|  |  |
| --- | --- |
| 교육 제목 | AI 프로젝트 기반 S/W 전문가 양성과정 |
| 교육 일시 | 22.01.19 |
| 교육 장소 | 집 |
| **교육 내용** | |
| 오전 | 1. 사전에 없는 단어의 문제 OOV ex) 코로나바이러스는 2019년 12월 중국 우한에서 처음 발생한 뒤 전 세계로 확산된, 새로운 유형의 호흡기 감염 질환입니다. 🡪 코로나바이러스, 우한을 모름, 이것이 OOV 2. Wordpiece Model(WPM) 🡪 preview, predict의 pre(미리, 이전)을 떼냄 한 단어를 여러 개의 Subword의 집합으로 보는 것이 WPM 3. Byte Pair Encoding(BPE) 데이터 압축을 위해 생겨남, Byte Pair를 새로운 단어로 치환하여 압축 aaabdaaabac # 가장 많이 등장한 바이트 쌍 “aa”를 “Z”로 치환… 계속 진행 4. Byte Pair Encoding(BPE) 2 BPE 압축 과정 ① 딕셔너리 생성(단어 : 빈도) ② Character 단위로 분절 후 pair 별 빈도 카운트 ③ 최빈도 pair를 골라, merge 수행 ④ pair별 빈도 카운트 업데이트 ⑤ 3번 과정 반복 5. Wordpiece Model(WPM) 구글에서 BPE를 변형해 제안한 알고리즘이 바로 WPM 공백을 복원하여 단어의 시작 부분에 언더바\_를 추가 빈도수 기반이 아닌 가능도(likelihood)를 증가시키는 방향으로 문자 쌍을 합침 가능도 🡪 내가 뽑을 확률 6. IMDB, NAVER 영화 리뷰 실습 |
| 오후 | 1. 과제 실습 |