Insight campus SeSAC

전처리(Preprocessing)

실무형 인공지능 자연어 처리

Insight campus SeSAC

텍스트 전처리란?

자연어 처리를 위해

용도에 맞도록

사전에 표준화 하는 작업

텍스트 전처리 왜 필요한가?

텍스트 내 정보를 유지하고,

분석의 효율성을 높이기 위해

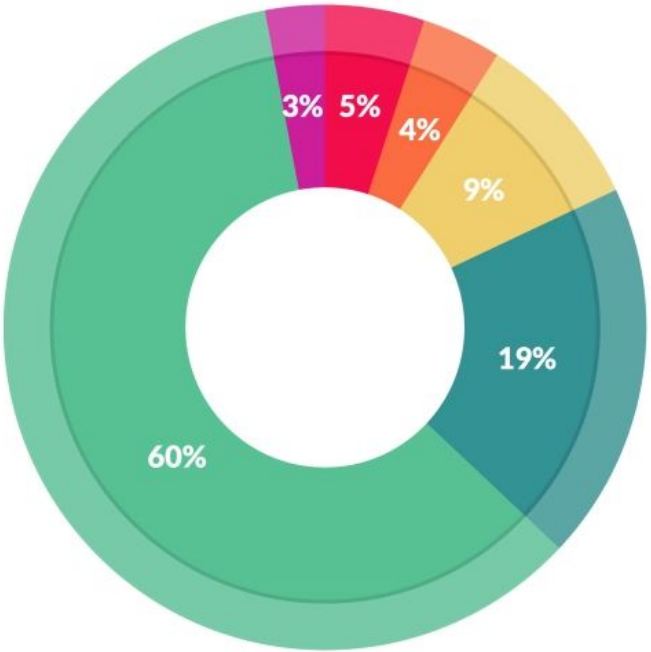
텍스트 전처리 개요

- 분석 하기 전 텍스트를 분석에 적합한 형태로 변환하는 작업
- 전처리 단계는 텍스트를 토큰화하고 자연어 처리에 필요없는 조사, 특수문자, 단어(불용어)의 제거과정을 포함
- 전처리는 분석결과와 모델 성능에 직접 영향을 미치기 때문에 전처리 단계는 매우 중요

1. 토큰화	구두점으로 문서를 문장으로 분리하는 문장 토큰화과 단어 단위로 분리하는 단어 토큰화로 구성된다.
2. 형태소 분석	뜻을 가진 가장 작은 단위인 형태소로 분리하는 과정이다.
3. 품사 태깅	분리된 토큰에 품사를 태깅하는 과정이다.
4. 원형 복원	단어의 원형을 복원하여 표준화하는 과정이다. Stemming방식과 Lemmatization방식이 있다.
5. 불용어 처리	분석에 불필요한 단어나 방해되는 단어를 제거하는 과정이다.

전처리 중요성

전처리 중요성



What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%

79%

- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

https://whatsthebigdata.com/2016/05/01/data-scientists-spend-most-of-their-time-cleaning-data/

텍스트 전처리 (Text Preprocessing)

텍스트 전처리 (Text Preprocessing)

1

토큰화
(Tokenization)

토큰화 (Tokenization)

- 텍스트를 자연어 처리를 위해 분리 하는 것
- 토큰화는
  문장별로 분리하는 "문장 토큰화(Sentence Tokenization)"와
  단어별로 분리하는 "단어 토큰화(Word Tokenization)"로 구분

문장 토큰화 (Sentence Tokenization)

- 문장(Sentence)를 기준으로 토큰화
- 온점(.), 느낌표(!), 물음표(?) 등으로 분류하면 해결 될 것으로 생각됨
- 하지만 단순하게 분리할 경우 정확한 분리가 어려움

Barack Obama likes fried chicken. He don't like spicy chicken.

Barack Obama likes fried chicken.

He don't like spicy chicken.

단어 토큰화(Word Tokenization)

- 단어(word)를 기준으로 토큰화
- 영문의 경우 공백을 기준으로 분리하면 유의미한 토큰화가 가능
- 반면 한글의 경우 품사를 고려한 토큰화(=형태소분석)가 필요

Insight campus SeSAC

단어 토큰화 고려사항

- 특수문자가 있는 경우
  (구두점 및 특수문자를 단순하게 제외해서는 안됨)

특수문자	원문	토큰화 예제1	토큰화 예제2
‘	Don’t	Do / n’t	Don / ‘ / t
-	State-of-the-art	State / of / the / art	State-of-the-art

- 단어 내 띄어쓰기가 있는 경우

	원문	토큰화 예제1	토큰화 예제2
공백	New York	New / York	New York

2

품사태깅
(PoS Tagging)

Insight campus SeSAC

품사 부착(PoS Tagging)

- 각 토큰에 품사 정보를 추가
- 분석시에 불필요한 품사를 제거하거나 (예. 조사, 접속사 등) 필요한 품사를 필터링 하기 위해 사용

Barack Obama likes fried chicken very much.

Barack / NNP 명사	Obama / NNP 명사	likes / VBZ 동사	fried / VBN 동사	chicken / JJ 형용사	very / RB 부사	much / RB 부사	.

3

개체명 인식
(NER, Named Entity Recognition)

개체명 인식 (NER, Named Entity Recognition)

- 사람, 조직, 지역, 날짜, 숫자 등 개체 유형을 식별
- 검색 엔진 색인에 활용

Barack Obama likes fried chicken very much.

Barack / NNP **PERSON**	Obama / NNP **ORGANIZATION**	likes / VBZ	fried / VBN	chicken / JJ	very / RB	much / RB	.

4

원형 복원
(Stemming & Lemmatization)

어간 추출 (Stemming)

- 각 토큰의 원형 복원을 함으로써 토큰을 표준화하여 불필요한 데이터 중복을 방지 (=단어의 수를 줄일수 있어 연산을 효율성을 높임)
- 어간 추출(Stemming) : 품사를 무시하고 **<u>규칙에 기반하여 어간을 추출</u>**
  규칙 : https://tartarus.org/martin/PorterStemmer/def.txt

원문	Stemming
running	run
beautiful	beauti
believes	believ
using	use
conversation	convers
organization	organ
studies	studi

표제어 추출 (Lemmatization)

- 각 토큰의 원형 복원을 함으로써 토큰을 표준화하여 불필요한 데이터 중복을 방지 (=단어의 수를 줄일수 있어 연산을 효율성을 높임)
- 표제어 추출 (Lemmatization) : **품사정보를 유지하여 표제어 추출 (사전 기반)**

원문	Lemmatization
running	running
beautiful	beautiful
believes	belief
using	using
conversation	conversation
organization	organization
studies	study

5

불용어 처리
(Stopwords)

불용어 처리(Stopwords)

- 불필요한 토큰을 제거 하는 작업
- 불필요한 품사를 제거 하기도 함

감사합니다.

Insight campus SeSAC

daniel@fins.ai