

2017 Contributhon



ankus crawler Team

- 장동훈 / 김태균 / 신홍중
- 조재훈 / 오시영 / 노유정

1. 아이디어 발생
2. 분석 구성 요소
3. 분석하고자 하는 목표 시스템
4. Ankus crawler Github 현황
5. 데이터 분석 절차
6. 분석 시나리오
7. 개발 사항
8. 분석 결과

Appendix

1. 아이디어 발생



How to use Ankus Crawler?



2. 분석 구성 요소

- 타켓 사이트 정의 : 음악 차트사이트
 - > Mnet, Melon, Gaon chart
- 데이터 수집 : 크롤링
 - > ankus Crawler를 이용하여 수집
 - > 언어 : Java
- 데이터 전처리 : Hadoop, MapReduce
 - > ankus analyzer ver 1.1
- 데이터 분석 : MapReduce기반 장르 군집분석
- 데이터 기반 예측 : 장르 순위 변화 통계 분석을 통한 예측
- 데이터 시각화 : D3.js을 통한 시각화

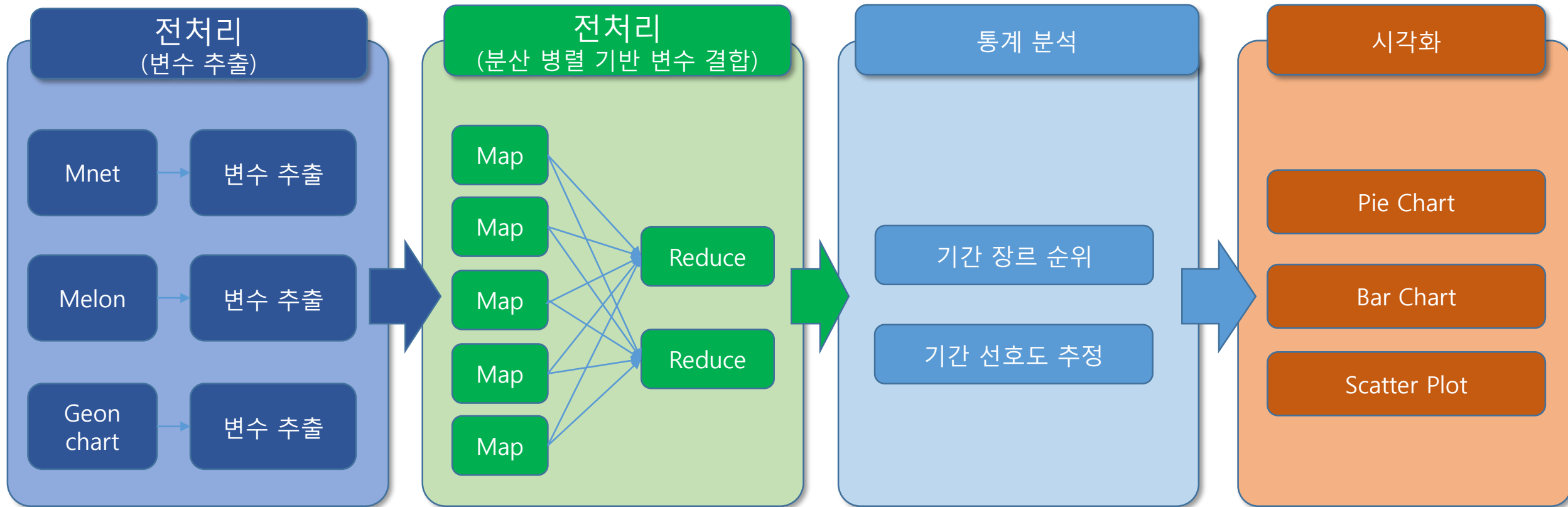
3. 분석하고자 하는 목표 시스템

<http://www.mnet.com/chart/TOP100/>

http://www.melon.com/genre/song_list.htm#params%5BgnrCode%5D=GN0100¶ms%5BdtlGnrCode%5D=¶ms%5BborderBy%5D=NEW¶ms%5BsteadyYn%5D=N&po=pageObj&startIndex=51

