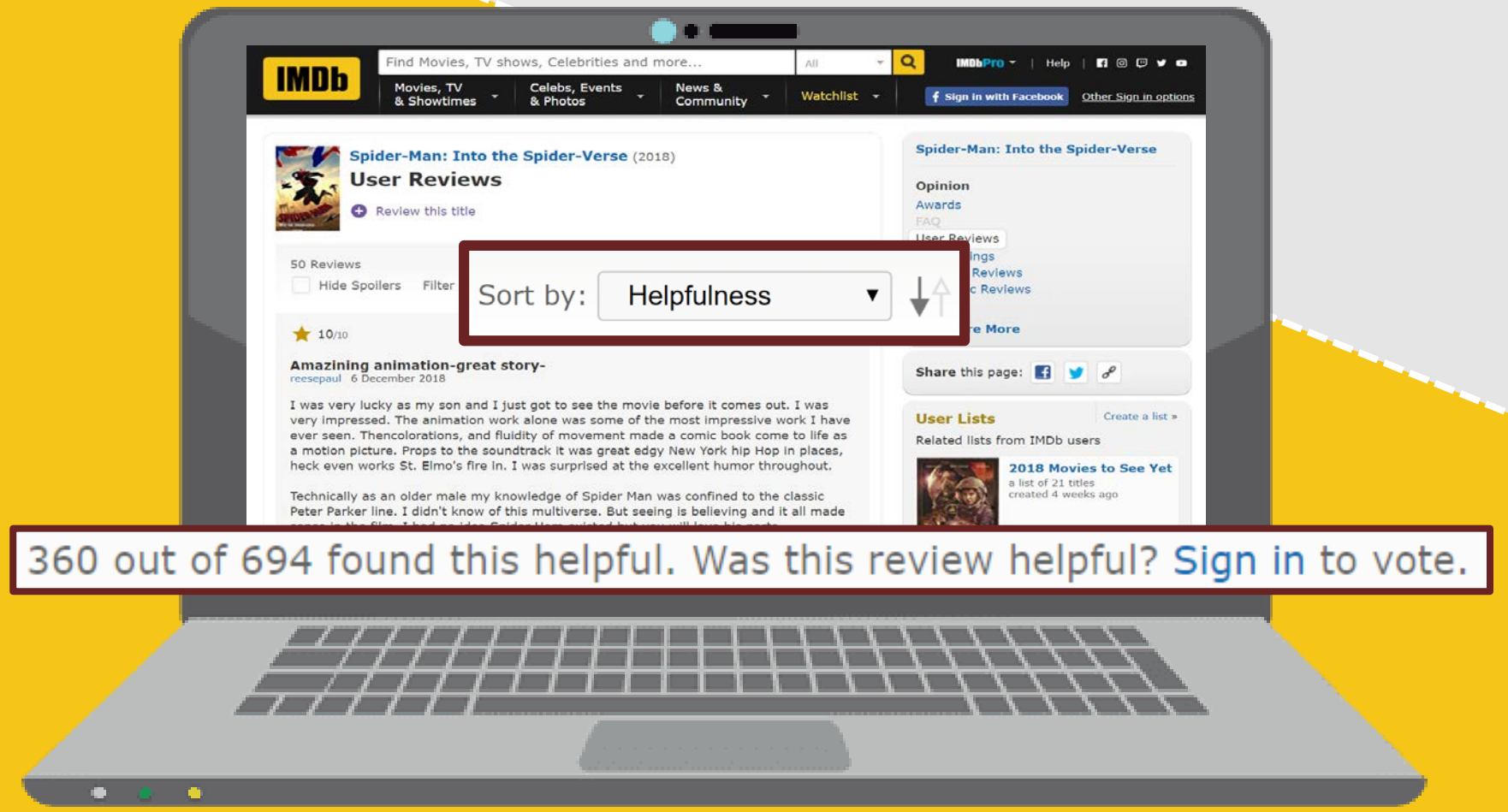

IMDb
박수영
현아연
정지원
변주연

	text	point
6	A very, very, very slow-moving, aimless movie about a distressed, drifting young man.	0
7	Not sure who was more lost - the flat characters or the audience, nearly half of whom walke...	0
8	Attempting artiness with black & white and clever camera angles, the movie disappointed - ...	0
9	Very little music or anything to speak of.	0
10	The best scene in the movie was when Gerardo is trying to find a song that keeps running t...	1
11	The rest of the movie lacks art, charm, meaning... If it's about emptiness, it works I guess bec...	0
12	Wasted two hours.	0
13	Saw the movie today and thought it was a good effort, good messages for kids.	1
14	A bit predictable.	0
15	Loved the casting of Jimmy Buffet as the science teacher.	1
16	nd those baby owls were adorable.	1
17	The movie showed a lot of Florida at it's best, made it look very appealing.	1
18	The Songs Were The Best And The Muppets Were So Hilarious.	1

IMDB_labelled.txt

What is Helpfulness?



" Helpfulness of Review "

리뷰의 유용성 지표 개발을 위하여 이에 영향을 미치는 요인 및 패턴 분석

"리뷰의 유용성"은 소비자의 입장에서는 의사결정에 도움이 되는 리뷰를 제공받을 수 있고
판매자의 입장에서는 잠재적 소비자들에게 리뷰를 제공할 수있고,
