

Statistician

통계분석학회 P-SAT

27기 신입학회원 모집

2007
P-Sat

2007
P-Sat



P-SAT

목차

1.

P-SAT 소개

P-SAT 이란?

팀 구성

학회 활동

2.

활동내역 및 수상내역

활동내역

수상내역

3.

리크루팅

리크루팅 지원 자격 및 일정

1. P-SAT 소개



"Power Statistical Analysis Technique"

성균관대학교 통계학과 소속 유일한 통계분석 학회로,
통계학적 전문성 확보를 통한 실제 데이터 분석능력 함양을 목표로 합니다.



14년째 활동을 이어오고 있으며, 그동안 축적해온 자료들로 분석 이론 공부와 분석을 해볼 수 있는 탄탄한 커리큘럼을 지니고 있습니다.



학회 내 프로젝트로 쌓은 실력을 기반으로 매년 많은 공모전 수상을 하고 있습니다.



데이터 분석 분야에 진출한 많은 선배님들과 유기적 네트워크가 형성되어 있습니다.

1. P-SAT 소개

P-SAT이란?
팀 구성
학회 활동

2. 활동내역 및 수상내역

3. 리쿠르팅



"TEAM"

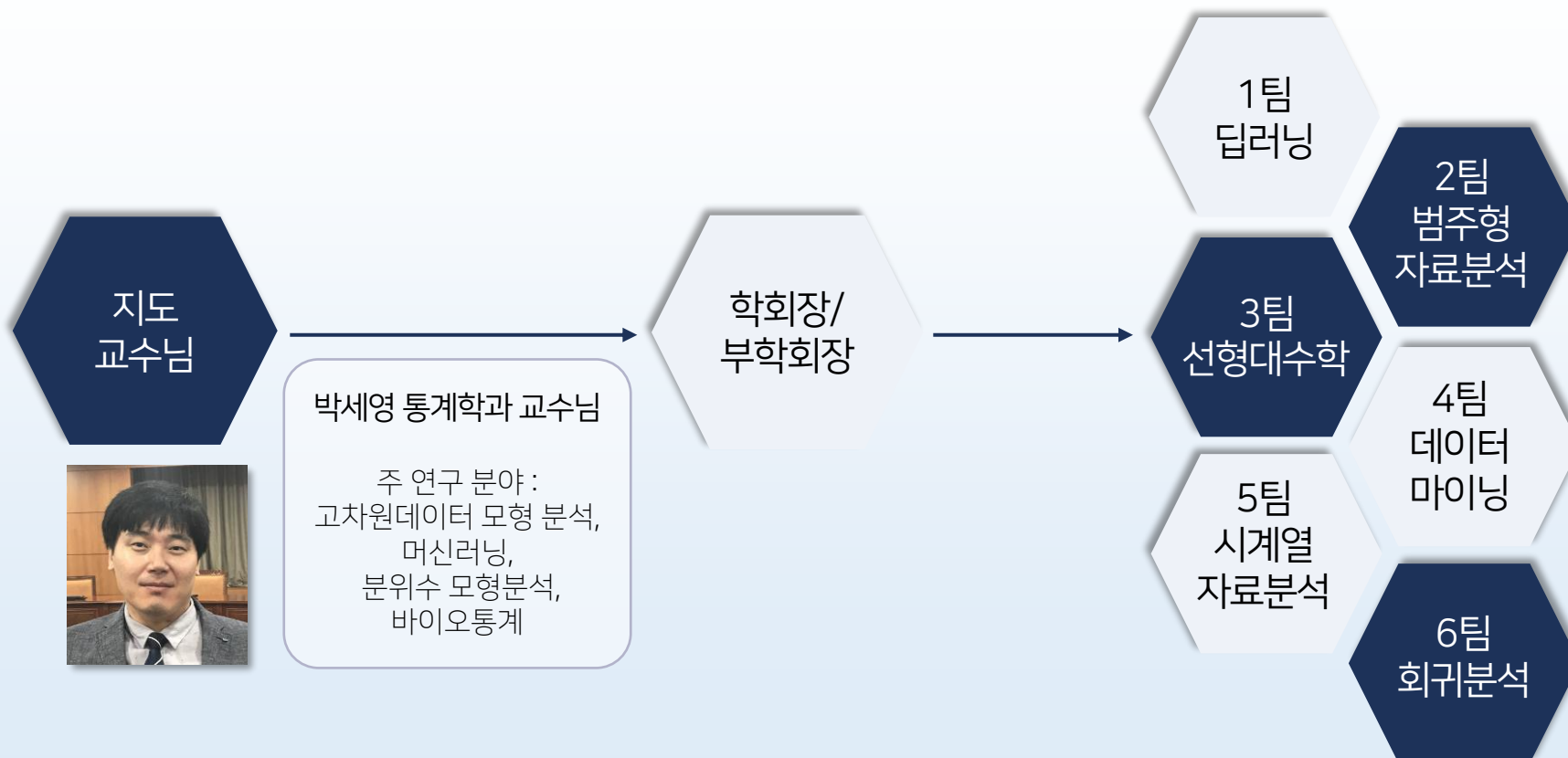
지도교수님, 학회장/부학회장, 그리고 6개의 팀으로 구성되어 있습니다.
전체 학회원은 32명입니다.