<http://gaonchart.co.kr/main/section/chart/album.gaon?nationGbn=T&serviceGbn=&targetTime=2011&hitYear=2011&termGbn=year>

4. 데이터 분석 절차



5. Ankus Crawler Github 현황

<https://github.com/ldbslovis/kosscontribution2017>

ldbslovis / kosscontribution2017
forked from kosslab-kr/kosscontribution2017

Watch 1 Star 1 Fork 13

Code Issues 20 Pull requests 8 Projects 0 Insights

2017 공개SW 컨트리뷰톤 Repository

110 commits3 branches0 releases4 contributors

Branch: masterNew pull requestCreate new fileUpload filesFind fileClone or download

This branch is 12 commits ahead, 21 commits behind kosslab-kr:master.

Pull requestCompare

ldbslovis Merge Data Visualization partLatest commit af1d19f 2 days ago

AnkusMelon Crawler 3rd commit14 days ago

ankus-crawlerAnkus crawler 9th commit3 days ago

webMerge Data Visualization part2 days ago

README.mdUpdate README.md2 months ago

2017-11-20

7

5. Ankus Crawler Github 현황

<https://github.com/ldbslovis/kosscontributhon2017>

① 20 Issues created by 3 people

- Opened #30 데이터 시각화 - scatter plot css.js 수정 7 hours ago
- Opened #28 데이터 시각화 - scatter plot js, css 최종본 15 hours ago
- Opened #26 데이터 시각화 - 메인 페이지 수정 및 라이선스 추가 15 hours ago
- Opened #24 데이터 수집 - 수집, 전처리 및 분석 과정 최종 문서화 16 hours ago
- Opened #21 결과 데이터 업로드 - 더미데이터 업로드 18 hours ago
- Opened #20 데이터 수집 - 연월 장르별 선호도 계산 20 hours ago
- Opened #19 데이터 수집 - 가온 크롤링 데이터 전처리 20 hours ago
- Opened #18 데이터 수집 - 멜론 크롤링 데이터 전처리 20 hours ago
- Opened #17 데이터 수집 - 엠넷 크롤링 데이터 전처리 20 hours ago
- Opened #16 데이터 수집 - 가온 사이트 크롤러 생성 21 hours ago
- Opened #15 데이터 수집 - 엠넷 사이트 크롤러 생성 21 hours ago
- Opened #14 데이터 수집 - 멜론 사이트 크롤러 생성 21 hours ago


5. Ankus Crawler Github 현황


<https://github.com/ldbslovis/kosscontribution2017>

The screenshot shows a GitHub pull request interface. At the top, a user 'ohahohah' is shown as the reviewer, with a 'View changes' button. Below this, a comment from '@NohYuJeong' states: '다른 차트와 한 페이지에서 합쳐 보여주기 위해 중들이 예상되는 부분에 대해 리뷰했습니다.' (I reviewed the parts where it was expected to be combined with other charts on one page). The main content area displays the diff for 'web/assets/css/scatterPlot.css'. The diff shows a change from '...' to '... @@ -0,0 +1,29 @@' and a new line '1 +body {' is added. Below the diff, there are two comments: one from 'ohahohah' 13 hours ago saying 'body 는 main.css 에서 관리합니다' (body is managed in main.css), and another from 'NohYuJeong' 7 hours ago saying '삭제하겠습니다!' (I will delete it!). There is a 'Reply...' input field below the second comment. To the right of the diff, there are sections for 'Labels' (None yet), 'Projects' (None yet), 'Milestone' (No milestone), and 'Notifications' (Unsubscribe button, with a note 'You're receiving notifications because you commented.'). Below these, it says '2 participants' and shows the avatars of 'NohYuJeong' and 'ohahohah'. At the bottom, there is a section for 'Commits' showing 'NohYuJeong added some commits 7 hours ago' with two commits: 'Modify scatter plot js,css' (13c697a) and 'Update scatter plot css' (11a23a0).


