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

과 제 번 호 (Project No.) :

5) 정규표현식 (Regular Expression)

-1. 정규표현식 문법과 모듈 함수

파이썬에서는 정규표현식 모듈 re를 지원.

1) 정규표현식 문법

ex) . , ? , * , + , ^ , $, {숫자} , {숫자1, 숫자2} , {숫자,}

2) 역슬래쉬 문자 규칙

ex) \ , \d , \D , \s , \S , \w , \W

3) 정규표현식 모듈 함수

ex) re.compile( ), re.search( ), re.match( )
re.split( ), re.findall( ), re.finditer( ), re.sub( )

-2. 정규표현식을 이용한 토큰화

NLTK에서는 정규표현식을 사용해서 단어 토큰화를 수행하는 RegexpTokenizer 지원.

6) 정수 인코딩 (Integer Encoding)

각 단어에 고유한 정수를 mapping 시키는 전처리 작업.

-1. 정수 인코딩

빈도수 순으로 정렬. → 빈도수 높은 순서대로 차례로 낮은 숫자부여

1) dictionary
- word_to_index에 'OOV'란 단어를 새롭게 추가하고
- 단어 집합에 없는 단어들은 'OOV'의 인덱스로 코딩.

2) Counter 사용하기

파이썬의 Counter( )의 입력으로 사용하면 중복을 제거하고 단어의 빈도수 기록.

3) NLTK의 FreqDist 사용하기

Counter와 같은 사용 방법

4) enumerate

순서가 있는 자료형을 입력 받아 인덱스를 순차적으로 리턴.

-2. 케라스 (Keras)의 텍스트 전처리

케라스 토크나이저를 사용할 때는 숫자 0도 단어 집합의 크기로 고려해야 한다.

케라스 토크나이저는 OOV에 대해서는 단어를 정수로 바꾸는 과정에서 아예 단어를 제거한다.

7) 패딩 (padding)

병렬 연산을 위해서 여러 문장의 길이를 임의로 동일하게 맞춰주는 작업

-1. Numpy로 패딩하기

데이터에 특정 값을 채워서 데이터의 크기 (shape)를 조정하는 것 = 패딩.

-2. 케라스 전처리 도구로 패딩하기

pad_sequences( ) 제공.

Numpy와 달리 문서의 앞에 0을 채운다.

8) 원-핫 인코딩 (One-Hot Encoding)

단어 집합 : 서로 다른 단어들의 집합

-1. 원-핫 인코딩이란?

단어의 크기를 벡터 차원으로 하고, 표현하고 싶은 단어의 인덱스에 1의 값을 부여하고, 다른 인덱스에는 0을 부여하는 단어의 벡터 표현 방식.

① 정수 인코딩 한다.

② 표현하고 싶은 단어에 1을 부여하고 나머지에 0을 부여.

-2. 케라스를 이용한 원-핫 인코딩

to_categorical( ) 지원.

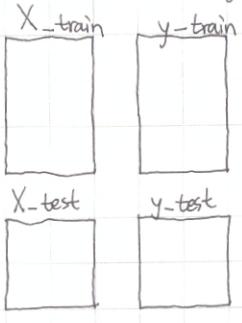
-3. 원-핫 인코딩의 한계

① 불필요한 저장공간이 늘어난다.

② 단어의 유사도를 표현하지 못한다.

→ LSA, HAL, NNLM, RNNLM, Word2Vec, FastText, GloVe로 해결

9) 데이터의 분리 (Splitting Data)



① Zip ② 데이터프레임 ③ Numpy 를 이용하여 데이터 분리. (X, y)

① 사이킷런을 통해 Test 데이터 분리

random_state에 따라 랜덤 배열

② 수동으로 Test 데이터 분리

num of test는 어느 한쪽을 먼저 계산하고 그 갯수만큼 제외하는 방식으로 계산.

분리하기 전 수동으로 섞는 과정 필요.

10) 한국어 전처리 패키지 (Text Preprocessing Tools for Korean Text)

① PyKoSpacing

띄어쓰기가 되어있지 않은 문장을 한 문장으로 변환.

② Py-Hanspell

네이버 한글 맞춤법 검사기

③ SOYNLP를 이용한 단어 토큰화

품사 태깅, 단어 토큰화, 비지도 학습

④ Customized KoNLPy

형태소 분석기에 사용자 사전 추가.

기 록 자 (Recorded by) 이상준	서 명 (Signature) 이상준	일 자 (Date) 2022.3.12	이 노트의 지적재산권은 숭실대학교 산학협력단에 있음 (Intellectual property right for this notebook belongs to Foundation of Soongsil)
점 검 자 (Reviewed by) 정선태	서 명 (Signature)	일 자 (Date) 2022.3.15	