

사용법

프로그램 기반: 자바




하는 일: 특정 갤러리의 모든 글을 모아 통계 데이터를 작성합니다.

하지 않는 일: 엑셀로 그래프짜내는건 사용자의 일

목차

- 구성요소
- 준비
- 사용방법과 주작의 이해

구성요소

이름	수성한 날짜	유형	크기
 Gallery Analyser.jar	2019-05-04 오후...	Executable Jar File	312KB
 run.bat	2019-05-04 오후...	Windows 배치 파일	1KB
 setting.txt	2019-05-04 오후...	텍스트 문서	1KB

- **Gallery Analyser.jar**

메인프로그램


- **run.bat**

이거로 실행

- **setting.txt**

UTF-8 인코딩의 설정파일

설정파일의 내용

 setting.txt - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

##설정파일입니다.

gallID,haruhiism,갤러리 고유 ID

major,false,메이저갤 =true / 마이너갤 = false

maxPage,1000,글어널 최대 페이지

GalleryYear,2016,갤러리 생성년도

CurrentYear,2019,현재 년도

minimumActiveDays,0,유저 필터링 기준 최소 활동기간

minimumFrequency,0,유저 필터링 기준 최소 글 작성수

kToken,3,키워드 끊어내는 글자수 (예: 3으로 설정 [스즈미]야하루히 = [스즈미]야 호루히)

crawlKeywords,true,키워드 분석도 한다 =true

collectOnlyAfterThisPage,-1

gallID

 <https://gall.dcinside.com/mgallery/board/lists/?id=haruhiism>

타겟 갤러리의 고유 ID입니다 URL의 이 부분을 보면 됩니다.

major

마이너갤은 false

메이저갤은 true

maxPage

갤어널 최대 페이지



여기서 100개 설정후

1 2 3 4 5 6 7 8 9 10 다음 끝

갤러리 목록 맨 아래의 끝을 눌러서

https://gall.dcinside.com/mgallery/board/lists/?id=haruhiism&page=80&list_num=100

최대 몇페이지 갤러리인지 확인합니다.

80페이지군요.

GalleryYear

• 개설일 2016-01-20

개설일에서 년도를 4자리로 기입합니다.

CurrentYear

현재 년도

minimumActiveDays,minimumFrequency

유저수가 너무 많은 갤러리는 모두 출력하지 않고 조건을 걸고 필터링이 가능합니다.

출력되지 않을 뿐 데이터는 모두 저장됩니다.

kToken

클러스터링 단위로 키워드 추출에 사용됩니다.

건드릴 필요 없습니다.

crawlKeywords

true= 키워드 분석도 합니다.

collectOnlyAfterThisPage

원지 모르겠습니다.

mergeGonick

true = 닉변한 고닉들은 동일인물로 취급

준비단계

0.자바 설치

PC에 JRE가 없는 유저는 먼저 JRE를 설치합니다.

<https://www.oracle.com/technetwork/java/javase/downloads/jre8-downloads-2133155.html>

여기에 가서 Accept License Agreement 동의후 적당히 맞는 OS의 .exe를 받아 설치합니다.

Java SE Runtime Environment 8 Downloads
Do you want to run Java™ programs, or do you want to develop Java programs? If you want to run Java programs, but not develop them, download the Java Runtime Environment, or JRE™.

If you want to develop applications for Java, download the Java Development Kit, or JDK™. The JDK includes the JRE, so you do not have to download both separately.

[JRE 8u201 Checksum](#)
[JRE 8u202 Checksum](#)

Java SE Runtime Environment 8u201
You must accept the [Oracle Binary Code License Agreement for Java SE](#) to download this software.

