



데이터/ AI/ ML 직무 포트폴리오

송정훈 (2023. 11)

목차

Item #01: 회귀분석/ 다변량정규분포 기반 건물 월별 에너지 사용량 database의 누락치 추정 알고리즘 개발	3
Item #02: 회귀분석/ 딥러닝 (DNN/ RNN) 기반 건물의 시간별 전기부하 학습 및 예측 모델 구현	6
Item #03: 심층강화학습 (DQN) 기반 태양광 + ESS 에너지시스템 경제적 제어 모델 제안 논문 구현	8
Item #04: 심층강화학습의 에너지시스템 제어변수를 연속변수로 모델링하는 기법 (DDPG/ TD3/ SAC) 추가 적용	10
Item #05: Clustering analysis 기반 도시 건물 신재생 보급정책 효과 분석 계산 시간 절감 framework 제안	12
Item #06: 비선형 제약이 포함된 에너지시스템최적화 문제를 선형화하는 flowchart 제안	14
Item #07: AWS EC2 + docker + MySQL로, client로부터 요청받은 건물 주소의 전기 사용량 반환 웹 서비스 구현	16
Item #08: AWS EC2 + docker + TF/Serving으로, 전달받은 영화 리뷰 문장의 긍/부정 여부를 판단하는 웹 서비스 구현	18
Item #09: 건물별 월별에너지사용량 dataset과 건축속성정보 dataset 결합 후 SQL DB화	20
Item #10: 웹사이트 크롤링을 통한 데이터 수집 및 저장 data pipeline 구현	22

※ 페이지 별 우상단에, 해당 item에 대한 상세 내용 설명 및 코드를 포함하는 블로그 페이지 링크가 있으니 필요 시 참고하시기 바랍니다. (클릭 시 해당 링크로 이동)

Item # 01

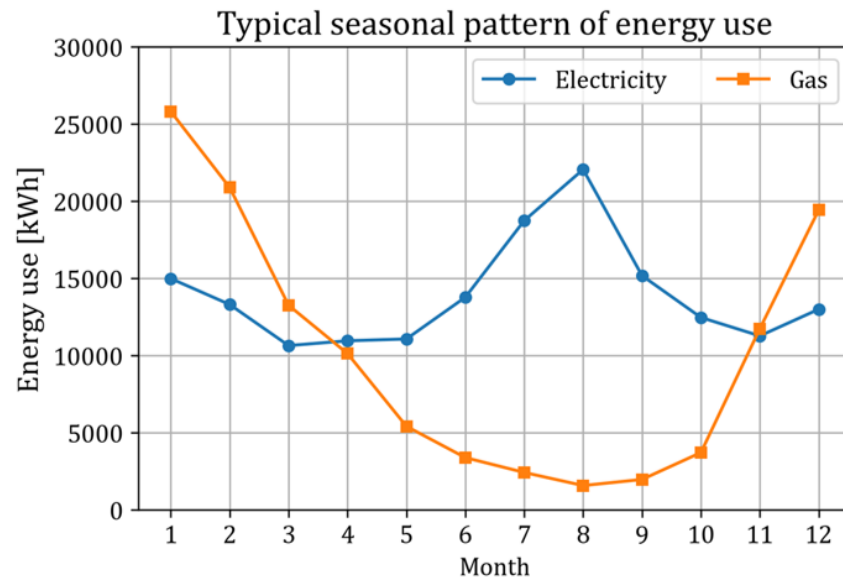
회귀분석/ 다변량정규분포 기반 건물 월별 에너지 사용량 database의 누락치 추정 알고리즘 개발 (SIAT 석사학위논문 주제)

다변량정규분포의 평균 벡터는 회귀분석으로, 공분산행렬은 회귀잔차들 간 공분산을 이용해 계산

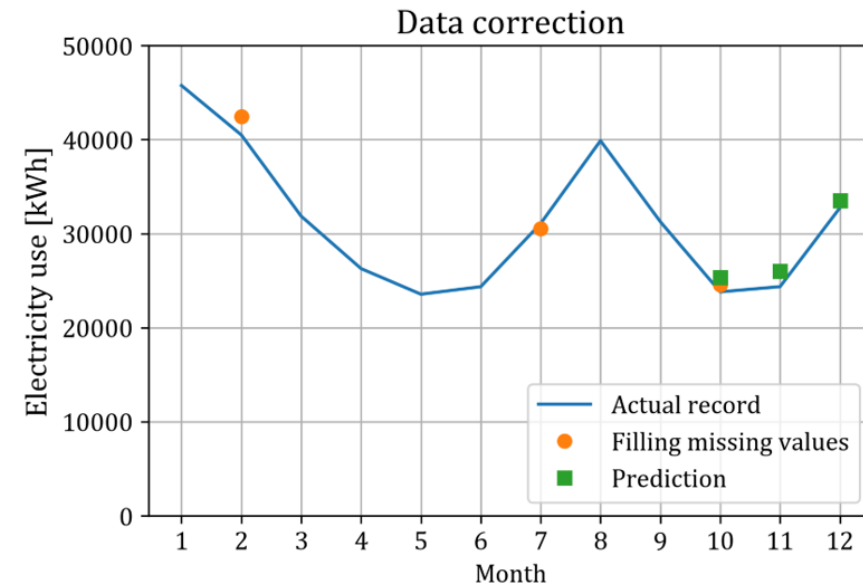
- ✓ 국내 비주거용 건물들 각각의 지번 별 월별 전기/ 도시가스 사용량 database에는 누락치 (missing value) 가 많음.

gas_202101	gas_202102	gas_202103	gas_202104	gas_202105	gas_202106	gas_202107	gas_202108	gas_202109	gas_202110	gas_202111	gas_202112
Filter	Filter	Filter 누락	Filter	Filter 누락	Filter	Filter	Filter	Filter	Filter	Filter	Filter
321.0	5002.0	NULL	4271.0	NULL	NULL	259.0	153.0	165.0	765.0	4969.0	5656.0

- ✓ 해당 database의 이상치를 판별 및 제거하고 누락치를 합리적으로 추정하는 알고리즘을 개발함.



<일반적인 건물의 월별 전기/ 도시가스 사용량 추이>



<개발된 알고리즘 기반으로 월별 사용량 누락치 추정 결과>
(파란색 선: 실제 월별 전기 사용량,
주황색 원: 2월/7월/ 10월 누락 가정 시의 추정값,
초록색 네모: 1~9월 사용량만 주어졌다고 가정 시의 예측값)

Item # 01

회귀분석/ 다변량정규분포 기반 건물 월별 에너지 사용량 database의 누락치 추정 알고리즘 개발 (SIAT 석사학위논문 주제)

다변량정규분포의 평균 벡터는 회귀분석으로, 공분산행렬은 회귀잔차들 간 공분산을 이용해 계산

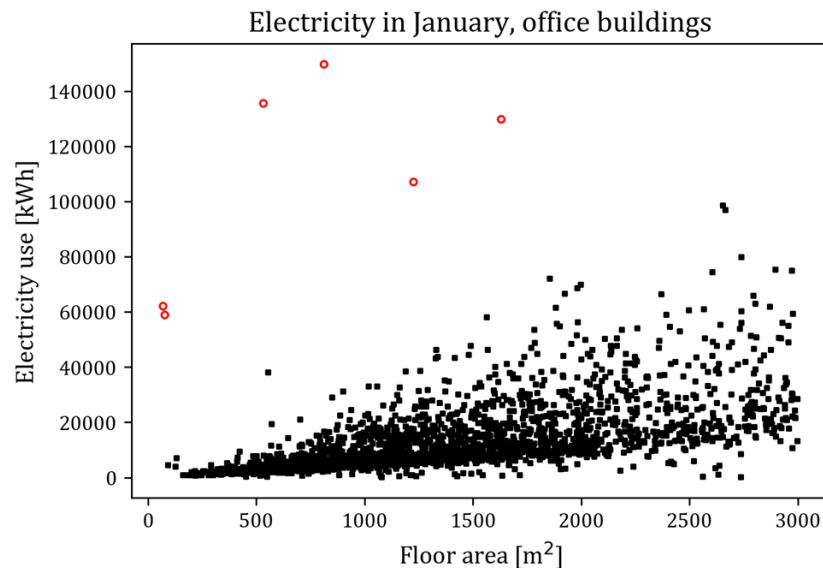
✓ 건물의 평균적인 에너지 사용량을 추정하는 다변량회귀모델을 각 월별/ 전기와 도시가스 각각에 대해 구성함.

✓ 설명변수는 건축물의 면적/ 층수/ 건축연도/ 용도

$$(y_i^{elec,m} = \beta_0^{elec,m} + \beta_1^{elec,m} x_i^{area} + \beta_2^{elec,m} x_i^{story} + \beta_3^{elec,m} x_i^{area} x_i^{story} + \beta_4^{elec,m} x_i^{year} + \beta_5^{elec,m} x_i^{area} x_i^{year} + \epsilon_i^{elec,m})$$

✓ 특정 월의 에너지 사용량이 비정상적으로 큰 값을 갖는 outlier는 Cook's Distance 계산으로 판별함.

✓ Cook's Distance는 해당 data point가 있을 때와 없을 때 각각에 대해 도출한 회귀모델 계수들 간 표준화된 차이이며, 시각화에 의존하지 않고 하나의 지표를 기준으로 outlier 제거를 자동화할 수 있음.



