Mini-projekt #2: PM4Py, ProM i Disco

Wykonaj raport z audytu szpitalnego procesu obsługi pacjentów z podejrzeniem sepsy na podstawie analizy logu Sepsis.xes.gz[1] z użyciem narzędzi PM4Py, ProM i Disco. Opis logu znajdziesz poniżej:

This real-life event log contains events of sepsis cases from a hospital. Sepsis is a life threatening condition typically caused by an infection. One case represents the pathway through the hospital.

The events were recorded by the ERP (Enterprise Resource Planning) system of the hospital. There are about 1000 cases with in total 15,000 events that were recorded for 16 different activities. Moreover, 39 data attributes are recorded, e.g., the group responsible for the activity, the results of tests and information from checklists.

Events and attribute values have been anonymized. The time stamps of events have been randomized, but the time between events within a trace has not been altered.

Attributes

==========

Name	*Description*
Age	*Age in 5-year groups*
*Diagnostic**	*Various checkboxes on the triage document*
DisfuncOrg	*Checkbox: Disfunctional organ*
Hypotensie	*Checkbox: Hypotension*
Hypoxie	*Checkbox: Hypoxia*
InfectionSuspected	*Checkbox: Suspected infection*
Infusion	*Checkbox: Intravenous infusion required*
Oligurie	*Checkbox: Oliguria*
SIRSCritHeartRate	*Checkbox: One of the SIRS criteria*
SIRSCritLeucos	*Checkbox: One of the SIRS criteria*

[1] Źródło logu: https://doi.org/10.4121/uuid:915d2bfb-7e84-49ad-a286-dc35f063a460

„Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)", projekt finansowany ze środków Programu Operacyjnego Polska Cyfrowa POPC.03.02.00-00-0001/20

SIRSCritTachypnea | Checkbox: One of the SIRS criteria

SIRSCritTemperature| Checkbox: One of the SIRS criteria

SIRSCriteria2OrMore | Checkbox: Two or more of the SIRS criteria

Leucocytes | Leucocytes measurement

CRP | CRP measurement

LacticAcid | Lactic-acid measurement

W raporcie uwzględnij następujące punkty:

- Opis problemu i opis pochodzenia logu.
- Analizę zawartości logu, opis typowego przebiegu procesu, rozkład częstości przypadków biznesowych, występowanie pętli, rozkład czasu trwania procesu.
- Analiza kompletności przypadków biznesowych zapisanych w logu. Jeśli występują niekompletne przypadki, to proszę obsłużyć problemy z nich wynikające.
- Wyniki przeprowadzenia dwóch analiz – pierwsza w narzędziu PM4Py lub ProM, druga w Disco.
- PM4Py i ProM: Zbuduj model procesu w co najmniej dwóch reprezentacja spośród: C-net, drzewo procesu, sieć Petriego, i wykonaj poniższe punkty:
    - Przefiltruj log i/lub dostrój parametry algorytmu uczącego tak, aby odrzucić zaszumione zależności oraz niekompletne przypadki, ale jednocześnie nie tracić pełnego obrazu na przebieg procesu. Przetestuj różne warianty algorytmów uczących i sprawdź czy uzyskane modele są w dobrym stanie (ang. sound), tj. są wolne od deadlock'a i livelock'a, nie posiadają martwych części, z każdego stanu pozwalają na przejście do stanu końcowego i po uzyskaniu stanu końcowego nie pozostają żadne tokeny wewnątrz modelu. Dokonaj analizy zgodności modeli z logiem. Wybierz model, który jest w dobrym stanie (albo bliski tego stanu), posiadający wysokie dopasowanie (ang. fitness) i precyzję (ang. precision) i wykonaj na jego podstawie kolejne kroki.
    - Opisz na podstawie modelu jak wygląda typowy przebieg procesu. Czy ta obserwacja jest zgodna ze wcześniejszą ręczną analizą logu?
    - Opisz na podstawie modelu i/lub symulacji jak wyglądają najczęstsze odchyłki od typowego przebiegu i czym mogą być spowodowane.
    - Dokonaj analizy czasowej procesu z użyciem modelu i logu, aby określić, które aktywności są głównymi przyczynami przestojów i spróbuj ocenić dlaczego.

