DACA

NGO 머신러닝 : 4주차

2021.1.15 ~ 2021.1.20

빅데이터 응용학과 석사 1기 양선욱

목차

0. 다항분포 나이브 베이즈 (수정)

* winemag-data

전처리 후 데이터 수 : 117,228

와인 지방 카테고리 수 : 48개

- 와인은 생산되는 지방에 맛을 특징이 다르다.
- 와인을 시음한 사람들의 설명이 지방의 특징을 설명하고 있다면 이것만으로 와인 생산 지방을 분류할 수 있다고 예측하였다.
- 와인 지방 390여 곳 중에 데이터의 수가 100이하인 곳을 제거하고 남은 곳이 48곳

7:3 분할 후 traindata에 언더샘플링 진행

train data : 6,288
test data : 35,169
와인 지방 카테고리 수 : 48개
train data의 카테고리 당 131개의 데이터

원본 데이터에 언더샘플링 진행 후 7:3 분할

train data : 6,249
test data : 2,679
와인 지방 카테고리 수 : 48개
train data의 카테고리 당 147~121개

* 모델평가&분석 결과

- 학습용 데이터 세트 정확도 : 0.471
- 평가용 데이터 세트 정확도 : 0.455
- 정밀도 평균 : 0.28
- 재현률 평균 : 0.07
- F1 스코어 평균 : 0.08

- 학습용 데이터 세트 정확도 : 0.835
- 평가용 데이터 세트 정확도 : 0.356
- 정밀도 평균 : 0.306
- 재현률 평균 : 0.453
- F1 스코어 평균 : 0.312

- 학습용 데이터 세트 정확도 : 0.800
- 평가용 데이터 세트 정확도 : 0.411
- 정밀도 평균 : 0.529
- 재현률 평균 : 0.421
- F1 스코어 평균 : 0.406

* 사용 데이터 설명

출처 : 월드비전 홈페이지

후원 구분	후원 종류	후원 설명
아동	해외아동후원	
	국내아동후원	
사업	해외사업후원	해외 취약아동이 사는 마을의 교육, 식수위생, 보건영양, 소득증대, 아동보호를 위한 통합적인 사업을 합니다.
	국내사업후원	국내 취약계층 아동의 영양지원, 위기지원, 꿈 지원을 위한 사업을 합니다.
	긴급구호사업후원	자연재해, 분쟁, 식량위기로 고통 받는 아이들을 신속하고 전문적으로 지원합니다.
	북한사업후원	북한 어린이들을 위한 영양지원, 농업개발, 식수지원 사업을 합니다.
	전체사업후원	월드비전의 모든 활동을 지속적이고 포괄적으로 지원합니다.

* 데이터 인구 정보

- 각 후원의 플릿지를 1개 이상 체결한 고객의 수
- 전체 1000명, 비이탈 800명, 이탈 200명

단위 : 명

후원 종류	비이탈	이탈	전체 고객
해외아동후원	477	114	591
국내아동후원	66	16	82
해외사업후원	59	15	74
국내사업후원	63	9	72
긴급구호사업후원	113	29	142
북한사업후원	9	1	10
전체사업후원	132	45	177

1. 연관규칙 분석

* 연관규칙 분석 계획

- 이탈여부 x 성별 x 연령대 x 7가지 후원 플릿지 여부

전처리 후 데이터 수 : 890

사용 모형 : Apriori

min_support = 0.05

* 연관규칙 분석 결과

no	antecedents	censequents	support	confidence	lift	개수
1	전체사업 제외 6가지 플릿지 비후원	전체사업 후원	0.124	1	6.40	143
2	긴급구호사업 후원	해외아동, 국내아동, 해외사업, 북한사업 비후원	0.05	0.70	3.96	50
3	해외아동, 국내아동, 해외사업, 긴급구호사업, 북한사업 비후원	국내사업 후원	0.1	0.37	2.55	92
4	해외아동 후원	나머지 6가지 비후원	0.51	0.84	1.63	510

* '전체사업 플릿지 ' 의 영향이 너무 커서 no2 결과 부터는 '전체사업 플릿지'를 제외한 연관규칙 분석의 lift순이다

- 이탈여부x플릿지, 성별x플릿지, 연령대x플릿지는 모두 lift 값이 0.999~1.001 사이의 값으로 연관성이 없다고 판단된다.