리뷰어 입장에서는 유용한 리뷰쓰기를 위한 가이드라인을 제공받을 수 있다는 데에 의의

분석 방법 소개

1. 리뷰의 특성 및 패턴 파악

2. 리뷰의 유용성에 영향을 미치는 요인 파악

3. 리뷰에 대한 유용성 지표화

리뷰 고찰

리뷰

- 리뷰 고유의 특징
- 텍스트 구조
- 서술적 의미
- 감성적 의미

영화 의 특성

리뷰어 특성

- 노출정도
- 평판

소비자 특성

리뷰 수용

데이터 고찰

관측불가능한 변수

- 리뷰어가 쓴 리뷰 총 개수
- 리뷰어의 리뷰의 평판
- 리뷰 작성 날짜
- 영화에 대한 평점

관측가능한 변수

- 영화에 대한 긍부정
- 리뷰의 표현적 측면
- 영화의 대표적 특성(3요소)에 대한 내용
- 감성적 서술을 담은 리뷰

분석 방법 구체화

- 리뷰의 유용성에 영향을 미칠 것이라고 예상되는 변수
- 우선, 그럴 것이라 가정

분석 방법 구체화

A1 평점	분류모형으로 분류해낸 긍부정의 평균X10 => 1~10점 척도	리뷰 감성 분석
A2 감정	감정 사전 내의 감성 단어가 포함된 비율	
A3 길이	리뷰에서 사용된 단어의 개수	리뷰의 표현적 측면
A4 강조	연속된 대문자의 수	
A5 설명	영화의 특성 혹은 3요소 설명 단어의 비율	리뷰의 기술적 측면

가중합 → " 유용성 지표화 "


ANALYSIS

데이터 전처리

전처리 전	전처리 후	원 텍스트	오류 텍스트
혯,헸,혚		… in all of them a true…	… in all of them 혯 a true…
챕	e	"movie business" of Québec	"movie business" of Qu챕bec
책	a	with Trond Fausa Aurv책g	with Trond Fausa Aurvåg

인코딩 오류 수정

활용 어미 단일화

복수형 단일화

특수 의미 단어 변형

before	movie	film	films	movies
after	제거			

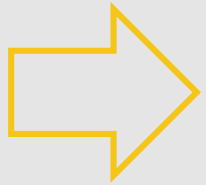
before	scenes	sucked	sucks	wasted	things	/10	out of 10	10.1	
after	scene	suck	suck	waste	thing	points	points	tenpoint	
before	makes	effects	characters	minutes	acting	acted	liked	likes	points
after	make	effect	character	minute	act	act	like	like	point

