**‘빅데이터 기반 댓글 분석 및 시각화 서비스’**

**상세 설계서**

**[팀 3S1C]**

| **요구 기능 ID** | **요구 기능 이름** |
| --- | --- |
| **Crawler\_002** | **웹 크롤링 모듈** |
| **Preprocessor\_002** | **데이터 전처리 모듈** |
| **Warehouse\_002** | **웨어하우스 모듈** |
| **Analyzer\_002** | **데이터 분석 모듈** |
| **DB\_002** | **데이터 베이스** |
| **Visualizer\_002** | **시각화 모듈** |

**참조 파일명 : [CSE-02] 요구분석정의서 ver 1.0\_팀 3S1C(2조)\_4월27일**

**문서 번호 :  
Version 0.1**

**개정 이력**

**제.개정내역**

| **버전** | **승인일자** | **개요** | **작성자** |
| --- | --- | --- | --- |
| 0.1 | 21.05.17 | 프로젝트 상세 설계서 최초 작성 | 팀 3S1C |
|  |  |  |  |

**배포이력**

| **버전** | **배포일자** | **배포처** |
| --- | --- | --- |
| 0.1 | 21.05.17 | 동국대학교 컴퓨터공학종합설계 1 이클래스 제출 |
|  |  |  |
|  |  |  |
|  |  |  |

**검토이력**

| **버전** | **검토일자** | **검토방법** | **검토자** |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

| **목 차** |
| --- |

[**1.**](#_heading=h.2s8eyo1) **요구기능 설명 4**

[**1.1.**](#_heading=h.17dp8vu) **주요 상세기능 설명 4**

[**1.2.**](#_heading=h.3rdcrjn) **요구기능 동작 절차 4**

[**1.3.**](#_heading=h.26in1rg) **동작 규칙 5**

[**1.4.**](#_heading=h.lnxbz9) **가정(Assumptions) 5**

[**2.**](#_heading=h.35nkun2) **기술 설계(internal) #1 7**

[**2.1.**](#_heading=h.1ksv4uv) **기능 설계 7**

[**2.2.**](#_heading=h.44sinio) **요소 설계 8**

[2.2.1. Class 9](#_heading=h.2jxsxqh)

[2.2.2. 저장소 9](#_heading=h.z337ya)

[**3.**](#_heading=h.3j2qqm3) **기술 설계(external) #2 11**

[**3.1.**](#_heading=h.1y810tw) **UI 설계 11**

[3.1.1. 입력 UI 11](#_heading=h.4i7ojhp)

[3.1.2. 출력 결과 11](#_heading=h.2xcytpi)

[**4.**](#_heading=h.1ci93xb) **주요상세기능 # 13**

[**5.**](#_heading=h.3whwml4) **기타사항 14**

1. **요구기능 설명**

## 1.1주요 상세기능 설명

* 웹 크롤링 모듈
  + ID : Crawler\_002
  + 정의: 웹 크롤링 모듈은 시각화 할 데이터 소스(기사 제목, 본문, 댓글)을 수집한다.
  + 상세 설명: ‘네이버 연합뉴스 속보’ 탭의 뉴스 기사(본문, 댓글)를 임의의 정해진 시간 간격마다 크롤링한다.
* 데이터 전처리 모듈
  + ID : Preprocessor\_002
  + 정의: 크롤링한 데이터를 데이터 분석 모듈에 맞게 변환한다.
  + 상세 설명: 띄어쓰기, 맞춤법 등을 교정하고 반복 문자열(ㅋㅋㅋ, ㅠㅠㅠ 등의 문자열 또는 이모티콘)을 정제하며 데이터베이스(+웨어하우스)와 데이터 분석 모듈에서 인식할 수 있는 형식으로 변환한다.
* 웨어하우스 모듈
  + ID : Warehouse\_002
  + 정의: 데이터 웨어하우스(임시저장소)
  + 상세 설명: 전처리가 된 데이터를 임시로 저장한다.
* 데이터 분석 모듈
  + ID : Analyzer\_002
  + 정의: 댓글 작성자의 연령대, 성별, 감정 및 찬반여부를 분석한다.
  + 상세 설명: 데이터베이스에 있는 데이터(파싱된 문자열)를 파이썬, R등에서 사용 가능한 라이브러리 및 자체 제작한 알고리즘을 적용하여 분석해 시각화 모듈에 호환되게끔 정제한다.
* 데이터베이스 모듈
  + ID : DB\_002
  + 정의: 데이터 분석 모듈을 거친 데이터를 저장한다.
  + 상세 설명: 데이터 분석 모듈을 거친 결과를 시각화 모듈에서 필요한 입력을 위한 데이터의 형식에 맞게 정규화 하여 저장한다.
* 시각화 모듈
  + ID : Visualizer\_002
  + 정의: UI 부분
  + 상세 설명: 데이터베이스 모듈에서 가져온 데이터를 이용하여 사용자에게 한 눈에 알아볼 수 있는 분석 결과를 도식화한 그림, 그래프 등으로 보여준다.

## 1.2요구기능 동작 절차

1. 단계 1 :: 웹 크롤링 모듈이 필요한 데이터(뉴스 본문, 댓글)를 수집한다.

2. 단계 2 :: 크롤링한 데이터를 분석 모듈에서 효율적으로 처리할 수 있도록 전처리 모듈에서 정제한다.

3. 단계 3 :: 정제된 데이터를 임시적으로 저장하기 위해 웨어하우스 모듈에 저장한다.

4. 단계 4 :: 웨어하우스 모듈에 저장한 데이터를 분석(데이터 분석 모듈)해 시각화 할 내용을 생성한다.

5. 단계 5 :: 분석된 데이터를 데이터베이스에 가능한 효율적인 형태로 저장한다.

6. 단계 6 :: 데이터베이스에 저장된 데이터를 시각화 모듈로 불러와 변환한 결과를 사용자에게 보여준다.  
  
각 단계의 상세 내용은 다음과 같다.

단계 1 :: 웹 크롤링 모듈은 데이터 수집 모듈에 포함되는 클래스로 구현되며 크롤링 객체는 각 데이터 종류(본문, 댓글, 성별 및 연령 등)별로 DTO를 만들어 데이터를 분리해 크롤링한다. 이렇게 불러온 데이터는 단계 2에서 활용한다.

단계 2 :: 데이터 전처리 모듈은 데이터 수집 모듈에 포함되는 클래스로 구현되며 전처리 객체는 단계 1에서 추출한 데이터를 파싱, 문법 수정, 반복 문자열 생략 및 축약, 불용어 삭제 등의 과정에 따라 전처리한다.

단계 3 :: 웨어하우스 모듈은 데이터 수집 모듈에 포함되는 클래스로 구현되며 웨어하우스 객체는 toCSV() 메소드를 호출해 전처리 된 데이터를 CSV 포맷으로 컴퓨터에 저장한다.

