2021311558 임형구

오토 인코더를 활용한 로그 데이터 이상탐지

1. 연구 배경

연구실에서는 주로 제조 관련 장비들에서 발생하는 시계열 수치 데이터를 바탕으로 이상탐지를 하는 연구를 했었습니다. 본 수업에서는 word2vec skipgram등 언어 모델에 대해 대해서 공부하였습니다. 따라서 이상탐지와 언어 모델을 결합한 분야를 연구 주제로 설정하였습니다.

2. 연구 주제

원자력 발전소에서 나오는 로그 데이터를 NLP로 접근하여 다양한 방식으로 전처리(불필요한 기호 및 숫자 제거)와 토큰화를 진행하여 이상탐지를 해보려고 합니다. 기존의 연구들은 랜덤포레스트와 같은 지도 학습으로 접근 하였지만, 이는 새로운 종류의 이상 상황이 발생했을 때는 취약한 것을 알 수 있습니다. 로그데이터를 다양한 방법으로 토큰화하여 중요 단어들을 뽑아내고 비지도 학습인 오토인코더를 통하여 이상탐지를 해보려고 합니다.

데이터 출처: https://dacon.io/competitions/official/235717/overview/description

3. 연구 목표 및 세부 내용

- 로그 데이터를 토큰화 하는 과정에서 다양한 토큰화 라이브러리 활용(countvectorizer, tfidvectorizer, n-gram 등)
- 오토 인코더 모델을 활용하여 reconstruction-error를 통한 이상탐지 접근 (오토인코더의 layer를 추가하는 과정에서 언어 모델에 적합한 layer들을 테스트)
- 추가적으로 그래프 임베딩을 사용하여 이상탐지가 가능할지에 대한 공부

4. 예상되는 결과물

- 전처리 방식에 따른 보안 등급 성능 변화 (숫자 문자 대소문자 제거 여부 등)
- 토큰화 방식에 따른 보안 등급 성능 변화 (f1-score)
- 오토 인코더의 layer의 구성에 따른 성능 변화 (f1-score)