1. P-SAT 소개

P-SAT이란?
팀 구성
학회 활동

2. 활동내역 및 수상내역

3. 리쿠르팅





"TEAM"

6개 팀으로 구성되어 있으며, 각 팀은 데이터 분석에 필요한 이론들을 공부하고 학회원들에게 발표하여 분석 지식을 전문적으로 학습합니다.

Deep Learning

비정형 데이터를 분석할 수 있는 딥러닝 모델들을 다룸. CNN, RNN 등 다양한 모델들의 이론을 학습하고 어텐션 기법을 배움. 학습한 이론들을 실습을 통해 구현함

Linear Algebra

통계의 근간이 되는 선형대수를 공부함. 선형대수 개념을 수식적, 공간적으로 익히고, 딥러닝, Least-Square 등의 원리를 선형대수의 응용 차원에서 이해함. 통계지식에 깊이를 더하는 것이 목표

Time Series

시간에 따라 관측되는 자료인 시계열 자료에 대한 분석방법을 다룸. 시계열분석의 개념과 다양한 정상, 비정상 시계열 모형을 학습하고 이를 R을 통해 실습.

1팀
딥러닝

2팀
범주형
자료분석

3팀
선형대수학

4팀
데이터
마이닝

5팀
시계열
자료분석

6팀
회귀분석

Categorical Data Analysis

범주형 자료를 분석하는 방법들을 이론과 실습을 통해 다룸. 오즈비, 로지스틱회귀를 비롯한 GLM, 모델 평가지표와 인코딩 방법, 비대칭 데이터를 해결하는 샘플링 방법들을 학습하고 R을 통해 실습.

Data Mining

데이터 분석의 대표적인 방법론인 지도학습과 비지도학습을 주로 다룸. Tree 모델 전반에 대해 익히고 더불어 추천시스템의 작동 원리를 이해하고자 함. R과 Python을 통해 실제 데이터에 적용하여 실습함.

Regression Analysis

데이터 분석의 가장 기본이 되는 회귀 모형의 이론과 기법에 대해 다룸. 회귀 모형의 특징과 가정, 변형에 대해 학습하고 R과 python을 통해 이를 구현하는 법을 실습.