단계 4 :: 데이터 분석 모듈은 통계분석, 군집화, 분류, 감성분석, 시계열 분석, 상관 분석 등의 통계 기법 및 분석 모델을 활용하여 이슈 키워드 발생 추이, 빈도, 키워드 간 연관성, 사용자 반응 등 시각화 모듈에서 사용할 데이터를 생성한다.

단계 5 :: 데이터 저장(DB) 모듈은 단계 4에서 생성한 데이터를 데이터베이스에 저장하고 이후에도 쿼리문을 통해 접근, 수정, 삭제할 수 있다.

단계 6 :: 시각화 모듈은 데이터베이스에 저장된 정보를 요청한 쿼리에 맞게 호출해 시각화 방법(그래프, 도표, 차트 등)에 따라 입력하고 화면에 표시한다.

## 1.3동작 규칙

동작 규칙 1 :: 웹 크롤링 모듈이 동작하기 위해서는 다음의 규칙을 만족해야 한다.  
 1) 크롤링 할 웹 사이트의 트래픽 제한을 넘어서 크롤링하지 않는다.  
 2) ‘네이버 뉴스’ 서버가 온라인 상태를 유지해야 한다.

동작 규칙 2 :: 데이터 용량을 조절하기 위해 일정 주기마다 또는 특정한 알고리즘에 따라 데이터베이스(및 웨어하우스)를 갱신한다.

동작 규칙 3 :: 인코딩 타입은 utf-8을 유지해야 한다.

## 1.4가정(Assumptions)

가정 1 :: 본문 하나에 수집되는 댓글이 100개 미만일 경우 데이터에서 배제해야 한다.

가정 2 :: 웹 크롤링 요청은 크롤링 대상 사이트의 최소 접근시간 제한을 넘을 수 없다

가정 3 :: 댓글 작성자의 닉네임은 서로 구분 가능할 정도로 공개되어 있다(네이버는 끝 네 자리 숨김, 다음은 전체 공개).

가정 4 :: 데이터 전처리 이후에도 감성분석 대상(entity)은 필터링 되지 않는다.

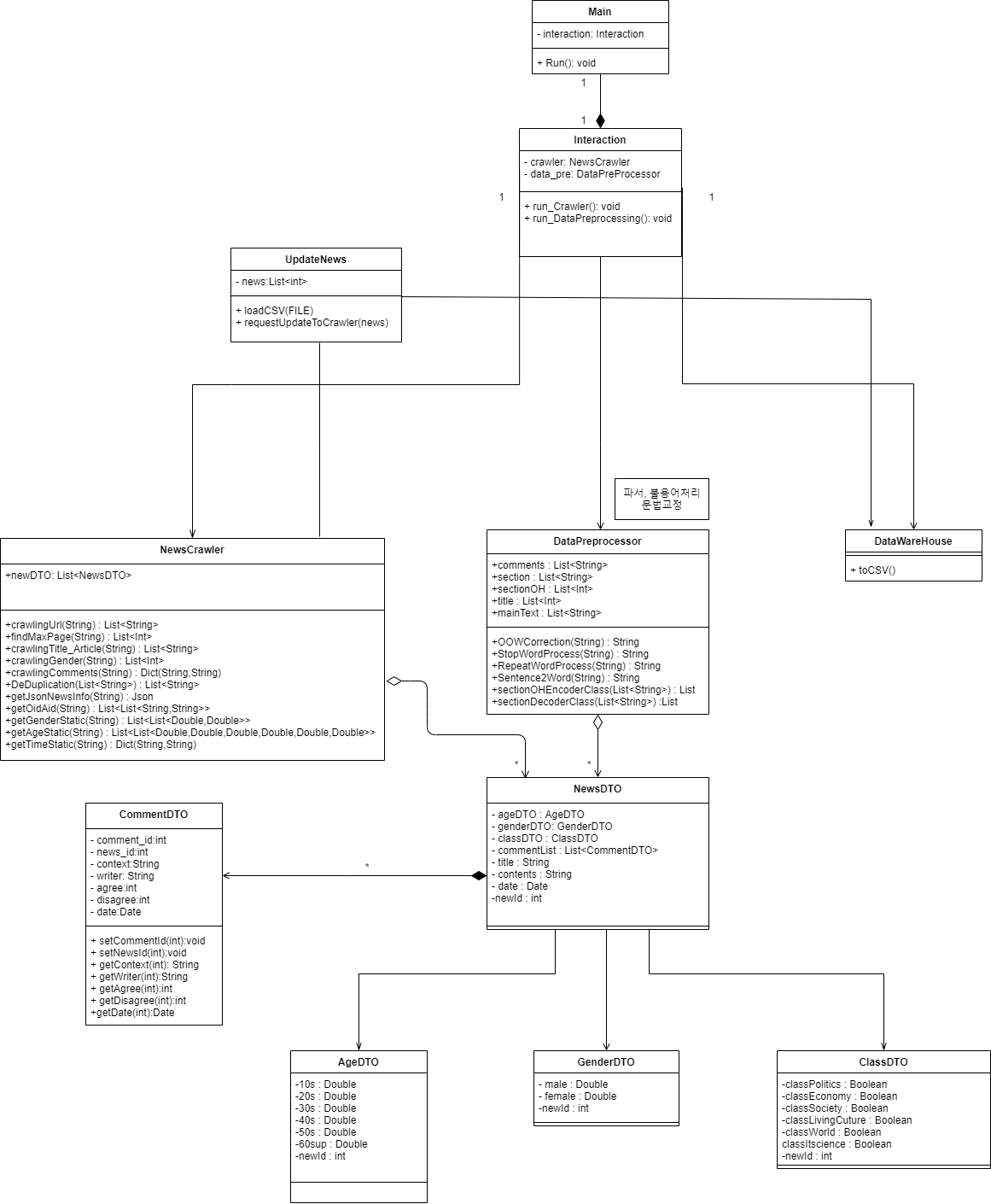
가정 5 :: 가정 4가 지켜지지 않은 데이터는 댓글 도메인(작성 본문)의 키워드라고 가정한다.

