**Ankus-lite 수집기**

* **공통 실행명령어:**

java –jar crawler-lite.jar BLOG(NEWS) 1 1

* java –jar crawler-lite.jar args[0] args[1] args[2]
* ‘**필수폴더 및 파일**  
  ../model/전달파일  
  역활: 사용자 단어사전 추출 및 기본 형태소분석 단어사전  
  ../logs/  
  역활: 문제발생 로그파일 작성
* **입력변수**

args[0] = 수집기 실행모드

args[1] = 일수(현재날짜 – 시작날짜) 예: 10일전 🡪 10

args[2] = 수집종료일(현재날짜 – 수집종료일) 예: 하루전 🡪 1

* **DB 설정**  
  - 실행파일과 동일폴더 내 CUBRIDDB.conf 파일  
  - driver: JDBC driver name  
  - url: JDBC address  
  - id: dba id  
  - pawword: dba password
* **BLOG core class**  
  - com.onycom.crawler.blog.ankusCrawlerNaverBlog.java   
  - 역활: 네이버에서 블로그 수집을 위한 절차 정의 및 동작  
  - com.onycom.crawler.blog.crawlerPage.java   
  - 역활: htmlunit.jar를 활용한 검색페이지 크롤링 및 파싱 클래스  
  - com.onycom.crawler.blog.NaverBlogVal.java  
  - 역활: 수집된 페이지에서의 정보 추출 규칙 정보 🡨 네이버에서 일부 수정할 경우 html 코드 분석후 정보변경  
  - com.onycom.crawler.blog.naverURL.java   
  - 역활: 검색 주소정의  
  **- 출력테이블: blog\_info**
* **NEWS core class**  
  - com.onycom.crawler.news.ankusCrawlerNaverNews.java  
  - 역활: 네이버에서 뉴스 수집을 위한 절차 정의 및 동작  
  - com.onycom.crawler. news.NaverNewsVal.java  
  - 역활: 수집된 페이지에서의 정보 추출 규칙 정보 정의 🡨 블로그랑 동일  
  **-출력 테이블:** news, rel\_word
* **BLOG 형태소 분석 클래스**  
  - com.onycom.crawler.pos.BLogPOS.java  
  - 역활: DB로드 후 제목 및 콘텐츠를 이용해서 형태소 분석 후 DB에 업데이트(뉴스의 경우는 키워드 테이블이 따로 존재하여 수집과 동시에 처리)
* **단어사전 로드 클래스**  
  - com.onycom.crawler.common.Dictionary.java  
  - 역활: DB의 브랜드, 아이템, 사전정보를 불러와서 형태소 분석 및 아이템 검색에 사용  
  - String newsSql 변수를 활용하여 단어 정보 로드