1. P-SAT 소개

P-SAT이란?
팀 구성
학회 활동

2. 활동내역 및 수상내역

3. 리쿠르팅

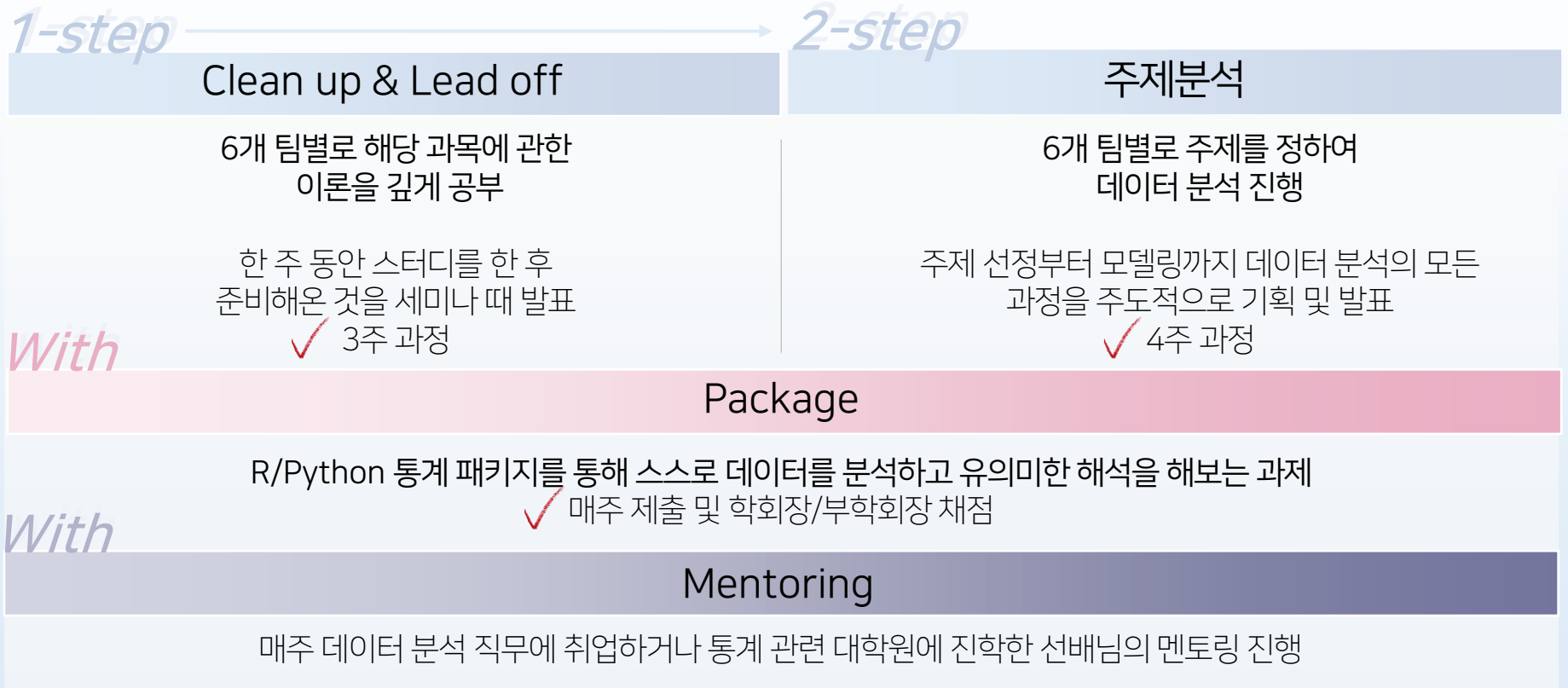
"Curriculum" Clean up&Lead off 이후 주제분석을 하는 2-step 커리큘럼으로 진행하고 있습니다. 더불어 Package 과제와 Mentoring을 병행합니다.

1. P-SAT 소개

P-SAT이란?
팀 구성
학회 활동

2. 활동내역 및 수상내역

3. 리쿠르팅



2. 활동내역 및 수상내역

2007 P-SAT 26기 주제분석

"Deep Learning"

1팀 딥러닝팀 주제분석 요약입니다.

1. P-SAT 소개

2. 활동내역 및 수상내역

활동내역
수상내역

3. 리쿠르팅



분석 내용

: 가사 데이터를 CNN + LSTM, Word2Vec, Doc2Vec, Fasttext 모델을 기반으로 임베딩 벡터로 만들고, 멜로디 데이터를 VGG16 모델을 기반으로 임베딩 벡터로 만들어 유사도 행렬을 기준으로 추천 시스템을 구축함

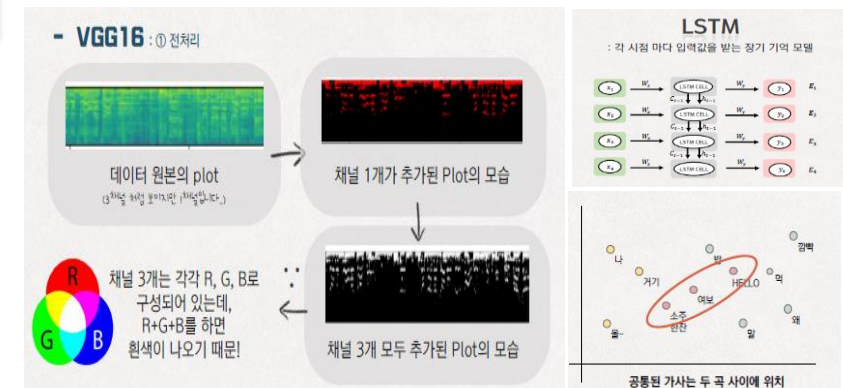
주제

: 현재 취향에 기반한 다른 시대의 음악 추천

사용 데이터

: 카카오 아레나의 멜로디 데이터, 멜론 가사 + 차트 크롤링 데이터

** PPT 예시



27th Recruiting
2021

PSat 2007

P-SAT 26기 주제분석

"Categorical Data Analysis"

2팀 범주형 자료분석팀 주제분석 요약입니다.

