1. 텍스트 정제 (Text Cleaning)

- 특수 문자 제거:
  텍스트에서 의미가 없는 특수 문자, 기호(예: @, #, & 등), 또는 HTML 태그를 제거합니다.

- 불필요한 공백 제거:
  문장의 앞뒤 공백이나 중복된 공백을 제거하여 텍스트를 정리합니다.

- 기타 잡음 제거:
  불필요한 이모티콘, URL, 이메일 주소 등의 정보를 제거합니다.

2. 소문자화 (Lowercasing)

- 대소문자 통일:
  텍스트의 모든 문자를 소문자로 변환하여 대소문자 차이에 따른 중복을 방지합니다. 예를 들어, "Python"과 "python"을 동일한 단어로 취급합니다.

3. 불용어 제거 (Stopwords Removal)

- 의미 없는 단어 제거:
  "the", "is", "in", "on"과 같은 문법적으로 중요한 의미는 있지만 분석에 도움이 되지 않는 단어들을 제거합니다. 불용어 제거는 분석을 더 효과적으로 만들 수 있습니다.

4. 토큰화 (Tokenization)

- 단어 단위 또는 문장 단위로 분리:
  텍스트를 단어(token) 또는 문장 단위로 분리하여 분석할 수 있는 형태로 변환합니다. 예를 들어, "Python is great!"을 "Python", "is", "great"로 분리합니다.

5. 어간 추출 (Stemming)

- 단어의 기본형 추출:
  단어에서 접미사를 제거하여 기본 형태로 바꿔주는 작업입니다. 예를 들어, "running", "runner", "ran"을 모두 "run"으로 변환합니다. 이 방법은 빠르지만 정확도는 다소 떨어질 수 있습니다.

6. 표제어 추출 (Lemmatization)

- 정확한 원형 단어 추출:
  어간 추출과 비슷하지만, 표제어 추출은 문맥을 고려하여 정확한 원형 단어로 변환합니다. 예를 들어, "better"는 "good"으로, "was"는 "be"로 변환됩니다. 어간 추출보다 더 정확한 방법입니다.

7. 숫자 및 날짜 처리

- 숫자 정규화:
  숫자는 텍스트 분석에서 중요한 정보일 수 있기 때문에 이를 정리합니다. 예를 들어, "5 years"는 숫자 "5"로 변환하거나, "년" 단위가 포함된 숫자는 일관되게 처리합니다.

- 날짜 처리:
  날짜 형식을 표준화하여 처리합니다. 예를 들어, "2023년 4월 16일"을 "2023-04-16"으로 변환합니다.

8. 개체명 인식 **(Named Entity Recognition, NER)**

- 중요한 정보 추출:
  텍스트에서 사람, 장소, 조직, 날짜와 같은 중요한 정보를 식별하여 추출합니다. 예를 들어, "삼성전자"는 `Organization`, "서울"은 `Location`으로 처리됩니다.

9. 텍스트 벡터화 **(Text Vectorization)**

- **TF-IDF (Term Frequency-Inverse Document Frequency)**:
  단어의 빈도뿐만 아니라 해당 단어가 문서 집합에서 얼마나 중요한지를 반영하여 벡터 형태로 변환합니다.

- **Word Embedding (**단어 임베딩**)**:
  Word2Vec, GloVe와 같은 모델을 사용하여 단어를 고차원 벡터로 변환하고, 단어 간 의미적 관계를 반영합니다. 예를 들어, "king"과 "queen"은 의미적으로 가까운 벡터로 변환됩니다.

10. 문장 길이 및 단어 길이 조정

- 불필요한 긴 문장 자르기:
  분석에 필요하지 않은 긴 문장을 자르고, 너무 짧거나 의미 없는 단어를 처리하여 문장을 정리합니다.