연구 대시보드 자연어 처리 알고리즘 조사

**공통적인 텍스트 전처리**

단어 토큰화(구두점 제거, 띄어쓰기로 토큰화 해도 괜찮을까)

노이즈 데이터(불용어) 를 제거한다. :

영어권 언어에서는 길이가 짧은 단어를 삭제하는 것만으로도 어느정도 자연어 처리에서 크게 의미가 없는 단어들을 제거하는 효과를 볼 수 있다고 알려져 있습니다. (it, at, to, on, in, by)

정규화(normalization) : 표현 방법이 다른 단어들을 통합시켜서 같은 단어로 만들어준다.

(ex: 단어 임베딩, 주의: 비슷한 단어로 생각하는 폭이 넓어 프로그램의 본질을 해칠 수 있음)

1. **Bag of Words (Bow)**

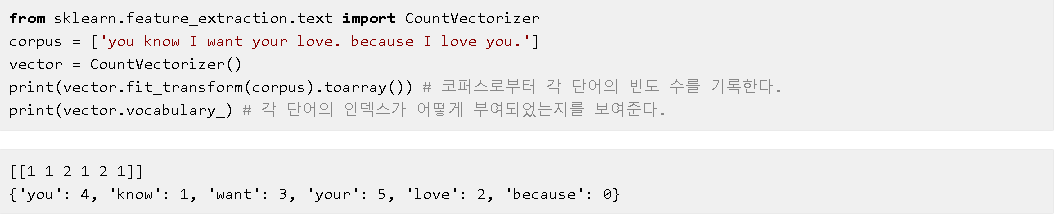
Bag of Words란 단어들의 순서는 고려하지 않고, 출현 빈도(frequency)에만 집중하는 텍스트 데이터의 수치화 표현 방법.

1. 우선, 각 단어에 고유한 정수 인덱스를 부여합니다.
2. 각 인덱스의 위치에 단어 토큰의 등장 횟수를 기록한 벡터를 만듭니다.

Example code



단어의 빈도를 Count하여 Vector로 만드는 CountVectorizer 클래스를 지원합니다





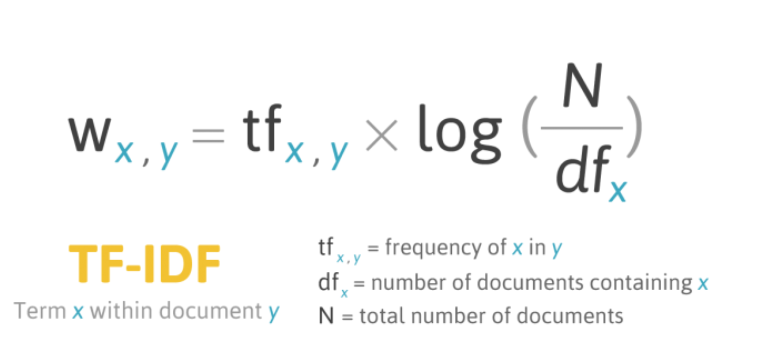
1. **TF-IDF(Term Frequenct-Inverse Document Frequency)**

TF(단어 빈도, term frequency): 는 특정한 단어가 문서 내에 얼마나 자주 등장하는지를 나타내는 값으로, 이 값이 높을수록 문서에서 중요하다고 생각할 수 있다.

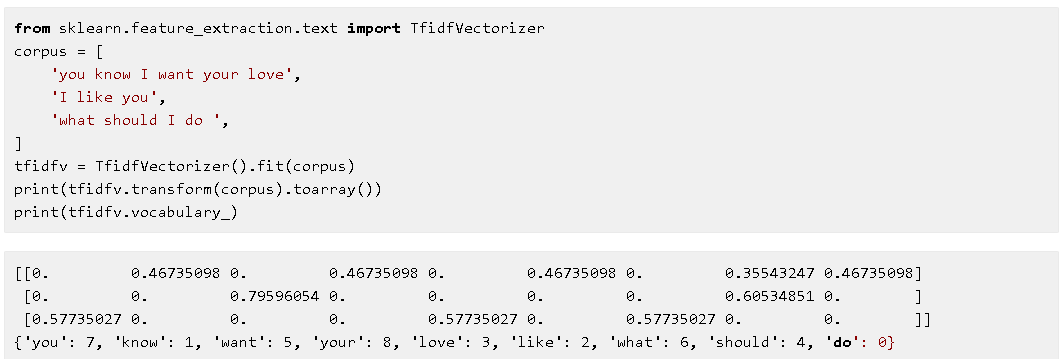
IDF(역문서 빈도, inverse document frequency), 하지만 단어 자체가 문서군 내에서 자주 사용되는 경우, 이것은 그 단어가 흔하게 등장한다는 것을 의미한다

TF-IDF는 TF와 IDF를 곱한 값이다.

TF-IDF는 모든 문서에서 자주 등장하는 단어는 중요도가 낮다고 판단하며, 특정 문서에서만 자주 등장하는 단어는 중요도가 높다고 판단합니다.



Example code



각 알고리즘의 단점

BoW: 불용어를 일일이 다 지정해서 삭제해야 하기 때문에, 프로그램의 유연성이 떨어진다.

TF-IDF: 각 과의 논문마다 specific한 주제는 다르겠지만, 공통적으로 사용하는 용어가 있을 텐데, 이 중요성을 떨어뜨리는 것은 문제가 될 수 있음.

Reference

<https://wikidocs.net/21667>