설문조사를 통한
지지정당 예측

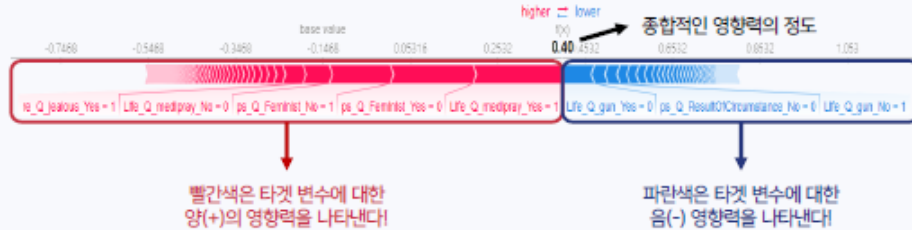
모델링 결과 해석

SHAP value

Shapley value

What is shap.force plot?

- 각 설문 참여자의 응답결과를 shap value를 활용하여 시각화한 plot
- 각 설문 참여자들의 예측 결과에 대한 해석이 가능해진다!



01. 1주차 피드백
02. DATA 정리
03. 모델링
04. 결과 해석
05. 한계와 의의

2학기 주제분석 | 범주형자료분석팀

분석 내용

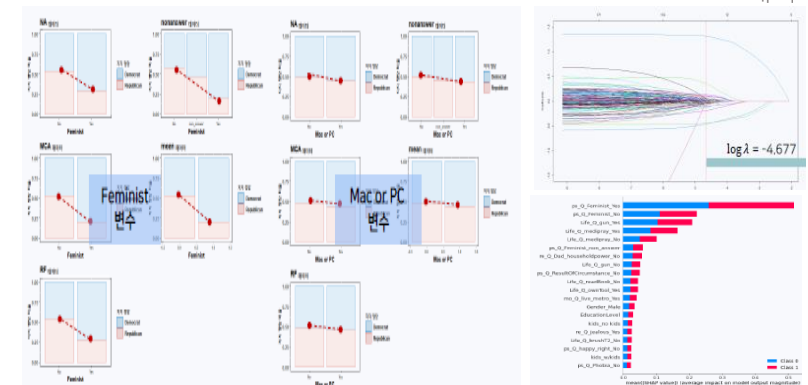
: 설문조사 데이터를 이용하여 지지정당 예측력을 높이는 모델을 만들.
MICE로 NA 를 처리하고 다양한 파생변수를 생성하여 데이터를 완성 했으며,
GLM, RandomForest, XGBoost, CATBoost, LGBM 모델을 만든 후
SHAP value를 이용하여 모델을 해석함

주제

: 설문조사를 통한 지지정당 예측

사용 데이터

: 101개 질문에 대한 설문조사 데이터



"Linear Algebra"

3팀 선형대수학팀 주제분석 요약입니다.

주제

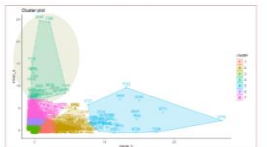
: 서울시 출근시간대 버스 노선 제안

사용 데이터

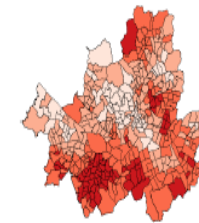
: 버스 승하차, 유동인구 등 출근시간대 관련
및 서울시 지도데이터

** PPT 예시

"클러스터 플롯"

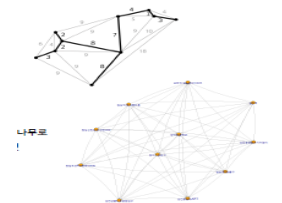
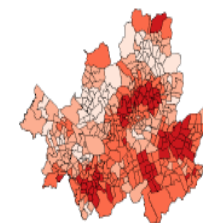