1 | 변수 정의 - A1 평점

분류모형으로 분류해낸 긍부정의 평균X10 => 1~10점 척도

감성 분석

로지스틱	TREE
나이브베이즈	BAGGING
SVM	FORESTS
GLMNET	MAXENTROPY
SLDA	LOGITBOOST



각 ML 알고리즘을
통해 분류된

텍스트의 긍정, 부정
1 0

분류모형으로 분류해낸 긍부정의 평균X10 => 1~10점 척도

감성 분석

로지스틱	TREE
나이브베이즈	BAGGING
SVM	FORESTS
GLMNET	MAXENTROPY
SLDA	LOGITBOOST

각 ML 알고리즘을
통해 분류된

텍스트의 긍정, 부정
1 0

각 모형에서 예측된 긍/부정의 평균값

평점 추정값

감성 단어 비율

데이터 전처리

- 소문자 변환
- 스테밍한 단어 이용

감성사전 단어

- AFINN 사전 이용
- 스테밍한 단어 이용

감성 사전 내 감성 단어 가 포함된 비율

리뷰가 얼마나 감성을 절제하여 자신의 느낌을 표현하는지

감성 단어 비율 예시

2407	waste	-1
2408	wasted	-2
2409	wasting	-2

Afinn

+

wasted two hours.

text

	txt	sent_cnt
1	a very, very, very slow-moving, aimless about a distressed, drifting young man.	0.153846153846154
2	not sure who was more lost - the flat character or the audience, nearly half of whom ...	0.210526315789474
3	attempting artiness with black & white and clever camera angles, the disappointed - ...	0.193548387096774
4	very little music or anything to speak of.	0
5	the best scene in the was when gerardo is trying to find a song that keeps running t...	0.0952380952380952
6	the rest of the lacks art, charm, meaning... if it's about emptiness, it works i guess bec...	0.4
7	wasted two hours.	0.333333333333333
8	saw the today and thought it was a good effort, good messages for kids.	0.333333333333333
9	a bit predictable.	0
10	loved the casting of jimmy buffet as the science teacher.	0.2

문자의 특성

단어의 개수

리뷰에서 사용된 단어의 수

강조 단어

리뷰에서 연속된 대문자의
사용으로 강조된 단어의 수

리뷰의 표현적 측면

리뷰에서 자신의 의견을 어떻게 표현하고 있는지

문자의 특성 예시

단어의 개수

연속 대문자 수

	V1	word_cnt	power_cnt
94	There was NOTHING believable about it at all.	8	1
95	The only suspense I was feeling was the frustration at ju...	16	0
96	MANNA FROM HEAVEN is a terrific that is both predictab...	17	3

리뷰에서 자신의 의견을 어떻게 표현하고 있는지

영화의 특성 혹은 3요소 설명 단어의 비율

지난주 토픽분석 결과

imdb_ctm			imdb_lda		
Topic.1	Topic.2	Topic.3	Topic.1	Topic.2	Topic.3
bad	good	like	good	bad	just
just	great	one	character	one	scene
script	character	really	really	time	love
time	wonderful	character	just	like	point
waste	make	act	act	best	suck
one	one	see	story	even	everything
can	watching	good	great	don	still