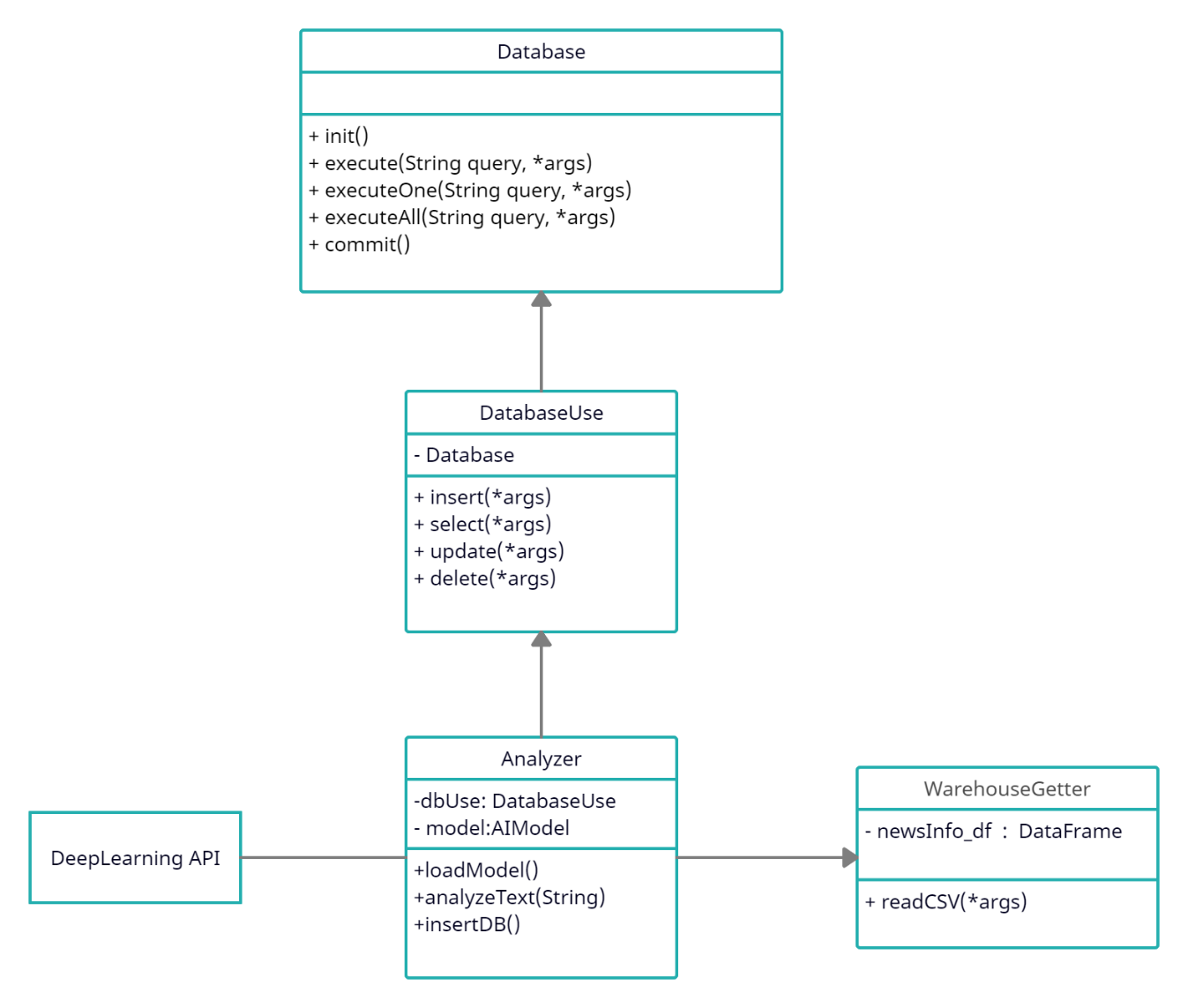
가정 6 :: 시각화된 자료로부터 특정 댓글 작성자를 유추할 수 없다.

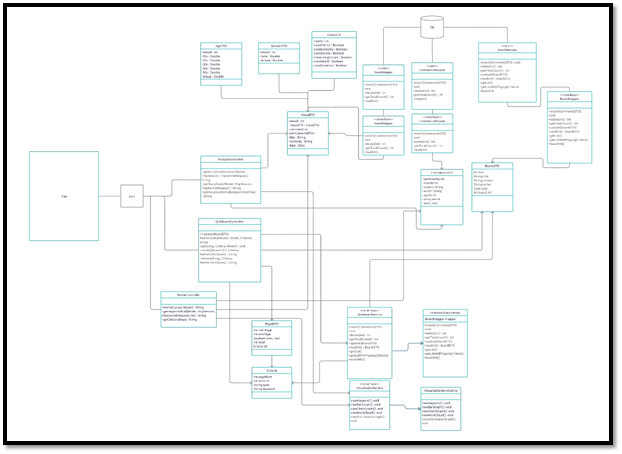
가정 7 :: 수집 및 분석하는 데이터는 ‘한국어’로써 의미가 있는 문장이다.