행정동별 버스승하차승객수평균



특히 Local G+ 지수에서 핫스팟이 뚜렷이 나타나며,
관악구·강남구·광진구·송파구에서 공간적 군집도가 높음

행정동별 버스하차승객수평균



최종 노선 선정

거리 알고리즘

최종 노선

최종 노선 4, 5 : 관악구

버스 승하차 수요 높았던 경류장과 버스 하차+지하철 승차 수요 높았던 역을 연결한 최종 버스 노선

2주차 복습

현행 노선 보완

신규노선제안 1-공간회귀

신규노선제안 2-군집분석

최종 노선 선정



의의

- 봉천역과 서울대입구역으로 각각 가고자 하는 주민들의 수요를 효율적으로 모두 반영
- 봉천역-서울대입구역 구간 혼잡인원 분담

한계

- 노선 최적화 과정에서 하나의 직선으로 연결되지 못할
- 도로 상황 및 도로의 종류, 양방향 고려하지 못할

분석 내용

: 서울시 출근시간대 버스인 '다람쥐버스'의 노선을 제안하기 위해 출근 시간대에 높은 버스 수요 요인과 지역을 공간회귀와 클러스터링을 통해 알아낸 후, 최소신장나무 알고리즘과 다익스트라 알고리즘을 이용하여 노선을 설계함

P-SAT 소개

활동내역 및 수상내역

활동내역
수상내역

리쿠르팅

"Data Mining"

4팀 데이터마이닝팀 주제분석 요약입니다.

주제

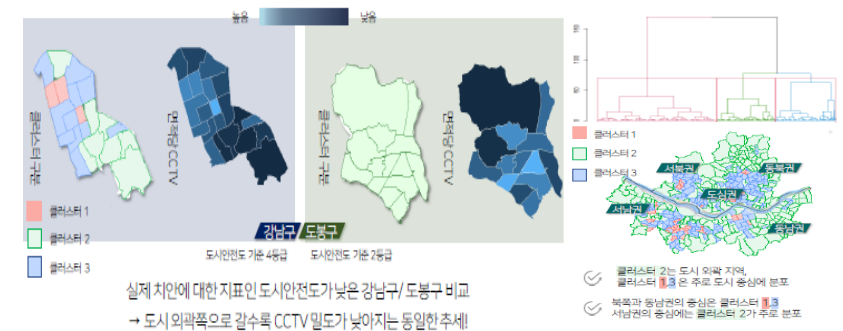
: 혼자 사는 청년들을 위한 서울시 살 곳 추천

사용 데이터

: 시세, 유동인구, 인프라, CCTV, 범죄 발생 등
서울시 내 주거 관련 데이터

** PPT 예시

잠깐! 면적당 CCTV 개수와 치안의 관계는?



1. P-SAT 소개

2. 활동내역 및 수상내역

활동내역
수상내역

3. 리쿠르팅

혼자사는 청년들을 위한 서울시 살 곳 추천

04 추천알고리즘 함수 구현

실제 구현

화려한 조명이 날 감싸네님을 위한 데마팀의 추천 결과는?

<서울특별시 마포구 합정동 월드오피스텔>

- ✓ 평수 20.52평
- ✓ 보증금 5천만원
- ✓ 월세 45만원

2020학년도 2학기 4월 데이터마이닝 4주차 주제분석

분석 내용

: 혼자 사는 청년이 살기 적절한 곳을 추천하기 위하여 주거 관련 데이터를 모아 PCA, Stepwise selection 등으로 다중공선성을 해결한 후, K-means, DBSCAN, Hierarchical 클러스터링 기법으로 지역 특성을 파악한 뒤 추천 알고리즘을 구현함

" Time Series "

5팀 시계열자료분석팀 주제분석 요약입니다.

주제

: 취향에 맞는 향수 추천

사용 데이터

: 향수 정보를 제공하는 웹사이트 크롤링 데이터

** PPT 예시

04. 추천시스템

완성된 향수추천시스템으로 시계열팀의 인생향수 찾기!

팀원B

성별 여성

main accord floral, fruity

싫어하는 노트 unusual

좋아하는 노트 musky, white flower

사용할 계절 겨울

사용할 시간 낮

floral향이 지배적이면서 musky와 white flower의 향을 가진 겨울에 쓰기 좋은 향수를 추천해 줘!

cold

main accords

floral

vanilla

almond

powdery

musky

iris

tuberose

sweet

white floral

nutty

분석 내용

: 향수 정보 웹사이트에서 색, 특징, 향과 관련한 데이터를 크롤링한 후 K-means로 향수 이미지에서 대표 색을 추출하고, Word2Vec을 통해 향과 관련한 리뷰 데이터를 가공하여 데이터를 완성 한 다음, 유사도 행렬을 구축하여 추천시스템을 구현함

Review Data

지난 주에 만들어 놔던 리뷰 태그들 (feat. TextRank)

문제점2. 태그 5개만으로도 w2v 모델 돌렸을 때 기대만큼 좋은 성능 ❌

레몬, 오렌지, 자몽 같은 가깝고 상충한 향글류 향 citrus

review_model.most_similar('citrus')