<에너지 사용량이 비정상적으로 큰 outlier들 (빨간 점들)>

$$D_i = \frac{(\hat{\beta}_{OLS,-i} - \hat{\beta}_{OLS})' X' X (\hat{\beta}_{OLS,-i} - \hat{\beta}_{OLS})}{p \cdot MS_{Residual}}$$

$$= \frac{e_i^2 h_{ii}}{p \cdot MS_{Residual} (1 - h_{ii})^2}$$

<Cook's Distance 계산 식>
(β 는 회귀계수, X 는 설명변수행렬,
 p 는 feature 개수, MS 는 잔차평균제곱합)

(Outlier 제거 방법 상세 포스팅 링크: <https://song4energydata.github.io/monthlyenergythree.html>)

Item # 01

회귀분석/ 다변량정규분포 기반 건물 월별 에너지 사용량 database의 누락치 추정 알고리즘 개발 (SIAI 석사학위논문 주제)

다변량정규분포의 평균 벡터는 회귀분석으로, 공분산행렬은 회귀잔차들 간 공분산을 이용해 계산

- ✓ 1년간의 월별 에너지 사용량을 다변량 벡터로 보고 다변량정규분포를 적합함.
다변량정규분포의 평균벡터는 회귀계수로 구하고, 공분산행렬은 잔차들 간의 공분산을 이용해 구함.

$$\underbrace{\begin{bmatrix} y_i^{elec,1} \\ y_i^{elec,2} \\ \vdots \\ y_i^{elec,12} \\ y_i^{gas,1} \\ \vdots \\ y_i^{gas,12} \end{bmatrix}}_{\substack{\text{각 월별} \\ \text{전기/가스 사용량} \\ \text{(설명변수 조건부)}}} \sim MVN \left(\underbrace{\begin{bmatrix} x_i^T \hat{\beta}^{elec,1} \\ x_i^T \hat{\beta}^{elec,2} \\ \vdots \\ x_i^T \hat{\beta}^{elec,12} \\ x_i^T \hat{\beta}^{gas,1} \\ \vdots \\ x_i^T \hat{\beta}^{gas,12} \end{bmatrix}}_{\substack{\text{평균 벡터} \\ (x \text{는 설명변수,} \\ \beta \text{는 회귀계수})}}, \underbrace{\begin{bmatrix} \hat{\alpha}_{(1,1)+}^{e,e} & \hat{\alpha}_{(1,2)+}^{e,e} & \cdots & \hat{\alpha}_{(1,12)+}^{e,e} & \hat{\alpha}_{(1,1)+}^{e,g} & \cdots & \hat{\alpha}_{(1,12)+}^{e,g} \\ \hat{\alpha}_{(2,1)+}^{e,e} & \hat{\alpha}_{(2,2)+}^{e,e} & \cdots & \hat{\alpha}_{(2,2)+}^{e,e} & \hat{\alpha}_{(2,1)+}^{e,g} & \cdots & \hat{\alpha}_{(2,12)+}^{e,g} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \hat{\alpha}_{(12,1)+}^{e,e} & \hat{\alpha}_{(12,2)+}^{e,e} & \cdots & \hat{\alpha}_{(12,12)+}^{e,e} & \hat{\alpha}_{(12,1)+}^{e,g} & \cdots & \hat{\alpha}_{(12,12)+}^{e,g} \\ \hat{\alpha}_{(1,1)+}^{g,e} & \hat{\alpha}_{(1,2)+}^{g,e} & \cdots & \hat{\alpha}_{(1,12)+}^{g,e} & \hat{\alpha}_{(1,1)+}^{g,g} & \cdots & \hat{\alpha}_{(1,12)+}^{g,g} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \hat{\alpha}_{(12,1)+}^{g,e} & \hat{\alpha}_{(12,2)+}^{g,e} & \cdots & \hat{\alpha}_{(12,12)+}^{g,e} & \hat{\alpha}_{(12,1)+}^{g,g} & \cdots & \hat{\alpha}_{(12,12)+}^{g,g} \end{bmatrix}}_{\substack{\text{공분산 행렬} \\ (\alpha \text{는 분산/공분산, } x_i^{area} \text{는 건물의 면적;} \\ \rightarrow \text{건물의 면적이 클수록 분산이 큼 (이분산) 을 반영})}} (x_i^{area})^2 \right)$$

- ✓ 어떤 건물의 1년간 월별 에너지 사용량 중 일부 월의 사용량만 누락되었을 때,
나머지 월들의 사용량을 조건부로, 누락된 월의 사용량들에 대한 조건부 다변량정규분포를 도출함.
해당 조건부 다변량정규분포의 평균이, 누락치에 대한 예측치로 적절함을 확인하였음.

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} | X}_{\substack{y_1: \text{누락된 월 사용량} \\ y_2: \text{관측된 월 사용량}}} \sim MVN \left(\underbrace{\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}}_{\substack{\text{사용량이 관측된 월의} \\ \text{전기 사용량 값}}}, \underbrace{\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}}_{\substack{\text{누락된 사용량의} \\ \text{조건부 분포의 평균 (예측치)}}} \right) \text{ 일 때, } P(y_1 | y_2 = a) = MVN \left(\underbrace{\mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (a - \mu_2)}_{\substack{\text{조건부 분포의} \\ \text{공분산}}}, \underbrace{\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}}_{\substack{\text{조건부 분포의} \\ \text{공분산}}} \right)$$

Item
02

회귀분석/ 딥러닝 (DNN/ RNN) 기반 건물의 시간별 전기부하 학습 및 예측 모델 구현

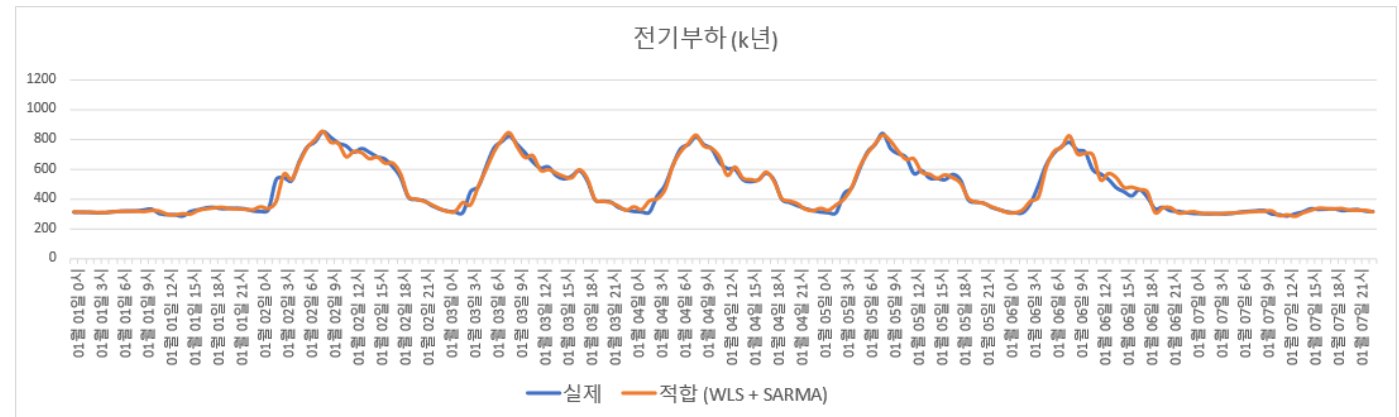
R로 전통적 선형회귀 및 SARIMA, Python Tensorflow로 Dense Neural Net/ Recurrent Neural Net 각각 적용 후 비교

✓ 특정 건물의 시간별 전기부하 데이터를 반응변수로, 해당 지역의 시간별 기온과 조도 데이터 (기상청 제공) 및 해당 시간의 계절과 시간대를 설명변수로 하는 부하 예측 모델을 적합 및 검증함.

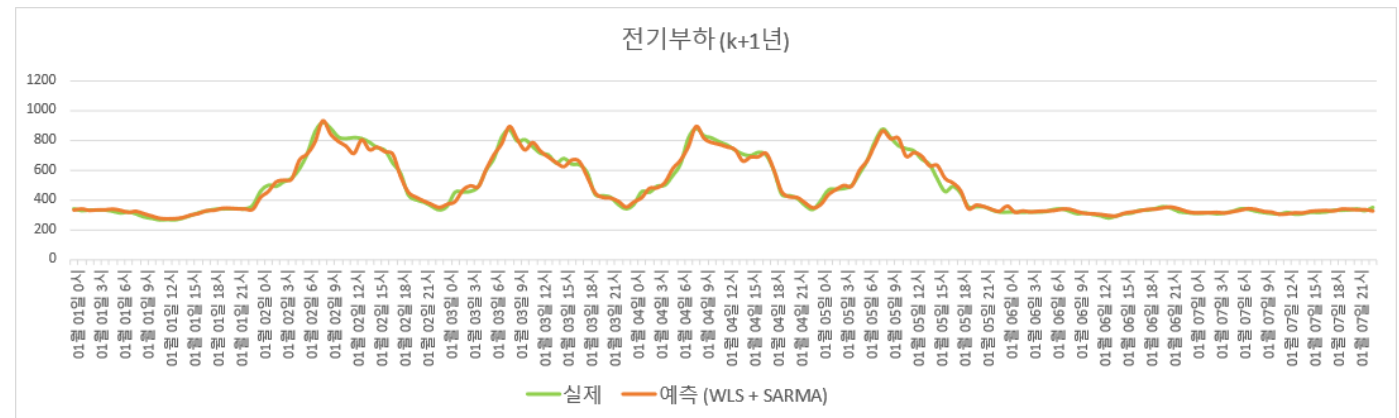
✓ 각 계절 및 평일/ 휴일 별로 하루 24시간 동안의 전기부하의 패턴이 다름 (non-stationarity) 을 고려해, 하절기 평일/ 동절기 평일/ 춘추절기 평일/ 휴일 각각에 대해 별도의 모델을 구성함.

