카운트 기반의 단어 표현(Count based word Representation) : 문서 단어 행렬 (Document-Term Matrix, DTM)

문서 단어 행렬을 통해 서로 다른 문서들을 비교할 수 있음

문서 단어 행렬의 표기법

- 문서 단어 행렬 : 다수의 문서에서 등장하는 각 단어들의 빈도를 행렬로 표현한 것. 즉, 서로 다른 문서들의 BoW들을 결합한 표현 방법.
- example : 4개의 문서를 문서 단어 행렬로 만들기

  문서1 : 먹고 싶은 사과

  문서2 : 먹고 싶은 바나나

  문서3 : 길고 노란 바나나 바나나

  문서4 : 저는 과일이 좋아요

-	과일이	길고	노란	먹고	바나나	사과	싶은	저는	좋아요
문서1	0	0	0	1	0	1	1	0	0
문서2	0	0	0	1	1	0	1	0	0
문서3	0	1	1	0	2	0	0	0	0
문서4	1	0	0	0	0	0	0	1	1

  각 문서에서 등장한 단어의 빈도를 행렬의 값으로 표기한 것이기에 **문서들을 서로 비교할 수 있도록 수치화할 수 있다는 점**에서 의의를 갖음. **형태소 분석기와 불용어 제거를 적용**한다면 더 정제된 문서 단어 행렬을 만들 수 있음.

문서 단어 행렬의 한계

- 희소 표현
  원-핫 벡터처럼 단어 집합의 크기가 벡터의 차원이 되기에, 전체 코퍼스가 커지면 커질수록 대부분의 값이 0인 희소 행렬(sparse matrix) (또는 희소 벡터(sparse vector))이 된다. 이러한 이유로 문서 단어 행렬을 사용할 때는 전처리를 통해 단어 집합의 크기를 줄이는 것이 중요 → **구두점 제거, 빈도수가 낮은 단어 제거, 불용어 제거 등이 희소 행렬 문제를 완화시킴**
- 단순 빈도 수 기반 접근
  단순하게 단어의 빈도 수를 이용하여 문서를 표현하기에 때때로 한계를 가짐. 예를 들어 'the'라는 단어는 어떤 문서든 자주 등장하므로 문서1, 문서2, 문서3에서 동일하게 'the'의 빈도수가 높다고 이 문서들이 유사한 것은 아님. 즉 불필요한 단어(자주 등장하는 단어)와 중요한 단어(몇몇 문서에만 등장하는 단어. 즉, 불용어에 비해 주제가 있는 단어)가 혼재되어 있음 → **해결책 : TF-IDF**. 불필요한 단어에 비해 중요한 단어에 가중치를 부여함

출처

- https://wikidocs.net/24559