['soft', 0.9990208148856299],
['leather', 0.9989205598831177],
['beautiful', 0.9987057447433472],
['original', 0.998691737651825],
['nit', 0.9986250400543213],
['sweet', 0.9986035823822021],
['little', 0.9985989328245544],
['powdery', 0.998575675964355]

ambrosia

unusual

Word	1	2	...	19
floral	0.5560168	-0.5489983		0.18829721
fruity				0.51
...				0.16
cocktail				0.16
moss				0.16
whiteflower				0.16

유사도 - Cosine

두 벡터 간의 코사인 각도를 이용한 유사도 측정 방법

코사인 각 = 0 → 코사인 유사도 = 1
코사인 각 = 90 → 코사인 유사도 = 0

1. P-SAT 소개

2. 활동내역 및 수상내역

활동내역
수상내역

3. 리쿠르팅

1. P-SAT 소개

2. 활동내역 및 수상내역

활동내역
수상내역

3. 리쿠르팅

"Regression Analysis"

6팀 회귀분석팀 주제분석 요약입니다.

Regression PLAYLIST

03. 모델링 MF(행렬 분해) CF(협업 필터링) 추천시스템 구현

• KNN 기반 추천시스템 추천 결과

리더보드

순위	팀명	평가	노출	노출	노출
1	우정	0.187287	0.290311	0.404031	
2		0.125128	0.188899	0.448811	
3	94P	0.023175	0.080453	0.112713	

카카오 아레나 홈페이지 리더보드에서 꽤 괜찮은 성능을 보임

Tag - Song이 예측된 결과를 보면 연관성 있게 예측 됨

'OST', '애니메이션', '디즈니' 태그에 대한 곡 예측 결과

최귀분석팀 | 108

주제

: 음원 사이트 멜론 플레이리스트 예측 및 추천

사용 데이터

: 멜론 곡 정보 및 플레이리스트 데이터

분석 내용

: 주어진 플레이리스트 수록곡과 태그를 예측하기 위해, 협업필터링 모델들을 사용해 추천시스템 구현함. 추천 과정에서 Matrix Factorization, Factorization Machine, Denoising AutoEncoder, Word2Vec, knn 등 다양한 모델들을 활용함

** PPT 예시

Word2Vec - 개념 및 활용

```
model_1 = Word2Vec(val_tag, size=100, window=4, min_count=5, workers=4, sg=1)
model_2 = Word2Vec(val_tag, size=100, window=4, min_count=1, workers=4, sg=1)
```

100개의 벡터
각 단어를
100차원으로 나타내 줌
일반적으로 size가 커지면 예측력 향상(보통 100)

PCA로 2차원으로 나타내면 이런 느낌..

Denoising AutoEncoder

1. Gaussian Noise Input 2. Dropout input

Matrix Factorization

ALS

플레이리스트 내의 Songs를 원형과 고정치로 나타내 줌

① 차등 고정치에서 차등

② 차등 고정치에서 차등



"Package" 실제 데이터분석 시 도움되는 시각화, 전처리, 모델링, 해석과 관련된 난이도 있는 문제들을 해결하면서 코딩 실력을 기를 수 있습니다.

1. P-SAT 소개

2. 활동내역 및 수상내역

활동내역
수상내역

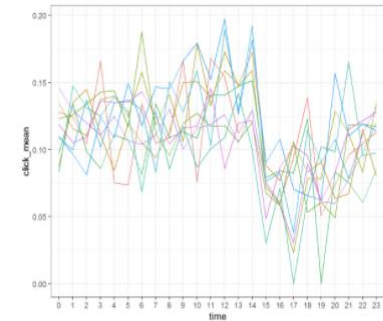
3. 리쿠르팅

** 과제 제출 코드 예시(R/PYTHON)

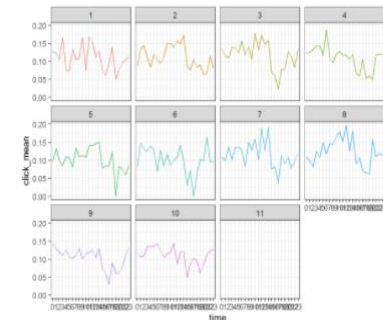
Ch.2 문제풀이

문제1

```
ggplot(data2, aes(time, click_mean)) + geom_line(aes(col = date)) + theme_bw()
```