**2. 기술 설계(Internal) #1**

## 2.1 기술 설계

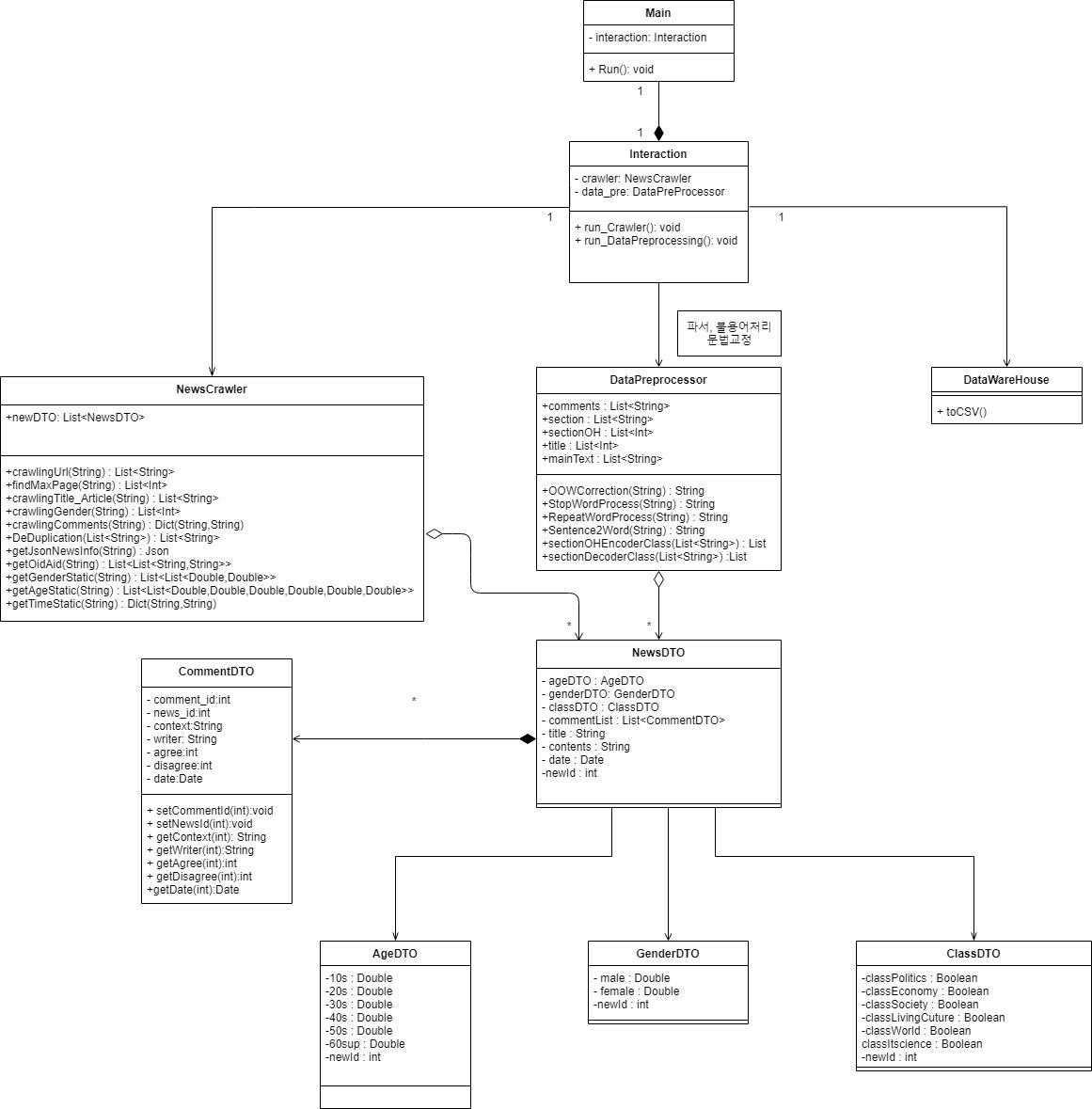