☐ Accept License Agreement ☒ Decline License Agreement

Product / File Description	File Size	Download
Linux x86	68.1 MB	jre-8u201-linux-i586.rpm
Linux x86	83.8 MB	jre-8u201-linux-i586.tar.gz
Linux x64	64.91 MB	jre-8u201-linux-x64.rpm
Linux x64	80.73 MB	jre-8u201-linux-x64.tar.gz
Mac OS X x64	76.18 MB	jre-8u201-macosx-x64.dmg
Mac OS X x64	67.77 MB	jre-8u201-macosx-x64.tar.gz
Solaris SPARC 64-bit	46.27 MB	jre-8u201-solaris-sparcv9.tar.gz
Solaris x64	50.14 MB	jre-8u201-solaris-x64.tar.gz
Windows x86 Online	1.87 MB	jre-8u201-windows-i586-iftw.exe
Windows x86 Offline	63.53 MB	jre-8u201-windows-i586.exe
Windows x86	66.51 MB	jre-8u201-windows-i586.tar.gz
Windows x64	71.44 MB	jre-8u201-windows-x64.exe
Windows x64	71.29 MB	jre-8u201-windows-x64.tar.gz

사용법

setting.txt - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

##설정파일입니다.

gallID,haruhiism,갤러리 고유 ID

major,false,메이저갤 =true / 마이너갤 = false

maxPage,1000,글어널 최대 페이지

GalleryYear,2016,갤러리 생성년도

CurrentYear,2019,현재 년도

minimumActiveDays,0,유저 필터링 기준 최소 활동기간

minimumFrequency,0,유저 필터링 기준 최소 글 작성수

kToken,3,키워드 글어내는 글자수 (예: 3으로 설정 [스즈미]야하루히 = [스즈미]야 호루히)

crawlKeywords,true,키워드 분석도 한다 =true

collectOnlyAfterThisPage,-1

1.setting.txt 를 열고 적당히 설정합니다.

보통은 gallID, major, maxPage 만 보면 됩니다.

```
[CORE #0] Completed: 10 | 66%
[CORE #1] Completed: 6 | 37%
[CORE #2] Completed: 10 | 62%
[CORE #3] Completed: 7 | 46%
[CORE #4] Completed: 11 | 73%
[CORE #5] Completed: 7 | 46%
[CORE #6] Completed: 6 | 40%
[CORE #7] Completed: 10 | 66%
[CORE #8] Completed: 3 | 20%
[CORE #9] Completed: 11 | 73%
```

완료되기를 기다립니다.

Status exception: Re-add

java.net.SocketTimeoutException: Read timed out

[NEW]http://gall.dcinside.com/mgallery/board/lists?id=haruhiism&list_num=100&sort_type=N&search_head=&page=56 Attempts:

at java.base/java.net.SocketInputStream.socketRead0(Native Method)

Attempt SIZE=2

at java.base/java.net.SocketInputStream.socketRead(SocketInputStream.java:115)

at java.base/java.net.SocketInputStream.read(SocketInputStream.java:168)

at java.base/java.net.SocketInputStream.read(SocketInputStream.java:140)

at java.base/java.io.BufferedInputStream.fill(BufferedInputStream.java:252)

at java.base/java.io.BufferedInputStream.read1(BufferedInputStream.java:292)

at java.base/java.io.BufferedInputStream.read(BufferedInputStream.java:351)

at java.base/sun.net.www.http.HttpClient.parseHTTPHeader(HttpClient.java:746)

at java.base/sun.net.www.http.HttpClient.parseHTTP(HttpClient.java:689)

at java.base/sun.net.www.protocol.http.HttpURLConnection.getInputStream0(HttpURLConnection.java:1604)

at java.base/sun.net.www.protocol.http.HttpURLConnection.getInputStream(HttpURLConnection.java:1509)

at java.base/java.net.HttpURLConnection.getResponseCode(HttpURLConnection.java:527)

at org.jsoup.helper.HttpConnection\$Response.execute(HttpConnection.java:429)

at org.jsoup.helper.HttpConnection\$Response.execute(HttpConnection.java:410)

at org.jsoup.helper.HttpConnection.execute(HttpConnection.java:164)

at org.jsoup.helper.HttpConnection.get(HttpConnection.java:153)

at gallery.Crawler.Crawler_DC_mx_title.lambda\$scrollRaw\$0(Crawler_DC_mx_title.java:42)

at java.base/java.lang.Thread.run(Thread.java:834)

중간중간 에러 메시지가 나옵니다.

이건 신경 쓸 필요가 없습니다.

