|  |  |
| --- | --- |
| 교육 제목 | 자연어처리 Embedding |
| 교육 일시 | 2021.11.30 |
| 교육 장소 |  |
| **교육 내용** | |
|  | 시퀀스 로 변환 -> 문장마다 길이가 다르다 -> 짧은 문장을 긴 문장으로 바꿔 줌 (padding) ->  **단어보다 더 작은 의미 단위: sub word**  언어별 사전이 필요하다  많은 언어 들에서, 단어는 더 작은 의미 단위들이 모여 구성됨  따라서, 이러한 작은 의미 단위로 분절할 수 있다면 좋을 것  하지만, 이를 위해선 언어별 sub word사전이 존재해야 할 것.  **서브 워드 분리 (Subworld segmentation)잡업**  : 하나의 단어는 더 작은 단위의 의미 있는 여러 서브 워드의 조합으로 구성된 경우가 많기 때문에 하나의 단어를 여러Subworld로 분리해서 단어를 인코딩 및 임베딩 하겠다는 의도를 가진전처리작업  = 이를 통해OOV나희귀단어, 신조어 같은 문제들을 완화  이런 작업을 하는 토크나이저 == **서브 워드 토크나이저(Sub word tokenizer)**  =>OOV문제를완화하는대표적인서브워드분리알고리즘인  **BPE (Byte Pair Encoding)**  - 학습 코퍼스를 활용하여BPE 모델을 학습 후, 학습/ 테스트 코퍼스에 적용  - 장점:  - 희소성을 통계에 기반하여 효과적으로 낮출 수 있다.  - 언어별 특성에 대한 정보 없이, 더 작은 의미 단위로 분절할 수 있다.  - OOV를없앨수있다. (seen character로만구성될경우)  - 단점:  - 학습데이터별로BPE 모델도 생성됨  **Word piece Model (WPM**)  구글에서 BPE를 변형해 제안한 알고리즘이 바로 WPM.  WPM은 BPE에 대해 **두 가지 차별성**을 가짐.  -공백 복원을 위해 단어의 시작 부분에 언더바 \_ 를 추가.  -빈도수 기반이 아닌 가능도(Likelihood)를 증가시키는 방향으로 문자 쌍을 합침  **Word Embedding**  One-hot 인코딩으로 표현된 값은 유사도나 모호성을 표현할 수 없다.   * Dense vector로 표현하는 것이 유리   Feature (특징)  - 샘플을 잘 설명하는 특징  - 특징을 통해 우리는 특정 샘플을 수치화 할 수 있다.  - Feature Vector  - 각 특징들을 모아서 하나의 vector로 만드는 것  WordNet(거리)  Thesaurus(단어 유사도 기반 어휘 사전), 지금은 잘 쓰지 않음  **Word Feature Vectors: Traditional Methods**  **-Bag of words (BOW)**  -단어들의 순서는 전혀 고려하지 않고,  -단어들의 출현 빈도(frequency)에만 집중하는 텍스트 데이터의 수치화 표현 방법  **-문서 단어 행렬(Document-Term Matrix, DTM)**  - 다수의 문서에서 등장하는 각 단어들의 빈도를 행렬로 표현한 것  - 각 문서에 대한 BOW를 하나의 행렬로 만든 것  - Bow표현을 다수의 문서에 대해서 행렬로 표현하고 부르는 용어  DTM의 한계  1) 희소 표현(Sparse representation)  - 각 문서 벡터의 차원은 원-핫 벡터와 마찬가지로 전체 단어 집합의 크기를 가짐  - 많은 문서벡터가 대부분의 값이 0을 가질 수도 있음 → 희소벡터, 희소행렬  - 희소벡터는 많은 양의 저장공간과 계산을 위한 리소스가 필요  2) 단순 빈도 수 기반 접근  - 예를 들어 영어에 대한 DTM을 만들었을 때, 불용어인 The는 어떤 문서 든 자주 등장한다.  - 그런데, 유사한 문서인지 비교하고 싶은 문서1, 문서2, 문서3에서 동일하게 the가 빈도수가 높다고 해서 이 문서들이 유사한 문서라고 판단 X  **TF-IDF**  • 텍스트 마이닝(Text Mining)에서 중요하게 사용  • 어떤 단어𝑤가문서𝑑 내에서 얼마나 중요한 지 나타내는 수치  • TF(Term Frequency)  • 단어 의문서내에 출현한횟수  • 숫자가 클수록 문서내에서 중요한단어  • 하지만, 'the'와 같은 단어도 TF값이 매우 클 것  • IDF(Inverse Document Frequency)  • 그 단어가 출현한 문서의 숫자의 역수(inverse)  • 값이 클수록 'the' 와 같이 일반적으로 많이 쓰이는 단어 |
|  | 소스코드: NLP\_day3.ipynb |