Group 2

Progress Report

01 Abusing classification model

Sampling

Sampling

어뷰징 기사 : 400개

정상적 기사 : 1100개

01 Abusing classification model
preprocessing

Category Integration

정치, 사회, 사설, 연예, 문화, 국제 등 15개의 카테고리로 통합

경향신문은 지나치게 세분화 -> 상위 카테고리로 통합

Ex)
-건강·의학 (사회), 라이프 (사회), 전국(사회)
-과학·환경 (과학), 테크(과학)
-마켓·비즈 (경제), 부동산 (경제)
-트래블(문화)

01 Abusing classification model

preprocessing

Target variable

어뷰징 : 0, 정상 : 1

Word extraction

본문, 제목에서 고유명사, 보통명사, 동사, 형용사를 추출

▼

제목, 본문에 사용된 단어의 수, 제목의 단어가 본문에 사용된 횟수 계산

01 Abusing classification model

preprocessing

01 Abusing classification model
Modeling

Result

Criteria	LDA	Decision Tree
Accuracy	0.8297	0.8366
TPR	0.9130	0.8750
TNR	0.5682	0.7015
Precision	0.8689	0.9116
F1 - measure	0.8905	0.8929

02 Politics & Entertainment
method

Association Rule

02 Politics & Entertainment
method

Step. 1

문건 분석을 통해 이슈, 특종 사건 추출

Step. 2

규칙 발견

Time	정치	사회	연예	....
Time 01	1	1	0	...
Time 02	0	0	0	...
Time 03	0	1	0	...
Time 04	1	0	0	...
...	...	...	...	...

02 Politics & Entertainment
method

Problem

수십 만개의 문건을 직접 구분하는 것은 불가능 : 분류모델을 통해 구분

▼

이슈, 특종 사건의 구분 기준은?

Solution

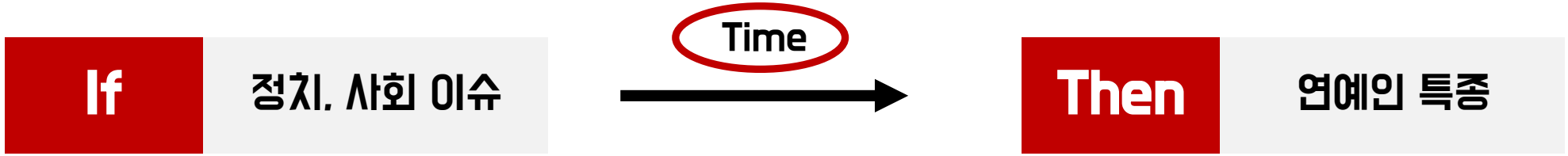
이슈, 특종이 발생했음을 간접적으로 암시하는 다른 변수 고려

1. 조회수
2. 기사 수

02 Politics & Entertainment
method

Association Rule

Project



02 Politics & Entertainment
Preprocessing

1. Aggregation

Daily ▶ Weekly

2 Data Reduction

"과학", "만화", "ESC" 등 1%미만 범주 삭제

▼

"정치", "사회", "국제", "사설", "연예" 등 9개 카테고리

02 Politics & Entertainment
Preprocessing

Data set

Time	C1	C2	…	C(i)
Time 1	A[1,1]	A[1,2]	…	A[1,i]
Time 2	A[2,1]	A[2,2]	…	A[2,i]
…	…	…	…	…
Time (j)	A[j1,]	A[j,2]	…	A[j,i]

3. Define Variable

	Diff(-1)			Diff(+1)		
Time	C1(-1)	C2(-1)	Ci(-1)	C1(+1)	C2(+1)	Ci(+1)
Time 1						
Time 2						
Time j			**B[j,i]**			**B[j,2i]**

1. B[j,i] = A[j,i] – A[j-1,i]
2. B[]j,2i] = A[j+1,2i] – A[j,2i]

4. Categorization

B[j,i] > 0 ▶ "U"

B[j,i] = 0 ▶ "S"

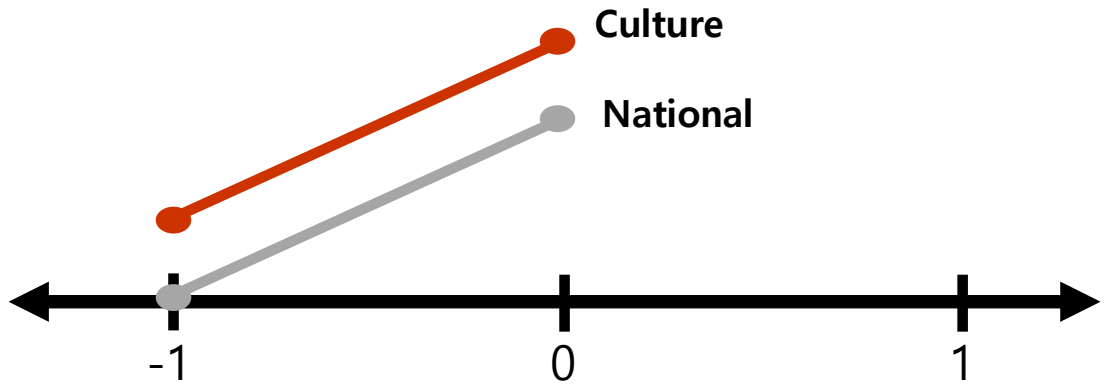
B[j,i] < 0 ▶ "D"