○○○○시 기상자료	기온(°C)	강수량(mm)	풍속(m/s)	습도(%)	일조(hr)	일사(MJ/m2)	전운량(10분위)
01-01 0:00	-1.5	0	0.3	99	0	0	0
01-01 1:00	-0.6	0	0.2	99	0	0	0
01-01 2:00	0	0	0.2	99	0	0	0
01-01 3:00	-0.5	0	0.2	98	0	0	0
01-01 4:00	-1.3	0	0	99	0	0	0
01-01 5:00	-2.2	0	0.1	99	0	0	0
01-01 6:00	-2.3	0	0	99	0	0	0
01-01 7:00	-2.6	0	0	99	0	0	0
01-01 8:00	-1.8	0	0.1	99	0	0	0
01-01 9:00	-1.4	0	0.9	99	0	0	0

<기상청 중관기상관측자료 예시>



<전통적 회귀분석을 이용한 예측: 훈련 데이터셋 (특정 연도)>



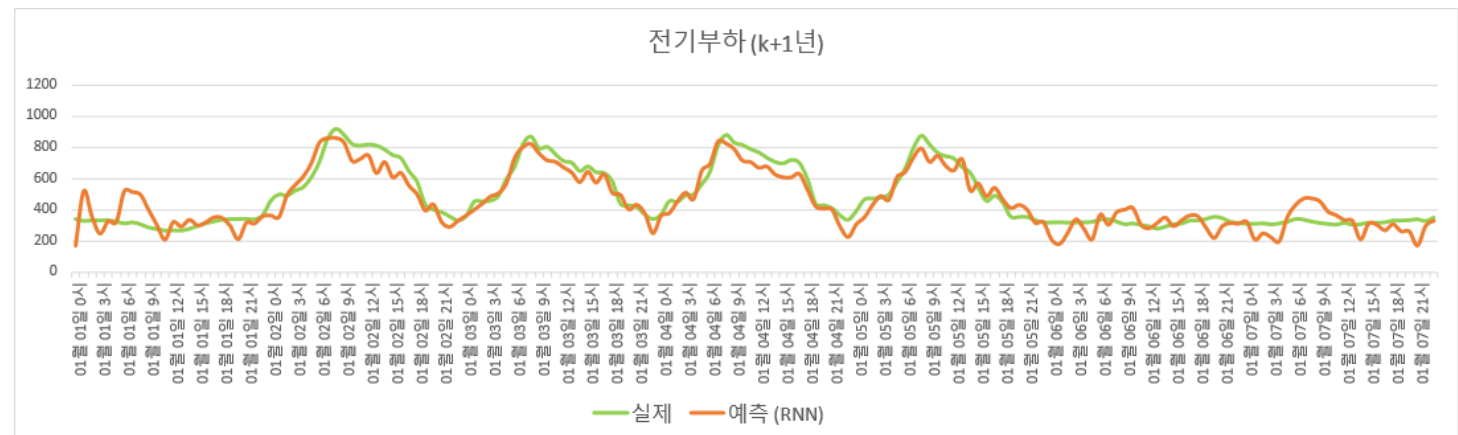
<전통적 회귀분석을 이용한 예측: 검증 데이터셋 (훈련 데이터의 바로 다음 연도)>

Item # 02

회귀분석/ 딥러닝 (DNN/ RNN) 기반 건물의 시간별 전기부하 학습 및 예측 모델 구현

R로 전통적 선형회귀 및 SARIMA, Python Tensorflow로 Dense Neural Net/ Recurrent Neural Net 각각 적용 후 비교

- ✓ 전통적 선형회귀 기반 예측 모델, 딥러닝 기반 예측 모델을 각각 구성함.
 - ✓ 전통적 선형회귀 (Weighted Least Squares) 를 적용. 잔차에 자기상관성이 있어, 시계열 모델 (SARIMA) 을 추가 적용. R로 구현.
 - ✓ 딥러닝의 경우 일반적인 Dense Neural Network (DNN) 적용 모델, Recurrent Neural Network (RNN) 적용 모델을 각각 별도로 구성. Python에서 Tensorflow로 구현, RNN 구현 시 은닉층은 SimpleRNN 적용.
- ✓ K년도의 시간별 데이터로 모델 적합 후 K+1년도의 시간별 데이터로 검증을 수행하였음.
 - ✓ 전통적 회귀분석 기반 모델은 검증 데이터셋에 대해 R-square 0.97의 우수한 성능을 보임.
 - ✓ 딥러닝 모델은 DNN 모델을 non-stationarity를 고려해 구성한 경우 R-square 0.83~0.85, RNN 모델은 R-square 0.8 이하로 더 낮음. 이는 전기부하 데이터의 비선형성은 크지 않으나 non-stationarity가 있는데, 딥러닝은 비선형성이 크면서 stationary한 데이터에 적절한 방법이기 때문임.



<Recurrent Neural Network를 이용한 예측: 검증 데이터셋>

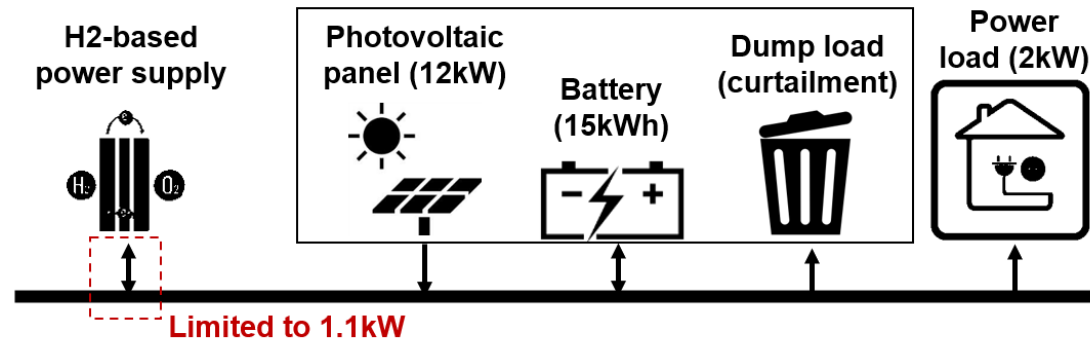
Item # 03

심층강화학습 (DQN) 기반 태양광 + ESS 에너지시스템 경제적 제어 모델 제안 논문 구현

Python Tensorflow로 Deep Q-Network

François-Lavet, V. (2017). Contributions to deep reinforcement learning and its applications in smartgrids.

- ✓ 심층강화학습 (Deep Reinforcement Learning) 기반으로, 아래와 같은 에너지시스템의 경제적 제어 모델이 제안됨.



<해당 논문의 case study 에너지시스템 도식도>

- ✓ 각 시간별로 전기부하를 수소기반 중앙급전계통(역송전 가능)/ 태양광/ 배터리/ 잉여전기 소모용 별도부하 의 조합으로 충족. 중앙급전계통과 에너지시스템 간 전력 흐름은 양방향이며 최대 1.1kW.
계통으로부터 전기를 받을 경우 비용 발생, 잉여 태양광 전기를 계통에 보낼 경우 적은 단가의 수익 발생.
별도부하로 잉여전기 소모 시 비용이나 수익 없음. 전기부하를 충족하지 못할 경우 큰 penalty 비용 발생.
- ✓ 제어 모델은, 미래의 태양광 발전량과 전기부하를 모르는 상태에서 과거 데이터에 기반해 시스템을 제어, 낮은 연간 총 비용으로 전기부하를 충족해야 함.
 - ✓ 상태변수 (state variable) 들은 직전 24시간의 시간별 전기부하와 태양광 발전량, 직전 시간의 배터리 내 저장된 에너지량.
 - ✓ 제어변수 (control variable) 은 각 시간 별 에너지시스템과 중앙급전계통 간 전력 흐름.

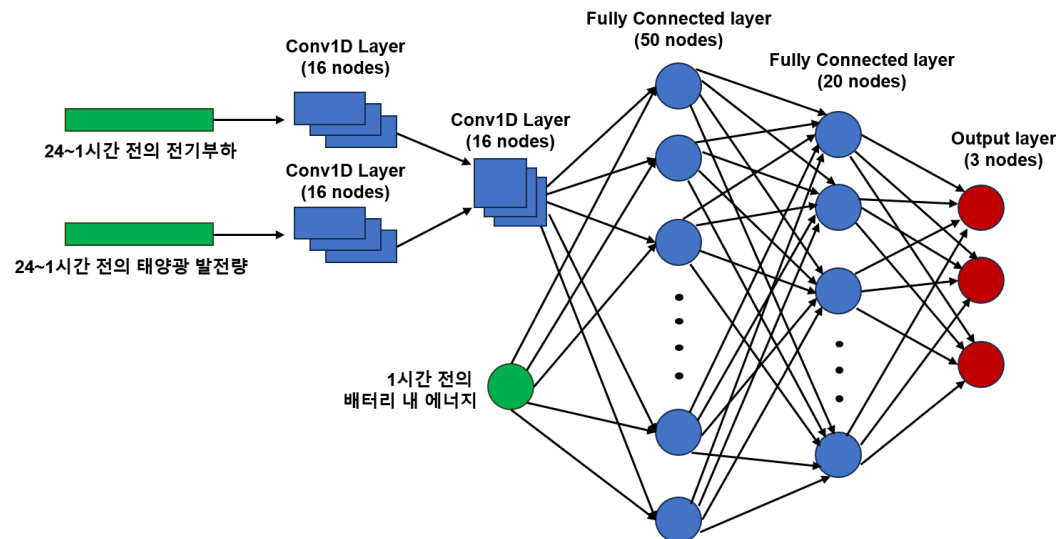
Item # 03

심층강화학습 (DQN) 기반 태양광 + ESS 에너지시스템 경제적 제어 모델 제안 논문 구현

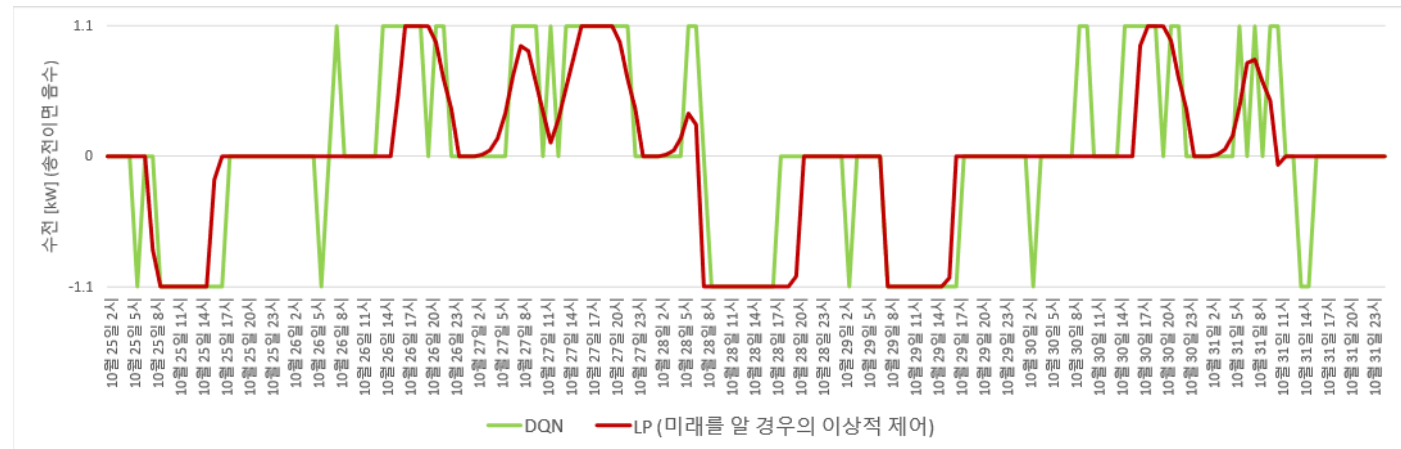
Python Tensorflow로 Deep Q-Network

François-Lavet, V. (2017). Contributions to deep reinforcement learning and its applications in smartgrids.