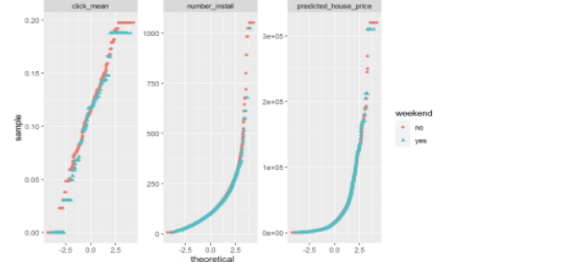


```
ggplot(data2, aes(time, click_mean)) + geom_line(aes(col = date)) + facet_wrap(~ date, ncol = 11) + theme_bw()
```

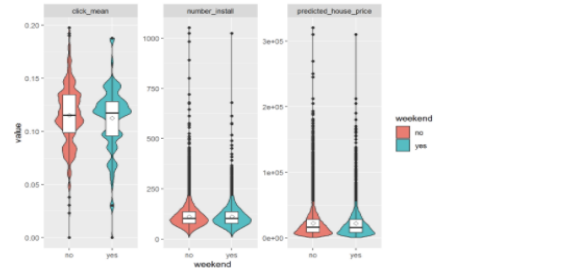


문제2 bubble은 num이라는 새로운 차

```
data1 %>% data2 %>% gather(key, value, number_instal, predicted_house_price, click_mean)
ggplot(data1, aes(sample, value, shape = weekend, colour = weekend)) + stat_qq() + facet_wrap(~ key, scales = "free_y")
```



```
ggplot(data1, aes(weekend, value)) + geom_qq() + facet_wrap(~ key, scales = "free_y")
ggplot(data1, aes(weekend, value)) + geom_qq() + facet_wrap(~ key, scales = "free_y")
ggplot(data1, aes(weekend, value)) + geom_qq() + facet_wrap(~ key, scales = "free_y")
```



```
set.seed(1)
n_split <- 5
cv <- createFolds(train$click, k = n_split)
tune_parameter <- expand_grid(k = c(3, 4, 5))
tune_parameter$logloss <- NA

for(k in 1:NROW(tune_parameter)){
  logloss_result <- c()
  for(i in 1:n_split){
    idx <- cv[[i]]

    train_x <- train[-idx, ]
    val_x <- train[idx, ]

    set.seed(1)
    model_rf_2 <- randomForest(click ~., data = train_x, stry = tune_parameter[k, 'k'])
    prediction <- predict(model_rf_2, newdata = val_x, type = "prob")

    logloss_temp <- LogLoss(prediction[, 2], as.numeric(val_x$click) - 1)
    logloss_result <- c(logloss_result, logloss_temp)
    result <- mean(logloss_result)
  }
  tune_parameter[k, 'logloss'] <- result
}
tune_parameter
```

```
## k logloss
## 1 3 0.2594496
## 2 4 0.2628726
## 3 5 0.2611207
```

```
Audience_rating = []
Audience_rating_num = []
Net_rating = []
Net_num = []

#관객의 평점, 참여 수, 넷izen 평점, 참여 수
for i in range(0, 142):
  driver.get(url + str(link_df.loc[i, 'movie_url']))
  driver.find_element_by_xpath('//*[@id="movieEndTabMenu"]/li[5]/a').click()
  time.sleep(1)

  source = driver.page_source
  soup = BeautifulSoup(source, 'html.parser')

  #관객의
  audience_content = soup.select('.st_sm')
  audience_num_content = soup.select('#actual_point_tab_inner > span')

  try:
    Audience_rating.append(audience_content[0].text)
    Audience_rating_num.append(audience_num_content[0].text)
  except:
    Audience_rating.append("")
    Audience_rating_num.append("")

  #넷izen
  content = soup.select('#netizen_point_tab_inner > ea')
  netizen_num_content = soup.select('#graph_area > div.grade_netizen > div.title_area_grade_tit > div.sc_area > span > ea')

  try:
    Net_rating.append(str(content[0].text) + str(content[1].text) + str(content[2].text) + str(content[3].text))
    Net_num.append(str(netizen_num_content[0].text))
  except:
    Net_rating.append("")
    Net_num.append("")
```



최우수상

빅콘테스트 혁신아이디어 분야

서울특별시장상 (대상)

서울특별시 빅데이터 캠퍼스 공모전

최우수상

한국정보화진흥원 데이터 크리에이터 캠프

최우수상 (1위)

제3회 상권분석 빅데이터 경진대회

"Prizes"

2020년에 총 7번 데이터분석
공모전 수상을 하였습니다.

