1 Übersetzung in relationale Algebra (8 Punkte)

Gegeben sei das Relationenschema aus der zweiten Übung.

[Personen] : {[PID: int, Name: string, Geburtsjahr: int, Wohnort: string]}
[Schüler*innen] : {[SID:(Personen→PID), Klassenstufe: int, Klassenraum: int, Schulform: string]}
[Klausuren] : {[KID: int, Fach: string, Thema: string, Dauer: int]}
[Lehrer*innen] : {[LID:(Personen→PID), Hauptfach:string, Gehalt:int, Dienstjahre:int]}
[unterrichten] : {[Schüler*in:(Schüler*innen→SID), Datum:date, Uhrzeit:time,
Lehrer*in:(Lehrer*innen→LID), Fach: string)]}
[korrigieren] : {[Klausur:(Klausuren→KID), Schüler*in:(Schüler*innen→SID),
Lehrer*in:(Lehrer*innen→LID), Note: int]}

Um die Lesbarkeit Ihrer Ausdrücke zu verbessern, dürfen Sie die Teilergebnisnotation aus der zweiten Übung benutzen. Benennen Sie zum Beispiel einen Ausdruck wie folgt

$$R_3 := R_1 \bowtie_{\mathrm{A=B}} R_2,$$

so kann R_3 nun in darauffolgenden Ausdrücken verwendet werden. Zudem dürfen Sie den syntaktischen Zucker für die gleichzeitige Umbenennung mehrerer Attribute aus der zweiten Übung nutzen

$$\rho_{\mathrm{A'\leftarrow A,\ldots,Z'\leftarrow Z}}\ R \ = \ \rho_{\mathrm{A'\leftarrow A}}\ (\ldots\ (\rho_{\mathrm{Z'\leftarrow Z}}\ R)).$$

(a) Übersetzen Sie die folgenden umgangssprachliche Anfragen in Ausdrücke der relationalen Algebra:

1. Die schlechteste (größte) Note, die in Klausuren über das Thema 'Quantenmechanik' erreicht wurde.
2. Die durchschnittlichen Noten je Klausur, die von Schüler*innen der Klassenstufe 12 geschrieben wurde.
3. Die Namen der Lehrer*innen, die schon mehr als 7 Schüler*innen bei einer Klausur über das Thema 'Analysis' korrigiert haben.
4. Die Schulformen der Schüler*innen, die noch nie an einer Klausur teilgenommen haben, in der mehr als 10 Schüler*innen mindestens die Note 3 erreicht haben.
5. Das Geburtsjahr der ältesten Schüler*innen, die genau eine Klausur geschrieben haben, in der sie von einem/einer Lehrer*in korrigiert wurden, der/die diese Schüler*innen bereits im Fach Informatik unterrichtet hat.

(b) Übersetzen Sie folgende Anfragen aus der relationalen Algebra in natürliche Sprache:

1. $R := (\sigma_{\mathrm{Klassenstufe\ =\ 10\ \wedge\ Schulform\ =\ 'Gymnasium'}}\ \mathrm{Schüler*innen})\ \bowtie_{\mathrm{SID\ =\ Schüler*in}}\ \mathrm{unterrichten}$
   $\pi_{\mathrm{Gehalt}}(\mathrm{R}\ \bowtie_{\mathrm{Lehrer*in=LID}}\ (\sigma_{\mathrm{Hauptfach='Sport'}}\ \mathrm{Lehrer*innen}))$
2. $R_1 := \gamma_{\mathrm{avg(Gehalt)}}\ ((\sigma_{\mathrm{Dienstjahre>10}}\ \mathrm{Lehrer*innen})\ \bowtie_{\mathrm{LID=Lehrer*in}}\ (\pi_{\mathrm{Lehrer*in}}\ \mathrm{unterrichten}))$
   $R_2 := (\mathrm{Lehrer*innen}\ \bowtie_{\mathrm{Gehalt<avg(Gehalt)}}\ R_1)\ \bowtie_{\mathrm{LID\ =\ Lehrer*in}}\ \mathrm{korrigieren}$
   $\pi_{\mathrm{Schulform}}\ (\mathrm{Schüler*innen}\ \bowtie_{\mathrm{SID=Schüler*in}}\ R_2)$
3. $R_1 := (\sigma_{\mathrm{Fach\ =\ 'Physik'\ \vee\ Fach\ =\ 'Biologie'}}\ \mathrm{Klausuren})\ \bowtie_{\mathrm{KID\ =\ Klausur}}\ \mathrm{korrigieren}$
   $R_2 := \mathrm{korrigieren}\ \bowtie_{\mathrm{Note\ =\ max(Note)\ \wedge\ Klausur\ =\ KID}}(\gamma_{\mathrm{KID,max(Note)}}\ R_1)$
   $R_3 := \sigma_{\mathrm{count(*)\ >\ 5}}\ (\gamma_{\mathrm{Schüler*in,count(*)}}\ R_2)$
   $\pi_{\mathrm{Name}}\ (R_3 \bowtie_{\mathrm{Schüler*in\ =\ PID}}\ \mathrm{Personen})$