- ✓ 심층신경망이 상태변수 (state variable) 49개를 입력받아, 제어변수 (control variable) 이 가질 수 있는 3개의 값들 각각에 대한 Q-value를 계산하여, 시간별 제어변수의 값을 결정함.
- ✓ Deep Q-Network (DQN) 적용을 위해, 불연속적인 3개 값들만 고려 (1.1kW 받음(수전)/ 1.1kW 보냄(역송전)/ 0).
- ✓ 2년치 데이터로 Python에서 Tensorflow로 DQN 모델을 훈련, 별도의 1년치 데이터로 검증함.
- ✓ Benchmark는 Linear programming (LP) 기반으로, '미래 모든 시간들의 태양광 발전량 및 전기부하를 안다' 는 이상적 가정 하의 결과임.
- ✓ DQN에서의 검증 set에서의 연간 총 비용이 benchmark에서의 비용에 근접하였으며, 시간별 수전/송전 패턴도 유사함.



<DQN에 사용된 심층신경망 구조>



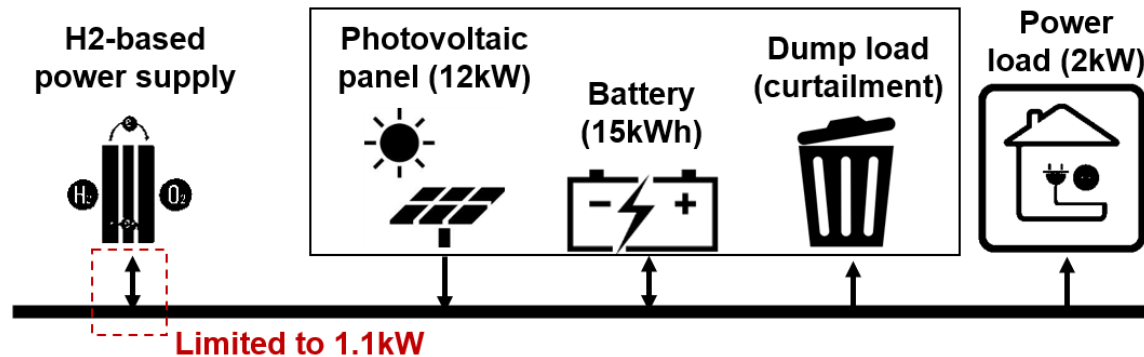
<시간별 수전/송전 패턴 (연두색: DQN, 빨간색: LP (Benchmark))>

Item # 04

심층강화학습의 에너지시스템 제어변수를 연속변수로 모델링하는 기법 추가 적용

Python Tensorflow로 Deep Deterministic Policy Gradient, Twin Delayed DDPG, Soft Actor-Critic

- ✓ Item #02에서의 심층강화학습 기반 경제적 제어 모델에서는, control variable (중앙계통과 에너지시스템 간 전력 흐름)의 값을 불연속적 (discrete) 값들로 제한하고 (1.1kW 수전/ 1.1kW 역송전/ 0) DQN 모델을 사용함.



<해당 논문의 case study 마이크로그리드 도식도>

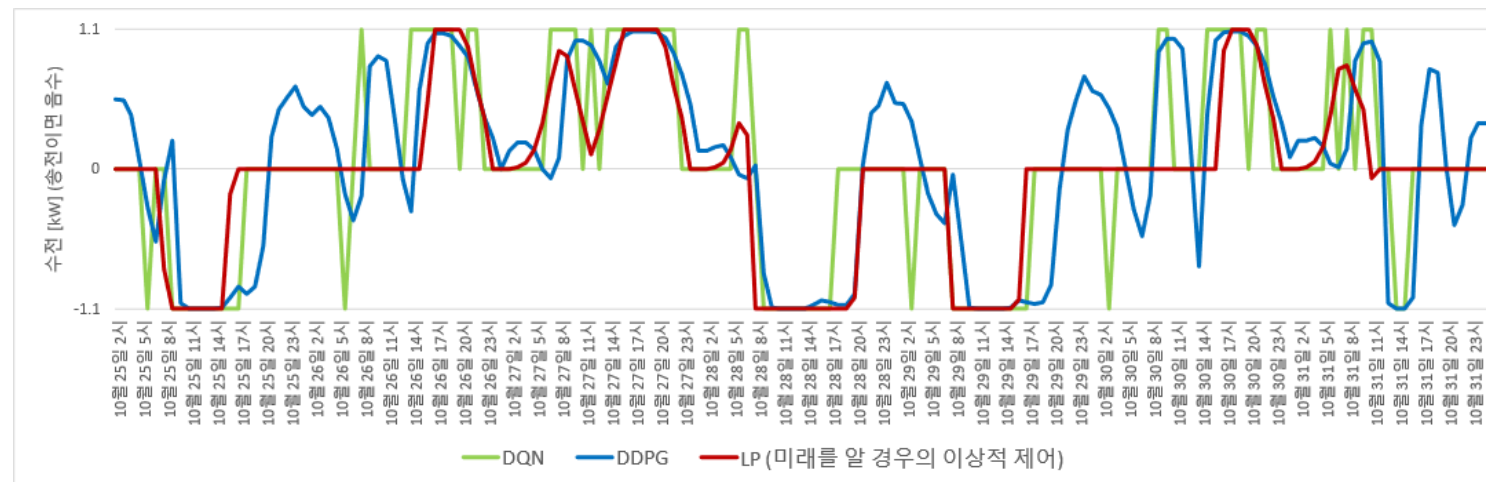
- ✓ 그러나 실제 전력 흐름은 -1.1에서 +1.1 범위 내의 연속적인 (continuous) 값임.
그러므로 제어 결과를 연속값으로 도출하는 심층강화학습 알고리즘 적용 시, 불연속값 기반 제어 대비 더 나은 경제급전이 가능한지 여부를 확인하고자 하였음.
- ✓ 제어 결과를 연속값으로 도출하는 심층강화학습 모델로는, Google Deepmind에서 발표한 Deep Deterministic Policy Gradient (DDPG), DDPG를 개량한 Twin Delayed DDPG (TD3), Soft Actor-Critic (SAC) 등이 있음.

Item # 04

심층강화학습의 에너지시스템 제어변수를 연속변수로 모델링하는 기법 추가 적용

Python Tensorflow로 Deep Deterministic Policy Gradient, Twin Delayed DDPG, Soft Actor-Critic

- ✓ Continuous control을 위한 DDPG/ TD3/ SAC 코드를 작성하고 각 모델을 훈련시킴.
그 결과 DDPG/ TD3/ SAC의 경제급전 성능이 DQN의 성능과 비슷한 정도로, DQN 대비 특별히 더 좋지 않았음.
- ✓ 이는 benchmark의 (미래를 안다는 이상적 조건 하에 도출된) 최적 제어 패턴이 DQN의 제어패턴과 더 비슷하기 때문임.
Benchmark의 제어 패턴에서 다수의 시간대에 1.1kW 수전/ 1.1kW 역송전/ 0 의 값을 가짐.



<시간별 수전/송전 패턴 (연두색: DQN, 파란색: DDPG, 빨간색: LP (Benchmark))>

- ✓ 최근 에너지기술연구원에서 수행 및 출판한 심층강화학습 기반 히트펌프 제어 연구에서도, 연구과정에서 SAC 적용을 시도하였으나 최종적으로 DQN 기반 대수 제어를 채택한 바 있음.¹⁾

1) 한국에너지기술연구원 Youtube 채널, “강화학습을 활용한 신재생에너지시스템 운영 최적화 사례”. <https://www.youtube.com/watch?v=L9Ye1gALFAk>

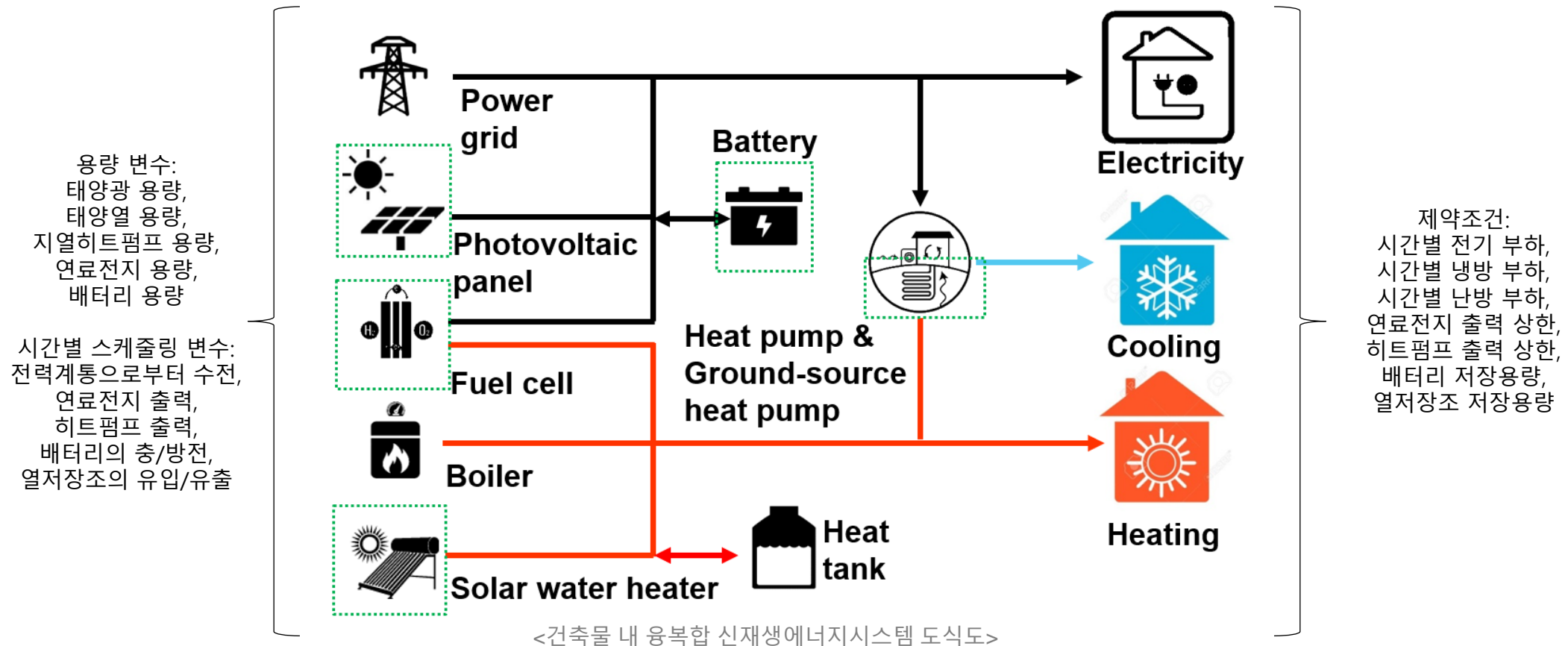
Item # 05

Clustering analysis 기반 도시 건물 신재생 보급정책 효과 분석 계산 시간 절감 framework 제안 (서울대학교 박사학위논문 주제, 국제학술지 Applied Energy에 논문 게재)

정수선형계획법 w/ GUROBI, *k*-Means, Genetic algorithm

Song, J., & Song, S. J. (2020). A framework for analyzing city-wide impact of building-integrated renewable energy. Applied Energy, 276, 115489.

