Data Mining & Business Intelligence

Aim: Mini Project

Requirements: Python 3x, WEKA, Google Chrome

Theory:

Topic: Programming courses to be taught in educational institutions

Problem Statement:

Keeping up with the latest technologies is difficult and time consuming for students looking for placements in todays world. This experiment aims to solve that problem by using data mining techniques to suggest popular languages wanted by the industry to be taught in educational institutions that can include it in their curriculum.

The Stack Overflow Developers survey dataset is used to find the programming languages that users want to learn. This data is then cleaned, visualized and then Apriori association rules are generated to find other related languages a student might want to learn based on a selected one. This grouping of itemsets can be used to provide discounted fee for students or to prepare workshops for these technologies.

Dataset:

Name: Stack Overflow Developer Survey (2017)

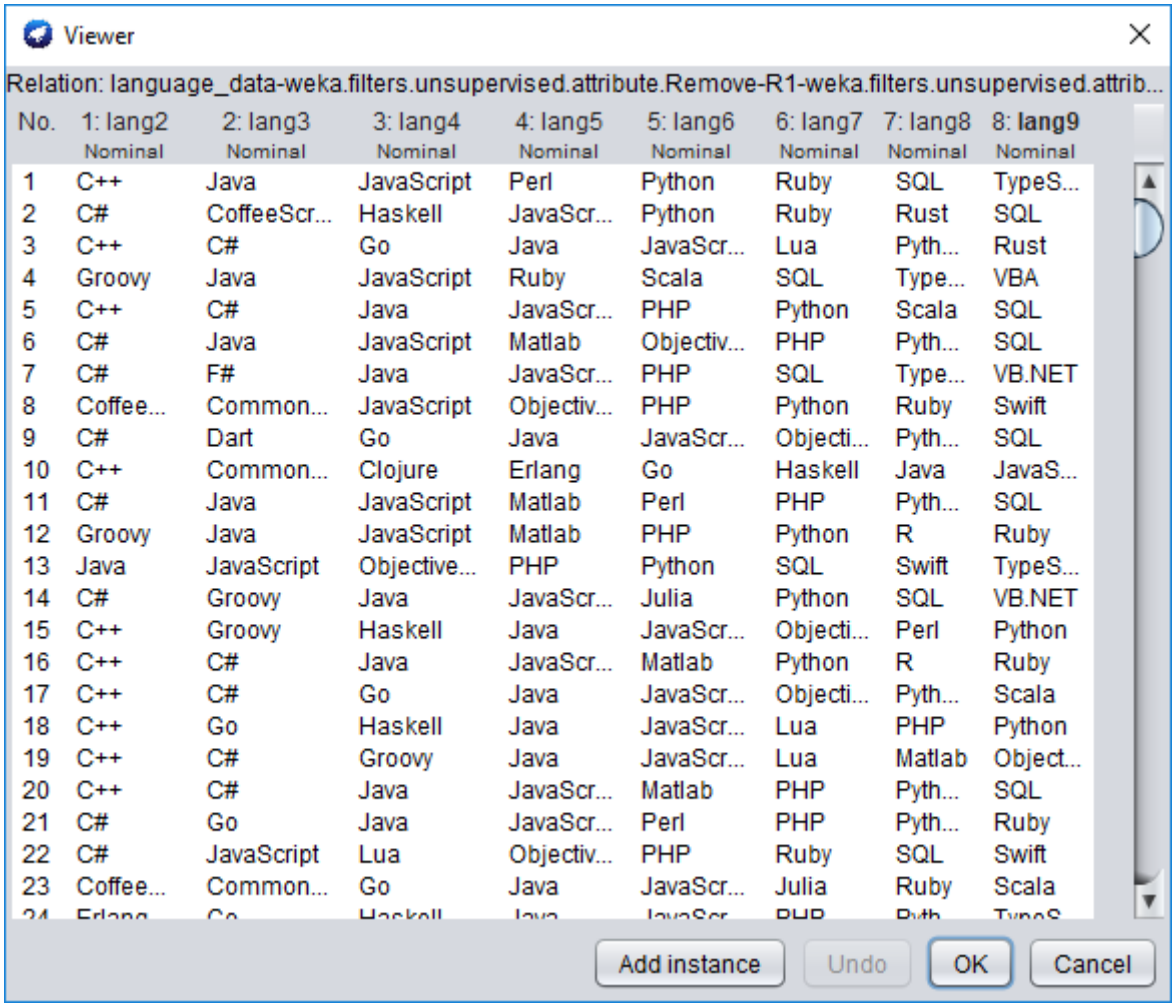
Records: 51,392 (3924 used)

