임베딩(Embedding)이란 무엇인가?

임베딩(Embedding)은 자연어 처리(NLP)에서 단어, 문장, 문서 등 텍스트 데이터를 고정된 크기의 실수 벡터로 변환하는 방법입니다. 임베딩을 통해 컴퓨터가 텍스트의 의미적, 문법적 특성을 수치적으로 이해할 수 있게 됩니다.

- **특징**
  - 고차원(One-hot, BoW 등) → 저차원(임베딩 벡터)으로 변환
  - 의미적으로 유사한 단어들은 임베딩 공간에서 가까운 위치에 배치됨
  - 대표적인 임베딩 방법: Word2Vec, GloVe, FastText 등

백오브워즈(Bag of Words, BoW) 가정

BoW는 문서를 단어들의 순서를 무시하고, 단어의 등장 빈도만을 고려하는 대표적인 텍스트 표현 방법입니다.

BoW의 주요 가정

가정 내용	설명
단어 순서 무시	문장 내 단어의 순서는 중요하지 않음. "나는 밥을 먹었다"와 "밥을 나는 먹었다"는 동일하게 처리됨.
단어 빈도만 고려	각 단어가 문서에 몇 번 등장했는지만 반영함. 단어의 위치, 문맥 등은 고려하지 않음.
문서의 길이 반영 가능	단어 빈도 벡터의 합을 통해 문서의 길이(단어 수)도 간접적으로 반영됨.

BoW 예시

아래는 두 개의 문서와 BoW 벡터화 과정을 표로 나타낸 예시입니다.

문서 번호	문장	단어 집합(어휘)	BoW 벡터 (나는, 밥을, 먹었다)
1	나는 밥을 먹었다	나는, 밥을, 먹었다	(1, 1, 1)
2	밥을 나는 먹었다	나는, 밥을, 먹었다	(1, 1, 1)
3	나는 밥을 먹지 않았다	나는, 밥을, 먹었다, 먹지, 않았다	(1, 1, 0, 1, 1)

- 위 표에서 볼 수 있듯이, BoW는 단어의 순서를 무시하고 각 단어가 몇 번 등장했는지만 벡터로 표현합니다.

BoW의 한계

- 단어의 순서와 문맥 정보를 반영하지 못함
- 의미적으로 유사한 단어(예: "먹었다", "식사했다")를 구분하지 못함
- 희소(sparse)하고 고차원 벡터가 생성됨

임베딩과 BoW의 관계

- BoW는 가장 단순한 임베딩(벡터화) 방법 중 하나입니다.
- 최근에는 단어 간 의미적 관계를 반영하는 임베딩(Word2Vec 등)이 BoW의 한계를 극복하기 위해 널리 사용됩니다.