1. 연관규칙 분석

* 연관규칙 분석 계획 2

- 두 종류의 플릿지를 후원 중인 고객의 연관분석

전처리 후 데이터 수 : 92

사용 모형 : Apriori

min_support = 0.05

* 연관규칙 분석 결과

no	antecedents	censequents	support	confidence	lift	개수
1	국내아동후원	해외아동후원	0.20	1	1.29	19
2	해외사업후원	해외아동후원	0.09	1	1.29	9
3	긴급구호사업후원	해외아동후원	0.09	0.9	1.16	9
4	국내사업후원	해외아동후원	0.18	0.65	0.84	17
5	전체사업후원	해외아동후원	0.16	0.62	0.80	15

* 결과 분석

국내아동만 후원 중인 비이탈 회원 : 33명

해외사업만 후원 중인 비이탈 회원 : 38명

협업 필터링 문제 정의

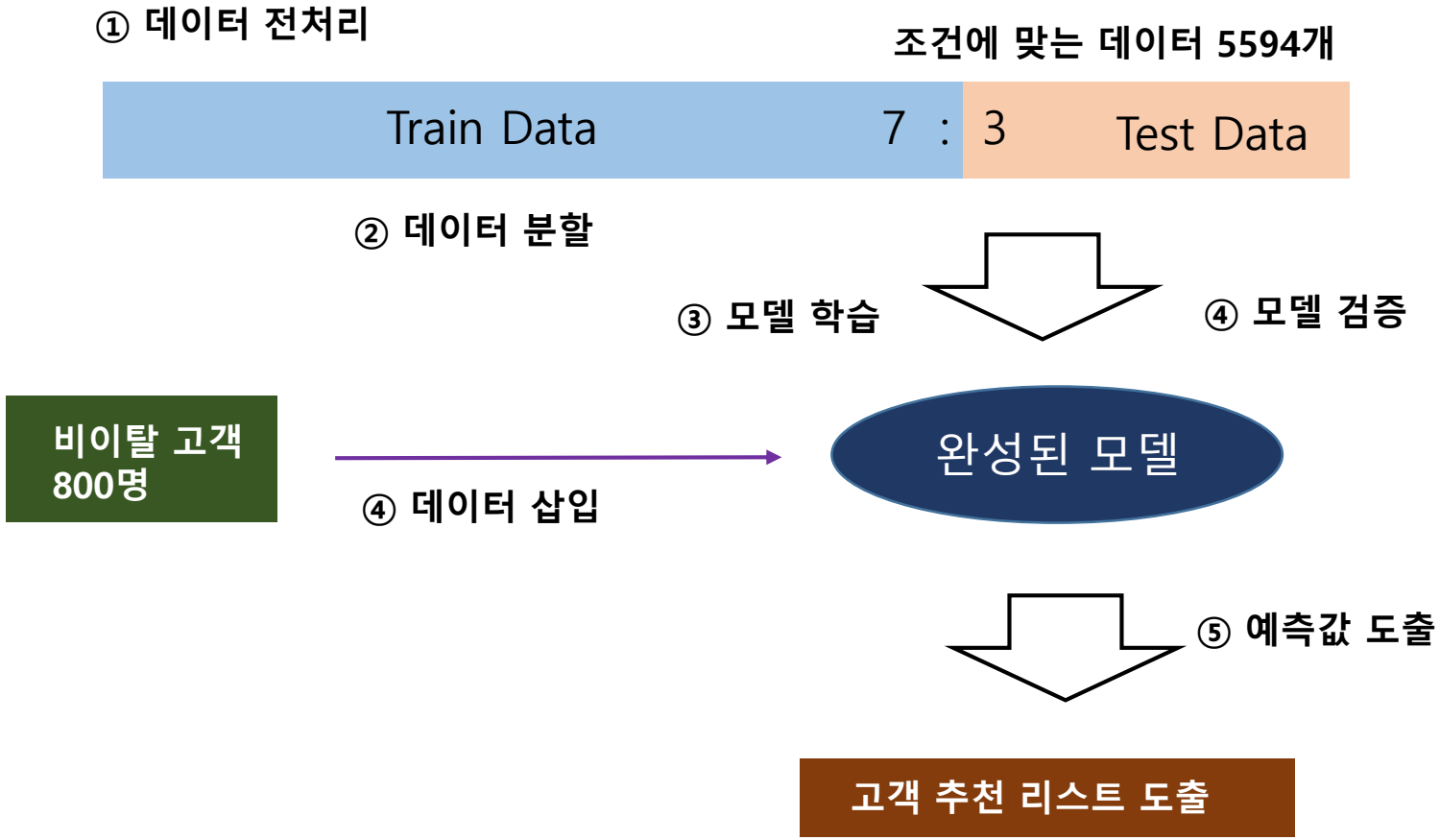
* 분석 목적

- 비이탈 고객들이 납부한 후원 종류와 플릿지 수를 바탕으로 아직 후원 하지 않은 후원의 플릿지 수를 예측한다.
- 예측된 후원의 플릿지 수가 예측 전 보다 높을 경우 해당 고객에게 추천한다.

* 사용 데이터

user : 고객 고유 ID
item : 후원 종류
rating : 후원 종류의 플릿지 수
전처리 후 : n=5594
고객 수 : 800명

* 분석 과정



협업 필터링 전처리

* 데이터 전처리

origin (800x8)

CONTACT	해외아동	국내아동	해외사업	국내사업	긴급구호사	북한사업	전체사업
DDB09260	1	0	0	0	1	0	0
F17E3A62-	1	0	0	0	0	0	0
06E8BD1D	1	0	0	0	0	0	0
578ED6D9	0	0	0	0	0	0	0
076AA8D3	1	0	0	0	0	0	0
993C2C74	0	0	0	0	0	0	1
761E2CD3	1	0	0	0	0	0	0
1204ABB8	1	0	0	0	0	0	0
6A3F867A	1	0	0	0	0	0	0
2D664C4B	1	0	0	0	0	0	0
