']).sum()
```

Wifi QR 코드 만들기

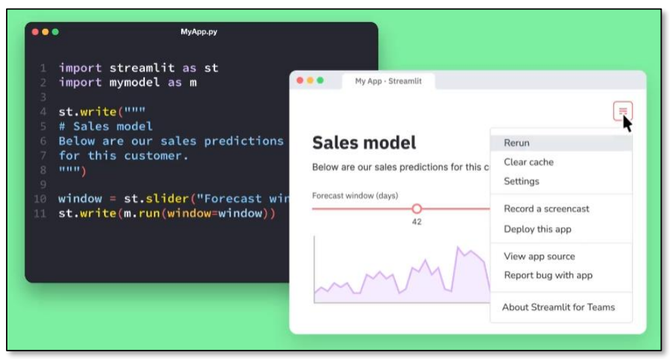
카페 같은 곳에서 와이파이 비밀번호를 받아적다 보면 여간 귀찮은 일이 아닙니다. Python에서 QR Code로 만들 수 있습니다.

```
!pip install wifi-qrcode-generator
import wifi_qrcode_generator as qr
qr.wifi_qrcode('WIFI_AP_NAME', False, 'WPA', 'PASSWORD')
```

Streamlit

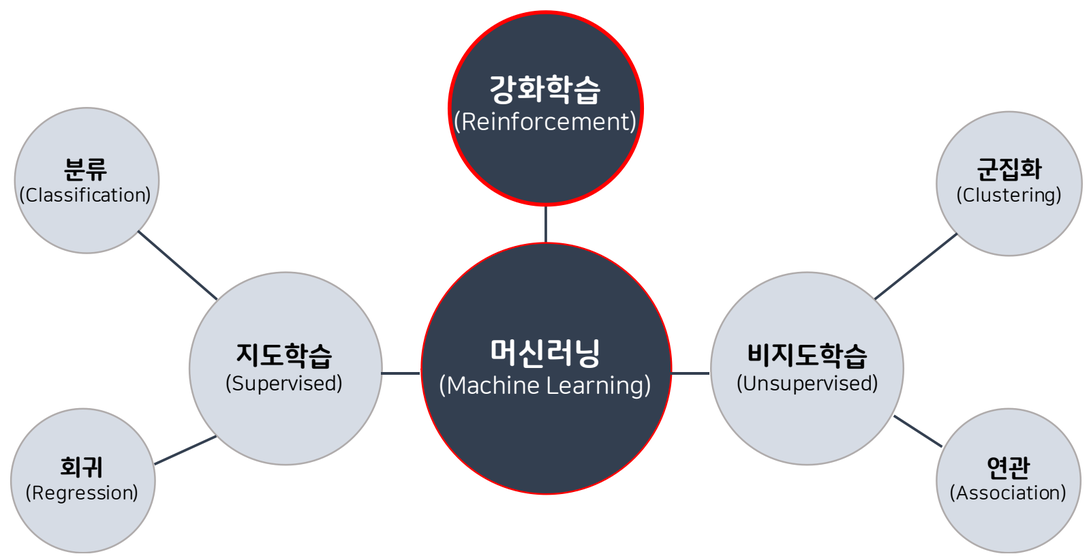
Streamlit은 python을 활용해 동적인 웹페이지를 만들 수 있는 도구입니다. K-water에서 운영중인 동파위험정보 서비스는 Streamlit으로 프로토타입을 만들고 실제 서비스로 이어진 좋은 예시입니다.

Streamlit



강화학습 (Reinforcement Learning)

강화학습은 알파고, 자율주행과 같이 인공지능이 시행착오를 통해 스스로 학습하는 알고리즘입니다.



간략히 말씀드리면, 외부 환경(environment)에서, 현재 상태(state)를 기준으로 agent가 액션을 취했을 때 보상(reward)가 최대가 되는 방향으로 정책(policy)를 수정해 나갑니다.

K-water AI Lab.

한국수자원학회 참석 (5/25-26)

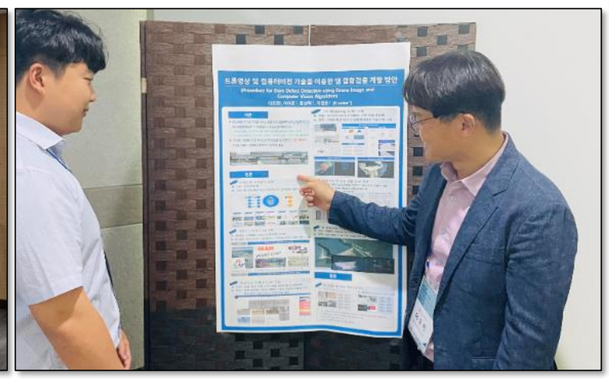
□ AI연구센터는 한국수자원학회 AI응용연구분과와 'AI 기술 융합을 통한 물관리 혁신 방안'을 주제로 기획세션을 개최하였습니다.

□ 1부 순서로 물관리 기술과 생성 AI 기술의 융합, 2부 순서로 AWS 클라우드 컴퓨터를 활용한 수자원 관련 실무 워크샵을 진행했습니다.

□ 또한 AI 연구센터의 논문 2건을 발표했습니다.

한국막학회, 한국정보통신학회 참석

□ 한국막학회(5/18)와 한국정보통신학회(5/25)에서 AI 기술을 활용한 막여과 정수장과 댐결함 자동검출 연구주제에 대해 발표했습니다.



AI 연구센터 세미나 개최

□ 지난 두 달간 AI 연구 관련 다양한 분야에 대해 5건의 세미나를 진행하였습니다.

날짜	3/30	4/14	4/17	4/19	5/23
발표자	이충성	장현준	이소령	김성훈	주경원
발표 키워드	수자원 계획관리	Docker 개념 및 활용법	Notion AI	연구인을 위한 실전 AI	장기시계열 예측 알고리즘

문의

주경원 선임 (7840)

이소령 사원 (7341)

깃허브 & 홈페이지