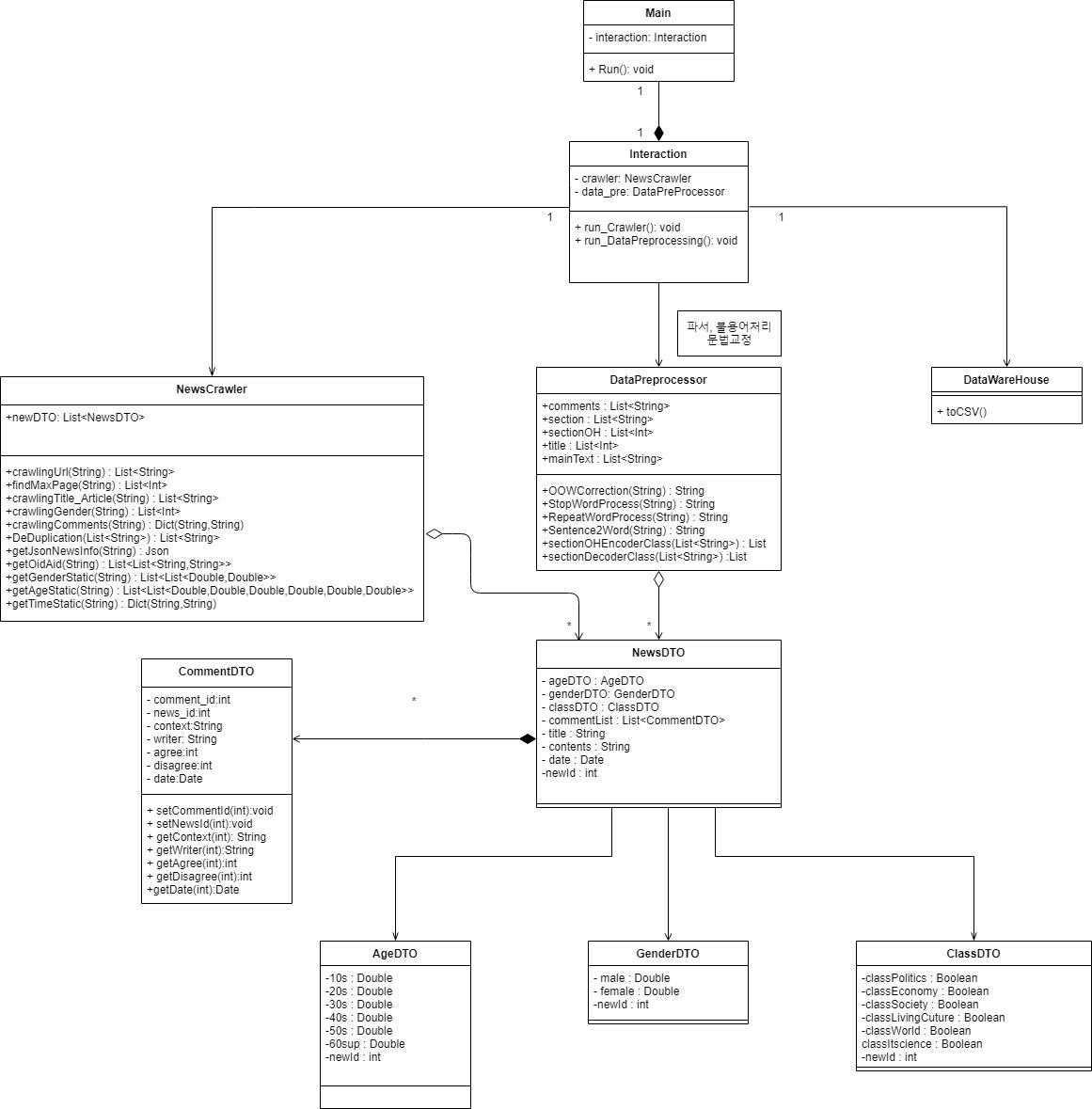
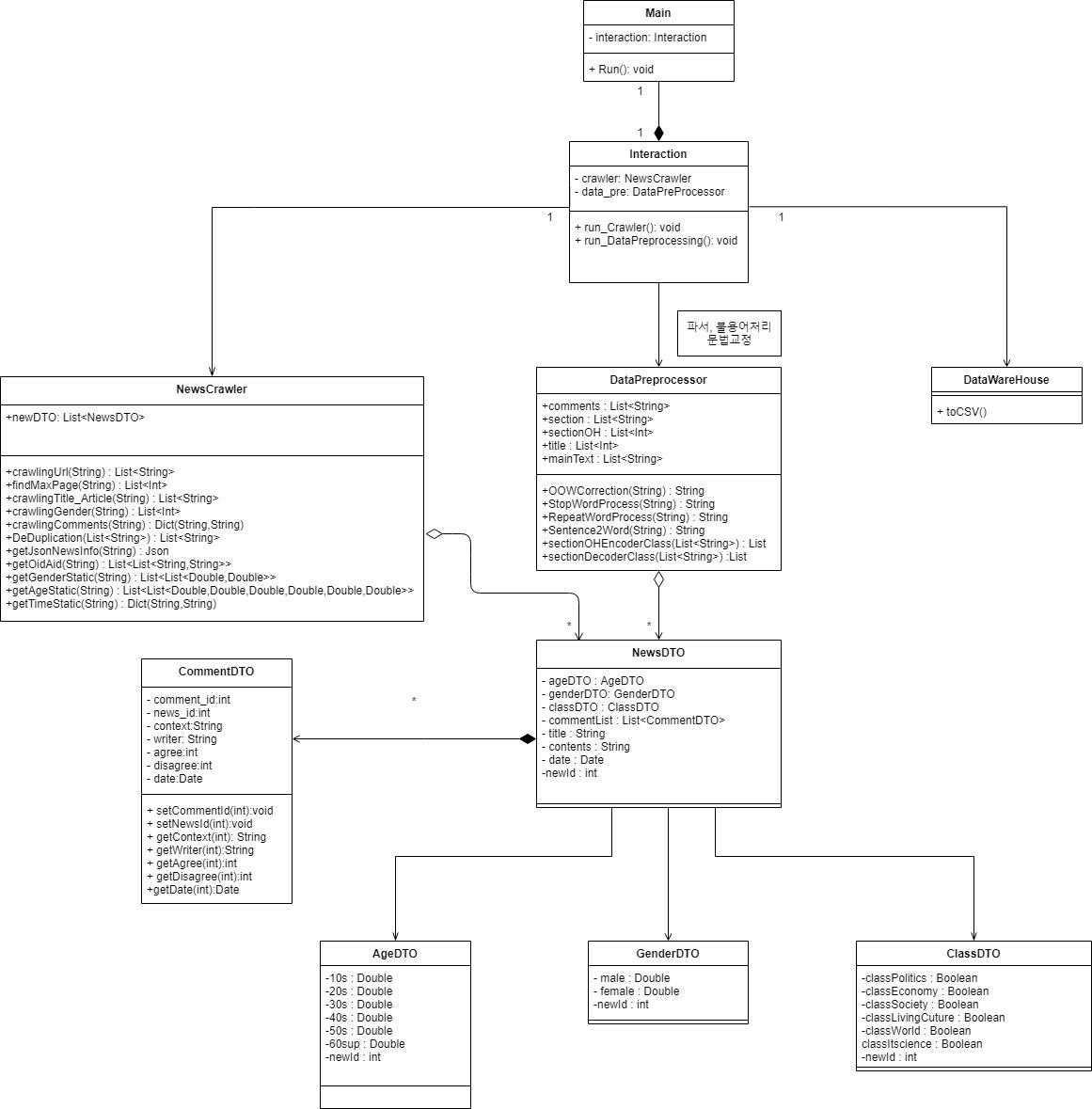
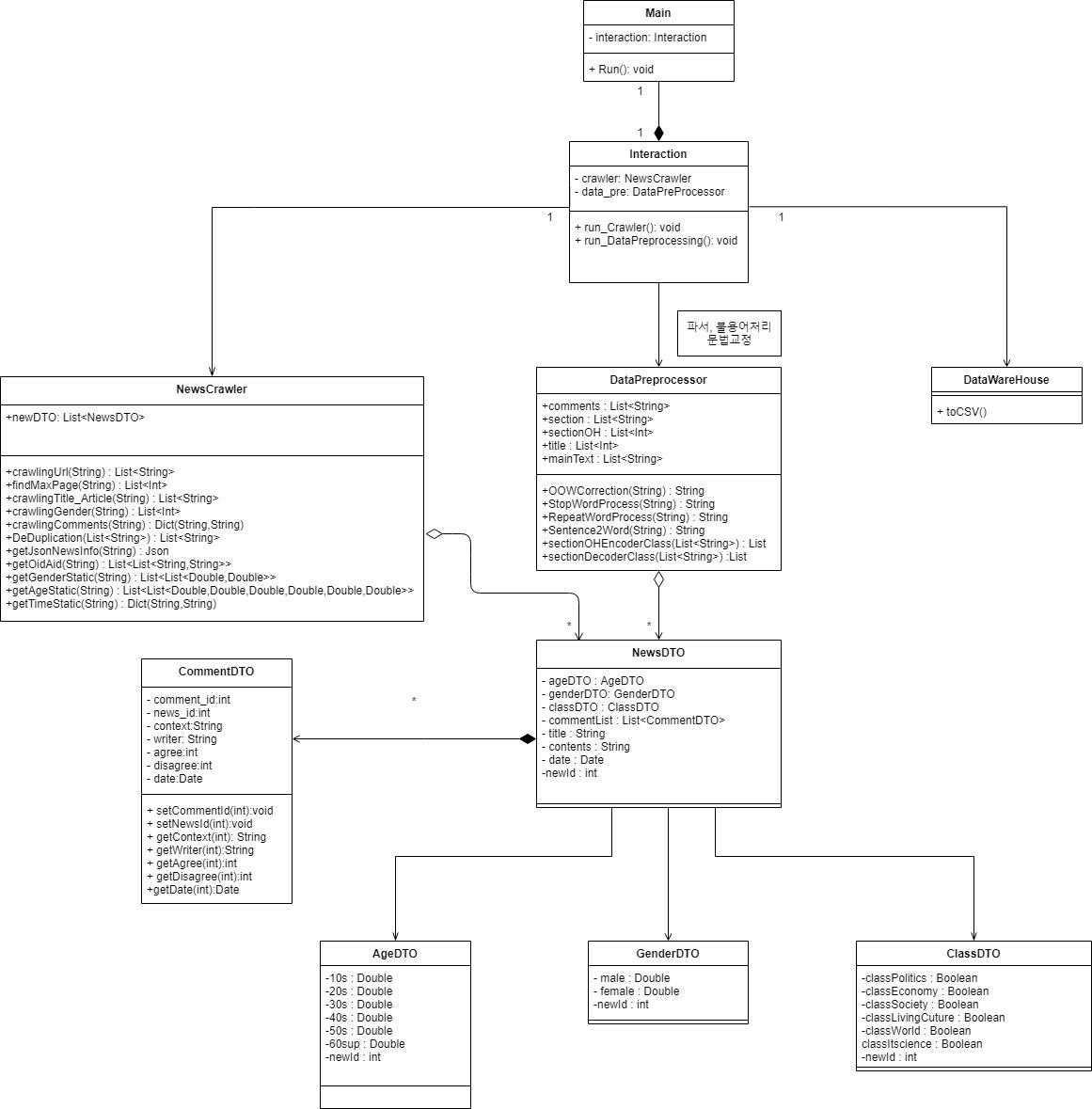