E47EA78F-	0	0	0	0	1	0	0
F673FA91-	1	0	0	0	0	0	0
2D99A0C3	1	0	0	0	0	0	0
9E8F2EEC-	3	0	0	0	0	0	0
8789D0C6	1	0	0	0	0	0	0
734AD57A	0	0	0	0	0	0	0
3D7969E8	0	1	1	1	0	1	0
BAF1BBEA	0	0	0	0	0	0	1
69B496DB	0	0	0	0	0	0	1
A339423F-	0	0	0	0	0	0	1
C77B0B07	2	0	0	0	0	0	0

고객 ID	후원종류A 플릿지 수	후원종류B 플릿지 수	후원종류C 플릿지 수	후원종류D 플릿지 수
01	a	b	0	d

data (5600x3)

uid	후원종류	플릿지수
DDB09260	해외아동	1
DDB09260	긴급구호사업	1
F17E3A62-	해외아동	1
06E8BD1D	해외아동	1
076AA8D3	해외아동	1
993C2C74	전체사업	1
761E2CD3	해외아동	1
1204ABB8	해외아동	1
6A3F867A	해외아동	1
2D664C4B	해외아동	1
E47EA78F-	긴급구호사업	1

고객ID	후원종류	납부금액
01	A	a
01	B	b
01	C	0
01	D	d

```
reader = Reader(rating_scale=(0, 5))
data = Dataset.load_from_df(data
[['uid', '후원종류', '플릿지수']], reader)
```

2. 사용자 기반 협업필터링

* 최적의 모델 찾기

유사도 평가 = cosine, msd

k : 30 ~ 50

min_k : 1 ~ 3

name	k	min_k	rmse
cosine	50	1	0.472189
msd	50	1	0.472189
cosine	50	2	0.472189
msd	50	2	0.472189
cosine	50	3	0.472189
msd	50	3	0.472189
cosine	49	1	0.472322
msd	49	1	0.472322
cosine	49	2	0.472322