Attributes: 155 (6 used)

Attribute Information:

NO.	Name	Type
1	Professional	Nominal
2	Country	Nominal
3	DeveloperType	Nominal
4	NonDeveloperType	Nominal
5	HaveWorkedLanguage	Nominal
6	WantWorkLanguage	Nominal

Apriori Association



Viewer

Relation: language_data-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attrib...

No.	1: lang2 Nominal	2: lang3 Nominal	3: lang4 Nominal	4: lang5 Nominal	5: lang6 Nominal	6: lang7 Nominal	7: lang8 Nominal	8: lang9 Nominal
1	C++	Java	JavaScript	Perl	Python	Ruby	SQL	TypeS...
2	C#	CoffeeScr...	Haskell	JavaScr...	Python	Ruby	Rust	SQL
3	C++	C#	Go	Java	JavaScr...	Lua	Pyth...	Rust
4	Groovy	Java	JavaScript	Ruby	Scala	SQL	Type...	VBA
5	C++	C#	Java	JavaScr...	PHP	Python	Scala	SQL
6	C#	Java	JavaScript	Matlab	Objectiv...	PHP	Pyth...	SQL
7	C#	F#	Java	JavaScr...	PHP	SQL	Type...	VB.NET
8	Coffee...	Common...	JavaScript	Objectiv...	PHP	Python	Ruby	Swift
9	C#	Dart	Go	Java	JavaScr...	Objecti...	Pyth...	SQL
10	C++	Common...	Clojure	Erlang	Go	Haskell	Java	JavaS...
11	C#	Java	JavaScript	Matlab	Perl	PHP	Pyth...	SQL
12	Groovy	Java	JavaScript	Matlab	PHP	Python	R	Ruby
13	Java	JavaScript	Objective...	PHP	Python	SQL	Swift	TypeS...
14	C#	Groovy	Java	JavaScr...	Julia	Python	SQL	VB.NET
15	C++	Groovy	Haskell	Java	JavaScr...	Objecti...	Perl	Python
16	C++	C#	Java	JavaScr...	Matlab	Python	R	Ruby
17	C++	C#	Go	Java	JavaScr...	Objecti...	Pyth...	Scala
18	C++	Go	Haskell	Java	JavaScr...	Lua	PHP	Python
19	C++	C#	Groovy	Java	JavaScr...	Lua	Matlab	Object...
20	C++	C#	Java	JavaScr...	Matlab	PHP	Pyth...	SQL
21	C#	Go	Java	JavaScr...	Perl	PHP	Pyth...	Ruby
22	C#	JavaScript	Lua	Objectiv...	PHP	Ruby	SQL	Swift
23	Coffee...	Common...	Go	Java	JavaScr...	Julia	Ruby	Scala