서버가 불안정해서 끊이지 않은 페이지는 모아줬다가 다시 끊게 구성해둔..듯 한데 제대로 되는지는
사실 저도 모름 지스

```
o o (122.254)
└─ 활동기간: 1일
└─ 글 작성:1
└─ 매일
```

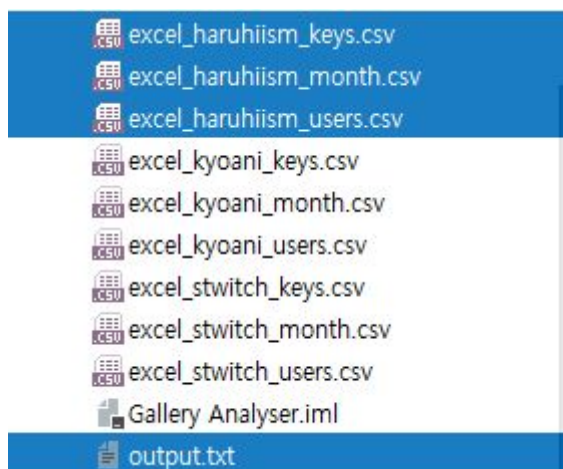
```
o o (180.229)
└─ 활동기간: 1일
└─ 글 작성:1
└─ 매일
```

```
_____19년 6월 자료없음
_____19년 7월 자료없음
_____19년 8월 자료없음
_____19년 9월 자료없음
_____19년 10월 자료없음
_____19년 11월 자료없음
_____19년 12월 자료없음
```

```
0
0
```

```
Process finished with exit code 0
```

이런게 나오면 작업이 끝입니다.



프로그램 실행 폴더에 가면
csv파일 세개와
txt파일 하나가 있습니다.

csv는 엑셀에서 가져와서 알아서 편집해서 씁니다.

excel_haruhiism_keys.csv

```
vocab,Freq
하루히,1563
스즈미야,262
나가토,238
하루히는,205
미쿠루,128
소설,117
3기,108
미거,106
애니,106
시발,100
입결,99
진짜,86
에미트,74
근대,69
하루히록 64
```

_keys.csv는 단어와 출현빈도가 정리되어있습니다.

```
Date,totalPost,totalReply,totalView,newUser,avgStay
2016-1-01,76,125,4727,5,717
2016-2-01,187,547,14184,4,396
2016-3-01,79,109,4619,0,0
2016-4-01,121,279,10876,1,893
2016-5-01,58,157,13033,1,47
2016-6-01,25,26,1627,2,0
2016-7-01,31,46,2346,1,0
2016-8-01,40,43,2186,0,0
2016-9-01,33,45,2136,3,8
2016-10-01,20,18,1361,1,939
2016-11-01,18,16,2187,1,0
2016-12-01,31,44,1759,5,687
2017-1-01,44,103,5076,5,95
2017-2-01,139,212,13957,2,532
2017-3-01,74,74,8650,3,314
2017-4-01,54,57,6493,0,0
```

_month.csv는 월별 데이터가 정리되어있습니다.

필드는 순서대로

월 / 총 글수 / 총 댓글수 / 총 조회수 / 신규 유저 /신규유저의 평균 체류기간

```

Name,fDate,lDate,activeDays,totalWrite
가나하하비키,2016-12-16,2019-4-28,859,478
스즈미야하루히,2018-1-10,2019-5-4,483,1023
ㅇㅇ,2016-1-25,2019-5-4,1198,490
dcoldbie,2018-11-6,2019-5-4,177,529
320,2019-4-28,2019-4-28,0,4
asnwer,2018-12-1,2019-4-28,144,4
49me,2018-11-27,2019-4-28,149,2
레병장,2019-4-28,2019-4-28,0,1
park2,2018-8-25,2019-4-28,244,109
ㅇㅇ (39.112),2018-9-2,2019-4-28,236,7
하라사와유미,2019-4-28,2019-4-28,0,3
김혜린,2019-4-28,2019-4-28,0,1
운영지,2019-3-25,2019-3-25,0,1
인천SK,2018-7-8,2019-5-4,299,221
-----,2019-4-28,2019-4-28,0,1

```

_users.csv는 유저정보가 들어있습니다.