* 결정된 모델

유사도 평가 = cosine

k : 50

min_k : 1

RMSE : 0.4722

* 분석 결과

- 기존플릿지 수는 0이고 예측플릿지 수가 1에 가깝게 예측된 고객 12명

uid	후원종류	기존플릿지수	예측플릿지수
4F2531CA-B62A-4065-946F-30738D6CBBE8	해외아동	0	0.94
6825AB8F-9E03-4912-AB67-B9301567D43D	해외아동	0	0.94
ABA28F06-82EC-4DE3-ACD8-06A61FD97299	해외아동	0	0.94
9576F31D-F262-4132-A554-9BBD4BC5B77F	해외아동	0	0.94
96D7B587-77B3-450F-8D43-F1E400F9F9DC	해외아동	0	0.94
CD819167-55D4-4137-9785-D04238F12BB3	해외아동	0	0.9
0AB5A6B8-E52A-4E46-BD92-B5ACA87466D8	해외아동	0	0.9
43783A7B-2573-4F25-B6A6-0AD9589848EA	해외아동	0	0.94
88005CE9-7CC7-450A-ADBE-863425F5D611	해외아동	0	0.9
A064D6F9-22ED-48B9-83D9-65242E8875C4	해외아동	0	0.94
506A3A2E-0F0D-4EC2-BDDA-DD5C25F5680B	해외아동	0	0.94
2A58DC42-07BE-4D39-A4D9-B4B3F699164A	해외아동	0	0.94

3. 아이템 기반 협업필터링

* 최적의 모델 찾기

유사도 평가 = cosine, msd

k : 30 ~ 50

min_k : 1 ~ 3

name	k	min_k	rmse
cosine	50	1	0.472189
msd	50	1	0.472189
cosine	50	2	0.472189
msd	50	2	0.472189
cosine	50	3	0.472189
msd	50	3	0.472189
cosine	49	1	0.472322
msd	49	1	0.472322
cosine	49	2	0.472322

* 결정된 모델

유사도 평가 = cosine

k : 50

min_k : 1

RMSE : 0.4722

* 분석 결과

- 기존플릿지 수는 0이고 예측플릿지 수가 1에 가깝게 예측된 고객 12명

uid	후원종류	기존플릿지수	예측플릿지수
4F2531CA-B62A-4065-946F-30738D6CBBE8	해외아동	0	0.94
6825AB8F-9E03-4912-AB67-B9301567D43D	해외아동	0	0.94
ABA28F06-82EC-4DE3-ACD8-06A61FD97299	해외아동	0	0.94
9576F31D-F262-4132-A554-9BBD4BC5B77F	해외아동	0	0.94
96D7B587-77B3-450F-8D43-F1E400F9F9DC	해외아동	0	0.94
43783A7B-2573-4F25-B6A6-0AD9589848EA	해외아동	0	0.94
A064D6F9-22ED-48B9-83D9-65242E8875C4	해외아동	0	0.94
506A3A2E-0F0D-4EC2-BDDA-DD5C25F5680B	해외아동	0	0.94
2A58DC42-07BE-4D39-A4D9-B4B3F699164A	해외아동	0	0.94
CD819167-55D4-4137-9785-D04238F12BB3	해외아동	0	0.9
0AB5A6B8-E52A-4E46-BD92-B5ACA87466D8	해외아동	0	0.9
88005CE9-7CC7-450A-ADBE-863425F5D611	해외아동	0	0.9

4. 잠재요인 모델 기반 협업필터링

* 최적의 모델 찾기

n_factors : 60 ~ 120 / 1간격

n_factor	rmse
119	0.468701
107	0.469597
118	0.470102
108	0.470143
89	0.470488
74	0.47173
87	0.471743
64	0.471867
106	0.471963
113	0.472238
120	0.472452
91	0.472533
110	0.472988

* 결정된 모델

n_factors : 119

RMSE : 0.4687

* 분석 결과

- 기존플릿지 수는 0이고 예측플릿지 수가 1에 가깝게 예측된 고객 12명

uid	후원종류	기존플릿지수	예측플릿지수
E62818E7-2B01-4853-BDAB-5C9FB19BBB4F	해외아동	0	1.03445545
65F836E4-54EE-44C2-9E30-54F76C49D1AD	해외아동	0	1.021401412
6E2F2D86-E11F-417F-A80E-21D887C32403	해외아동	0	0.993955253
7D9217AB-88DF-4407-9195-28B39D33AC16	해외아동	0	0.986628502
3FF42989-B67C-4D8E-9B2C-0528BE81A363	해외아동	0	0.986593376
61CDE553-66AA-4CE0-BE66-8ED0D0AB8144	해외아동	0	0.933379748
915A9D23-7F55-4D7B-93D1-BD5E38C7C82A	해외아동	0	0.90768758
E7D8D02D-8DE8-4242-8EA7-B0C733E9BAD2	해외아동	0	0.90725288
D27E6524-B53D-4344-B50C-28B55B6786C2	해외아동	0	0.898519197
EE72DDB1-E328-4C3D-8C5B-8CFA3DC41A1A	해외아동	0	0.895220213
C4D9C18F-F4E2-4C58-BA83-3829C7F13555	해외아동	0	0.894115447
69D4914C-5FF5-494F-ADB1-C3C4DEBC24D1	해외아동	0	0.892050026

