Εξόρυξη δεδομένων με χρήση τεχνικών μηχανικής μάθησης

Τμ. Μηχανικών Πληροφορικής και Υπολογιστών

Παράδειγμα

Data set ξενοδοχείων, εστιατορίων κ.λπ.

Αρχίζουμε με απλοποιημένο παράδειγμα εξαγωγής κανόνων συσχέτισης από το σύνολο δεδομένων των αξιολογήσεων του υποθετικού ξενοδοχείου La Strada (hotel_strada)

Περιγραφή του απλοποιημένου dataset.

Κανονικά θα πρέπει το δείγμα μας να έχει τουλάχιστον 40 παρατηρήσεις – παραδείγματα. Για λόγους απλοποίησης θα πειραματιστούμε με ένα δείγμα που έχει 19 documents και 11 χαρακτηριστικά, τα παρακάτω:

Clean, comfortable, good location, extra facilities, liked breakfast, bad breakfast, friendly staff, rude staff, good value for money, Wi-fi, bad Wi-fi

Μοντελοποίηση με κανόνες συνάφειας

	clean	comfortable	good location	extra facilities	liked breackfast	bad breakfast	friendly staff	rude staff	good value for money	Wi-fi	bad Wi-fi
Doc1	yes	no	no	No	no	no	yes	no	no	yes	no
Doc2	no	yes	yes	No	no	no	yes	no	yes	yes	no
Doc3	yes	yes	yes	No	no	yes	no	yes	no	no	yes
Doc4	no	yes	yes	No	no	no	yes	no	no	yes	no
Doc5	yes	yes	no	No	yes	no	yes	no	yes	yes	no
Doc6	yes	no	yes	No	no	no	no	yes	yes	no	yes
Doc7	no	no	yes	yes	no	no	no	no	yes	yes	no
Doc8	no	no	no	No	no	no	yes	no	yes	yes	no
Doc9	yes	no	yes	yes	yes	yes	no	yes	yes	no	no
Doc10	no	no	yes	no	no	no	yes	no	no	no	no
Doc11	no	no	yes	no	no	no	yes	no	no	yes	no
Doc12	no	yes	yes	no	no	no	yes	no	yes	yes	no
Doc13	yes	yes	yes	yes	no	yes	yes	no	yes	no	yes
Doc14	yes	no	no	no	no	no	no	yes	no	yes	no
Doc15	no	no	yes	no	no	no	no	no	no	no	yes
Doc16	yes	yes	yes	yes	yes	no	yes	no	yes	yes	no
Doc17	yes	yes	yes	yes	yes	no	yes	no	yes	no	no
Doc18	yes	yes	yes	yes	Yes	no	yes	no	yes	no	no
Doc19	yes	yes	yes	no	Yes	no	yes	no	yes	yes	no

Format your columns.

Replace errors with missing values

	A *polynominal*	clean *binominal*	comfortable *binominal*	good locati... *binominal*	extra facilit... *binominal*	liked breac... *binominal*	bad breakf... *binominal*	friendly staff *binominal*
1	Document 1	yes	no	no	no	no	no	yes
2	Document 2	no	yes	yes	no	no	no	yes
3	Document 3	yes	yes	yes	no	no	yes	no
4	Document 4	no	yes	yes	no	no	no	yes
5	Document 5	yes	yes	no	no	yes	no	yes
6	Document 6	yes	no	yes	no	no	no	no
7	Document 7	no	no	yes	yes	no	no	no
8	Document 8	no	no	no	no	no	no	yes
9	Document 9	yes	no	yes	yes	no	yes	no
10	Document 10	no	no	yes	no	no	no	yes
11	Document 11	no	no	yes	no	no	no	yes
12	Document 12	no	yes	yes	no	no	no	yes
13	Document 13	yes	yes	yes	yes	no	yes	yes
14	Document 14	yes	no	no	no	no	no	no
15	Document 15	no	no	yes	no	no	no	no
16	Document 16	yes	yes	yes	yes	yes	no	yes
17	Document 17	yes	yes	yes	yes	yes	no	yes
18	Document 18	yes	yes	yes	yes	yes	no	yes

PREVIEW

no problems.

Previous Next Cancel

Where to store the data?

