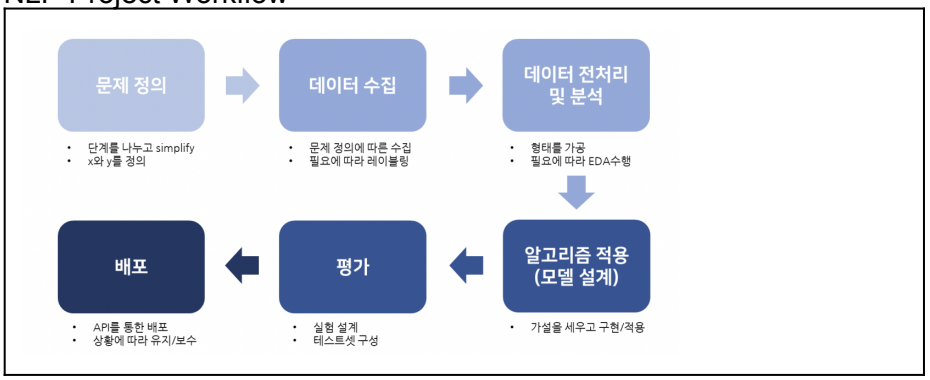
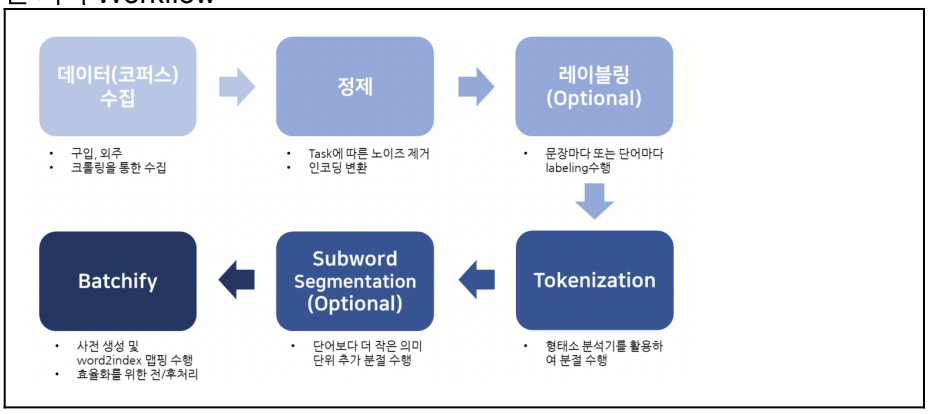
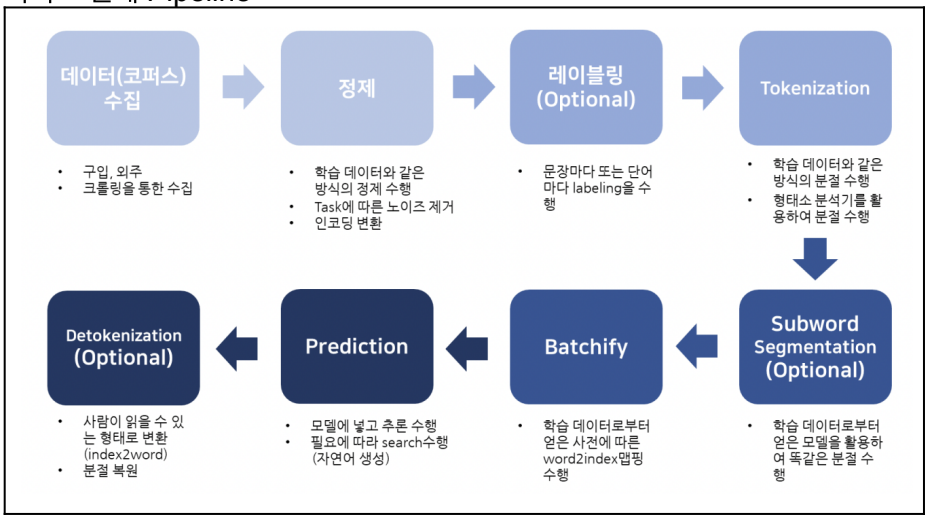
교육 제목	데이터 기반 인공지능 시스템 엔지니어 양성 과정
교육 일시	2021년 11월26일
교육 장소	YGL C-6 학과장 & 자택(디스코드 이용한 온라인)
교육 내용	
오전	자연어 처리 1. NLP Project Workflow  2. 전 처리 Workflow  3. 서비스 전체 Pipeline 

4. Data Cleaning

정규식(Regular expression)을 활용하면 복잡한 규칙의 노이즈도 제거/치환 가능
코딩 없이 단순히 텍스트 에디터(Sublime Text, VSCode등)도 가능

```
>>> # 주민등록번호 형식을 변경
>>> re.sub("-", "@", "901225-1234567")
'901225@1234567'
>>> # 필드 구분자를 통일
>>> re.sub(r"[:,|\s]", ", ", "Apple:Orange Banana|Tomato")
'Apple, Orange, Banana, Tomato'
>>> # 문자열의 변경 횟수를 제한
>>> re.sub(r"[:,|\s]", ", ", "Apple:Orange Banana|Tomato", 2)
'Apple, Orange, Banana|Tomato'
```

5. Tokenization

한국어의 경우
1) 접사를 분리하여 희소성을 낮추고
2) 띄어쓰기를 통일하기 위해 tokenization을 수행

굉장히 많은 POS Tagger가 존재하는데,
- 전형적인 쉬운 문장(표준 문법을 따르며, 구조가 명확한 문장)의 경우, 성능이 비슷함
- 하지만 신조어나 고유명사를 처리하는 능력이 다름
- 따라서, 주어진 문제에 맞는 정책을 가진 tagger를 선택하여 사용해야 함.

6. Tokenization Style의 특성

토큰 길이가 짧을 수록

- Vocabulary 크기 감소
  - 희소성 문제 감소
- OOV가 줄어듬
- Sequence의 길이가 길어짐
  - 모델의 부담 증가
- 극단적 형태
  - Character단위

토큰 길이가 길 수록

- Vocabulary 크기 증가
  - 희소성 문제 증대
- OOV가 늘어남
- Sequence의 길이가 짧아짐
  - 모델의 부담 감소

토큰 길이에 따른 Trade off가 존재

- 빈도가 높을 경우, 하나의 token으로 나타내고,
- 빈도가 낮을 경우 더 잘게 쪼개어, 각각 빈도가 높은 token으로 구성한다.

압축 알고리즘?

오후	별도 실습 파일 참조