협업필터링 모델 비교

* 모델 품질 비교

	RMSE
사용자	0.4722
아이템	0.4722
잠재요인	0.4687

* 분석 결과 비교

- 기존플릿지 수는 0이고 예측플릿지 수가 1에 가깝게 예측된 고객 12명

	사용자 기반		아이템 기반		잠재요인 모델 기반	
no	고객index	예측플릿지수	고객index	예측플릿지수	고객index	예측플릿지수
1	638	0.94	638	0.94	238	1.03
2	74	0.94	74	0.94	306	1.02
3	392	0.94	392	0.94	236	0.99
4	285	0.94	285	0.94	446	0.98
5	455	0.94	455	0.94	42	0.98
6	789	0.94	789	0.94	302	0.93
7	517	0.94	517	0.94	770	0.90
8	140	0.94	140	0.94	232	0.90
9	776	0.94	776	0.94	294	0.89
10	692	0.9	692	0.9	476	0.89
11	328	0.9	328	0.9	311	0.89
12	399	0.9	399	0.9	107	0.89

* 데이터 수집

사이트 : 트위터

검색어 : '월드비전' 포함

해시태그 : 미포함

댓글 : 제외

링크 : 제외

기간 : 2020.01.01 ~ 2020.12.31

* 수집 목적

- 젊은 여성층의 수요가 많은 트위터에서 2020년 월드비전의 모습을 알아본다.

주요 사건 :

1) 5월 정의연 기부금 횡령 논란

2) 11월 월드비전 70주년 콘서트

3) 코로나-19 상황

* 데이터 분포

총 : 264

date	content
2020-03-27	그동안 우리나라에 도움 받은 국가(아프리카와
2020-09-25	오늘은 너무 힘든 날입니다 3년동안 월드비전
2020-12-03	월드ㅂㅣ전 섬세해ㅠㅠ
2020-06-08	'정의연 지지성명' 330개 단체 공동명의도 거
2020-09-30	"성관계하면 일자리 줄게"..콩고서 WHO 직원
2020-07-10	하성운, MBC-월드비전 'World is ONE' 참가...
2020-12-03	월드비전 고맙습니다
2020-08-07	worldvisionkorea [월드비전X김세정] 가정 밖
2020-03-27	세계 각 국에서 마음을 담아 선물한 마스크는
2020-12-21	" B.I & IOK COMPANY Donated goods worth
2020-12-18	엄유민법 월드비전

* 분석 과정

1) 2020년 전체에 대한 분석

2) 트윗 수가 가장 많은 12월에 대한 분석

3) 70주년 콘서트가 있었던 11월에 대한 분석

4) 정의연 사건이 있었던 5월에 대한 분석

월	트윗 수
1	2
2	5
3	19
4	14
5	24
6	23
7	24
8	22
9	12
10	14
11	23
12	82

6. 빈도 분석

* 2020년에 대한 빈도 분석 - 264

단어	빈도
월드비전	257
후원	94
기부	38
생각	28
아이	24
단체	23
광고	21

```
stopwords = ['그냥','정말','진짜']
```

* 12월에 대한 빈도 분석 - 82

단어	빈도
월드비전	82
기부	15
후원	14
비아이	6
생각	6

* 12월에 등장한 비아이는?

B.I (김한빈) 가수
출생 1996년 10월 22일
수상 2018년 제10회 멜론뮤직어워드 송라이터상
사이트 인스타그램, 트위터

비아이-IOK컴퍼니, 월드비전에 기부물품 전달

<서울 여의도 월드비전 사옥에서 비아이(왼쪽)와 월드비전 한상호 본부장이 기부물품 전달식 후 기념촬영을 하고 있다. 제공=IOK컴퍼니>

KHU

* 202년 11월에 대한 빈도 분석 - 23

단어	빈도
월드비전	23
전쟁	8
한국	6
후원	6
나라	4
당시	4

"한국전쟁x월드비전 70년 역사 되새긴다"…'월드 이즈 원 콘서트' 열려

월드비전이 한국전쟁, 월드비전 70주년을 맞아 '월드 이즈 원' 기부 콘서트를 개최한다. (자료제공=월드비전)