Name hotel_strada_corr

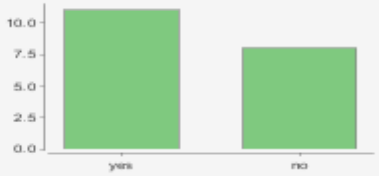
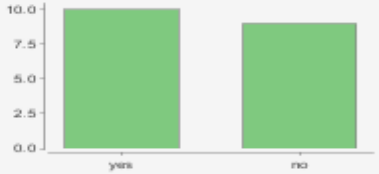
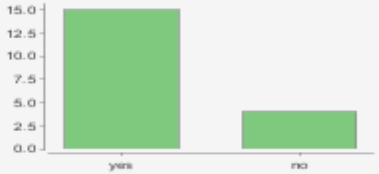
Open in Turbo Prep Auto Model

Filter (19 / 19 examples): all

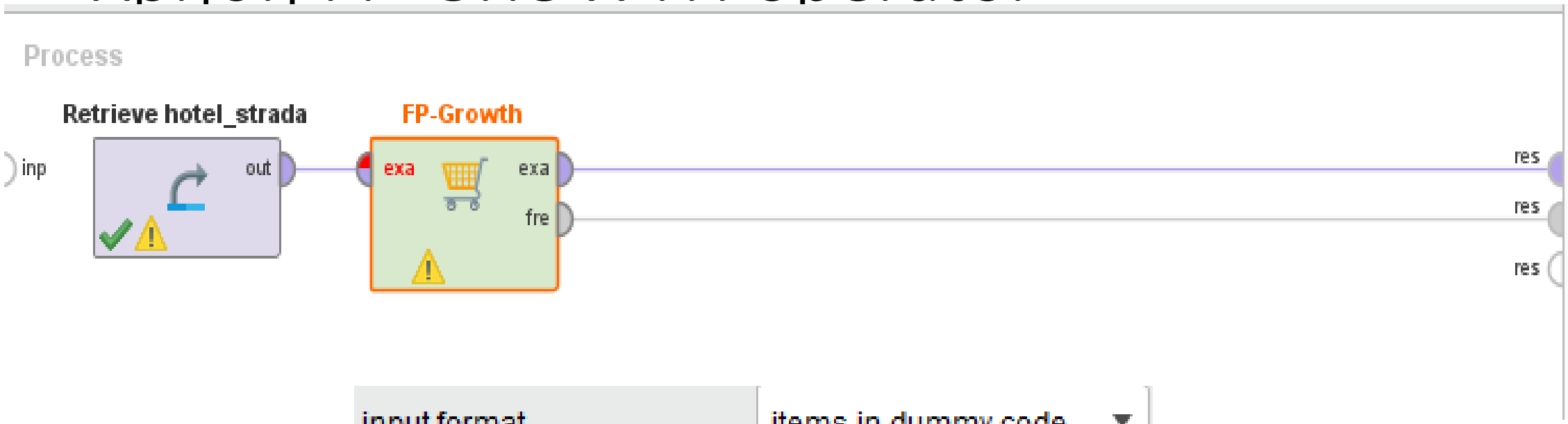
Row No.	A	clean	comfortable	good location	extra faciliti...	liked breack...	bad breakfast	friendly staff	r
1	Document 1	yes	no	no	no	no	no	yes	n
2	Document 2	no	yes	yes	no	no	no	yes	n
3	Document 3	yes	yes	yes	no	no	yes	no	y
4	Document 4	no	yes	yes	no	no	no	yes	n
5	Document 5	yes	yes	no	no	yes	no	yes	n
6	Document 6	yes	no	yes	no	no	no	no	y
7	Document 7	no	no	yes	yes	no	no	no	n
8	Document 8	no	no	no	no	no	no	yes	n
9	Document 9	yes	no	yes	yes	no	yes	no	y
10	Document 10	no	no	yes	no	no	no	yes	n
11	Document 11	no	no	yes	no	no	no	yes	n
12	Document 12	no	yes	yes	no	no	no	yes	n
13	Document 13	yes	yes	yes	yes	no	yes	yes	n
14	Document 14	yes	no	no	no	no	no	no	y
15	Document 15	no	no	yes	no	no	no	no	n
16	Document 16	yes	yes	yes	yes	yes	no	yes	n
17	Document 17	yes	yes	yes	yes	yes	no	yes	n

ExampleSet (19 examples, 0 special attributes, 12 regular attributes)

