**REPORT**

(CHAPTER 8)

로고, 상징, 등록 상표, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

|  |  |
| --- | --- |
| 이름 | 김민서 |
| 학과 | 컴퓨터인공지능공학부 |
| 학번 | 202213083 |
| 과목명 | 인공지능입문 |
| 교수님 | 진예지 교수님 |
| 분반 | 105분반 |
| 제출일 | 2025.06.12 |

1. **이 장에서 자연어 처리에 대해 배웠다. 아래에 제시된 위젯은 Text Mining 카테고리에 속해있는 위젯이다. 각 위젯의 이름과 기능을 설명하시오.**

|  |  |  |
| --- | --- | --- |
| **화이트, 원이(가) 표시된 사진  AI 생성 콘텐츠는 정확하지 않을 수 있습니다.** | Corpus | 오렌지에서 텍스트 데이터를 불러오는 위젯 |
| **화이트, 원이(가) 표시된 사진  AI 생성 콘텐츠는 정확하지 않을 수 있습니다.** | Preprocess Text | 텍스트 데이터의 정제, 정규화, 토큰화 작업 |
| **화이트, 원이(가) 표시된 사진  AI 생성 콘텐츠는 정확하지 않을 수 있습니다.** | Bag of Words | 코퍼스의 등장 어휘별 빈도로 수치화한 임베딩 |
| **화이트, 원이(가) 표시된 사진  AI 생성 콘텐츠는 정확하지 않을 수 있습니다.** | Word cloud | 코퍼스의 등장 어휘별 빈도에 따른 시각화 |
| **화이트, 원이(가) 표시된 사진  AI 생성 콘텐츠는 정확하지 않을 수 있습니다.** | Document Embedding | 문서별 특징 값들을 추출한 임베딩 |
| **화이트, 원이(가) 표시된 사진  AI 생성 콘텐츠는 정확하지 않을 수 있습니다.** | Sentiment Analysis | 코퍼스 내의 등장 어휘 통계 기반 주제 분류 |

1. **텍스트 전처리의 4가지 단계는 무엇인가?**

* 1. 정제(Cleaning) & 정규화(Normalization)
  + 정제 : 불필요한 텍스트를 제거하는 작업
  + 정규화 : 표현 방법이 다르지만 같은 뜻을 가지는 단어를 하나로  
     통합하는 작업
* 2. 토큰화(Word Tokenization)
  + 코퍼스를 토큰으로 나누는 작업
* 3. 어간 추출(Stemming) & 표제어 추출(Lemmatization)
  + 어간 추출 : 각 토큰의 핵심 텍스트를 뽑아, 단어의 사소한 변형을  
     무시하려는 목적으로 활용되는 절차
  + 표제어 추출 : 어간 추출과 비슷한 작업이지만, 단어의 모양이 완전히  
     바뀌는 경우 활용되는 절차
* 4. 불용어(Stopword)

1. **다음 제시된 그림은 Preprocess Text 위젯 창의 모습이다. 해당 위젯 창에 배치되어 있는 각 요소들이 텍스트 전처리 과정에서 각각 어떤 역할을 수행하는지 작성하시오.**

|  |  |
| --- | --- |
| 텍스트, 영수증, 스크린샷, 평행이(가) 표시된 사진  AI 생성 콘텐츠는 정확하지 않을 수 있습니다. | * 1. 소문자로 변환 * 2. 단어 단위로 토큰화 * 3. Porter 알고리즘을 사용한 어간 추출 * 4. 영어 불용어 제거 |