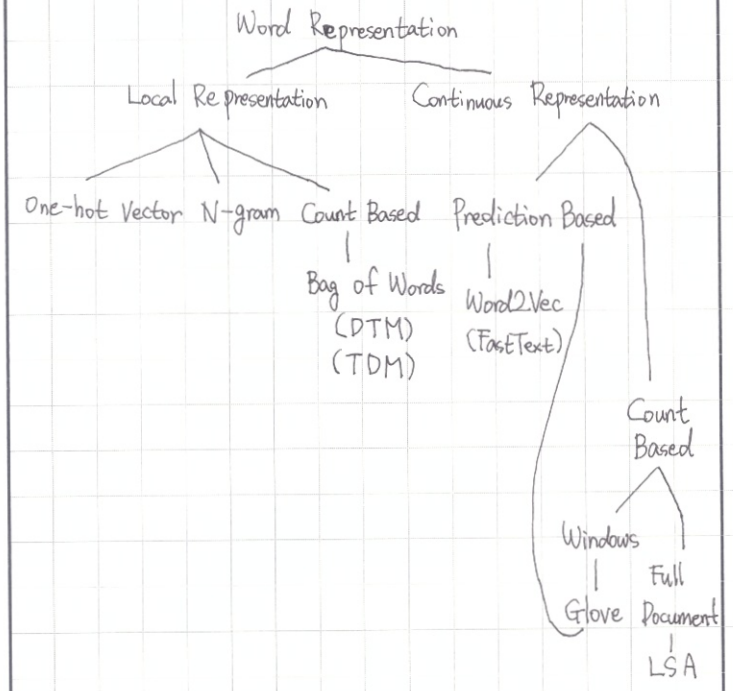
과제명 (Title) : 지식 그래프 기반 음식점 광고 콘텐츠 추천 알고리즘 개발

과 제 번 호 (Project No.) :

4. 카운트 기반의 단어 표현 (Count based word Representation)

1) 다양한 단어의 표현방법



2) Bag of Words (BoW)

-1. BoW란?

: 단어들의 순서는 전혀 고려하지 않고 단어들의 출현 빈도에 만 집중하는 텍스트 데이터의 수치화 표현 방법.

① 각 단어에 고유한 정수 인덱스를 부여한다. (단어 집합 생성)

② 각 인덱스 위치에 단어 토큰의 등장 횟수를 기록한 벡터를 만든다.

-2. CountVectorizer 클래스로 BoW 만들기

사이킷 런에서는 단어의 빈도를 Count 하여 Vector로 만드는 CountVectorizer 클래스를 지원.

영어에 대해 BoW 만들기 유용.

한국어에 대해서는 유용하지 않다.

-3. 불용어를 제거한 BoW 만들기

① 사용자가 직접 정의한 불용어 사용

```
CountVectorizer(stop_words=["the","a"])
```

② CountVectorizer 에서 제공하는 자체 불용어 사용

```
CountVectorizer(stop_words="english")
```

③ NLTK에서 지원하는 불용어 사용

```
stop_words = stopwords.words("english")
CountVectorizer(stop_words=stop_words)
```

3) 문서 단어 행렬 (Document-Term Matrix, DTM)

-1. 문서 단어 행렬 (Document-Term Matrix, DTM)의 표기법

: 다수의 문서에 등장하는 각 단어들의 빈도를 행렬로 표현. BoW를 하나의 행렬로 만든 것.

-2. 한계

① 희소 표현 (Sparse representation)

DTM은 원-핫 벡터와 마찬가지로 공간적 낭비와 계산 리소스를 증가시키는 단점이 있다.

원-핫벡터나 DTM 같은 대부분의 값이 0인 표현을 희소벡터 또는 희소행렬이라 한다. → 해결방법: 텍스트 전처리

② 단순 빈도수 기반 접근

"the" 같은 불용어로 인해 두 문서가 유사하다고 판단 할 수있다.

→ 해결 방법 : TF-IDF

4) TF-IDF (Term Frequency-Inverse Document Frequency)

-1. TF-IDF (단어 빈도-역 문서 빈도)

: 단어의 빈도와 역 문서 빈도를 사용하여 DTM내에 각 단어들 마다 중요한 정도를 가중치로 주는 방법

(1) $tf(d,t)$: 특정 문서 d에서의 특정 단어 t의 등장 횟수

(2) $df(t)$: 특정 단어 t가 등장한 문서의 수

(3) $idf(d,t)$: $df(t)$에 반비례 하는 수

$$idf(d,t) = \log\left(\frac{n}{1+df(t)}\right)$$

log를 씌우면 희귀 단어들과 다른 단어들 간의 격차를 줄여 준다.

분모가 0이 되는 것을 막기 위해 1을 더해준다.

TF-IDF는 TF와 IDF를 곱한 값이다.

$$TF\text{-}IDF = TF \times \ln\left(\frac{\text{문서의 총 개수}}{1+DF}\right)$$

* $n=4$ 고, $df(t)=3$ 인 경우 가중치의 역할을 수행하지 못한다. 그래서 머신러닝 패키지마다 식이 상이하다.

-2. 사이킷 런을 이용한 DTM과 TF-IDF 실습

```
from sklearn.feature_extraction.text import CountVectorizer
CountVectorizer().fit_transform(corpus).toarray()
```

CountVectorizer를 사용하면 DTM을 만들 수 있다.

TfidfVectorizer를 제공.

기 록 자 (Recorded by) 이상준	서 명 (Signature) 이상준	일 자 (Date) 2022.3.26	이 노트의 지적재산권은 숭실대학교 산학협력단에 있음 (Intellectual property right for this notebook belongs to Foundation of Soongsil)
점 검 자 (Reviewed by) 정 선 태	서 명 (Signature)	일 자 (Date) 2022.3.29	