6. 빈도 분석

* 2020년 5월에 대한 빈도 분석 - 24

단어	빈도
월드비전	25
후원	21
유니세프	8
정의	7
단체	6

"회계 불투명, 정의연에 기부하기 싫어졌어요"...이건 친일인가요

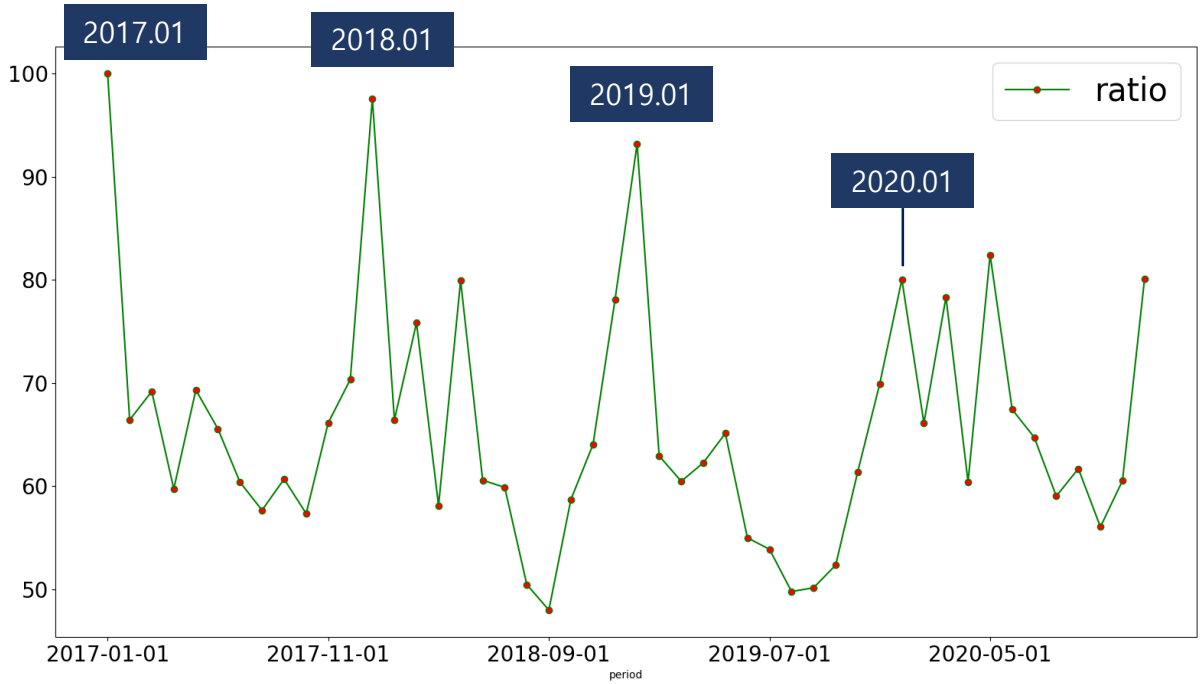
월드비전 31쪽 vs 정대협 1줄...기부금 공시, 이렇게 달랐다

최근 정의기억연대(정의연)의 부실 공시 논란으로 시민단체의 불투명한 운영 실태가 그대로 드러났다는 지적이다. 대다수 시민단체는 사회복지 등을 목적으로 설립된 비영리 공익법인에 속해 기부금에 대한 공시 의무를 가진다. 정의연과 전신인 정대협은 국고보조금 수익을 국세청 공시에서 누락하고, 수십 곳에 지출한 기부금을 한 곳에서 쓴 것처럼 기재했다. 수혜 인원으로 99명, 999명, 9999명이 반복적으로 등장한다.