3위

KB 국민은행 제 2회 Future Finance Ai Challenge

우수상

포스트 코로나 데이터 시각화 경진대회 공모전

최우수상

빅데이터 활용 정책 아이디어 공모전

1. P-SAT 소개

2. 활동내역 및
수상내역

활동내역
수상내역

3. 리쿠르팅



1. P-SAT 소개

2. 활동내역 및 수상내역

활동내역
수상내역

3. 리쿠르팅

2019 IGAWorks BIG DATA Competition 입상
2019 제6회L.POINT Big Data Competition 우수상
2019 미래에셋대학생디지털금융페스티벌 대상/ 은상/ 동상
2019 디지털금융혁신과금융보안공모전 장려상
2019 KCB x 데이콘: 금융스타일시각화대회 5위
2019 기상청날씨빅데이터콘테스트 우수상
2019 제5회L.POINT Big Data Competition 대상/ 우수상(2)
2019 프로농구데이터활용경진대회 최우수상/ 우수상
2018 빅콘테스트 최우수상(신한은행)
2018 네이버Data Science Competition 3위
2018 미래에셋대우빅데이터페스티벌 대상

"Prizes"

2015-19년 5년간 총 24번
데이터분석 공모전 수상을 하였습니다.

2018 기상청날씨빅데이터콘테스트 우수상, 장려상
2018 한국수자원공사빅데이터공모전 입선
2018 성균C-School 프로젝트성과발표회 대상
2017 빅콘테스트SK텔레콤데이터사업 본부장상
2017 기상청날씨빅데이터콘테스트 우수상
2016 기상청날씨빅데이터콘테스트 최우수상
2015 SAS 분석챔피언십공모전 입선

3. 리쿠르팅

"Recruiting"

아래 세 조건을 만족하시는 분은 언제든지 지원이 가능합니다.
조건을 반드시 숙지해 주시길 바랍니다.

통계학과
원전공/
복수전공

&

1년의
의무기간

&

열정과
관심

✓ 이번학기는 학과차원의 요구로
통계학과 원전공 및 복수전공생만
지원 가능합니다.

✓ 1년간 활동이 반드시 가능한 분
✓ 금요일 저녁 6시-9시 세미나 참가 가능한 분

**온라인으로 진행할 예정이며,
팀 별로 3시부터 미리 세미나를 준비합니다.

✓ 주 2회 팀별 스터디 참가 가능한 분
**확진자 추이 및 사회적거리두기 정책 상 소규모 스터디가
가능할 경우 해화에서 오프라인 스터디를 진행합니다.

✓ 데이터분석에 관심있는 분

1. P-SAT 소개

2. 활동내역 및
수상내역

3. 리쿠르팅

지원자격 및
일정



"Recruiting" 아래 일정을 확인하시어 지원해주시길 바랍니다.

MARCH						
Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
2.21	2.22 ● 개강 1주차	2.23	2.24	2.25	2.26	2.27 면접
2.28 면접	1 결과 발표 ● 개강 2주차	2	3	4	5 ● O.T	6
7	8 ● 개강 3주차	9	10	11	12 ● 클린업 1주차	13
14	15 ● 개강 4주차	16	17	18	19 ● 클린업 2주차	20
21	22 ● 개강 5주차	23	24	25	26 ● 클린업 3주차	27
28	29 ● 개강 6주차	30	31			

지원서 접수 기간

: 2021년 2월 22일(월) -
2021년 2월 25일(목) 15시

온라인면접 기간

: 2021년 2월 27일(토) -
2021년 2월 28일(일)

** 서류합격자에 한해 개별적으로 면접 일정을
알려드릴 예정입니다.

최종 발표

: 2021년 2월 28일(일) 밤 /
2021년 3월 1일(월) 예정

1. P-SAT 소개

2. 활동내역 및
수상내역

3. 리쿠르팅

지원자격 및
일정

27th Recruiting
2021



카페 주소

: <https://cafe.naver.com/powersat>

인스타 아이디

: skku_psat2007

깃헙 주소

: <https://github.com/P-Sat>

카카오 채널 이름

: SKKU P-SAT (혹은 "피셋"으로 검색)

문의

: 학회장 권남택 010-7518-8810

The background is a light gray with various faint, white line-art icons representing business and data concepts, such as bar charts, pie charts, line graphs, and organizational charts. On the right side, there are three diagonal stripes in blue, light blue, and gold. The main text is centered in a large, bold, dark blue font.

**27기 여러분을
기다립니다.**