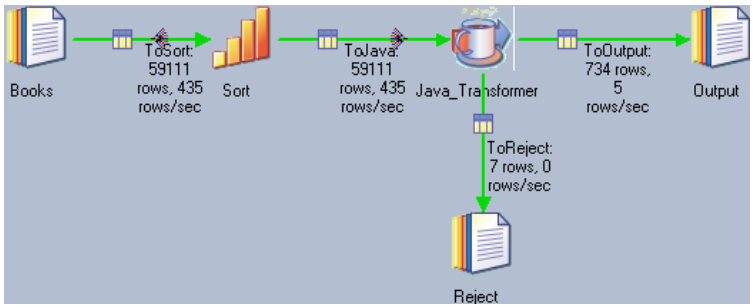
Aufgabenblatt 6
IBM Information Server

- Letzter Abgabetermin: **Sonntag, 20. Juli 2008**
- Abgabe: Per E-Mail an Alexander Albrecht, Fachgebiet Informationssysteme.

Gegeben ist ein für diese Übung vorbereiteter IBM WebSphere QualityStage Job, der in einer Relation von Büchern Duplikate findet. Der zugrundeliegende Algorithmus verwendet den bekannten Ansatz der Partitionierung (Blocking). Das Attribut, nach dem partitioniert wird, muß in der Java Transformer Stage bei den User's Properties angegeben werden. Als Resultat wird eine Liste der gefundenen Duplikate ausgeben. Das im Algorithmus verwendete Ähnlichkeitsmaß klassifiziert zwei Tupel als Duplikate, wenn ihre Wertepaare in keinem der Attribute die edit-distance 2 übersteigt.



Aufgabe

Optimiere den in Java implementierten Algorithmus zur Duplikaterkennung mit dem Ziel, die Effektivität der Duplikaterkennung zu erhöhen. Die Qualität der Lösung messen wir mit den Maßen PRECISION und RECALL.

PRECISION misst die Richtigkeit der gefundenen Duplikate, indem der Quotient aus den korrekt gefundenen Duplikaten und allen gefundenen Duplikaten gebildet wird.

RECALL misst die Vollständigkeit der gefundenen Duplikate, indem der Quotient aus den korrekt gefundenen Duplikaten und allen korrekten Duplikaten gebildet wird.

Für beide Maße wird das harmonische Mittel (also die F-Measure) berechnet, das maximiert werden soll.

Abgabe

Erstelle in dem Projekt deiner Gruppe den vorgestellten IBM WebSphere QualityStage Job. Die Daten und das Java Programm findest du im Übungsverzeichnis. Das Verwenden eigener Java Programme in IBM WebSphere QualityStage Jobs ist in der Dokumentation im Java Pack Guide beschrieben und wird in der Übungsveranstaltung vorgestellt.

Schicke deine Java-Klasse sowie eine kurze Dokumentation deines Ansatzes bis zum 20. Juli an Alexander Albrecht, Fachgebiet Informationssysteme.