Lösung:

(a) Vorgesehen ist ein Punkt pro Anfrage (je 0,5 Punkte Abzug für inkorrekte (Join-)Prädikate oder Aggregate, unterschiedliche Schemata bei Mengenoperationen, nicht-disjunkte Schemata bei Joins, uneindeutige Bezeichner, inkorrekte Bezeichner durch Verwechslung von Umbenennung und Teilergebnisnotation, usw.). Wichtig ist hierbei nur, dass die Anfragen das korrekte Ergebnis produzieren, nicht, dass sie mit der Musterlösung übereinstimmen.

1. $\gamma_{\text{max(Note)}}\ ((\sigma_{\text{Thema = 'Quantenmechanik'}}\ \text{Klausuren})\ \bowtie_{\text{KID = Klausur}}\ \text{korrigieren})$
2. $R := \text{korrigieren}\ \bowtie_{\text{Schüler*in = SID}}\ (\sigma_{\text{Klassenstufe=12}}\text{Schüler*innen})$
   $\pi_{\text{avg(Note)}}(\gamma_{\text{Klausur,avg(Note)}}\ R)$
3. $R_1 := \text{korrigieren}\ \bowtie_{\text{Klausur = KID}}\ (\sigma_{\text{Thema='Analysis'}}\ \text{Klausuren})$
   $R_2 := \sigma_{\text{count(*)} > 7}(\gamma_{\text{Klausur, Lehrer*in,count(*)}}\ R_1)$
   $\pi_{\text{Name}}(R_2\ \bowtie_{\text{Lehrer*in = PID}}\ \text{Personen})$
4. $R_1\ := \rho_{\text{Klausur'}\leftarrow\text{Klausur}}(\sigma_{\text{count(*)} > 10}\ (\gamma_{\text{Klausur,count(*)}}(\sigma_{\text{Note} \leq 3}\ \text{korrigieren})))$
   $R_2\ := \text{Schüler*innen}\ \bowtie_{\text{SID = Schüler*in}}\ (\text{korrigieren}\ \bowtie_{\text{Klausur = Klausur'}}\ R_1)$
   $\pi_{\text{Schulform}}\ (\pi_{\text{SID, Schulform}}\text{Schüler*innen - }\pi_{\text{SID, Schulform}}\ R_2)$
5. $R_1 := \rho_{\text{Lehrer*in'}\leftarrow\text{Lehrer*in,Schüler*in'}\leftarrow\text{Schüler*in}}\text{korrigieren}$
   $R_2 := (\sigma_{\text{Fach = 'Informatik'}}\ \text{unterrichten})\ \bowtie_{\text{Lehrer*in = Lehrer*in'}\ \wedge\ \text{Schüler*in = Schüler*in'}}\ R_1$
   $R_3 := \sigma_{\text{count(*)} = 1}(\gamma_{\text{Schüler*in,count(*)}}\ (\pi_{\text{[korrigieren]}}\ R_2))$
   $\gamma_{\text{min(Geburtsjahr)}}(R_3\ \bowtie_{\text{Schüler*in = PID}}\ \text{Personen})$

(b) Vorgesehen ist ein Punkt pro Übersetzung.

1. Die Gehälter der Lehrer*innen, die Sport als Hauptfach haben und einmal ein/e Schüler*in in der 10. Klasse auf dem Gymnasium unterrichtet haben.
2. Die Schulformen der Schüler*innen, deren Klausur mindestens einmal von einem/einer Lehrer*in korrigiert wurde, deren Gehalt geringer ist als das Durchschnittsgehalt der Lehrer*innen, die bereits mehr als 10 Dienstjahre haben und unterrichtet haben.
3. Die Namen der Schüler*innen, die bereits mehr als fünf mal in einer Klausur in den Fächern Physik oder Biologie Klassenschlechteste*r waren, also die schlechteste Note hatten.

