Business Process Mining

Practice 9: 
Pre-processing & Filtering practice

Manuel Camargo

Phd student, Junior research fellow @ University of Tartu
Preliminaries

Tools:

• Anaconda → (Download)
• Jupyter Notebook

• Google Colaboratory → (Notebook)

Libraries (Environment):

• Numpy
• Pandas
• Matplotlib
• Seaborn
• PM4PY
Common quality issues

- Lack of Case ID or ID changing during the process
  - Case ID generation
  - Case ID consolidation
- Timestamp granularity or formatting
  - Change the logging frequency
  - Auto format detection
- Incomplete transitions (Start + End Timestamps or Only End Timestamp)
- Labels of activities not meaningful, too many activities, or not relevant for analysis
  - Relabeling, aggregation
  - Grouping, clustering
  - Filtering
- Other issues...
Event-log characteristics

Process attributes

Case attributes

Events attributes

Data attributes
<table>
<thead>
<tr>
<th>creator</th>
<th>Case ID</th>
<th>Activity</th>
<th>Resource</th>
<th>time:timestamp</th>
<th>lifecycle:transition</th>
<th>variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluxicon Disco</td>
<td>Case1</td>
<td>Start</td>
<td>Start</td>
<td>30/01/2012 04:24:00</td>
<td>start</td>
<td>7</td>
</tr>
<tr>
<td>Fluxicon Disco</td>
<td>Case1</td>
<td>Start</td>
<td>Start</td>
<td>30/01/2012 04:24:00</td>
<td>complete</td>
<td>7</td>
</tr>
<tr>
<td>Fluxicon Disco</td>
<td>Case1</td>
<td>Turning &amp; Milling</td>
<td>ID4932</td>
<td>30/01/2012 04:24:00</td>
<td>start</td>
<td>7</td>
</tr>
<tr>
<td>Fluxicon Disco</td>
<td>Case1</td>
<td>Turning &amp; Milling</td>
<td>ID4932</td>
<td>30/01/2012 10:43:00</td>
<td>complete</td>
<td>7</td>
</tr>
<tr>
<td>Fluxicon Disco</td>
<td>Case1</td>
<td>End</td>
<td>End</td>
<td>17/02/2012 06:00:00</td>
<td>start</td>
<td>7</td>
</tr>
<tr>
<td>Fluxicon Disco</td>
<td>Case1</td>
<td>End</td>
<td>End</td>
<td>17/02/2012 06:00:00</td>
<td>complete</td>
<td>7</td>
</tr>
<tr>
<td>Fluxicon Disco</td>
<td>Case102</td>
<td>Start</td>
<td>Start</td>
<td>27/02/2012 21:40:00</td>
<td>start</td>
<td>11</td>
</tr>
<tr>
<td>Fluxicon Disco</td>
<td>Case102</td>
<td>Start</td>
<td>Start</td>
<td>27/02/2012 21:40:00</td>
<td>complete</td>
<td>11</td>
</tr>
<tr>
<td>Fluxicon Disco</td>
<td>Case102</td>
<td>Packing</td>
<td>ID4820</td>
<td>14/03/2012 05:00:00</td>
<td>start</td>
<td>11</td>
</tr>
<tr>
<td>Fluxicon Disco</td>
<td>Case102</td>
<td>Packing</td>
<td>ID4820</td>
<td>14/03/2012 06:00:00</td>
<td>complete</td>
<td>11</td>
</tr>
<tr>
<td>Fluxicon Disco</td>
<td>Case102</td>
<td>End</td>
<td>End</td>
<td>14/03/2012 06:00:00</td>
<td>start</td>
<td>11</td>
</tr>
<tr>
<td>Fluxicon Disco</td>
<td>Case102</td>
<td>End</td>
<td>End</td>
<td>14/03/2012 06:00:00</td>
<td>complete</td>
<td>11</td>
</tr>
</tbody>
</table>