dbproject3 보고서[[1]](#footnote-1)

6조- 남재현, 문현지, 박경린, 심규민

|  |
| --- |
| **목차**   1. 서론 2. association 함수의 설명    1. (R1) 테이블 및 Tagmatrix 뷰 생성 코드 설명    2. (R2) 연관 분석 코드 설명    3. 프로그램 실행 결과       1. 실행 결과의 정량적 분석       2. 실행 결과의 정성적 분석 3. decisiontree1 함수의 설명    1. (R4) ReputStatMatrix 뷰 생성 코드 설명    2. (R5) 의사결정나무 분석 코드 설명    3. 프로그램 실행 결과       1. 의사결정나무 모델의 분석       2. 추가 데이터 분류 결과 4. decisiontree2 함수의 설명    1. feature가 추가된 ReputStatMatrix2 뷰 생성 코드 설명    2. 의사결정나무 분석 코드 설명    3. 프로그램 실행 결과       1. 의사결정나무 모델의 분석       2. R5의 의사결정나무 모델과의 비교 5. 결론 |

1. 서론
2. association 함수의 설명
   1. (R1) 테이블 및 Tagmatrix 뷰 생성 코드 설명
   2. (R2) 연관 분석 코드 설명

(R2)에서의 연관 분석을 실시하는 코드는 다음과 같다:

data = np.array(df)[:, 1:]

X = data.astype(int)

print(frequent\_itemsets(X, 0.01))

itemsets = dict(frequent\_itemsets(X, 0.01))

print(itemsets)

rules = association\_rules(itemsets, 0.05)

rules = list(rules)

print(rules)

Orange3의 메서드인 frequent\_itemsets를 통해 Support>=0.01인 태그를 가져오고, 이를 dictionary 형태로 itemsets에 저장하여 출력한다. 또 메서드 association\_rules를 통해 confidence>=0.05 인 룰을 가져오고, 이를 list 형태로 rules에 저장하여 출력한다.

* 1. 프로그램 실행 결과

프로그램의 실행 결과는 다음과 같다:

**itemsets)** {frozenset({0}): 7244, frozenset({1}): 5413, frozenset({0, 1}): 1059, frozenset({2}): 2737, frozenset({0, 2}): 607, frozenset({3}): 2564, frozenset({4}): 2081, frozenset({5}): 2000, frozenset({6}): 1880, frozenset({7}): 1784, frozenset({8}): 1624, frozenset({8, 1}): 576, frozenset({9}): 1585, frozenset({10}): 1362, frozenset({11}): 1342, frozenset({12}): 1338, frozenset({3, 12}): 431, frozenset({13}): 1333, frozenset({14}): 1165, frozenset({15}): 1111, frozenset({16}): 982, frozenset({17}): 939, frozenset({18}): 911, frozenset({19}): 885, frozenset({20}): 875, frozenset({21}): 872, frozenset({22}): 838, frozenset({23}): 836, frozenset({24}): 786, frozenset({25}): 774, frozenset({26}): 766, frozenset({27}): 756, frozenset({28}): 743, frozenset({29}): 681, frozenset({30}): 668, frozenset({31}): 667, frozenset({32}): 636, frozenset({33}): 624, frozenset({34}): 614, frozenset({35}): 608, frozenset({36}): 602, frozenset({37}): 590, frozenset({38}): 586, frozenset({39}): 581, frozenset({40}): 568, frozenset({41}): 559, frozenset({42}): 535, frozenset({43}): 520, frozenset({44}): 514, frozenset({45}): 514, frozenset({46}): 504, frozenset({47}): 490, frozenset({48}): 489, frozenset({49}): 487, frozenset({50}): 473, frozenset({51}): 463, frozenset({52}): 460, frozenset({53}): 457, frozenset({54}): 454, frozenset({55}): 452, frozenset({56}): 448, frozenset({57}): 441, frozenset({58}): 439, frozenset({59}): 437}

**rules)** [(frozenset({1}), frozenset({0}), 1059, 0.19564012562349897),

(frozenset({0}), frozenset({1}), 1059, 0.1461899503036996),

(frozenset({2}), frozenset({0}), 607, 0.2217756667884545),

(frozenset({0}), frozenset({2}), 607, 0.08379348426283821),

(frozenset({1}), frozenset({8}), 576, 0.10641049325697395),

(frozenset({8}), frozenset({1}), 576, 0.35467980295566504),

(frozenset({12}), frozenset({3}), 431, 0.3221225710014948),

(frozenset({3}), frozenset({12}), 431, 0.16809672386895474)]

* + 1. 실행 결과의 정량적 분석

다음과 같이 64개의 태그가 Support가 0.01 이상으로 나타나는 것으로 나타난다. 대표적으로 높은 Support를 보인 태그는 r, regression, time-series, machine-learning 등이 있다.

또, Confidence>=0.05를 만족하는 rule은 다음과 같은 8개로 나타난다:

regression → r: confidence= 0.1956

r → regression: confidence= 0.1462

time-series → r: confidence= 0.2218

r → time-series: confidence= 0.0838

regression → logistic: confidence= 0.1064

logistic → regression: confidence= 0.3547

classification → machine-learning: confidence= 0.3221

machine-learning → classification: confidence= 0.1681

* + 1. 실행 결과의 정성적 분석

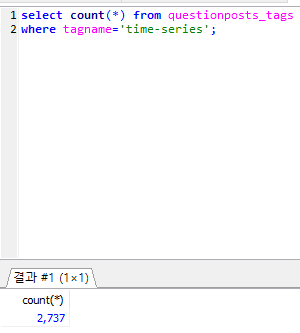
Support>= 0.01 인 태그 중 특히 Support가 높은 태그를 중심으로 볼 때, 현재 해당 사이트의 주된 관심사와 질문 주제는 r, regression, machine-learning 등 기계학습과 관련된 분야이다. 특히 이러한 태그들은 2500~7200개 수준의 질문에서 등장하여 여타 태그들(평균적으로 500~1000개의 질문에서 나타났음)에 비해 압도적인 수준으로 나타났다.

Confidence>= 0.05인 rule 역시 이러한 태그들을 중심으로 나타났다. 특히 rule은 특정한 분야의 용어를 중심으로 높게 나타나는 경향을 보였다. 예를 들어, logistic 태그는 regression에서 활용되는 함수로 이러한 두 태그 간의 Confidence가 매우 높게 나타났다. 이는 classification과 machine-learning 간에도 유사하게 나타났다. 또한 통계분석 언어인 r과 기계학습 용어 간의 Confidence가 높은 수준으로 나타났다.

1. decisiontree1 함수의 설명
   1. (R4) ReputStatMatrix 뷰 생성 코드 설명
   2. (R5) 의사결정나무 분석 코드 설명
   3. 프로그램 실행 결과
      1. 의사결정나무 모델의 분석
      2. 추가 데이터 분류 결과
2. decisiontree2 함수의 설명
   1. feature가 추가된 ReputStatMatrix2 뷰 생성 코드 설명
   2. 의사결정나무 분석 코드 설명
   3. 프로그램 실행 결과
      1. 의사결정나무 모델의 분석
      2. R5의 의사결정나무 모델과의 비교
3. 결론

select count(\*) from questionposts\_tags

where tagname='time-series';



count가 2737개로 전체 tuple이 42921개임을 고려할 때 time-series 이후의 tag들은 frequent itemset이 되거나, 포함될 수 없다.

1. 안 쓰거나 보충이 필요한 부분은 빨간 색으로 표시해 두었습니다! 참고해서 적어 주시면 됩니다. [↑](#footnote-ref-1)