2 Relationale Anfragen (7 Punkte)

Betrachten Sie das vereinfachte IMDb-Schema aus der Vorlesung und den Notebooks.

$$
\begin{aligned}
[\text{actors}] &: \{[\underline{\text{id:int}}, \text{first\_name:varchar}, \text{last\_name:varchar}, \text{gender:varchar}]\} \\
[\text{movies}] &: \{[\underline{\text{id:int}}, \text{name:varchar}, \text{year:int}, \text{rank:float}]\} \\
[\text{directors}] &: \{[\underline{\text{id:int}}, \text{first\_name:varchar}, \text{last\_name:varchar})]\} \\
[\text{movies\_directors}] &: \{[\underline{\text{director\_id:(directors}\rightarrow\text{id}), \text{movie\_id: (movies}\rightarrow\text{id})}]\} \\
[\text{movies\_genres}] &: \{[\underline{\text{movie\_id:(movies}\rightarrow\text{id})}, \text{genre:varchar}]\} \\
[\text{roles}] &: \{[\underline{\text{actor\_id:(actors}\rightarrow\text{id}), \text{movie\_id:(movies}\rightarrow\text{id})}, \text{role:varchar}]\} \\
[\text{directors\_genres}] &: \{[\underline{\text{director\_id:(directors}\rightarrow\text{id})}, \text{genre:varchar}, \text{prob:float}]\}
\end{aligned}
$$

Um die Lesbarkeit Ihrer Ausdrücke zu verbessern, dürfen Sie wieder die Teilergebnisnotation und den syntaktischen Zucker für die gleichzeitige Umbenennung mehrerer Attribute nutzen, die in den vorherigen Aufgaben eingeführt wurden.

Übersetzen Sie die folgenden umgangssprachlichen Anfragen in Ausdrücke der relationalen Algebra:

1. Die Jahre, in denen eine Komödie veröffentlicht wurde.
2. Die Vor- und Nachnamen der Schauspieler*innen, die schon einmal eine Rolle im Film 'Star Wars' gespielt haben.
3. Die Nachnamen und die Anzahl an Filmen, die Regisseur*innen mit dem gleichen Nachnamen gedreht haben.
4. Die Genre der Filme mit der besten Bewertung.
5. Die Nachnamen der Schauspieler*innen, die die Rolle des/der Superman/Superwoman in einem Actionfilm gespielt haben.
6. Die Anzahl an Horrorfilmen, die vom Regisseur James Wan nach 2000 gedreht wurden.
7. Die höchste Anzahl an Filmen die von einer/einem Regisseur*in gedreht worden sind, dessen/deren Filme vor 1995 mindestens eine Durchschnittswertung von 3 haben.

Lösung:

Vorgesehen ist ein Punkt pro Anfrage (je 0,5 Punkte Abzug für inkorrekte (Join-)Prädikate oder Aggregate, unterschiedliche Schemata bei Mengenoperationen, nicht-disjunkte Schemata bei Joins, uneindeutige Bezeichner, inkorrekte Bezeichner durch Verwechslung von Umbenennung und Teilergebnisnotation, usw.). Wichtig ist hierbei nur, dass die Anfragen das korrekte Ergebnis produzieren, nicht, dass sie mit der Musterlösung übereinstimmen.

1. $\pi_{\text{year}}\ ((\sigma_{\text{genre = 'Komödie'}}\ \text{movies\_genres}) \bowtie_{\text{movie\_id = id}}\ \text{movies})$
2. $R_1 := (\sigma_{\text{name = 'Star Wars'}}\ \text{movies}) \bowtie_{\text{id = movie\_id}}\ \text{roles}$
   $\pi_{\text{first\_name, last\_name}}\ (R_1 \bowtie_{\text{actor\_id = id'}}\ (\rho_{\text{id'}\leftarrow\text{id}}\ \text{actors}))$
3. $\gamma_{\text{last\_name,count(*)}}(\pi_{\text{last\_name, movie\_id}}(\text{directors} \bowtie_{\text{id = director\_id}}\ \text{movie\_directors}))$
4. $\pi_{\text{genre}}\ (\text{movies\_genres} \bowtie_{\text{movie\_id = id}}\ (\text{movies} \bowtie_{\text{rank = max(rank)}}\ (\gamma_{\text{max(rank)}}\ \text{movies})))$
5. $R_1 := \rho_{\text{movie\_id'}\leftarrow\text{movie\_id}}(\sigma_{\text{genre='Action'}}\ \text{movies\_genres})$
   $R_2 := R_1 \bowtie_{\text{movie\_id'=movie\_id}}\ (\sigma_{\text{role='Superman'} \vee \text{role='Superwoman'}}\ \text{roles})$
   $\pi_{\text{last\_name}}\ (R_2 \bowtie_{\text{actor\_id=id}}\ \text{actors})$