Add instance Undo OK Cancel

Fig 1

Dataset is imported in WEKA, selecting only 8 languages per record

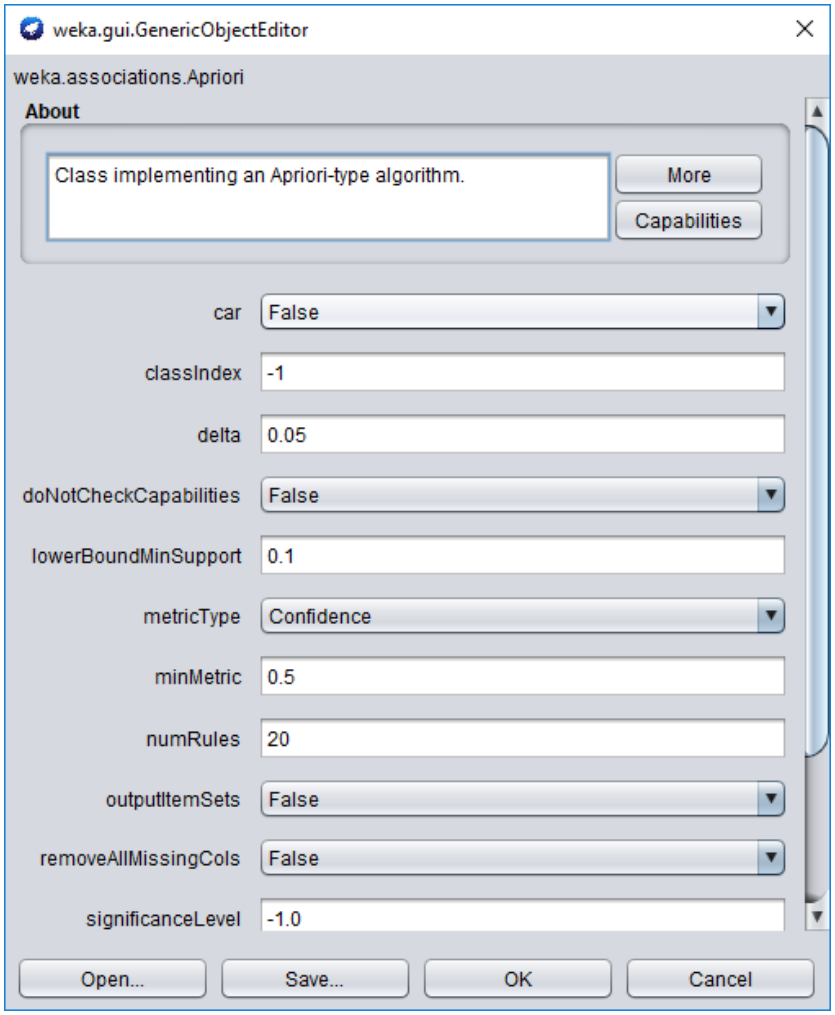


Fig 2

Apriori Association is applied to the dataset, with support 0.1 and confidence 0.5