1 | 변수 정의 - A5 설명

영화의 특성 혹은 3요소 설명 단어의 비율

영화의 3요소

sound

- Melody
- Music
- Accent
- Noise
- …

composition

- Plot
- Script
- Story
- Scene
- …

play

- Act
- Perform
- Character
- Extra
- …

The act was decidely wooden,

FACTOR 사전 구축

1 | 변수 정의 - A5 설명

영화의 특성 혹은 3요소 설명 단어의 비율

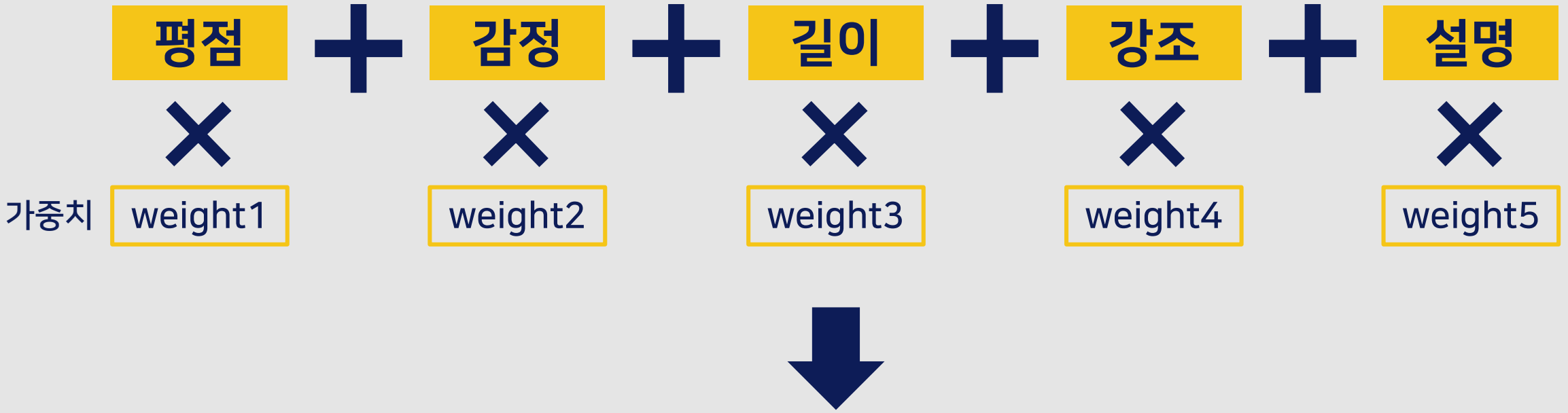
영화의 3요소

sound + composition + play

각 요소를 설명하는 단어의 비율

리뷰가 얼마나 각 영화의 특성에 대하여 설명하고 있는지

리뷰 유용성의 지표화


평점
감정
길이
강조
설명
가중치
weight1
weight2
weight3
weight4
weight5
리뷰의 유용성 지표

실제 데이터 크롤링

all.review	all.point	helpful_per
Yes, Only Lovers Left Alive is another vampire movie.	10	75.7377
Wonderful imagery. style and atmosphere in the extr	6	69.44444
One knows that a Jim Jarmusch movie about vampir	8	66.06335
This is Jim at his utter best. The balance between em	10	65.44715
Only Lovers Left Alive is one of the most breath-taki	10	67.44186
Jim Jarmusch's delicious new comedy is a vampire m	9	62.99213
ONLY LOVERS LEFT ALIVE pulls no punches with its a	8	58.75
In the abandoned Detroit, the depressed musician A	8	56.66667
It's pretty easy to fancy this film as cool and arty and	2	49.41176
This is a story how sid & nancy wanted to be va	5	50.9434
This film was a drag, very slow and quiet. The music	1	50.9434
This is the first Jarmusch film I've seen, so did not kn	1	50
Adam, the main character, is a perfect example: A ce	3	56
I am all vampired out. Hate to admit it, but it's true. A	5	52.77778
I don't know how I managed to watched the whole r	2	52.94118
The origin story of The Mighty "Thor" is revealed of I	7	95.45455
The idea of a "classic" director like Kenneth Branagh	8	64.60177

- 807개의 데이터
- 평점
- 유용성
  사람들로부터 얼마나 많은 유용하다 평가를 받았는지(%)

실제 데이터에 모형 적합

	text	point	helpful	word_cnt	factor_cnt	sent_cnt	power_cnt
1	Yes, Only Lovers Left Alive is another vampire . Yes, the charac...	10	0.7573770	213	5	48	1
2	Wonderful imagery. style and atmosphere in the extreme. gre...	6	0.6944444	150	9	33	0
3	This is Jim at his utter best. The balance between emotive writ...	10	0.6544715	179	7	42	0
4	Only Lovers Left Alive is one of the most breath-taking s I ha...	10	0.6744186	369	13	81	0
5	Jim Jarmusch's delicious new comedy is a vampire unlike any ...	9	0.6299213	270	4	55	1
6	ONLY LOVERS LEFT ALIVE pulls no punches with its audience; ...	8	0.5875000	245	6	63	4
7	In the abandoned Detroit, the depressed musician Adam (Tom...	8	0.5666667	281	9	58	0
8	It's pretty easy to fancy this as cool and arty and whatnot, bu...	2	0.4941176	517	15	107	3
9	This was a drag, very slow and quiet. The music featured is gr...	1	0.5094340	126	4	27	0
10	This is the first Jarmusch I've seen, so did not know what to e...	1	0.5000000	131	2	25	0
11	The idea of a "classic" director like Kenneth Branagh making a ...	8	0.6460177	503	31	121	0
12	In the pantheon of Marvel Superheroes, from my vantage poi...	8	0.6776860	971	43	190	0
13	Straight to the point, this was not only one of the better ones,...	9	0.5233333	129	8	37	0
14	Okay, so Thor is pretty darn decent. Phew. Without revealing ...	9	0.5254902	216	8	58	0