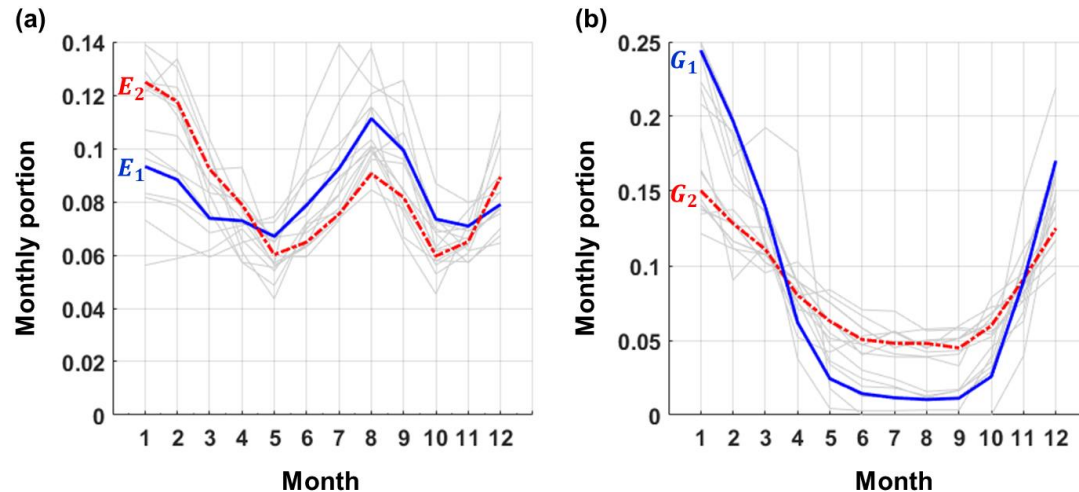
- ✓ 도시 내 건물 신재생에너지 (태양광/ 태양열/ 지열/ 연료전지) 보급 정책의 효과 추정을 위해서는, 개별 건물 내 신재생에너지 용량 결정 최적화 문제를 수만~수십만 회 풀어야 해 계산비용이 높음 (각 건물이 1개 문제에 대응)



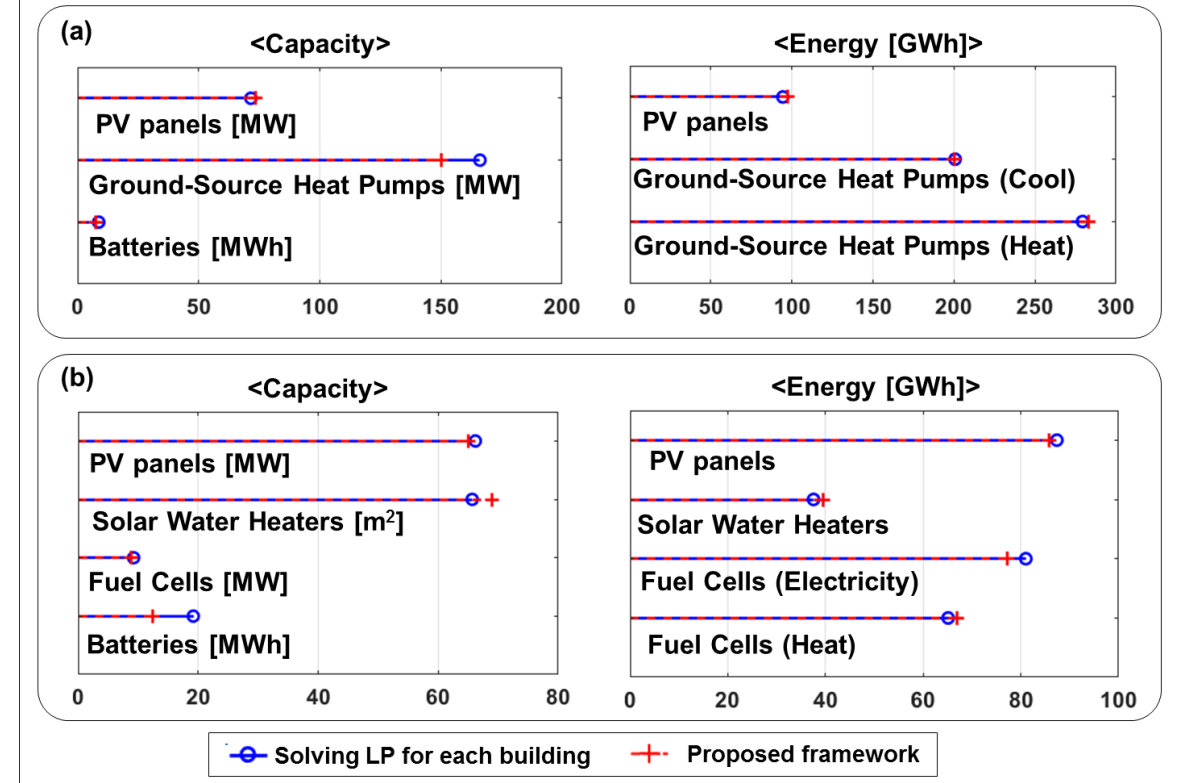
Item
05Clustering analysis 기반 도시 건물 신재생 보급정책 효과 분석 계산 시간 절감
framework 제안 (서울대학교 박사학위논문 주제, 국제학술지 Applied Energy에 논문 게재)정수선형계획법 w/ GUROBI, k -Means, Genetic algorithm

Song, J., & Song, S. J. (2020). A framework for analyzing city-wide impact of building-integrated renewable energy. Applied Energy, 276, 115489.

- ✓ 비슷한 건물들을 clustering analysis로 grouping한 후 각 group 별 centroid에 대응하는 건물에 대해서만 최적화 문제를 푸는 방식으로, 보급효과 추정을 위한 계산량을 80~90% 절감하는 framework를 제안하였음.
- ✓ 유사도 기준 및 방법은 표준화된 월별 사용량 패턴 벡터 간 거리 (k -Means), 전기 사용량 대비 가스 사용량 비율 및 지붕 면적 (Genetic algorithm) 등.
- ✓ 제안된 framework로 계산된 에너지원 별 용량과 공급량의 '도심지 내 건물들의 총 합' 이, 모든 건물 각각에 대해 최적화 문제를 풀어 얻는 결과와 유사함을 확인하였음.



<표준화된 건물 내 월별 전기/ 가스 사용량 패턴>

(연회색: 개별 data point들, 파란색/ 빨간색: k -Means로 도출한 각 group의 centroid)

<에너지원 별 용량과 공급량 총 합 계산 결과 비교>

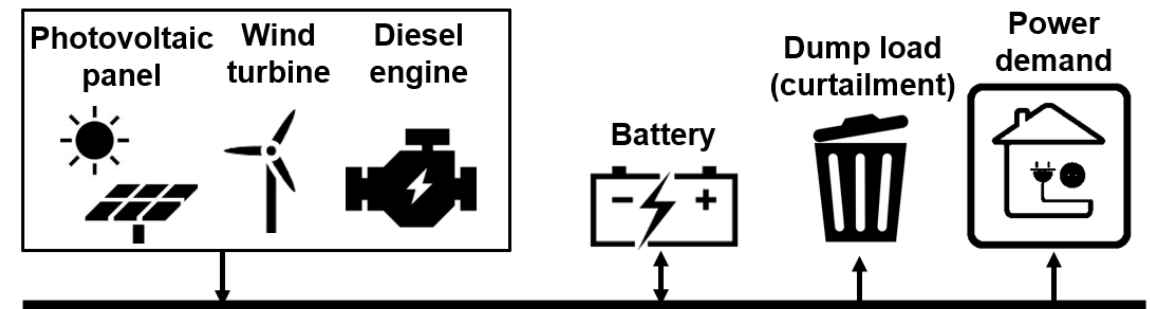
Item # 06

비선형 제약이 포함된 에너지시스템최적화 문제를 선형화하는 flowchart 제안 (국제학술지 Applied Energy에 논문 게재)

정수선형계획법 w/ GUROBI

Song, J., Oh, S. D., Yoo, Y., Seo, S. H., Paek, I., Song, Y., & Song, S. J. (2018). System design and policy suggestion for reducing electricity curtailment in renewable power systems for remote islands. Applied energy, 225, 195-208.