7. 버즈 분석

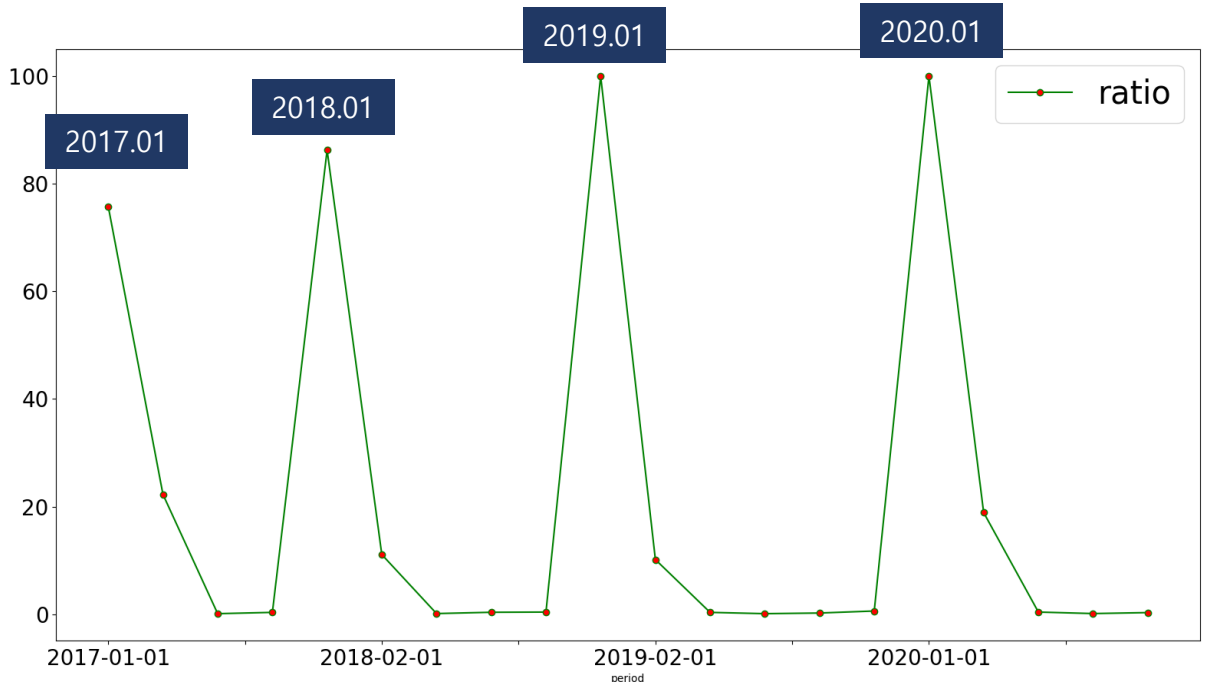
* '월드비전 검색 '

- 검색 포털 : 네이버
- 검색 기간 : 17.01.01 ~ 20.12.31
- 검색량 200 이상
- 최대값을 100으로 하는 비율 그래프
- 매년 1월에 높아지는 검색량



* '월드비전 연말정산 ' 검색

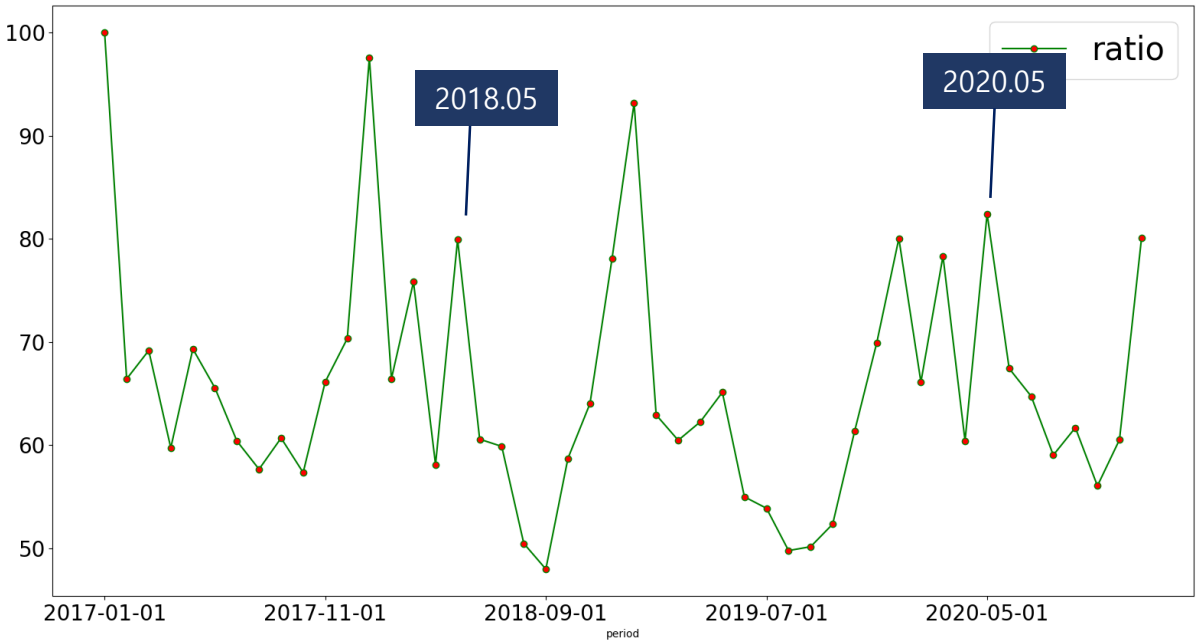
- 매년 1월에 높아지는 계절성은 연말정산을 위한 검색 때문이라고 추정된다.



