Word Vectors, Advanced RNN, and Embedding Visualization

Introduction to Word Embeddings

텍스트를 처리하는 전통적인 NLP 접근법 중 하나는 문장의 각 단어를 정수 ID로 표시하는 방식이다. 앞 장에서 했던 것처럼 예를 들어 ‘agriculture’라는 단어는 정수 3452로, ‘farm’은 12, ‘AI’는 150, ‘deep-learning’은 0으로 매핑할 수 있다.

이런 표현을 사용함으로써 기초적인 NLP 작업에서는 실제로 훌륭한 결과를 얻었지만 몇 가지 중요한 문제가 내제되어 있다. 먼저 이러한 고립된 표현을 사용함으로써 단어 내에 숨어 있는 모든 의미를 잃어버리게 되고 따라서 단어 사이의 의미적 근접성과 관련 정보를 놓치게 된다. 예를 들어 한 문장 안에서 agriculture와 farm, AI와 deep-learning은 서로 밀접한 단어인 반면, deep-learning과 farm은 서로 별 관련이 없다. 이러한 내용은 임의의 정수 ID에 반영되지 않는다.