- ✓ 수학적최적화 기반 에너지시스템 용량산정/ 스케줄링 문제에서 배터리가 있을 경우, 같은 시간에는 충전 또는 방전 둘 중 하나만 수행함. 즉 임의의 시간 t 에 대해 충전 $P_{charge}[t]$ 와 방전 $P_{discharge}[t]$ 의 곱 $P_{charge}[t]P_{discharge}[t] = 0$ 이어야 함.
- ✓ 제약조건 $P_{charge}[t]P_{discharge}[t] = 0$ 은 비선형 등호 조건으로, 문제를 풀기 어렵게 함 (non-convex).
 - ✓ 시간별 이진수 $u[t]$ 를 도입해 선형 제약 ($P_{charge}[t] \leq Mu[t], P_{discharge}[t] \leq M(1 - u[t])$)으로 바꿀 경우, 이진수가 많아 풀기 어려움.
- ✓ 위 제약조건을 없애는 대신 목적함수에 $c(P_{charge}[t] + P_{discharge}[t])$ 를 추가하면 (c 는 아주 작은 양수), 목적함수 최소화를 위해 $P_{charge}[t]$ 와 $P_{discharge}[t]$ 둘 중 하나는 0이 되며, 선형 문제가 되어 풀기 쉬워짐.
- ✓ 별도로 특정 유형의 제약조건이 추가된 문제에서는, 목적함수에 위 항을 추가해도 모든 시간에 대해 $P_{charge}[t]P_{discharge}[t] = 0$ 이 성립하지 않는 것을 발견함.
 - ✓ 재생발전기가 포함된 에너지시스템에서 재생발전기의 잉여전력을 버리는 (curtailment) 총량의 상한 제약 추가 시, $P_{charge}[t]P_{discharge}[t] > 0$ 인 시간 t 가 존재하고 이 때 배터리 최적용량 및 목적함수 (총 비용)의 값이 더 낮아짐.



<재생발전기, 배터리, 잉여전력 curtailment가 포함된 에너지시스템 도식도>

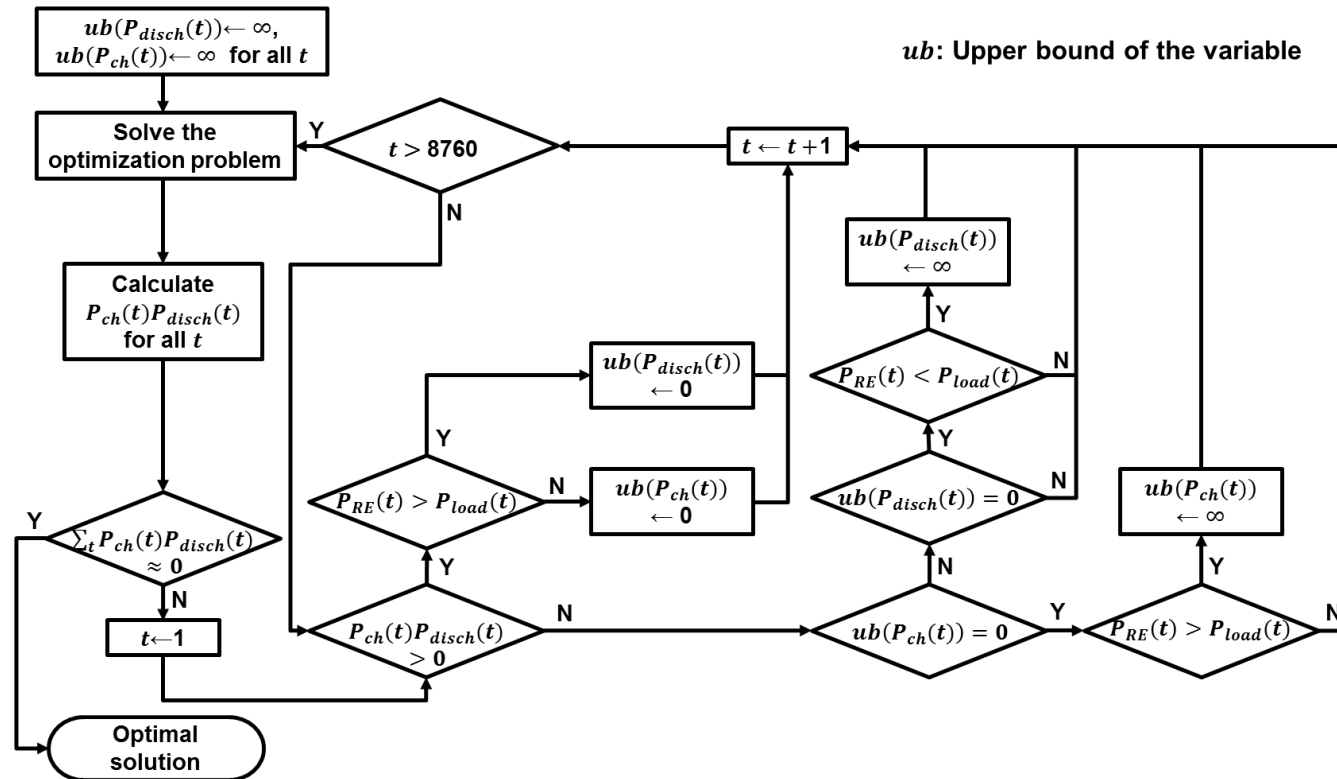
Item # 06

비선형 제약이 포함된 에너지시스템최적화 문제를 선형화하는 flowchart 제안 (국제학술지 Applied Energy에 논문 게재)

정수선형계획법 w/ GUROBI

Song, J., Oh, S. D., Yoo, Y., Seo, S. H., Paek, I., Song, Y., & Song, S. J. (2018). System design and policy suggestion for reducing electricity curtailment in renewable power systems for remote islands. Applied energy, 225, 195-208.

- ✓ $\sum_t P_{charge}[t]P_{discharge}[t] = 0$ 을 만족하는 near-optimal solution을 빠르게 도출하는 flowchart를 제안하였음.
- ✓ 제약조건 $P_{charge}[t]P_{discharge}[t] = 0$ 및 시간별 이진수가 없는 선형 문제를 여러 번 풀되 (iteration), $P_{charge}[t]P_{discharge}[t] > 0$ 인 시간 t 에 대해 $P_{charge}[t]$ 와 $P_{discharge}[t]$ 둘 중 하나를 0으로 강제로 설정하는 과정을 반복함.

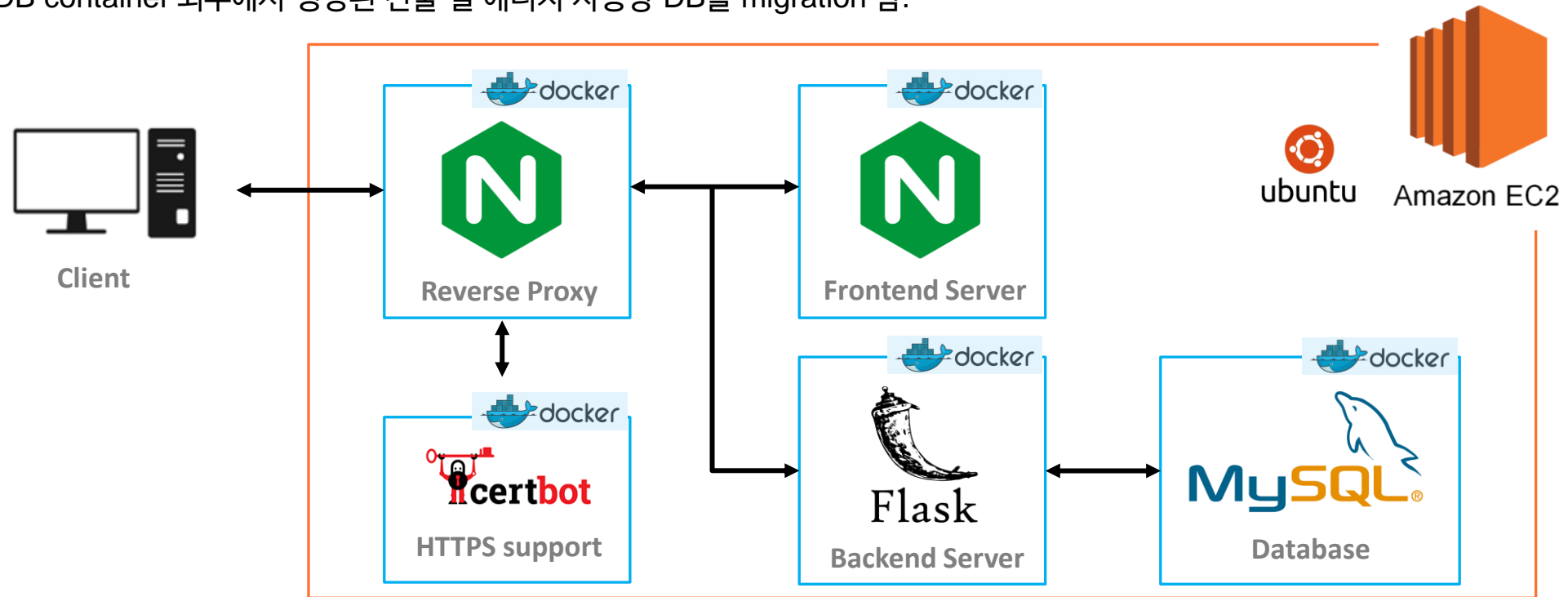


Item # 07

AWS EC2 인스턴스에 docker로 서버 구축, client가 입력한 주소에 대응하는 건물의 에너지 사용량 기록을 반환하는 마이크로 웹 서비스 구현

Client의 요청을 backend container (Flask) 에서 select query로 변환해 DB container에 전달, 조회된 레코드를 client에 반환

- ✓ AWS EC2와 docker-compose로, 공개된 HTTPS 도메인 주소로 접속한 클라이언트로부터 건물 주소를 요청받으면 해당 건물의 2021년의 각 월 별 전기 사용량을 클라이언트에게 보여주는 마이크로 서비스를 구현함.
- ✓ Proxy server는 Nginx, backend server는 Flask (Python), DB server는 MySQL임 (각각이 docker container).
- ✓ DB container 외부에서 생성된 건물 별 에너지 사용량 DB를 migration 함.



<서버 개요도>

Item # 07

AWS EC2 인스턴스에 docker로 서버 구축, client가 입력한 주소에 대응하는 건물의 에너지 사용량 기록을 반환하는 마이크로 웹 서비스 구현

Client의 요청을 backend container (Flask) 에서 select query로 변환해 DB container에 전달, 조회된 레코드를 client에 반환

- ✓ Client가 서울특별시 내 비주거용 건물의 주소를 입력하면, backend server에서 select query가 생성되고 해당 query가 DB container에 전달됨.
- ✓ 해당 건물 주소가 DB에 존재할 경우, 아래와 같이 해당 건물의 전기 사용량을 기재한 HTML 페이지를 반환함.