실제 데이터에 모형 적합

```
> summary(regression)

Call:
lm(formula = helpful ~ point + word_cnt + sent_cnt + factor_cnt +
    power_cnt)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33527 -0.07345 -0.00605  0.06574  0.28206

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.964e-01  1.157e-02  42.885   <2e-16 ***
point        1.958e-02  1.294e-03  15.133   <2e-16 ***
word_cnt     1.599e-04  8.668e-05   1.845   0.0656 .
sent_cnt    -3.501e-04  3.930e-04  -0.891   0.3733
factor_cnt  -9.983e-04  7.364e-04  -1.356   0.1757
power_cnt    4.669e-04  1.178e-03   0.396   0.6919
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1025 on 628 degrees of freedom
Multiple R-squared:  0.2834,    Adjusted R-squared:  0.2777
F-statistic: 49.66 on 5 and 628 DF,  p-value: < 2.2e-16
```

```
> summary(stp_reg)

Call:
lm(formula = helpful ~ point + word_cnt + factor_cnt)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33359 -0.07318 -0.00706  0.06425  0.28280

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.964e-01  1.144e-02  43.398   <2e-16 ***
point        1.947e-02  1.273e-03  15.292   <2e-16 ***
word_cnt     9.497e-05  3.688e-05   2.575   0.0102 *
factor_cnt  -1.095e-03  7.235e-04  -1.514   0.1305
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1024 on 630 degrees of freedom
Multiple R-squared:  0.2823,    Adjusted R-squared:  0.2789
F-statistic:  82.6 on 3 and 630 DF,  p-value: < 2.2e-16
```

$$helpful = 0.496 + 0.02 * point + 0.001 * word_{cnt} - 0.001 * factor_{cnt}$$

추정된 모형에 적합 후 유용성 지표화

	text	point	avrg	sent_cnt	word_cnt	power_cnt	factor_cnt	helpful.x
6	A very, very, very slow-moving, aimless about a distresse...	0	5.00	0.15384615	13	0	0	0.5190
7	Not sure who was more lost - the flat character or the a...	0	3.75	0.21052632	19	0	2	0.5205
8	Attempting artiness with black & white and clever camer...	0	1.25	0.16129032	31	0	2	0.5275
9	Very little music or anything to speak of.	0	6.25	0.00000000	8	0	1	0.5155
10	The best scene in the was when Gerardo is trying to fin...	1	10.00	0.09523810	21	0	1	0.5360
11	The rest of the lacks art, charm, meaning... If it's about e...	0	8.75	0.35000000	20	0	0	0.5335
12	Wasted two hours.	0	7.50	0.33333333	3	0	0	0.5140
13	Saw the today and thought it was a good effort, good ...	1	10.00	0.33333333	15	0	0	0.5310
14	A bit predictable.	0	8.75	0.00000000	3	0	0	0.5165
15	Loved the casting of Jimmy Buffet as the science teacher.	1	10.00	0.20000000	10	0	0	0.5260

```
> summary(model$helpful)
       x
 Min.   :0.4970
 1st Qu.:0.5100
 Median :0.5185
 Mean   :0.5190
 3rd Qu.:0.5260
 Max.   :0.5825
```

$$helpful = 0.496 + 0.02 * point + 0.001 * word_{cnt} - 0.001 * factor_{cnt}$$

분석의 한계 및 제언점

리뷰의 유용성에 대한 기준의 모호함

- 리뷰의 유용성의 기준 모호
- 리뷰의 유용성에 영향을 미치는 변수가 작위적

SMALL DATA

- 데이터 작으며 따라서 요소에 대한 언급 빈도가 낮음
- 위에서 scale이 다른 변수들 간의 적절하지 않은 모형 적합

LARGE DATA

- IMDB의 축적된 방대한 양의 데이터와 측정 가능한 변수들로 알고리즘을 적용시키면 더 유의한 결과물이 나타날 것이다.

TITLE

IMDb 리뷰의 유용성 분석

Production

7 조

Director

PARK SU YEONG
Department of Statistics

BYUN JU YEON
Department of Public Administration

HYUN AH YEON
Department of Computer Science

JEONG JI WON
Department of Economics