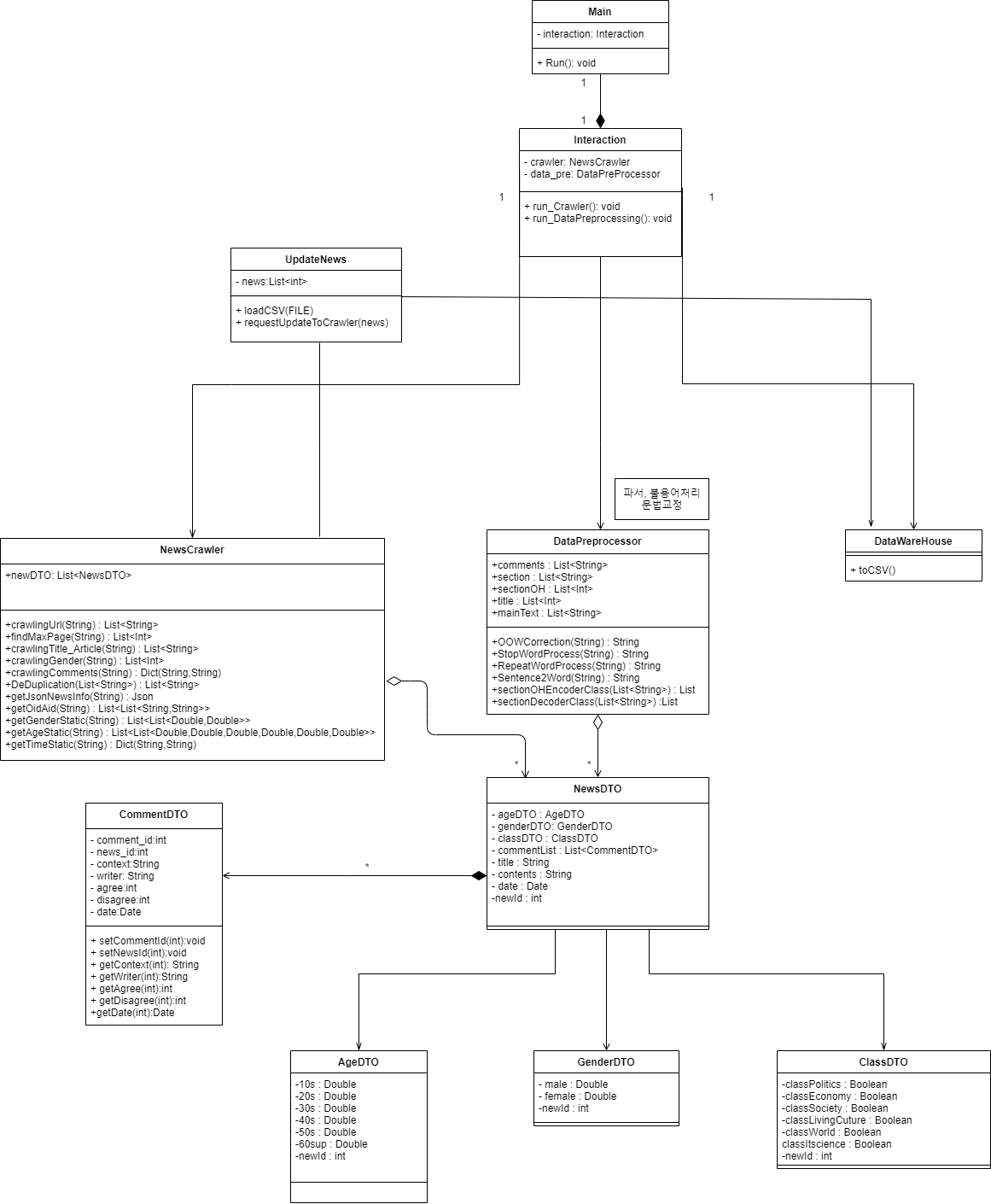
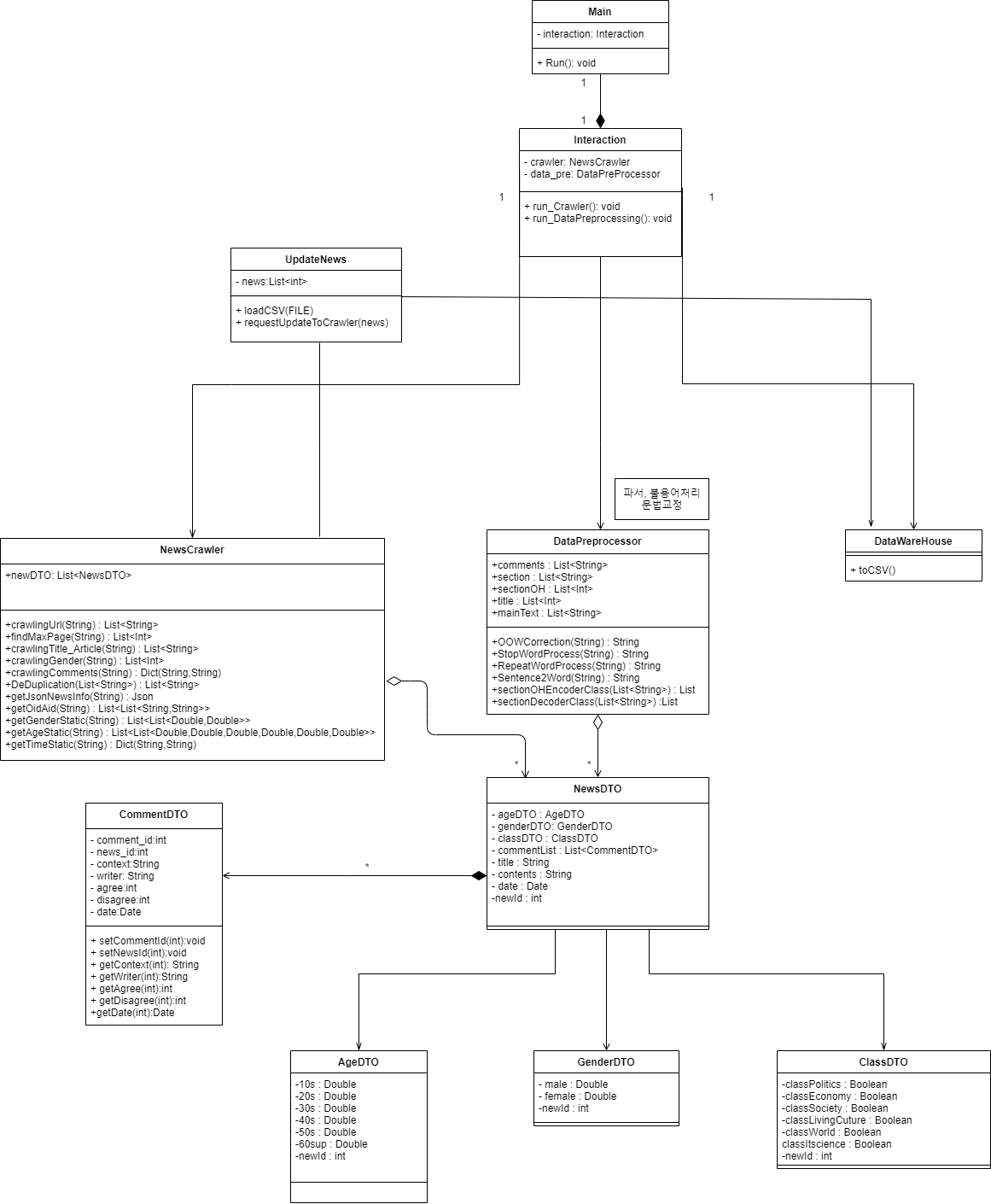
  