🔄 🔒 <https://song4energy.shop/util/bldgelectricity> 🔍 ↗ ☆ ⚙ □

공개 도메인

**서울특별시의 비주거용 건물 주소를 입력하세요
(서울특별시 제외, 새주소 아닌 지번주소로),
2021년의 월별 전기 사용량 기록을 조회합니다.**

예시: 송파구 잠실동 40-1

주소:



서울특별시 송파구 신천동 7-13의 2021년 월별 전기 사용량 (kWh)

1월	868379.0
2월	754938.0
3월	699952.0
4월	664736.0
5월	765950.0
6월	861476.0
7월	968335.0
8월	966590.0
9월	839454.0
10월	713692.0
11월	709494.0
12월	875455.0

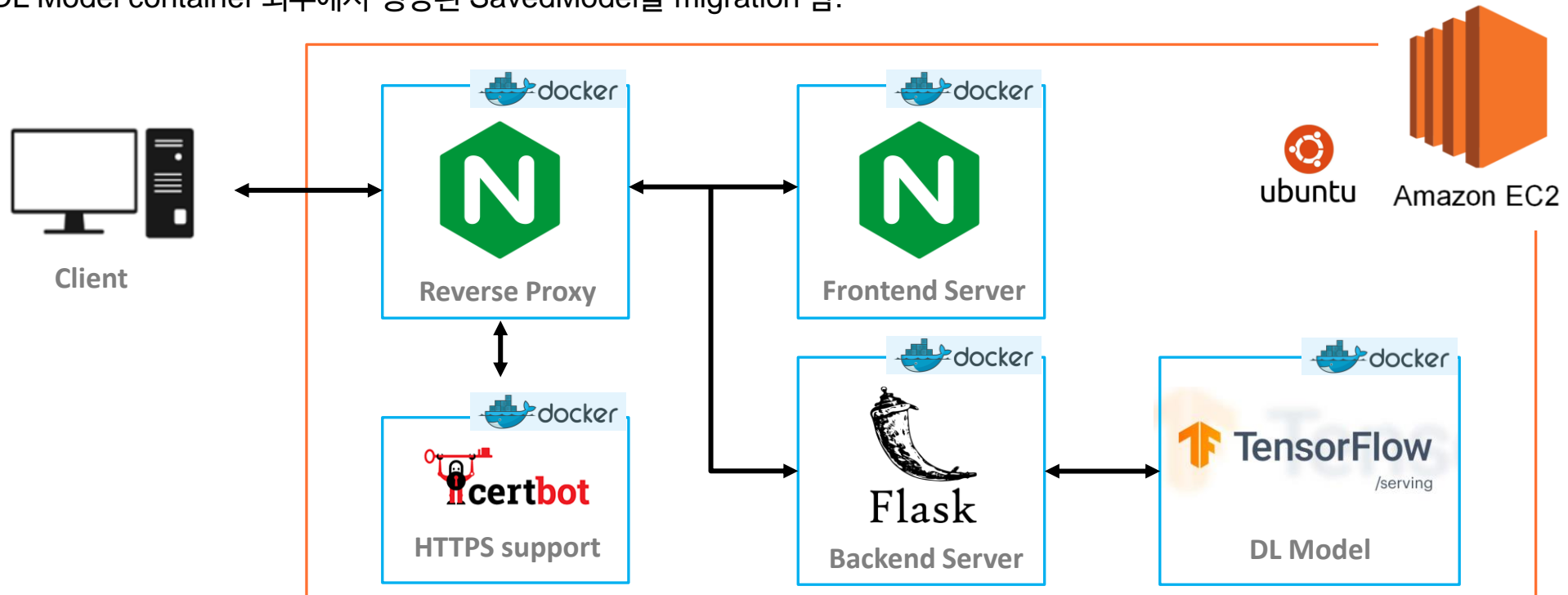
용도: 업무시설
연면적: 61723.15

Item # 08

AWS EC2 인스턴스에 docker로 서버 구축, client가 입력한 영화 리뷰 문장의 긍/부정 여부를 반환하는 마이크로 웹 서비스 구현

Client가 입력한 문장을 Tensorflow/Serving container에 전달, 자연어처리 모델이 긍정리뷰일 확률을 client에 반환

- ✓ AWS EC2와 docker-compose로, 공개된 HTTPS 도메인 주소로 접속한 클라이언트로부터 영화 리뷰 문장을 전달받으면 자연어처리 DL 모델 기반으로 해당 문장의 긍정/부정 여부를 클라이언트에게 보여주는 마이크로 서비스를 구현함.
- ✓ Proxy server는 Nginx, backend server는 Flask (Python), DL Model server는 Tensorflow/Serving임.
- ✓ DL Model container 외부에서 생성된 SavedModel을 migration 함.



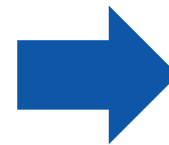
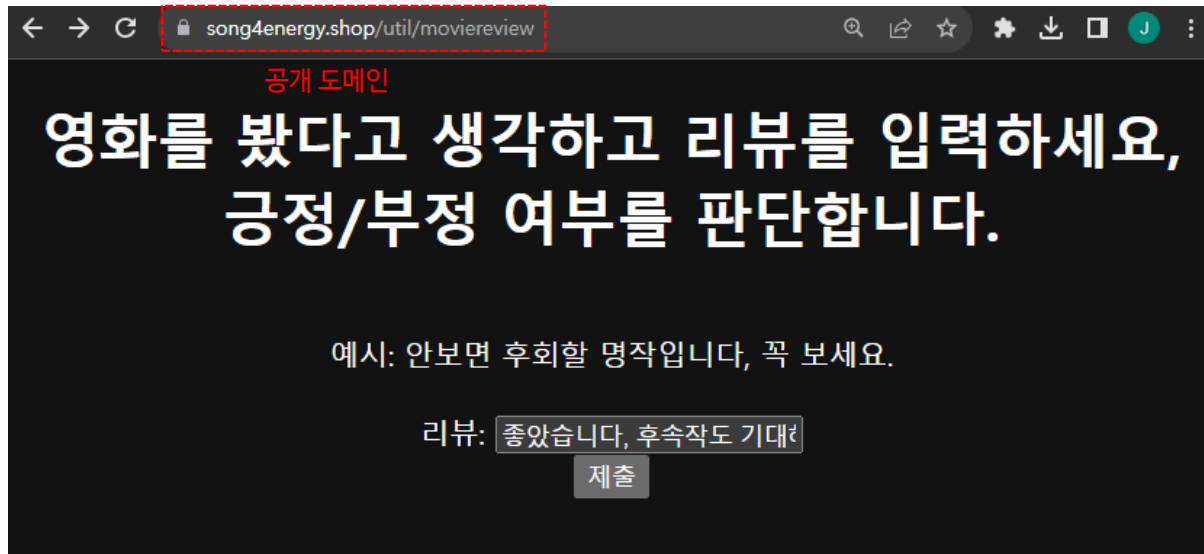
<서버 개요도>

Item # 08

AWS EC2 인스턴스에 docker로 서버 구축, client가 입력한 영화 리뷰 문장의 긍/부정 여부를 반환하는 마이크로 웹 서비스 구현

Client가 입력한 문장을 Tensorflow/Serving container에 전달, 자연어처리 모델이 긍정리뷰일 확률을 client에 반환

- ✓ Client가 영화리뷰 문장을 입력하면, backend server에서 해당 문장의 각 글자를 tokenize하고 token들의 모음 (tokenize된 문장) 을 HTTP POST request로 Tensorflow/Serving container에 전달함.
- ✓ 따라서 단어-token 디렉터리를 담은 JSON 파일을 Flask container의 volume에 저장해야 함.
- ✓ Tensorflow/Serving container 내의 자연어 처리 모델이 긍정 리뷰일 확률을 계산하고 backend server에 반환함.
- ✓ 해당 DL 모델 구조는 token들의 모음 (tokenize된 문장) 을 입력받는 Embedding layer, LSTM layer, Fully connected layer, sigmoid activation function을 갖는 output layer로 구성됨.



"좋았습니다, 후속작도 기대해요" 는
긍정적 리뷰입니다.
(긍정적 리뷰일 확률: 0.975390)
돌아가기

Item
09

건물별 월별에너지사용량 dataset과 건축속성정보 dataset 결합 후 SQL DB화

Python Pandas로 두 dataset을 inner join 후 SQLite DB로 저장

- ✓ Item #01의 누락치 추정 알고리즘 개발을 위해서는, 각 건물의 월별 에너지사용량 dataset을 건물별 건축속성정보 (면적/ 층수/ 건축연도/ 용도/ 재질 등) 의 dataset과 결합해야 함.

B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
addresstxt	sigunguCd	bjdongCd	bun	ji	elec_2022	elec_2022	elec_2022	elec_2022	elec_2022	elec_2022	elec_2022	elec_2022	elec_2022	elec_2022	elec_2022	elec_2022	gas_20220	gas_20220	gas_20220
서울특별시 종로구 청운동 6번지	11110	10100	6	0	1131	1226	899	346	266	304	379	2600	683	612	528	253			
서울특별시 종로구 청운동 6-3번지	11110	10100	6	3	1054	919	688	380	200	118	1126	1641	825	513	484	849	7221	5881	5819
서울특별시 종로구 청운동 30-7번지	11110	10100	30	7	3157	3515	3036	1823	812	405	396	380	432	355	23	199	668	465	513
서울특별시 종로구 청운동 30-13번지	11110	10100	30	13	13314	14170	11890	7706	5714	6398	7511	9168	6675	4942	5988	8804			
서울특별시 종로구 청운동 49-2번지	11110	10100	49	2	862	772	761	732	741	743	1700	2103	1754	1369	1425	1314	9909	10306	7882
서울특별시 종로구 청운동 49-3번지	11110	10100	49	3	810	659	613	578	564	605	66	133	95	43	9	12			
서울특별시 종로구 청운동 53-24번지	11110	10100	53	24	1433	2103	1458	826	749	728	909	1336	1234	751	669	645	4607	3735	2467
서울특별시 종로구 청운동 53-56번지	11110	10100	53	56	777	821	735	783	692	766	1346	1737	1596	1197	802	723	19079	17052	8840
서울특별시 종로구 청운동 57-8번지	11110	10100	57	8	24262	24685	21467	21949	19093	19497	24742	27818	22580	19357	20417	21069			
서울특별시 종로구 청운동 66번지	11110	10100	66	0	5806	5618	5274	5664	5283	5608	5609	6215	6344	5813	5602	5387	9599	8981	6345
서울특별시 종로구 청운동 67-1번지	11110	10100	67	1	104	129	93	94	102	122	92	171	156	142	89	91			
서울특별시 종로구 청운동 89번지	11110	10100	89	0	1	1	2	1	1	2	1	2		2	2		15951	14736	50672
서울특별시 종로구 청운동 89-109번지	11110	10100	89	109	2485	2805	767	92		83							8946	8528	1771
서울특별시 종로구 청운동 94-2번지	11110	10100	94	2	2977	2874	2634	3145	3409	3805	4765	6192	5304	4236	2752	2753	19268	18200	17304
서울특별시 종로구 청운동 108-21번지	11110	10100	108	21	8628	9485	8370	6738	5335	5537	6367	8070	7805	6648	5939	6189	8098	8146	5237
서울특별시 종로구 청운동 123번지	11110	10100	123	0	19000	19407	17381	18534	17519	17991	19923	21157	20192	17947	18623	18648	148279	114236	134369

