Ankus crawler v1.0 사용 매뉴얼

- 기능 개요
  - 지정된 주소의 페이지 수집 가능
  - 스크립트를 이용한 주소입력 및 페이지 내에서 추출할 정보 정의 가능

- 사용 라이브러리
  - httpComponent v4.5(httpclient, httpcore)
  - Jsoup v1.10.3

- 클래스 설명
  - CrawlerDriver.class: 메인 실행 클래스. 각 사이트에 맞는 다양한 설정 및 수집 절차등을 수집대상 사이트에 맞게 정의하여 사용
  - PageArgument.class: 스크립트파일 변수 클래스
  - ParseAddr.class: http 주소 분석 클래스. 주소를 분석하여 주소내의 변수값을 분리
  - ParseScript.class: 스크립트 파일 분석 클래스. 스크립트 파일을 분석하여 추출하고자 하는 정보의 내용 및 주소등을 변수에 저장
  - Crawler.class: 입력받은 주소의 페이지를 수집하는 클래스. http 데이터 전송방식에 맞게 post, get 타입 사용
  - ParseHTML.class: 스크립트 내용에 맞게 html 페이지에서 원하는 정보를 추출
  - ElementDEF.class: 스크립트 파일을 위한 Element 정의 클래스
  - ScriptDEF.class: 스크립트 파일 정보 저장 클래스
  - GithubParser: Github 서브페이지 추출을 위한 주소정보 추출을 위한 파서클래스

- 스크립트 문법
  - `id:0`
    스크립트로 추출할 정보 아이디

- addr:https://github.com/Netflix?page=1
  id에 해당하는 수집 페이지 주소
- info-1:TAG,poll-include-fragment,link,html
  페이지에서 추출하고자 하는 정보내용
  info-N 형식으로 다양한 내용 작성가능.
  TAG: 추출하려는 정보가 위치에 정의된 타입(TAG,CLASS)
  poll-include-fragment: TAG나 CLASS의 변수명
  link: 해당 내용에 정의된 정보의 이름
  html: 해당 내용에서 가져올 정보의 타입(html, text)

- Github 프로젝트별 수집기 스크립트 예시
  - 스크립트 형태로 탐색과 추출내용을 정의

id:0
addr:https://github.com/Netflix?page=1
#프로젝트 1번 페이지 주소
info-1:TAG,poll-include-fragment,link,html
#html 페이지내에 서브 프로젝트 링크가 저장된 곳의 정보
#아래의 태그에서 세부 링크주소를 가져오는 부분을 정의

```
<div class="col-3 float-right text-right">
  <poll-include-fragment
    src="/Netflix/lemur/graphs/participation?h
  </poll-include-fragment>
</div>
```

id:1
info-1:CLASS,repository-meta,Description,text
#서브프로젝트 페이지에서 프로젝트 설명정보가 위치한 곳의 정의
#아래의 태그내 class부분의 정보가 저장된 내용을 수집

```
<div class="repository-meta mb-0 mb-3 js-repo-meta-edit js-details-
  <div class="repository-meta-content col-11 mb-1">
        <span class="col-11 text-gray-dark mr-2" itemprop="about">
          A distributed in-memory data store for the cloud
        </span>
  </div>
```

info-2:CLASS,text-emphasized,info,text
#서브프로젝트 페이지에서 프로젝트 정보가 위치한 곳의 정의 1(commit 등)
#아래의 태그내 class부분의 정보가 저장된 내용을 수집

```
H6V6h5v2H8v5zM7 1C4.81 1 2.87 2.02 2.02 1.59 3.
.34.03-.67.09-1H.08C.03 7.33 0 7.66 0 8c0
    <span class="num text-emphasized">
      1,301
    </span>
    commits
</a>
```

info-3:CLASS,social-count,subInfo,text
#서브프로젝트 페이지에서 프로젝트 정보가 위치한 곳의 정의 2(watch 등)

#메인함수에서 0 번 스크립트로 수집할 링크 정보를 수집후, 세부 페이지에서 1 번 스크립트에 해당하는 정보 수집 가능하도록 플로우 구성

- 실행 화면 예시

```
Randolui-MacBook-Pro:ac randol$ ls
ac.jar          script.prj
Randolui-MacBook-Pro:ac randol$ java -jar ac.jar
log4j:WARN No appenders could be found for logger (org.apache.http.client.protocol.RequestAddCookies).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
Netflix/chaosmonkey/
Description     Chaos Monkey is a resiliency tool that helps applications tolerate random instance failures.
Commits:        90
Branches:       8
Releases:       3
Contributors:   9
Watch:  239
Star:   2,561
Fork:   150
Netflix/sstable-adaptor/
Description     No description, website, or topics provided.
Commits:        18
Branches:       2
Releases:       0
Contributors:   0
Watch:  146
Star:   1
Fork:   1
Netflix/lemur/
Description     Repository for the Lemur Certificate Manager
Commits:        927
Branches:       52
Releases:       11
Contributors:   50
Watch:  233
Star:   640
Fork:   108
```

- 데모사이트(2017 10 월 http://www.openankus.com 사이트에 구현 예정)