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

과 제 번 호 :
(Project No.)

1. 자연어 처리 : 자연어 처리란 일상생활에서 사용하는 언어의 의미를 분석하여 컴퓨터가 처리 할 수 있도록 하는 일

2. 텍스트 전처리 : 풀고자 하는 문제의 용도에 맞게 텍스트를 사전에 처리하는 작업

1) 토큰화 : 주어진 코퍼스에서 토큰이라 불리는 단위로 나누는 작업

-1. 단어 토큰화 : 토큰의 기준이 단어

토큰화 작업은 단순히 구두점이나 특수문자를 전부 제거하는 정제(Cleaning) 작업을 수행하는 것만으로 해결되지 않음.

-2. Don't와 Jone's는 어떻게 토큰화 할까?

NLTK는 영어 코퍼스를 토큰화 하기 위한 도구를 제공.

① word_tokenize

Don't = Do + n't , Jone's = Jone + 's

② WordPunctTokenizer

구두점을 별도로 분류하는 특징을 갖고 있다.

Don't = Don + ' + t , Jone's = Jone + ' + S

③ 케라스의 text_to_word_sequence

기본적으로 모든 알파벳을 소문자로 바꾸고 구두점 제거. ' 는 보존.

-3. 토큰화에서 고려할 사항

① AT&T 처럼 단어 자체에 구두점이 있는 경우

② New York 처럼 한 단어로 인식할 줄 알아야 함.

③ Penn Treebank Tokenization

-4. 문장 토큰화 (= 문장 분류)

NLTK에서 sent_tokenize : 영어 코퍼스

KSS : 한국어 코퍼스

-5. 한국어에서의 토큰화 어려움.

영어와 달리, 어절 토큰화와 단어 토큰화가 같지 않다.

그리고 띄어쓰기가 잘 지켜지지 않는다.

-6. 품사 태깅 : 단어의 품사 구분

2) 정제 , 정규화

정제 : 노이즈 데이터 제거

정규화 : 표현 방법이 다른 단어를 같은 단어로 통합.

-1. 규칙에 기반한 표기가 다른 단어들의 통합.

-2. 대소문자 통합

소문자 변환을 사용하면 automobile과 Automobile 찾을 수 있다.

미국을 뜻하는 US와 우리를 뜻하는 us는 구분해야 함.

-3. 불용어 제거

(1) 등장빈도가 적은 단어

(2) 길이가 짧은 단어

-4. 정규 표현식

코퍼스 내에서 계속 등장하는 글자들을 규칙에 기반하여 한번에 제거하는 방식으로서 매우 유용.

3) 어간 추출 , 표제어 추출

눈으로 봤을 때는 서로 다른 단어들이지만, 하나의 단어로 일반화 시킬 수 있다면 하나의 단어로 일반화시켜서 문서 내 단어 수를 줄이는 목적.

-1. 표제어 추출 (Lemmatization)

: 단어들이 다른 형태를 갖더라도, 그 뿌리의 단어를 찾아서 단어의 개수를 줄일 수 있는지 판단.

: 단어의 형태학적 파싱을 우선 진행.

NLTK에서 표제어 추출을 위해 WordNetLemmatizer 제공.

단어의 형태가 적절히 보존되는 양상이 특징.

본래 단어의 품사를 알아야 정확한 결과를 얻을 수 있기에 의미를 알 수 없는 적절치 못한 단어도 출력 가능.

lemmatizer.lemmatize('dies', 'v')

단어가 동사 품사라는 사실을 알려줄 수 있다.

-2. 어간 추출 (stemming)

품사 정보가 보존되지 않으며 사전에 존재하지 않는 단어일 수 있다.

NLTK에서 포터 어간 추출기와 Lancaster Stemmer 지원.

두 알고리즘은 용도 제대로 된 일반화를 수행하지 못할 수 있다.

4) 불용어 (stopword)

-1. 불용어 확인하기 (NLTK로)

stopwords.words('english')

: 영어 불용어 리스트

-2. NLTK로 불용어 제거하기

-3. 한국어에서 불용어 제거하기

한국어에서 불용어를 제거하는 방법으로는 간단히 토큰화 후에 조사, 접속사 등을 제거하지만

사용자가 직접 불용어 사전을 만들게 되는 경우가 많다.

기 록 자 (Recorded by)	서 명 (Signature)	일 자 (Date)	이 노트의 지적재산권은 숭실대학교 산학협력단에 있음 (Intellectual property right for this notebook belongs to Foundation of Soongsil)
이 상 준	이 상 준	2021. 3. 5	
점 검 자 (Reviewed by) 정 선 태	서 명 (Signature)	일 자 (Date) 2021. 3. 8	