<건물 지번별 각 월의 전기/ 도시가스 사용량 dataset>

순번	대지위치	시군구코드	법정동코드	대지구분코드	번	지	관리건축물대장PK	대장구분코드	대지면적(㎡)	건축면적(㎡)	건폐율(%)	연면적(㎡)	용적률산정연면적(㎡)	용적률(%)	구조코드	구조코드명	
74	서울특별시 종로구 청운동 30-7번지	11110	10100	0	0030	0007	11110-20597	1	354.7	106.33	29.98	171.52	171.52	48.36	11	벽돌구조	
75	서울특별시 종로구 청운동 30-13번지	11110	10100	0	0030	0013	11110-100208025	1	1114.9	321.28	28.82	998.3	829.98	74.44	21	철근콘크리트구조	철
76	서울특별시 종로구 청운동 30-15번지	11110	10100	0	0030	0015	11110-10746	1	791.2	227.02	28.69	472.23	440.26	55.65	21	철근콘크리트구조	

<건물 지번별 건축속성정보 dataset (표제부)>

Python Pandas로 두 dataset을 inner join 후 SQLite DB로 저장

- | Table: seoul_2022 Filter in any column | | | | | | | | | | | | |
|--|------------------------|----------|--------------|-----------|--------------|---------------|----------|-------------|-----------|--------------|--|--|
| | addresstxt | sitearea | buildingarea | floorarea | netfloorarea | num_household | num_gagu | structureCd | structure | structuresub | | |
| | Filter | Filter | Filter | Filter | Filter | Filter | Filter | Filter | Filter | Filter | | |
| 1 | 서울특별시 강남구 개포동 1055번지 | 3017.0 | 507.18 | 11729.79 | 1568.56 | 0 | 0 | 21.0 | 철근콘크리트구조 | 철근콘크리트조 | | |
| 2 | 서울특별시 강남구 개포동 1057-9번지 | 3062.0 | 697.07 | 2657.63 | 2215.79 | 0 | 0 | 21.0 | 철근콘크리트구조 | 철근콘크리트구조 | | |
| 3 | 서울특별시 강남구 개포동 1163-2번지 | 0.0 | 335.72 | 1125.96 | 671.44 | 0 | 0 | 21.0 | 철근콘크리트구조 | 철근콘크리트조 | | |

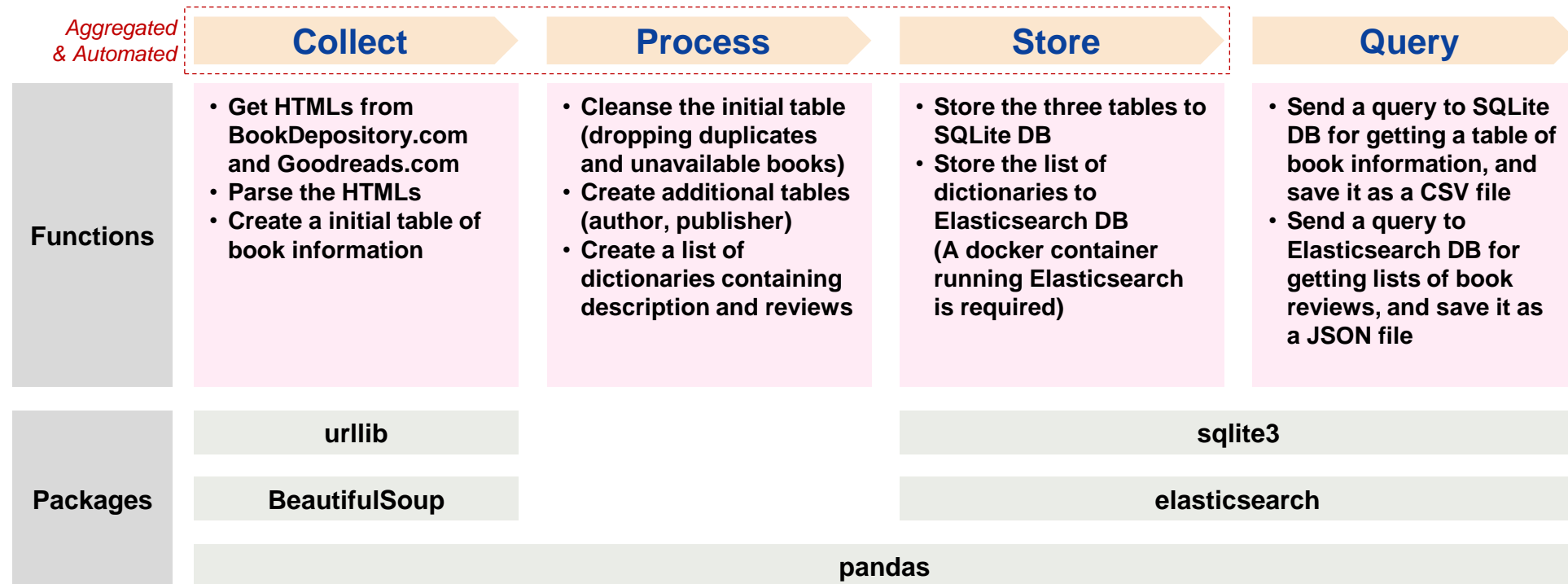
21

Item # 10

웹사이트 크롤링을 통한 데이터 수집 및 저장 data pipeline 구현 (SIAI Data management 교과목 프로젝트)

Python urllib/ BeautifulSoup로 수집, 책 속성 정보는 SQLite DB화, 소개/ 리뷰 등 텍스트 정보는 NoSQL DB (Elasticsearch) 화

- ✓ 해외 온라인서점 웹사이트¹⁾에 특정 검색어 검색 시 검색된 책들 각각의 상세페이지를 크롤링하고 데이터 정제 후 DB에 저장하는 data pipeline을 구축함.



<Data pipeline 개요: 단계별 기능 및 사용된 패키지들>

1) bookdepository.com. 프로젝트 수행 당시 (2022년) 에는 영업하였으나 2023년 4월에 폐업함.

Item
10

웹사이트 크롤링을 통한 데이터 수집 및 저장 pipeline 구현

(SIAI Data management 교과목 프로젝트)

Python urllib/ BeautifulSoup로 수집, 책 속성 정보는 SQLite DB화, 소개/ 리뷰 등 텍스트 정보는 NoSQL DB (Elasticsearch) 화

- ✓ 크롤링 및 파싱에는 Python urllib/ BeautifulSoup 라이브러리를 사용함.
- ✓ 책 속성 정보 (제목/ 저자/ 가격/ 서적고유번호(ISBN-13)/ 출판사/ 페이지수 등) 는 SQLite DB에 저장함.
 - ✓ 각 책을 row로 하는 table은 서적고유번호를 primary key로, 저자와 출판사를 foreign key로 가짐.
 - ✓ 저자 및 출판사를 각 row로 하는 별도의 table들이 추가 생성되며, 각각 저자와 출판사를 primary key로 가짐.
- ✓ 책 소개글 및 독자 리뷰 등 장문의 텍스트 정보는 JSON schema 기반 NoSQL DB (Elasticsearch) 에 저장함.
 - ✓ Python Elasticsearch 라이브러리에서 기본 제공하는 put 메서드 대신 helpers 라이브러리의 bulk 메서드를 사용해, 다수의 책들에 대한 텍스트 정보를 더 빠르게 저장하도록 함.
 - ✓ Elasticsearch는 Docker for Windows로 구동하였음.

Table: Booksinfoeconometrics 검색어

	ISBN_13	Title	Author	Publisher	Rating	NumofRatings
	Filter	Filter	Filter	Filter	Filter	Filter
1	978069101...	Econometrics	Fumio Hayashi	Princeton University Press	4.08	108
2	978069112...	Mostly Harmless Econometrics : An Empiricist's ...	Joshua Angrist	Princeton University Press	4.15	855
3	978140518...	A Guide to Econometrics	Peter Kennedy	John Wiley and Sons Ltd	4.17	291

Primary key

<SQLite DB로 저장된 책 속성 정보>

```
{
  "_index" : "reviews_econometrics",
  "_type" : "_doc",
  "_id" : "s_DaDYMBjmOGRA1XShfI",
  "score" : 0.25247142,
  "_ignored" : [
    "Review.keyword"
  ],
  "_source" : {
    "title" : "Econometrics",
    "URL" : "https://www.bookdepository.com/Econometrics-Fumio-Hayashi/9780691010182?ref=grid-view&qid=1662384228079&sr=1-1",
    "Review" : "I have some misgivings about the field of econometrics, fundamental elements are missing, and traditionally even if GMM estimation, is within the realm of admissible methods in statistics, it is quite fringe. This is also true of MA, ARMA and ARIMA for me, because in the case of economics they are not backed by fundamentals, nor do they elucidate them, given a theoretical mismatched. This is on some level, the same misgivings I have for Robert Lucas Jr.'s attempt to bridge micro to macro, through the use of statistical methods. There is some intuition or concept gap, where certain things are not being bridged. At least from what limited I have read. Perhaps, this is worthy of some further exploration."
  }
}
```

<Elasticsearch에 JSON으로 저장된 리뷰 정보>