# [ Tabular Data for 장기 예측 ]

## (1) 데이터 전처리 / 군집화

### 3개의 스크립트 파일

- `clustering.py` : (NLP 기반) 부품명 기준 군집화
  - output : 부품별 cluster 인덱스
- `preprocess.py` : tabular 데이터 전처리
  - `preprocess_utils.py` : tabular 데이터 전처리 위한 함수/모듈

## (2) 모델링

### 5개의 스크립트 파일

- `train_utils.py` : VIME 알고리즘 위한 함수 및 데이터로더
  - Data loader 관련 함수
  - SSL task 관련 함수
  - Early stopping 등
- 모델 관련 파트
  - `model_backbone.py` : backbone 모델
  - `model_SSL.py` : SSL task 관련 모델 ( = VIME 알고리즘 )
- 학습 관련 파트
  - `pretrain_VIME.py`
- 추론 관련 파트
  - `eval_VIME.py`

**산출물 : 부품별 tabular 정보를 담고 있는 128/256차원 임베딩벡터**

# Process

- step 1) `clustering.py` : 부품명 군집화
- step 2) `preprocess.py` : 부품 데이터 전처리 ( + 부품명 군집화 정보 붙이기 )

- step 3) `pretrain_VIME.py` : 부품 데이터 바탕으로 VIME 알고리즘 학습
- step 4) `eval_VIME.py` : VIME 사용하여, 부품 별 임베딩 벡터 추출