Visualization

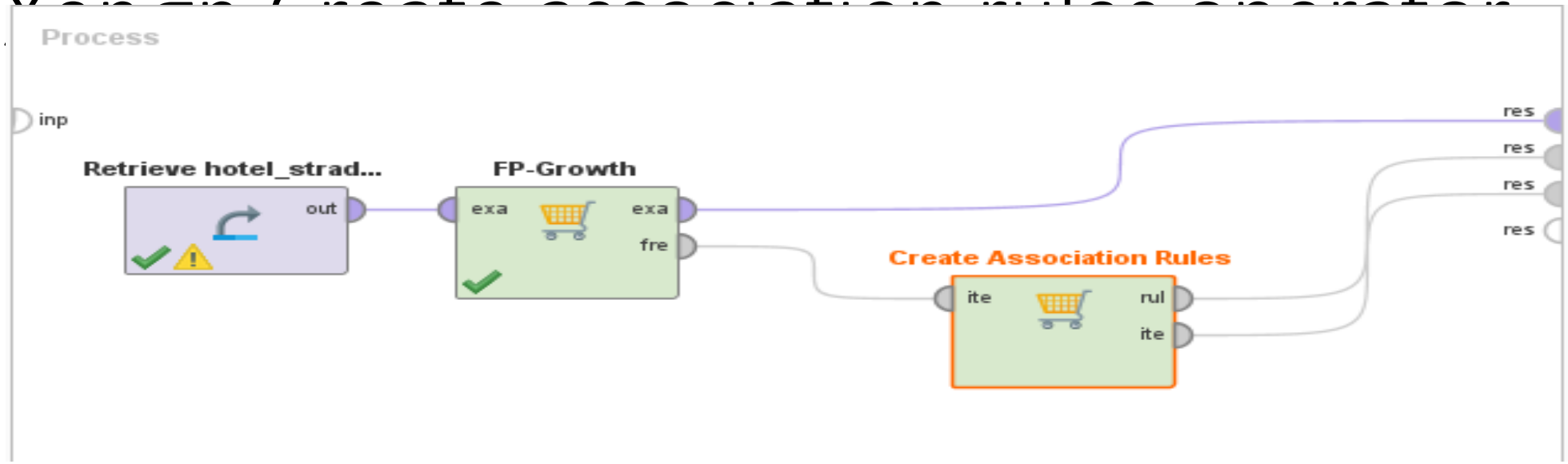
clean	Binominal	0	 Open visualizations	Least no (8)	Most yes (11)
comfortable	Binominal	0	 Open visualizations	Least no (9)	Most yes (10)
good location	Binominal	0	 Open visualizations	Least no (4)	Most yes (15)

Χρήση FP-GROWTH operator



input format	items in dummy code...
min requirement	support
min support	0.5

Χρήση Create association rules operator



Show rules matching

all of these conclusions:

good location
good value for money
comfortable
Wi-fi
extra facilities
friendly staff
bad Wi-fi
rude staff
bad breakfast

Show rules matching

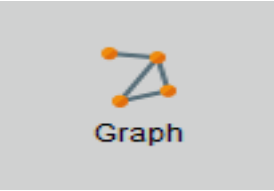
all of these conclusions:

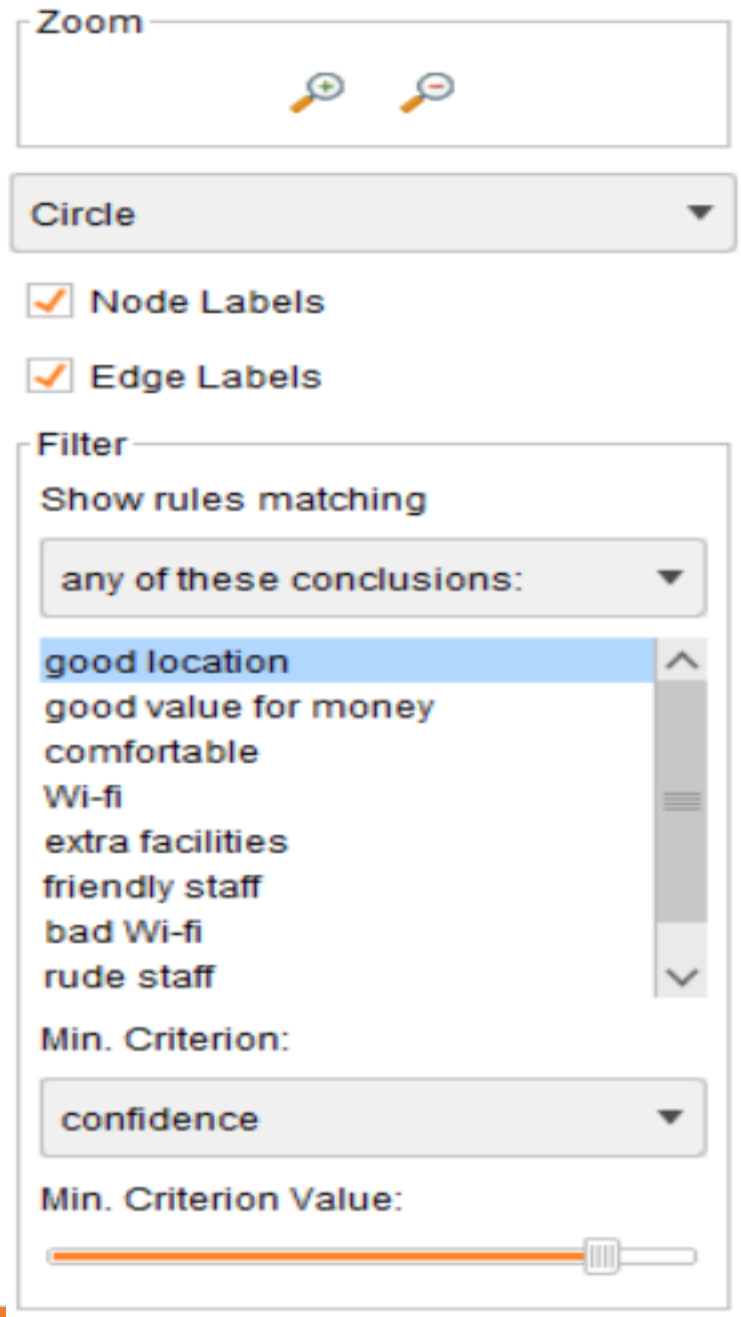
good location
good value for money
comfortable
Wi-fi
extra facilities
friendly staff
bad Wi-fi
rude staff
bad breakfast