6. $R_1 := \sigma_{\text{first\_name='James'} \wedge \text{last\_name='Wan'}}\ \text{directors}$
   $R_2 := (\sigma_{\text{genre = 'Horror'}}\ \text{movies\_genres}) \bowtie_{\text{movie\_id=movie\_id'}} (\rho_{\text{movie\_id'} \leftarrow \text{movie\_id}}(\text{movie\_directors}))$
   $\gamma_{\text{count(*)}}((R_2 \bowtie_{\text{director\_id=id}}\ R_1) \bowtie_{\text{movie\_id = m\_id}} (\rho_{\text{m\_id} \leftarrow \text{id}}(\sigma_{\text{year}>2000}\ \text{movies})))$

7. $R_1 := \text{movie\_directors} \bowtie_{\text{movie\_id=id}} (\sigma_{\text{year} < 1995}\ \text{movies})$
   $R_2 := \rho_{\text{director\_id'} \leftarrow \text{director\_id}}(\sigma_{\text{avg(rank)} \geq 3}\ (\gamma_{\text{director\_id,avg(rank)}}\ R_1))$
   $\gamma_{\text{max(count(*))}}\ (\gamma_{\text{director\_id,count(*)}}\ (\text{movie\_directors} \bowtie_{\text{director\_id = director\_id'}}\ R_2))$

3 Implementierung Co-Group-Join (5 Punkte)

Implementieren Sie den Co-Group-Join in den beiden mit *Exercise* markierten Zellen im beigefügten Notebook. Dieses finden Sie zusammen mit den anderen zur Ausführung benötigten Dateien im Anhang der Übung. Hierbei müssen die Klassen `Equi_Join` und `Co_Group_Join` vervollständigt werden.

Die Klasse `Equi_Join` erweitert dabei die Klasse `BinaryOperator` für logische, binäre Operatoren. Vervollständigen Sie hier die folgende Methode:

- `get_schema`: Gibt das Ausgabeschema des Operators zurück. Hier müssen Sie die Schemata der beiden Kinder konkatenieren.

Die Klasse `Co_Group_Join` erweitert die Klasse `Equi_Join` um die physische Funktionalität des Auswertens. Hierbei soll ein Equi_Join mit Hilfe einer Co-Gruppierung ausgeführt werden. Dabei soll eine Auswertung in *linearer Laufzeit* $\mathcal{O}(n)$ gewährleistet sein. Beachten Sie, dass diese Laufzeitbeschränkung nur gelten soll, sofern man über die Schlüssel beider Relationen joint. Ihre Implementierung soll trotzdem alle Fälle abdecken. Implementieren Sie die folgende Funktion:

- `evaluate`: Wertet den Operatorbaum rekursiv aus. Beachten Sie, dass ihr Join in *linearer Laufzeit* $\mathcal{O}(n)$ in der Anzahl an Elementen der Eingaberelationen durchgeführt werden soll, sofern über die Schlüssel beider Relationen gejoint wird. Es dürfen keine anderen Operatoren wie z.B. Theta Join als Subroutine verwendet werden. Eine Ausnahme bildet hier das Kreuzprodukt.

Für einen einfachen Einstieg haben wir Ihnen bereits einige Beispielanwendungen beider Klassen sowie einen Unit Test mit mehreren Testfällen im Notebook zur Verfügung gestellt, der `evaluate` von `Co_Group_Join` testet. Ihre finale Abgabe muss auf allen gültigen Eingaben das korrekte Ergebnis berechnen. Die Qualität Ihrer Implementierung wird bei der Punktevergabe berücksichtigt.

Abgabe

Lösungen sind in Teams von 2 bis 3 Studierenden bis zum 12. Mai 2022, 10:15 Uhr über Ihre persönlichen Statusseite im CMS einzureichen. Nutzen Sie hierfür die Team Groupings Funktionalität im CMS.

Ihre Abgabe muss dem folgenden Format entsprechen:

```
abgabe.zip
├── abgabe.pdf
└── jupyter.txt
```

Hierbei enthält `abgabe.pdf` Ihre Lösungen zu Aufgabe 1 und 2 und `jupyter.txt` Ihre Lösung zu Aufgabe 3. Achten Sie darauf, dass Sie nur die von Ihnen zu ergänzenden Jupyter Zellen so kopieren, dass Einrückung und Formatierung korrekt sind.

Abgaben, die nicht den oben angegeben Vorgaben entsprechen, führen zu Punktabzug. Einzelabgaben werden nicht mehr korrigiert und mit 0 Punkten bewertet.