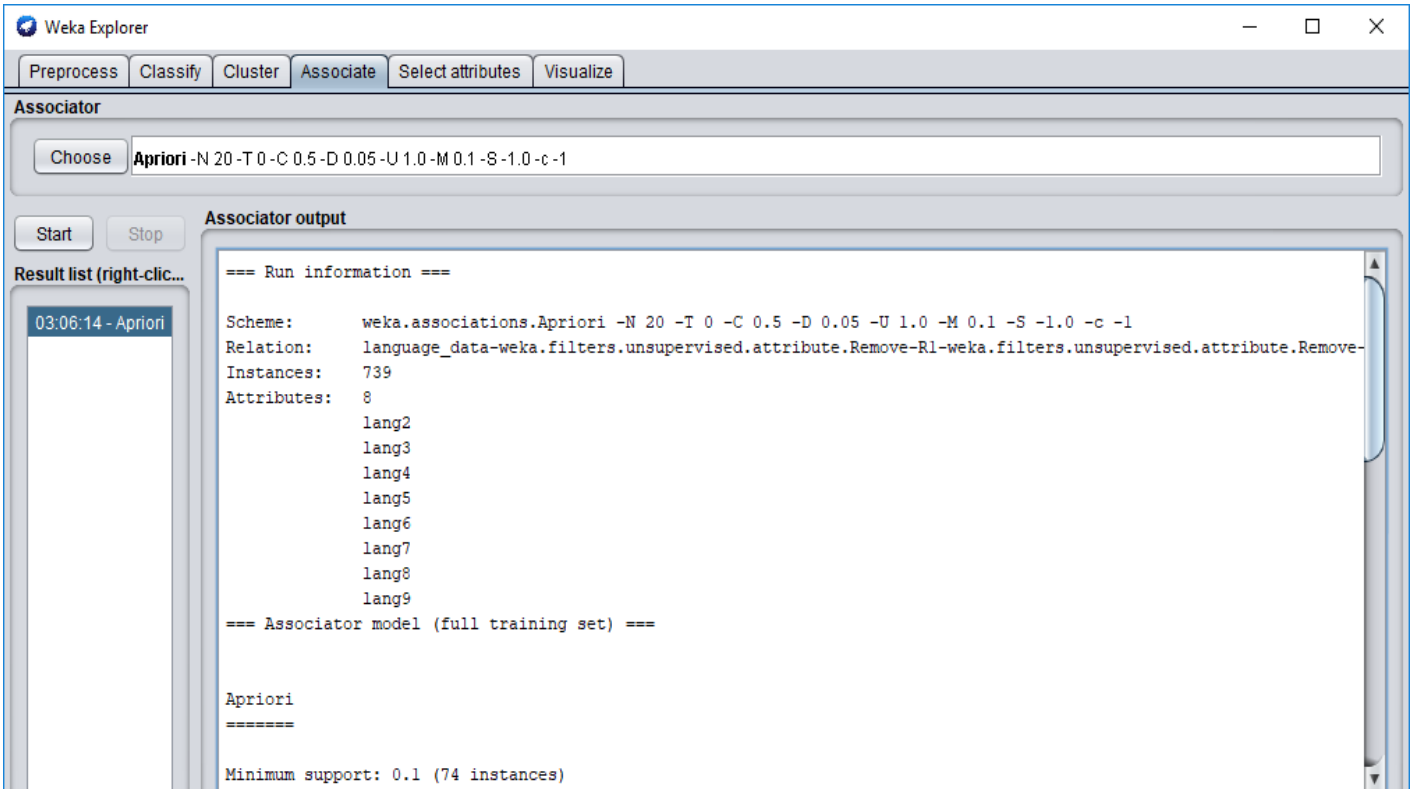


Fig 3

Results of association are shown

SimpleCLI

```
Apriori
=======

Minimum support: 0.1 (74 instances)
Minimum metric <confidence>: 0.5
Significance level: 1
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 19

Large Itemsets L(1):
lang2= C++ 248
lang2= C# 211
lang3= Java 194
lang3= C# 164
lang3= JavaScript 77
lang4= JavaScript 202
lang4= Java 201
lang5= JavaScript 202
lang5= Java 99
lang5= PHP 79
lang6= Python 95
lang6= JavaScript 100
lang6= PHP 157
lang7= Python 169
lang7= PHP 137
lang8= SQL 152
lang8= Python 146
lang9= SQL 175
lang9= Swift 78
```

```
java weka.associations.Apriori -N 20 -T 0 -C 0.5 -D 0.05 -U 1.0 -S 1.0 -I -t "F:\DMBI\data\language_data.arff"
```

SimpleCLI

```
Size of set of large itemsets L(2): 13

Large Itemsets L(2):
lang2= C++ lang3= C# 164
lang2= C++ lang4= Java 87
lang2= C++ lang5= JavaScript 86
lang2= C# lang3= Java 83
lang2= C# lang4= JavaScript 80
lang3= Java lang4= JavaScript 177
lang4= JavaScript lang6= PHP 78
lang4= Java lang5= JavaScript 191
lang4= Java lang7= PHP 78
lang5= JavaScript lang7= PHP 75
lang5= Java lang6= JavaScript 94
lang6= PHP lang7= Python 121
lang7= PHP lang8= Python 113
```