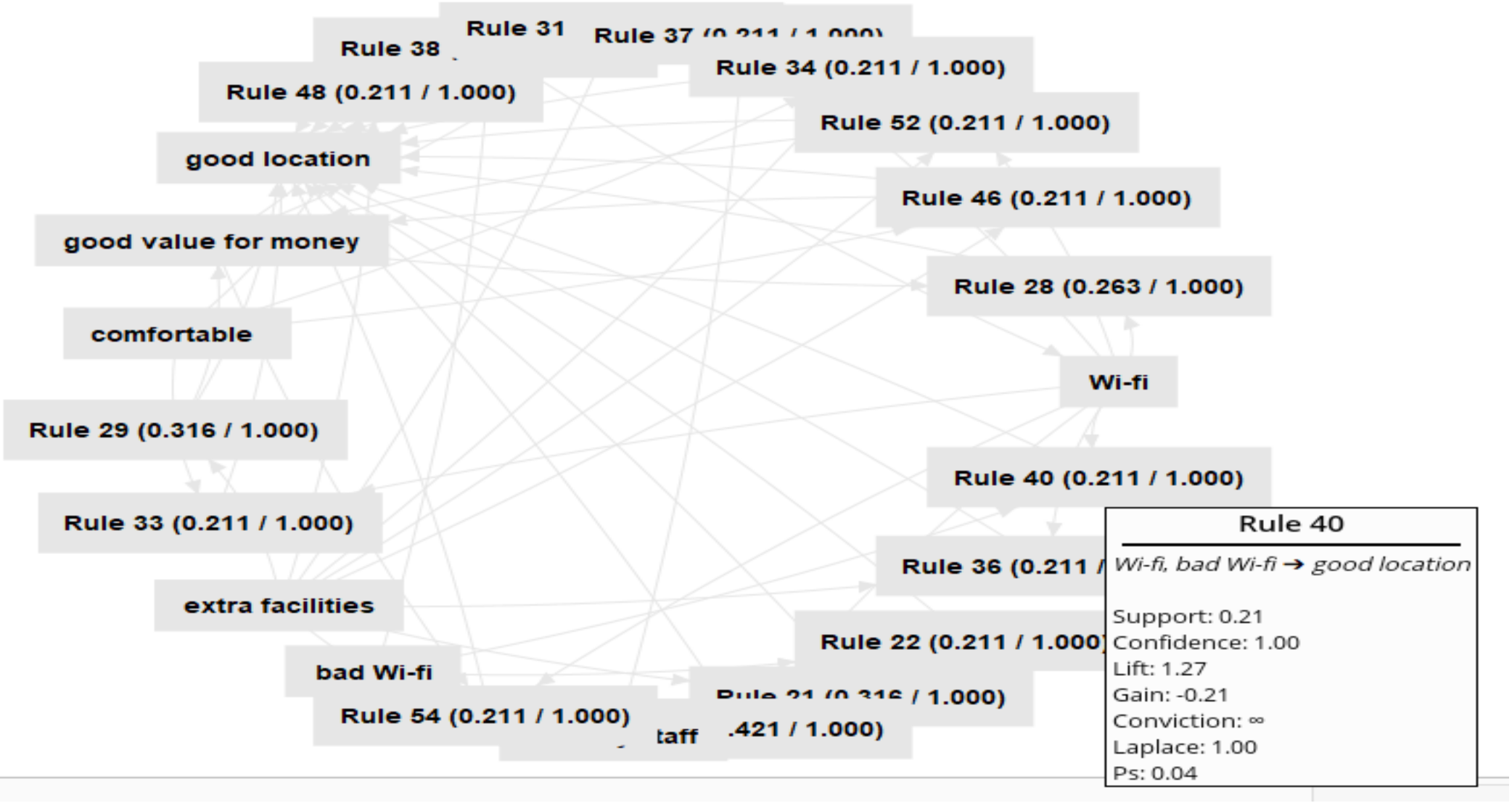
No	Premises	Conclusion	Support	Confidence	Lift
15	good value for money	good location	0.526	0.833	1.056
16	friendly staff	good location	0.263	0.833	1.056
17	liked breackfast	good location	0.263	0.833	1.056
20	good value for money, liked breackfast	good location	0.263	0.8333	1.056
23	Clean	good location	0.368	0.875	1.108
24	good value for money, comfortable	good location	0.368	0.875	1.108
25	comfortable	good location	0.474	0.9	1.14
26	Wi-fi	good location	0.421	1.0	1.267
27	extra facilities	good location	0.316	1.0	1.267
28	bad Wi-fi	good location	0.211	1.0	1.267
29	bad breakfast	good location	0.158	1.0	1.267
35	good value for money, Wi-fi	good location	0.263	1.0	1.267

