인공지능기초응용II

9주차 과제

인공지능응용

K2025029 금동환

목차

1. [문서벡터 구성의 문제점](#문서벡터)
2. [워드임베딩이란](#워드임베딩)
3. [word2vec, fastText, GloVe 특징 및 차이점](#word2vecfastTextGloVe)
4. [ELMo와 BERT의 특징 및 차이점](#ELMoBERT)

1. [문서벡터 구성의 문제점](#목차)
   1. 문서 벡터를 구성할 때 키워스가 수심만~수백만개로 늘어나면서 메모리, 저장 공간문제가 발생하며, 연산효율이 떨어짐
      * + 저장 공간의 문제
   2. PCA, LDA, LSA 등의 차원 축소 기법을 활용하여 고차원 데이터를 저차원으로 투영해야 함.
      * + 무엇을 feature 로 할것인가, feature selection 문제

1. [워드임베딩이란](#목차)

* 단어를 고차원 벡터가 아닌 저차원 밀집 벡터로 표현하는 기법

1. [word2vec, fastText, GloVe 특징 및 차이점](#목차)
   1. word2vec의 특정
      1. CBOW
         * 좌우 문맥으로부터 현재 단어 예측
         * 학습속도 빠름, 대규모 학습말뭉치에 적합
      2. Continuous skip-gram
         * 빈도가 높은 단어 벡터를 잘 표현
         * 학습속도 느림
   2. fastText의 특징
      1. 단어응 서브워드(subword, 문자 n-gram) 단위로 분해하여 벡터를 구성
         * 학습되지 않은 단어(OOV, Out of Vocabulary)를 효과적으로 처리
      2. 형태소가 풍부한 교착어에 강점, 새로운 단어나 희귀 단어의 의미를 보다 잘 반영
   3. GloVe의 특정
      1. 전체 말뭉치의 행렬을 기반으로 하는 행렬 분해 방식
      2. 단어 간의 관계를 더욱 포괄적으로 표현 가능
   4. 차이점

|  |  |  |  |
| --- | --- | --- | --- |
| 기법 | 학습 방법 | 장점 | 단점 |
| word2vec | 로컬 문맥 기반 | 학습속도 빠름, 간단한 구조 | OOV 처리 한계 |
| fastText | 서브워드 단위 | 형태소 언어 강점, | 모델의 크기 증가, 학습속도 느림 |
| GloVe | 글로벌 행렬 | 단어의 관계 반영 우수 | 메모리 사용 증가, 계산 복잡성 |

1. [ELMo와 BERT의 특징 및 차이점](#목차)
   1. ELMo의 특징
      1. 랭귀지 모델을 좌>우로 반영
   2. BERT의 특정
      1. 랭귀지 모델을 양항뱡으로 반영
   3. 차이점

|  |  |  |  |
| --- | --- | --- | --- |
| 기법 | 학습 방법 | 장점 | 단점 |
| ELMo | LSTM 기반 동적 언어 모델 | 문맥별로 동적인 의미 포착, 형태학적 특징 활용 우수 | 긴 문장 처리 시 속도 느림, 순차적 연산으로 병렬처리 제한 |
| BERT | Transformer 기반 Masked Language Model, Next Sentence Prediction | 양방향으로 깊은 문맥 이해 가능, 다양한 NLP 작업에서 높은 성능 | 모델의 크기, 자원 소모 큼. 사전 학습과 파인튜닝에 자원 소모 큼. |