5. Ankus Crawler Github 현황


<https://github.com/ldbslovis/kosscontribution2017>


 Reply...


web/assets/css/scatterPlot.css  Hide outdated

```
9 + font-family: sans-serif;
10 + font-size: 14px;
11 +}
12 +svg{
```


 ohahahah 13 hours ago
다른 chart에서도 쓰이는 svg, text 같은 tag 와 공통으로 쓰는 class들의 이름을 바꾸어 scatterPlot에만 적용하도록 바꾸는 건 어떨까요? chart 합칠때 충돌이 날 것 같네요.


 NohYujeong 7 hours ago
네 이부분은 수정해서 다시 PR 넣도록 하겠습니다!

 Reply...

web/assets/js/scatterPlot.js  Hide outdated

```
1 +// Tutorial of (Scatterplot d3 beautiful) by Jaydev
2 +
3 +// insert csv raw url
4 +d3.csv("https://raw.githubusercontent.com/NohYujeong/practice_d3_js/master/result.csv")
```

 ohahahah 13 hours ago
프로젝트 폴더안의 result.csv 바인딩으로 바꿔주세요.

 NohYujeong 7 hours ago

5. Ankus Crawler Github 현황

<https://github.com/ldbslovis/kosscontribution2017>

👤 2 Pull requests merged by 1 person

Merged #1 Clean Data mr part 6 hours ago

Merged #23 Calculate Preference 6 hours ago

👤 8 Pull requests proposed by 2 people

Proposed #4 Delete un used file 2 days ago

Proposed #9 Upload Description of DataVisualization page 2 days ago

Proposed #12 Create Chart View 22 hours ago

Proposed #13 Create genreYearInfo 21 hours ago

Proposed #22 Upload dummy data 18 hours ago

Proposed #25 Create Genre Info Chart 15 hours ago

Proposed #27 Upload Img File 15 hours ago

Proposed #29 Modify scatter plot 15 hours ago

🔔 20 Issues created by 3 people

5. Ankus Crawler Github 현황

<https://github.com/ldbslovis/kosscontributhon2017>

데이터 수집 - 연월 장르별 선호도 계산 #20

 Open jae0303 opened this issue 22 hours ago · 4 comments



jae0303 commented 22 hours ago



No description provided.



ohahohah commented 19 hours ago



@jae0303 File encoding set UTF-8로 확인부탁드립니다.



1



jae0303 referenced this issue 18 hours ago

Calculate Preference #23

Merged



jae0303 commented 18 hours ago



UTF-8 확인했습니다

5. Ankus Crawler Github 현황

<https://github.com/ldbslovis/kosscontribution2017>

The screenshot shows a GitHub pull request interface. At the top, the repository name is 'ldbslovis / kosscontribution2017', noted as 'forked from kosslab-kr/kosscontribution2017'. It has 1 Watch, 1 Star, and 13 Forks. The pull request title is 'Modify scatter plot #29'. A green 'Open' button indicates the pull request is ready for review. The pull request was created by 'NohYujeong' and wants to merge 7 commits into 'ldbslovis:web' from 'NohYujeong:modifyScatterPlot'. The interface shows 8 Conversations, 7 Commits, and 4 Files changed, with a net change of +1,124 lines and -0 deletions. A comment from 'NohYujeong' 15 hours ago mentions issue #28 and the modification of js and css for the scatter plot chart. A review by 'ohahohah' 13 hours ago is shown, with a 'View changes' button. The right sidebar lists 'Reviewers' (ohahohah), 'Assignees' (None assigned), 'Labels' (None yet), and 'Projects' (None yet).

5. Ankus Crawler Github 현황

- 토즈 오프라인 미팅
- Github Flow

6. 분석 시나리오

데이터 수집

데이터 전처리

데이터 결합/분석

분석 시각화

Mnet

노래 제목, 앨범명,
가수명, 일별 순위
(2011.08.01 ~ 2017.10.31)