그림 1 데이터 수집 모듈(웹 크롤링, 데이터 전처리, 웨어하우스)의 클래스 다이어그램

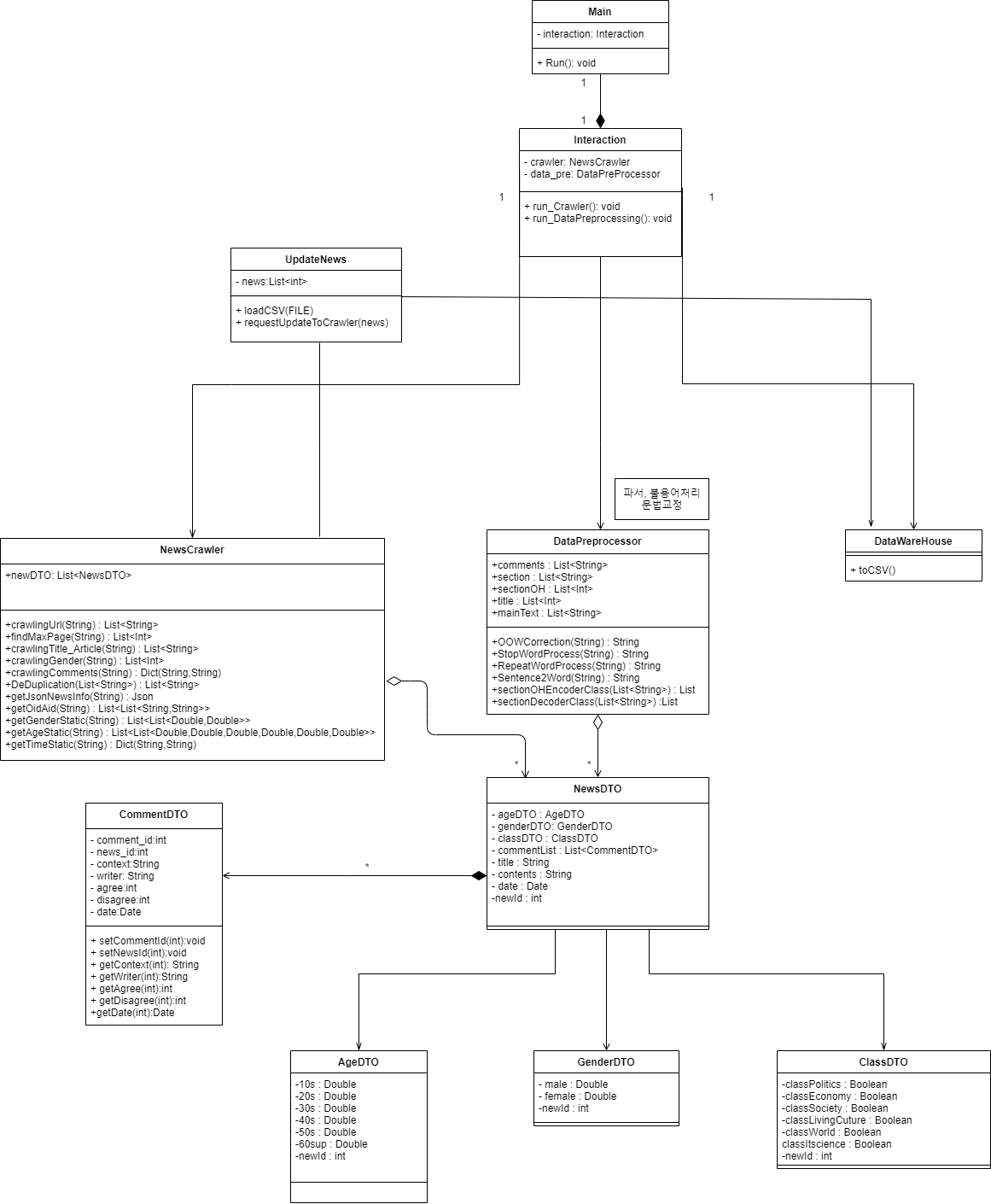
  
그림 2 데이터 분석 모듈(데이터 분석, 데이터베이스)의 클래스 다이어그램

그림 3 웹 모듈의 클래스 다이어그램

## 요소 설계

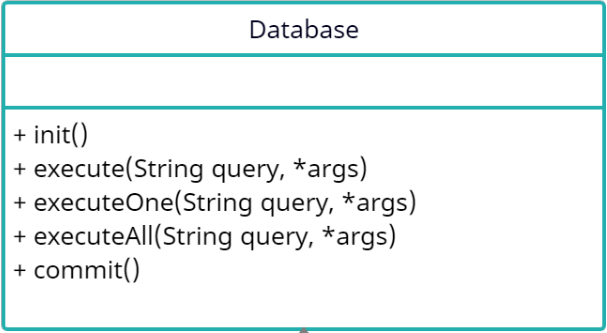
### 2.2.1 데이터 수집 모듈 요소

* Class 1  
  
* Class 2  
  
* Class 3  
  
* Class 4  
  
* Class 5  
  
  + 이름: DTOs
* Class 6  
  
* Class 7

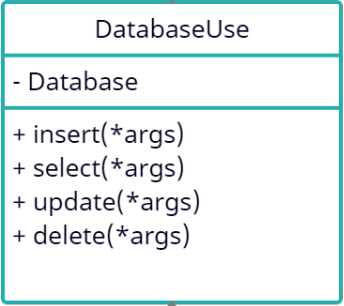


### 데이터 분석 모듈 요소

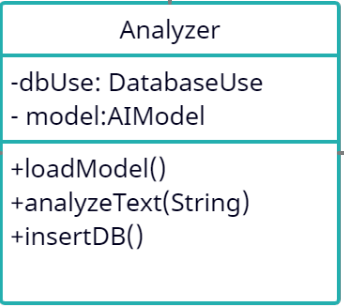
* Class 1



* Class 2



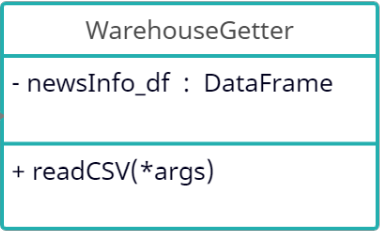
* Class 3

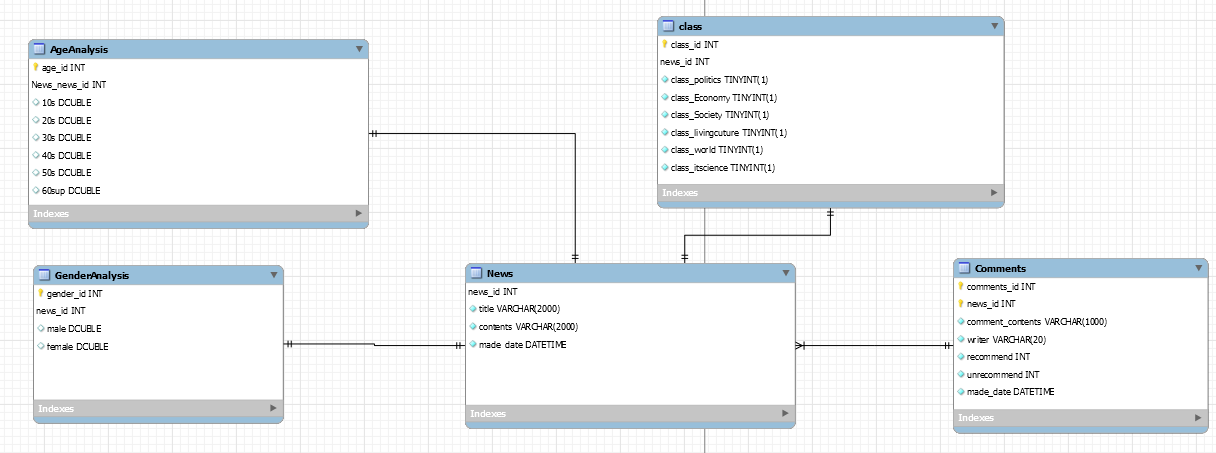


* Class 4



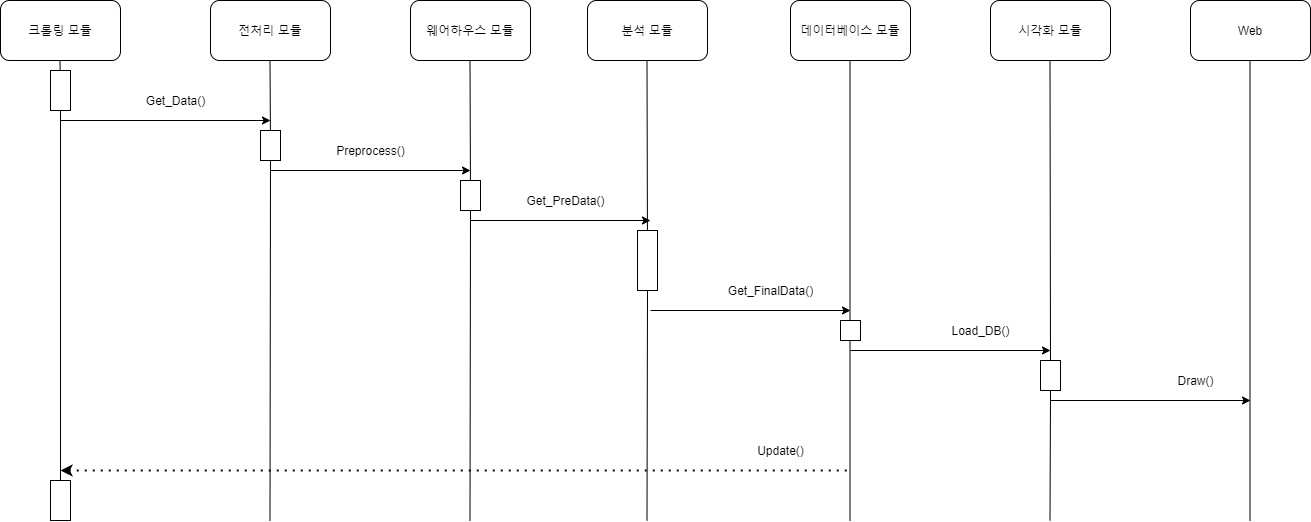
* Class 5





- 유저의 성별, 연령은 ‘모든 뉴스 기사’로부터 분석하기 때문에 따로 클래스 분리

- 섹션 분류 클래스 테이블을 활용해 각 분야(정치면, 사회면 등)별 분석을 용이하게 함



2.2.3 웹 모듈 요소

* Class 1



* Class 2



* Class 3



