|  |  |
| --- | --- |
| 교육 제목 | **Word2Vec** |
| 교육 일시 | 2021.12.02 |
| 교육 장소 | C6 |
| **교육 내용** | |
|  | **Word Embedding**  = 자연어를 컴퓨터가 이해하고 효율적으로 처리하게 하기 위해서는 컴퓨터  가 이해할 수 있도록 자연어를 적절히 변환할 필요가 있다.  단어 표현 방법에 따라서 자연어 처리의 성능이 크게 달라짐 = 다양한 연구  **희소표현 (Sparse Representation**)  - 원\_핫 인코딩을 통해서 나온 원\_핫 벡터들은 표현하고자 하는 단어의 인덱스의 값만 1이고, 나머지 인덱스에는 전부 0으로 표현되는 벡터 표현 방법  - 벡터 또는 행렬(matrix)의 값이 대부분이 0으로 표현되는 방법을 희소 표현(  - 문제점  -단어의 개수가 늘어나면 벡터의 차원이 한없이 커진다는 점  **-밀집표현 (Dense Representation)**  : 밀집 표현은 벡터의 차원을 단어 집합의 크기로 상정x.  사용자가 설정한 값으로 모든 단어의 벡터 표현의 차원을 맞춤   * 이 과정에서 더 이상 0과 1만 가진 값이 아니라 실수값 을 가짐   **워드 임베딩 (Word Embedding)**  : 단어를 밀집 벡터(dense vector)의 형태로 표현하는 방법을 워드 임베딩  - 이 밀집 벡터를 워드 임베딩 과정을 통해 나온 결과라고 하여  == 임베딩 벡터(embedding vector)  - **워드 임베딩 방법론으로는( LSA, Word2Vec, FastText, Glove )**등이 있다  -Word2Vec  Objective:  : 주변(context window)에 같은 단어가 나타나는 단어일 수록 비슷한 벡터 값을 가져야 한다.  -문장의 문맥에 따라 정해지는 것이 아님  - context window의 사이즈에 따라 embedding의 성격이 바뀔 수 있다  **Skip-gram: Basic Concept**  기본 전략: 주변 단어를 예측하도록 하는 과정에서 적절한 단어의 임베딩(정보의 압축)을 할 수 있다.  - Non-linear activation func. 이 없음  - 기본적인 개념은 오토 인코더와 굉장히 비슷함  - y를 성공적으로 예측하기 위해 필요한 정보를 선택/압축  - 장점(at that time):  • 쉽다. • 빠르다. • 비교적 정확한 벡터를 구할 수 있다.  - 단점(currently)::  • 출현 빈도가 적은 단어일 경우 벡터가 정확하지 않다  **Glove**  - 단어x와 윈도우 내에 함께 출현한 단어들의 출현빈도를 맞추도록 훈련  - 출현빈도가 적은 단어에 대해서는loss의기여도를낮춤  - 따라서 출현빈도가 적은 단어에 대해 부정확 해 지는 단점을 보완  - 장점:  • 더 빠르다 • 출현빈도가 적은 단어도 벡터를 비교적 정확하게 잘 구할 수 있다  •전체 코퍼스에 대해 각 단어 별co-occurrence를 구한 후, regression을 수행  **Fast Tex**t: (Upgrade Version of Skip-gram)  Motivation:  기존의word2vec은 저 빈도 단어에 대한학습과OOV에대한대처가어려웠음  - Fast Text는학습시,  1) 단어를sub word로 나누고,  2) Skip-gram을 활용하여, 각sub word에 대한embedding vector에 주변단어의context vector를 곱하여 더한다.  3) 이 값이 최대가 되도록 학습을 수행한다.  = 최종적으로 각sub word에 대한embedding vector의 합  = word embedding vector가된다 |
|  |  |