|  |  |
| --- | --- |
| 교육 제목 | **NLP,** **Tokenization** |
| 교육 일시 | 2021.11.26 |
| 교육 장소 | 비대면 |
| **교육 내용** | |
|  | 정규표현식 중요 = 전 처리 중요  영어: nltk, 한국어: ko nltk  **Tokenization**  문장을 어떤 기준으로 쪼겠 을 때, 쪼개진 각 단어들을 **토큰(Token)**   쪼개진 기준이 **토큰화(Tokenization) 기법**에 의해 정해진다  - Sentence Segmentation  -보통 훈련 시 우리가 원하는 데이터는sentence/line  -우리가 수집한 corpus는 한 라인에 여러 문장이 들어있거나,  한 문장이 여러 라인에 들어있음  - Sentence Segmentation을 통해 원하는 형태로 변환  - 마침표 등을 단순히 문장의 끝으로 처리하면 안됨!  - NLTK를 활용하여 변환 가능  Multiple sentence/line, /sentence  -Tokenization:  두 개이상의 다른 token들의 결합으로 이루어진 단어를 쪼개어,  Vocabulary 숫자를 줄이고, 희소성(sparseness)를 낮추기 위함  형태소 분석 및 품사 태깅 (Part of Speech Tagging)  형태소 분석: 형태소를 비롯하여, 어근, 접두사/접미사,  품사(POS, part-of-speech) 등 다양한 언어적 속성의 구조를 파악하는 것  품사 태깅: 형태소의 뜻과 문맥을 고려하여 그것에 마크업을 하는 일  토큰 길이가 짧을수록 토큰 길이가 길 수록  - Vocabulary 크기 감소 - Vocabulary 크기 증가  -희소성 문제 감소 - 희소성 문제 증대  - OOV가 줄어 든다 - OOV가 늘어남  - Sequence의 길이가 길어진다 -Sequence의 길이가 짧아짐  - 모델의 부담 증가 -모델의 부담 감소  - 극단적 형태  - Character단위 |
|  | -한국어의 경우  - 접사를 분리하여 희소성을 낮추고  - 2띄어쓰기를 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  **형태소 토큰화**    **단어기반 인코딩**    **텍스트를 시퀀스로 변환하기**    **패딩 설정**    **한글-영어 코퍼스(실제 데이터로)로 토큰화 수행하기**        **중복제거**    **가장 긴 데이터 기준으로 padding**    **공백 기반 토큰화**    **형태소 기반 토큰화**    **문제** |