Fig 4

Apriori Association shows frequent itemsets in L1, L2 and L3

SimpleCLI

```
Size of set of large itemsets L(3): 3

Large Itemsets L(3):
lang2= C++ lang4= Java lang5= JavaScript 83
lang2= C# lang3= Java lang4= JavaScript 75
lang4= Java lang5= JavaScript lang7= PHP 75

Best rules found:

 1. lang3= C# 164 ==> lang2= C++ 164    <conf:(1)> lift:(2.98) lev:(0.15) [108] conv:(108.96)
 2. lang5= JavaScript lang7= PHP 75 ==> lang4= Java 75    <conf:(1)> lift:(3.68) lev:(0.07) [54] conv:(54.6)
 3. lang2= C++ lang5= JavaScript 86 ==> lang4= Java 83    <conf:(0.97)> lift:(3.55) lev:(0.08) [59] conv:(15.65)
 4. lang4= Java lang7= PHP 78 ==> lang5= JavaScript 75    <conf:(0.96)> lift:(3.52) lev:(0.07) [53] conv:(14.17)
 5. lang2= C++ lang4= Java 87 ==> lang5= JavaScript 83    <conf:(0.95)> lift:(3.49) lev:(0.08) [59] conv:(12.64)
 6. lang4= Java 201 ==> lang5= JavaScript 191    <conf:(0.95)> lift:(3.48) lev:(0.18) [136] conv:(13.28)
 7. lang5= Java 99 ==> lang6= JavaScript 94    <conf:(0.95)> lift:(7.02) lev:(0.11) [80] conv:(14.27)
 8. lang5= JavaScript 202 ==> lang4= Java 191    <conf:(0.95)> lift:(3.48) lev:(0.18) [136] conv:(12.25)
 9. lang6= JavaScript 100 ==> lang5= Java 94    <conf:(0.94)> lift:(7.02) lev:(0.11) [80] conv:(12.37)
10. lang2= C# lang4= JavaScript 80 ==> lang3= Java 75    <conf:(0.94)> lift:(3.57) lev:(0.07) [53] conv:(9.83)
11. lang3= Java 194 ==> lang4= JavaScript 177    <conf:(0.91)> lift:(3.34) lev:(0.17) [123] conv:(7.83)
12. lang2= C# lang3= Java 83 ==> lang4= JavaScript 75    <conf:(0.9)> lift:(3.31) lev:(0.07) [52] conv:(6.7)
13. lang4= JavaScript 202 ==> lang3= Java 177    <conf:(0.88)> lift:(3.34) lev:(0.17) [123] conv:(5.73)
14. lang7= PHP 137 ==> lang8= Python 113    <conf:(0.82)> lift:(4.17) lev:(0.12) [85] conv:(4.4)
15. lang8= Python 146 ==> lang7= PHP 113    <conf:(0.77)> lift:(4.17) lev:(0.12) [85] conv:(3.5)
16. lang6= PHP 157 ==> lang7= Python 121    <conf:(0.77)> lift:(3.37) lev:(0.12) [85] conv:(3.27)
17. lang7= Python 169 ==> lang6= PHP 121    <conf:(0.72)> lift:(3.37) lev:(0.12) [85] conv:(2.72)
18. lang2= C++ 248 ==> lang3= C# 164    <conf:(0.66)> lift:(2.98) lev:(0.15) [108] conv:(2.27)
19. lang7= PHP 137 ==> lang4= Java 78    <conf:(0.57)> lift:(2.09) lev:(0.06) [40] conv:(1.66)
20. lang7= PHP 137 ==> lang5= JavaScript 75    <conf:(0.55)> lift:(2) lev:(0.05) [37] conv:(1.58)

=== Evaluation ===

Elapsed time: 0.032s
```

```
java weka.associations.Apriori -N 20 -T 0 -C 0.5 -D 0.05 -U 1.0 -S 1.0 -I -t "F:\DMBI\data\language_data.arff"
```

Fig 5

Apriori Association Mining generates frequent itemsets and association rules that show frequently grouped together Programming languages