AssociationRules

```
Association Rules
[liked breackfast] --> [good location] (confidence: 0.800)
[comfortable ] --> [good value for money] (confidence: 0.800)
[liked breackfast] --> [good location, good value for money] (confidence: 0.800)
[good value for money, liked breackfast] --> [good location] (confidence: 0.800)
[liked breackfast] --> [good location, comfortable ] (confidence: 0.800)
[comfortable , liked breackfast] --> [good location] (confidence: 0.800)
[good location, friendly staff] --> [Wi-fi] (confidence: 0.800)
[good value for money, Wi-fi] --> [extra facilities] (confidence: 0.800)
[liked breackfast] --> [good location, good value for money, comfortable ] (confidence: 0.800)
[good value for money, liked breackfast] --> [good location, comfortable ] (confidence: 0.800)
[comfortable , liked breackfast] --> [good location, good value for money] (confidence: 0.800)
[good value for money, comfortable , liked breackfast] --> [good location] (confidence: 0.800)
[good value for money, Wi-fi] --> [good location, extra facilities] (confidence: 0.800)
[good location, good value for money, Wi-fi] --> [extra facilities] (confidence: 0.800)
[good value for money] --> [good location] (confidence: 0.833)
[friendly staff] --> [good location] (confidence: 0.833)
[clean] --> [good location] (confidence: 0.875)
[good value for money, comfortable ] --> [good location] (confidence: 0.875)
[comfortable ] --> [good location] (confidence: 0.900)
[Wi-fi] --> [good location] (confidence: 1.000)
```


Graph


Zoom
Circle
Node Labels
Edge Labels
Filter
Show rules matching
any of these conclusions:
good location
good value for money
comfortable
Wi-fi
extra facilities
friendly staff
bad Wi-fi
rude staff
Min. Criterion:
confidence
Min. Criterion Value:


Rule 31
Rule 38
Rule 34 (0.211 / 1.000)
Rule 48 (0.211 / 1.000)
Rule 52 (0.211 / 1.000)
good location
Rule 46 (0.211 / 1.000)
good value for money
Rule 28 (0.263 / 1.000)
comfortable
Wi-fi
Rule 29 (0.316 / 1.000)
Rule 40 (0.211 / 1.000)
Rule 33 (0.211 / 1.000)
Rule 40
Wi-fi, bad Wi-fi → good location
Support: 0.21
Confidence: 1.00
Lift: 1.27
Gain: -0.21
Conviction: ∞
Laplace: 1.00
Ps: 0.04
extra facilities
Rule 22 (0.211 / 1.000)
bad Wi-fi
Rule 54 (0.211 / 1.000)
taff

ΤΕΛΟΣ ΕΝΟΤΗΤΑΣ