필드는 순서대로 이름/fDate첫글날짜/lDate마지막글날짜/activeDays활동기간(일)/totalWrite총글작성입니다.

```

92. 오늘의 [21]회
93. 아직도 [21]회
94. 아직 [21]회
95. 씨발 [21]회
96. 무슨 [21]회
97. 나는 [21]회
98. 그냥 [21]회
99. 겔주님 [21]회
100. 이렇게 [20]회
-----16년 1월 통계
----- 신규 유입 유저: 17 명
1. 하루히 (갤러리 공통 키워드)
2. 오늘은 (갤러리 공통 키워드)
3. 갤러리 (갤러리 공통 키워드)
4. 스즈미야 (갤러리 공통 키워드)
5. 날입니다
6. 귀엽지
7. 하루히가
8. 저저개크

```

output.txt는 콘솔에 출력된 결과가 복사되어있습니다.

유지보수 쟁점

```

Thread go = new Thread() -> {
    while (!CORES.get(finalC).isEmpty()) {
        String tempPage = CORES.get(finalC).poll();
        try {
            Document doc = Jsoup.connect(tempPage).get();
            Elements posts = doc.select( cssQuery: "tbody").first().select( cssQuery: "tr[class=ub-content],tr[class=ub-conti
            //System.out.println("Frequency: "+posts.size());
            assert posts != null;
            for (Element post : posts) {
                String type = post.select( cssQuery: "td[class=gall_num]").first().text();
                if (!isValidIndex(type)) continue;
                String writer = post.select( cssQuery: "td[class=gall_writer ub-writer]").attr( attributeKey: "data-nick");
                writer = cleanseNick(writer);
                String ip = post.select( cssQuery: "span[class=ip]").text();
                String view = post.select( cssQuery: "td[class=gall_count]").text();
                //Replies
                Element ReplyBox = post.select( cssQuery: "span[class=reply_num]").first();
                String replies = "0";
                if (ReplyBox != null) {
                    replies = ReplyBox.text();
                    replies = removeChars(replies).split( regex: "[ ]")[0];
                }
                String date = post.select( cssQuery: "td[class=gall_data]").text();
            }
        } catch (Exception e) {
            e.printStackTrace();
        }
    }
}

```

Crawler_mx_title.java -> 갤러리 긁기 담당

문의

```

reading settings.txt done!
Initiating Chrome Driver
Starting ChromeDriver 2.39.562718 (9a2698cbe08cf5a471a29d90c8b3e12becabb0e9) on port 36717
Only local connections are allowed.
1월 25, 2019 5:40:04 오후 org.openqa.selenium.remote.ProtocolHandshake createSession
정보: Detected dialect: OSS
Selenium connected to tps://gall.dcinside.com/board/view?id=comic_new&no=9232991
시작
Exception in thread "Thread-4" Exception in thread "Thread-5" java.lang.NoClassDefFoundError: org/
    at com.company.ViewAttack.attack(ViewAttack.java:46)
    at com.company.ViewAttack.lambda$startThreads$0(ViewAttack.java:20)
    at java.lang.Thread.run(Unknown Source)
Caused by: java.lang.ClassNotFoundException: org.jsoup.Jsoup
    at java.net.URLClassLoader.findClass(Unknown Source)
    at java.lang.ClassLoader.loadClass(Unknown Source)
    at sun.misc.Launcher$AppClassLoader.loadClass(Unknown Source)
    at java.lang.ClassLoader.loadClass(Unknown Source)
    ... 3 more
Exception in thread "Thread-8" Exception in thread "Thread-7" Exception in thread "Thread-6" java.
    at org.jsoup.Jsoup
    at com.company.ViewAttack.attack(ViewAttack.java:46)
    at com.company.ViewAttack.lambda$startThreads$0(ViewAttack.java:20)

```

문제 발생시 cmd 화면을 찍어 프로그래밍 갤러리에 문의하면 프잘알이 튀어나와서 프로그램 새로 만들어줌