Ex) Ent(-1) = "U", Culture(+1) = "D"

02 Politics & Entertainment
Preprocessing

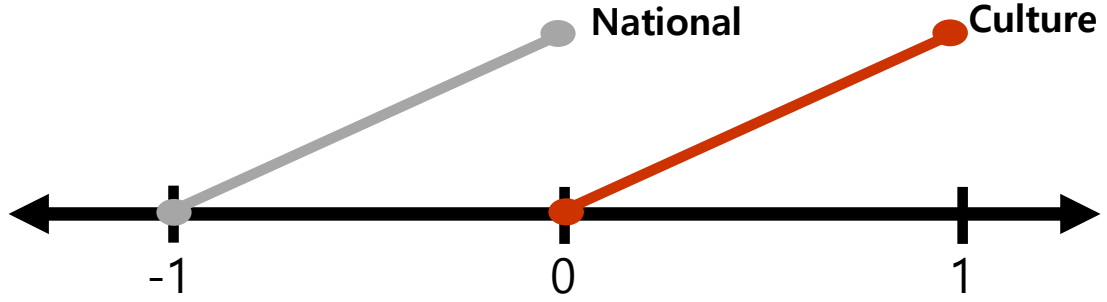
Ex 1)

If) National(-1)="U" ▶ Then) Culture(-1)="U"



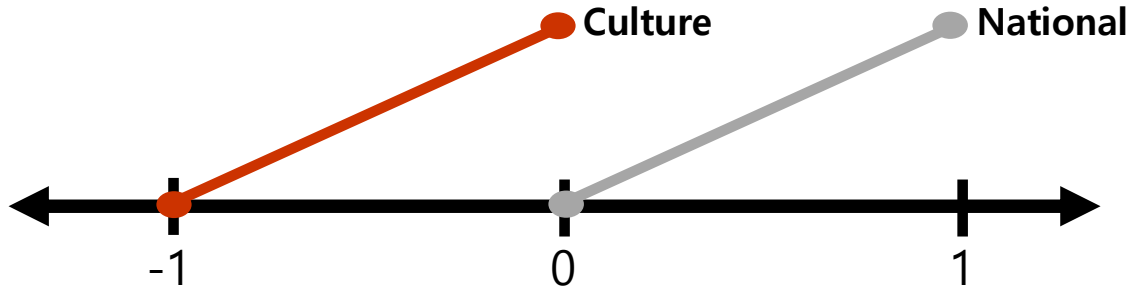
Ex 2)

If) National(-1)="U" ▶ Then) Culture(+1)="U"



Ex 3)

If) National(+1)="U" ▶ Then) Culture(-1)="U"



02 Politics & Entertainment
Preprocessing

Problem

변화량의 차이를 반영하지 못함

Time	Culture	Ent
Time 04	100	100
Time 05	120	180
Δ	+20	+80

Solution

1. 변화량에 따라 구간을 세분화
2. 선규칙 후검증

02 Politics & Entertainment

Rule discovery

조건) Support ≥ 0.1 Confidence ≥ 0.8

Lhs	Rhs	Support	Confidence	Lift
Economy(-1) = D Ent(-1) = D Editorials(+1) = U Politics(+1) = U	Ent(+1) = U	0.1007	0.9375	2.149
Economy(-1) = D Ent(-1) = D Editorials(-1) = D Politics(+1) = U	Ent(+1) = U	0.1007	0.9375	2.149
Culture(-1) = D Ent(-1) = D Editorials(-1) = D Politics(+1) = U	Ent(+1) = U	0.1007	0.9375	2.149
Editorials(-1) = D Ent(-1) = D politics(-1) = D Politics(+1) = U	Ent(+1) = U	0.121	0.9	2.063

02 Politics & Entertainment

Rule discovery

조건) Support ≥ 0.1 Confidence ≥ 0.8

Lhs	Rhs	Support	Confidence	Lift
Economy(-1) = D Ent(-1) = D Editorials(+1) = U Politics(+1) = U	Ent(+1) = U	0.1007	0.9375	2.149
Economy(-1) = D Ent(-1) = D Editorials(-1) = D Politics(+1) = U	Ent(+1) = U	0.1007	0.9375	2.149
Culture(-1) = D Ent(-1) = D Editorials(-1) = D Politics(+1) = U	Ent(+1) = U	0.1007	0.9375	2.149
Editorials(-1) = D Ent(-1) = D politics(-1) = D Politics(+1) = U	Ent(+1) = U	0.121	0.9	2.063

02 Politics & Entertainment

Rule discovery

1. Normality check

Anderson Darling test -> Not normal

2. Rank sum test

$$H_0: \Delta\mu_{Rule} = \Delta\mu_{Others}$$
$$H_1: \Delta\mu_{Rule} > \Delta\mu_{Others}$$
$$\alpha = 0.1$$

	Rule 1	Rule 2	Rule 3	Rule 4
Ent	0.0122	0.0127	0.0249	0.018
Politics	0.0091	0.0299	0.0251	0.2024
Editorials	0.0032	NA	NA	NA

02 Politics & Entertainment
Conclusion

1

조건절이 발생하면 매우 높은 확률로 연예 기사 수가 상승

2

모든 규칙에 공통적으로 연예, 사설, 정치 카테고리가 포함

3

규칙에 해당하는 시점의 평균 증가량은 일반적인 평균 증가량보다 크다