    - Zbuduj i oceń klasyfikatory decyzji podejmowanych w węzłach typu XOR-split/OR-split, które na podstawie atrybutów przypadku i/lub zdarzenia podejmą decyzję, którą ścieżkę przetwarzania obrać.[2]
    - Przeanalizuj zgodność odkrytego modelu z wiedzą dziedzinową, np.: [1], [2].
- Disco: Zbuduj mapę procesu i dokonaj następujących analiz:
    - Filtrowanie logu i strojenie parametrów algorytmu.
    - Opis typowego przebiegu procesu wg mapy procesu. Czy jest on zgodny z obserwacjami w logu?
    - Analiza częstości wykonywania aktywności: najdłuższe aktywności, najczęściej powtarzane. Dlaczego?
    - Czy występują pętle długości 1 (aktywność A -> aktywność A)? Jeśli tak, to uzasadnij jaka może być ich przyczyna?
    - Analiza okresowości pracy. Jeśli występuje, to uwzględnij ją w analizie.
    - Co jest kluczowym czynnikiem wpływającym na całkowity czas trwania procesu (ang. total active time), co wpływa na przestoje (patrz: waiting time), a co na wzrost liczby zdarzeń/aktywności w procesie?
- Dokonaj analizy porównawczej uzyskanych wyników w obu narzędziach. Które narzędzie tworzy lepsze modele i dlaczego? Czy jedno z narzędzi pozwoliło wyjaśnić działanie procesu lepiej niż inne? Opisz swoje wnioski z pracy z tymi narzędziami.
- Możliwe są rozszerzenia o więcej analiz i obserwacji niż wymienione wyżej.[3]
- Zapisz do plików częściowe wyniki analizy, modele itp. tak, aby można je było uruchomić w PM4Py, ProM, lub Disco na potrzeby weryfikacji wyników przedstawionych w raporcie.

Raport powinien być dokumentem tekstowym złożonym zgodnie z dobrymi zasadami składu tekstu i zasadami redagowania raportu. Tekst należy opatrzyć tabelami, np.: prezentującymi dane i/lub statystyki; ilustracjami np.: modele, wykresy i bibliografią, np.: referencjami na użyte algorytmy i/lub fakty z wiedzy dziedzinowej. Mile widziane jest wykorzystanie LaTeXa lub nakładki LyX na niego, ale nie jest to wymagane.

Raport należy dostarczyć w formie elektronicznej (PDF) do prowadzącego. Nie ma potrzeby wydruku. Długość dokumentu nie ma wpływu na ocenę, więc proszę unikać nadmiernego

[2] Rozszerzenie punktowane dodatkowo, max 1p. W PM4Py odtwórz log na modelu metodą Token-Based Replay (TBR) i utwórz klasyfikatory dla punktów decyzji. W ProM należy wykorzystać modele posiadające „Data" w nazwie, np.: „Data Causal net", „Data Petri net".
[3] Rozszerzenia punktowane dodatkowo, max 1p.

rozdmuchiwania tekstu formatowaniem i/lub zbyt wielkimi ilustracjami. Liczy się kompletność treści i zwięzłość prezentacji.

Omówienie/wyjaśnienie terminów z wiedzy dziedzinowej:

- Triage – procedura szybkiej klasyfikacji pacjentów, patrz: [1],
- ER – Emergency Room, ostry dyżur [3],
- IV – Intravenous therapy, wlew dożylny [4],
- CRP – C-Reactive Protein, białko C-reaktywne [5],
- NC – Normal Care ward, „normalny" oddział szpitalny/inny niż intensywna terapia,
- IC – Intensive Care ward, intensywna terapia [6],
- Release A-E – sposoby wypisania pacjenta (przypisanie liter do typów nie jest znane, ale w niektórych przypadkach można wydedukować z modelu procesu):
    - Zwolnienie bez przyjęcia (Discharge without admission),
    - Zwolnienie na „normalny" oddział (Admission normal ward),
    - Zwolnienie do domu (Discharge home),
    - Przyjęcie na intensywną terapię (Admission ICU),
    - Przyjęcie na „normalny" oddział i przejście pacjenta na intensywną terapię w ciągu 72h.

Więcej szczegółów można znaleźć w dokumencie: [7].