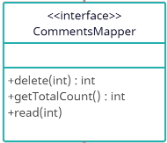
* Class 4



* Class 5



* Class 6



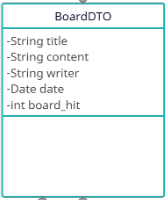
* Class 7



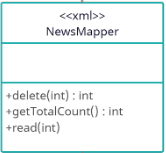
* Class 8



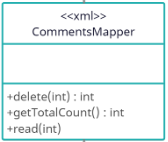
* Class 9



* Class 10



* Class 11



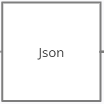
* Class 12



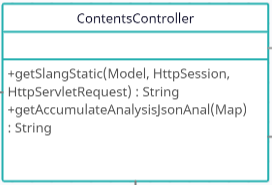
* Class 13



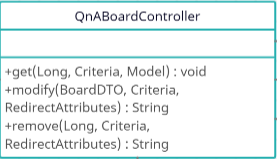
* Class 14



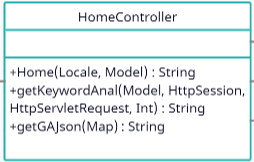
* Class 15



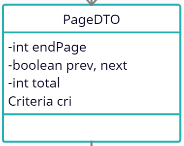
* Class 16



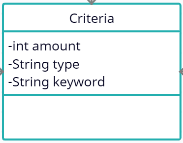
* Class 17



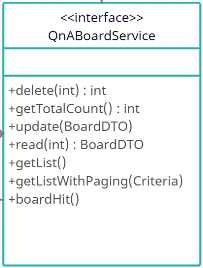
* Class 18



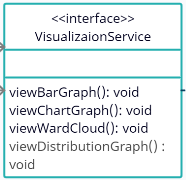
* Class 19



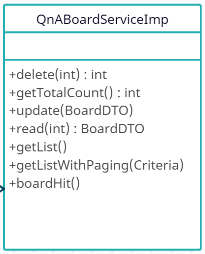
* Class 20



* Class 21



* Class 22



* Class 23