7. 버즈 분석

* '월드비전 검색 '

- 검색 포털 : 네이버
- 검색 기간 : 17.01.01 ~ 20.12.31
- 검색량 200 이상
- 최대값을 100으로 하는 비율 그래프
- 불규칙적으로 검색량이 높아지는 월



* 영향을 준 사건들(추정)

2018.05

8. 토픽 모델링

* 모델 설정

```
stopword = ['하지만','근데','대한','이게','없는','ㅋㅋ','내내','그냥
','내가','그리고','진짜','정말','너무','나는','있는','가장','ㅎㅎ
','#’,’아니’]
```

* TfidfVectorizer

max_df : .15

ngram_range : (1,4)

min_df : 2,

* LatentDirichletAllocation

n_components : 임의조정

learning_method : 'batch'

max_iter : 25

random_state : 0

* 2020년 전체에 대한 토픽 모델링 3개

Topic 1

['월드비전에', '후원', '기부', '기부금', '먼저', '해서', '볼때마다', '다른', '월드비전에서', '하고']

Topic 2

['월드비전은', '앨범', '정기후원', '정기', '언제', '월드비전 콘서트', '콘서트', '영상', '드디어', '월드비전을']

Topic 3

['월드비전에서', '광고', '굿네이버스', '월드비전도', '마음을', '위해', 'world', 'world is one', 'is one', 'one']

- Topic 1 : 월드비전의 후원과 기부금에 대한 주제이다.
- Topic 2 : 월드비전 콘서트에 대한 주제이다.
- Topic 3 : 월드비전의 광고에 대한 주제이다. 월드비전의 광고문구도 나왔다.

8. 토픽 모델링

* 2020년 12월에 대한 토픽 모델링 2개

Topic 1

['후원', '월드비전에', '월드비전에서', '기부', '기부금', '유니세프', '광고', '하고', '굿네이버스', '먼저']

Topic 2

['월드비전은', '앨범', '정기후원', '월드비전도', '캠페인', '라비던스', '언제', '월드비전 콘서트', '위해', '콘서트']

- Topic 1 : 타 자선단체를 언급하며 월드비전과 비교하는 것으로 생각된다.
- Topic 2 : 월드비전의 콘서트와 앨범에 대한 주제이다.

topic을 3개로 늘리면 의미 불명의 topic이 되어 버린다.

* 2020년 11월에 대한 토픽 모델링 2개

Topic 1

['유엔', '전쟁', '생긴', '12월에', '월드비전도', '돕기위해 생긴 것도 그렇고', '돕기위해 생긴 것도', '난민기구가 1950년', '난민기구가', '나라는 돕는다 라는 개념이']

Topic 2

['월드비전에서', '월드비전에', '그걸', '나보고', '같은', '아이', '하는데', '어떤', '12월에', '월드비전도']

- Topic 1 : 70주년 콘서트의 주제인 한국 전쟁과 유엔에 관한 주제이다.
- Topic 2 : 의미를 파악하기 힘들다. topic 수를 1개로 줄이면 이 주제가 살아남는다.

8. 토픽 모델링

* 2020년 5월에 대한 토픽 모델링 2개

Topic 1

['대신', '후원을', '다현이', '없나', '월드비전이', '하는', '어느', '그나마도', '나오는', '받으려면']

Topic 2

['후원', '기부금', '유니세프 굿네이버스', '굿네이버스', '보면', '굿네이버스 월드비전', '유니세프 굿네이버스 월드비전', '기부금 받는', '받는', '이건']

- Topic 1 : 월드비전에 받는 후원금에 대한 주제라 추정. 2020.02 부터 광고에 나온 다현이가 언급됨.
- Topic 2 : 정의연 사태 때문인지 타 기부단체들과 비교하는 주제라 추정


감사합니다
THANK YOU
gettyimagesBank