Melón

노래 제목, 해당 장르

gaon

앨범명, 가수명,
월 누적판매량, 배급사
(2011.01 ~ 2017.09)

ankus
ankus crawler



순위#날짜#제목#가수#앨범 이름

장르#제목#가수#앨범 이름

번호#제목#가수#월 판매량 /누적 판매량



날짜#장르#가수#제목#순위

날짜#장르#가수#제목#순위

날짜#장르#가수#제목#순위



종합 Scatter 차트

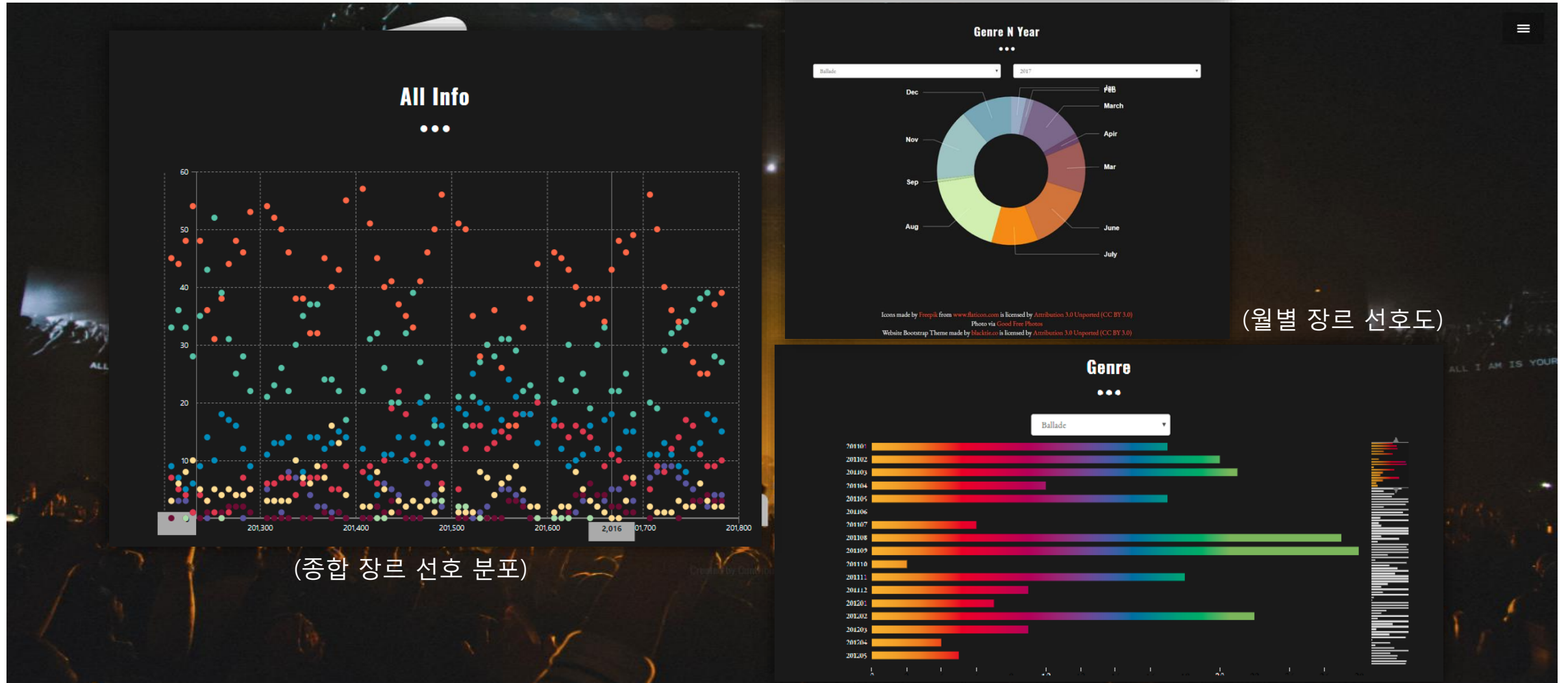
월별 Pie 차트

월별 Bar 차트

7. 개발 사항

- 데이터 수집 및 전처리
 - > 엠넷, 멜론, 가온 사이트 수집 / 공통 값 전처리
 - > 전처리에 대한 보완 필요
- 데이터 분석
 - > 선호도에 대한 더 정확한 알고리즘 필요
- 데이터 시각화
 - > Scatter Plot

8. 분석 결과



Appendix