개요

자연어의 처리는 텍스트 임베딩을 통해 수행함

텍스트 임베딩

- 텍스트 형태의 자연어를 기계가 이해할 수 있도록 숫자를 나열하여 벡터로 나타내는 방식, 다시 말해 데이터의 특징을 추출하여 정형 데이터로 변형하는 것
- 종류 : 단어 가방, 워드2벡 등

개요

자연어

사람들이 일상생활에서 사용하는 언어

- 자연어 처리의 예 : 번역, 요약, 표절 검사, 감성 분석, 챗봇 등

텍스트 **분류**

미리 정해진 범주에 맞추어 텍스트 자료 분류

텍스트 **비교**

텍스트 자료의 비슷한 정도를 판별

텍스트 **번역**

어떤 언어로 되어 있는 텍스트를 자동으로 다른 언어로 바꾸는 작업

텍스트 **요약**

텍스트 자료에서 중요한 단어를 추출하고 주제어 검출

〈출처4〉

개요

자연어의 구성

Document(문서)	문장들의 집합
Sentence(문장)	단어들의 집합
Word=Term(단어)	텍스트 분석의 기본 단위

개요

- Corpus : 말뭉치, 즉 텍스트 분석을 위한 Data set
- 전처리 요소

1. Transformation(변환)
2. Tokenization(토큰화)
3. Filtering(필터링)
4. Normalization(정규화)
5. N-grams Range(N-그램)
6. POS(Part Of Speech) Tagger(품사 태깅)

개요

전처리 요소

1 Transformation(변환)

- 입력 데이터를 변환하는 것으로, 텍스트 데이터에서 불필요한 요소를 제거하고 수정함

구분	내용
Lowercase	모든 텍스트를 소문자로 바꿈
Remove accents	텍스트의 모든 악센트/악센트가 제거됨
Parse html	html 태그를 탐지하고 텍스트만 구문 분석함
Remove urls	텍스트에서 URL이 제거됨

개요

전처리 요소

2 Tokenization(토큰화)

- 텍스트를 더 작은 구성 요소(단어, 문장, 빅그램)로 나누는 방법

구분	내용
Word Punctuation	텍스트를 단어별로 나누고, 구두점 기호를 유지함
Whitespace	텍스트를 공백으로만 분할함
Sentence	전체 문장만 유지한 채 텍스트를 완전히 중지하여 분할함
Regexp	제공된 정규식별로 텍스트를 분할하며, 기본적으로 단어 단위로만 분할됨
Tweet	해시태그, 이모티콘 및 기타 특수 기호를 보관하는 사전 훈련된 트위터 모델에 의해 텍스트를 분할함

개요

전처리 요소

3 Filtering(필터링)

- 단어를 제거하거나 유지함

구분	내용
Stopwords	▪ 분석에 필요하지 않은 단어를 제거함 예 'and', 'or', 'in', … ▪ 필터링할 언어를 선택함(기본값 : 영어) ▪ 간단한 *.txt 파일에 제공된 중지 단어 목록을 한 줄에 하나씩 로드할 수도 있음 ▪ 제공된 중지 단어만 필터링하려면 '영어'를 '없음'으로 변경함 ▪ 중지 단어 목록을 다시 로드하려면 '재로드' 아이콘을 클릭함

(1/3)

개요

전처리 요소

3 Filtering(필터링)

- 단어를 제거하거나 유지함

구분	내용
Lexicon	▪ 파일에 제공된 단어만 보관함 ▪ 어휘로 사용할 *.txt 파일을 한 줄에 하나씩 로드한 'Reload' 아이콘을 클릭하여 어휘를 다시 로드함
Regexp	▪ 정규식과 일치하는 단어를 제거함 ▪ 기본값은 구두점을 제거하도록 설정되어 있음

(2/3)

개요

전처리 요소

3 Filtering(필터링)

- 단어를 제거하거나 유지함

구분	내용
Document frequency	▪ 지정된 문서 수 또는 퍼센트 이상에 나타나는 토큰을 유지함 ▪ Relative는 문서 백분율에 나타나는 토큰만 보관함 ▪ Absolute는 지정된 문서 수에 나타나는 토큰만 보관함
Most frequent tokens	▪ 지정된 개수의 가장 자주 사용하는 토큰만 유지함 ▪ 기본값은 가장 자주 사용하는 100개의 토큰임

(3/3)

개요

전처리 요소

4 Normalization(정규화)

표제어 추출 (Lemmatization)	표제어는 한글로는 '표제어' 또는 '기본 사전형 단어' 정도의 의미 단어들로부터 표제어를 찾아가는 과정
어간 (Stem)	단어의 의미를 담고 있는 단어의 핵심 부분

예 am, are, is는 서로 다른 스펠링이지만 그 뿌리 단어는 be라고 볼 수 있음 → 이 때 이 단어들의 표제어는 be

개요

전처리 요소

4 Normalization(정규화)

구분	내용
Porter Stemmer	▪ 원래 포터 스템머를 적용함
Snowball Stemmer	▪ 개선된 포터 스템머(Porter 2)를 적용함 ▪ 표준화를 위한 언어를 설정함 (기본값 : 영어)
WordNet Lemmatizer	▪ 영어의 큰 어휘 데이터베이스를 기반으로 하는 토큰에 인지 동의어 네트워크를 적용함
UDPipe	▪ 데이터를 정규화하기 위해 사전 훈련된 모델을 적용함

개요

전처리 요소

5 N-grams Range(N-그램)

- 텍스트 단위를 연속적인 N개의 토큰으로 구성함
- 토큰에서 N-gram을 생성함(기본값 : 1그램, 2그램)

예 How are you? Fine, thank you and you?

개요

전처리 요소

6 POS(Part Of Speech) Tagger(품사 태깅)

- 각 단어의 품사를 태깅하여 단어의 의미를 파악함

  예 I can fly! Because I am a fly.

구분	내용
Averaged Perceptron Tagger	Matthew Honnibal의 평균 Perceptron Tagger와 함께 POS 태깅을 실행함
Treebank POS Tagger (MaxEnt)	훈련된 Penn Treebank 모델로 POS 태깅을 실행함

〈출처5〉