Let Σ an alphabeth of events $\Sigma = \{o_1 \dots o_n\}$

a transaction w over o is an element of Σ^*

$\boxed{\text{example} \quad \Sigma = \{A, T, C, G\} \text{ and a word } w \text{ is AACCTTG}}$

given $w, \overline{w} \in \overline{\Sigma}$ we say:

w is a subsequence of w' (written $w \sqsubseteq w'$)

if and only if there exists a sequence of indexes $i_1 < \dots < i_n$ with $n = |w'|$

such that $\overline{w}[i_1] \dots \overline{w}[i_n] = w$

example: $ATT \sqsubseteq C\underline{A}\underline{T}G\underline{T}$ (take 1, 2, 4 as indexes)

while

$TAT \not\sqsubseteq CATGT$ (there is no A following a T in CATGT)

Given a word w we define its preceding subsequent as the words w' such that $\underline{w' \sqsubseteq w \text{ and } |w'| = |w| - 1}$

we denote them with $w' < w$

$$\text{Apriori\_for\_seq}(\mathcal{T}, \varepsilon) \longrightarrow \left\{ \overline{w} : \underbrace{\frac{\sum_{w \in \mathrm{Dom}(\mathcal{T}),\, \overline{w} \sqsubseteq w} w}{\sum_{w \in \mathrm{Dom}(\mathcal{T})} \mathcal{T}(w)}}_{Sup_{\mathcal{T}}(w)} \geqslant \varepsilon \right\}$$

ε: support

$\mathcal{T}$: multiset of sequences

$r_1 = \{o : Sup_{\mathcal{T}}(o) \geqslant \varepsilon\}$

$k = 1$

while $r_k \neq \emptyset$ **do**:

  $r_{k+1} = \emptyset$

  for each $w \in r_k$, $\overline{w} \in r_1$ **do**:

    if $\left(\{w'.\overline{w} : w' < w\} \subseteq r_k \text{ and } Sup_{\mathcal{T}}(w.\overline{w}) \geqslant \varepsilon \right)$ **then**: $r_{k+1} = r_{k+1} \cup \{w.\overline{w}\}$

  $k = k+1$

return $\bigcup_{i=1}^{k-1} r_k$

Exercise 1

Using the Notebook Apriori Sepsis

Ⓐ Implement the Apriori_for_seq explained above,

Ⓑ Test the code Ⓐ on the sepsis sequence extraction provided in the notebook. ($\varepsilon = 0.05$)

Ⓒ Extract the association rules based on sequences on the result of algorithm Ⓐ with confidence $\delta = 0.8$

(hint: Join the mined patterns x and y with $x \sqsubseteq y$ and $x \neq y$)