Process Discovery Using Python (Praktikum) WS 2020

CPN Models

Project Initiation Document

Version 1.0

12.11.2020

Team Members: Mehmood Ahmed, Younes Müller, Shubham Balyan, Sonam Chugh, Deekshith Shetty

Project Supervisor: Madhavi Shankar and Mahsa Pourbafrani

Contents

List of Tables

List of Figures

1 Overview

1.1 Project Overview

The aim of the project is to allow a user to generate a simulation model represented as a Coloured Petri net (CPN) [1] based on event logs. The discovered simulation models would be enriched with all of the relevant information such as time, the flow of activities, resources, etc. All of these perspectives can be tracked down using existing process mining techniques and as a result they can be combined to form a single simulation model. Moreover, the user can also configure information such as service time of the activities or arrival rate of the process and generating the CPN model for a specific period of time.

1.2 Project Background

As per the market research done by International Data Corporation (IDC), most companies are completely unaware of their processes not performing at the maximum capacity and are losing 20-30% of their revenue [2]. Process mining provides a way to detect such problems early and allows companies to use it for monitoring and optimizing their process. Providing a tool to do computer simulations is very useful and versatile and allows us to gain insight into the operation of systems.

The idea of applying process mining in the context of workflow management has been proposed before, however, most developments were done using the ProM framework. ProM (which is short for Process Mining) framework is an Open Source Java-based framework that offers a variety of plug-ins that enable you to apply the latest developments in process mining research. It has a plugin to simulate CPN models. However, it lacks good user experience and documentation. Furthermore, developing the application using popular languages such as Python allows for a wide community reach.

2 Business Use Case

The business use case of this project is to derive a simulation model that reflects the real process as precisely as possible. This project will follow a business process management approach. Information systems deployed in any industry usually logs the events happening in a business process, now these event logs can be transaction logs or audit logs. Further analysis of these event logs can reveal important domain knowledge and patterns that can help the organizations to improve the quality of their business process, fulfill customer expectations, and increase profits. As an added advantage, the pain points and the delays within organizational processes can also be identified in this way, and hence the allocation of resources to various tasks can be performed efficiently. This project aims to fulfill the business requirement of creating a simulation model for a business process, with some additional information to have a better understanding of the process and its various aspects.

2.1 Project Scope

Our project scope is to use the process mining algorithms to generate a simulation model, also known as CPN model, which can then be used to generate simulated event logs that follow the real process as given in the original event logs. As a first step, it is important

to get the algorithm right to bring about the feature to completion. In addition to this, there would be an option for the user to change some of the information such as service time of activities or arrival rate of the process as a configurable parameter. Understanding the intricate details behind the various process discovery algorithms to generate process models in the form of Petri nets is out of scope for this project.

2.2 Key Benefits

The project has many benefits from all perspectives, namely, theoretical foundations, tool support, and practical applications. Traditionally, the simulation models were created manually but with the advent of process mining techniques, it is a lot faster to arrive at a simulation model. As an instance, simulation can help to estimate the benefit of an expected process redesign or to predict flow times for an increasing number of incoming cases. It is also regarded as a useful tool to gain insights into the operation of systems. In addition to this, generating simulated logs from a CPN model can be very useful to evaluate the performance of process mining algorithms. This project is expected to have an overall positive impact on the people working in the process mining domain.

3 Feasibility Study

3.1 Theoretical View

We make use of process mining techniques on the information (event logs) recorded during the process enactment to (semi-)automatically discover a simulation model. Typical formats to store event logs are XES and CSV. We will apply several process mining algorithms to gain insight into different perspectives.

We start with a control-flow discovery algorithm such as Alpha Miner to create a process model (Petri net model). This is followed by a decision point analysis using decision trees to discover decision rules. A performance analysis is then carried out to enhance the process model with information. Furthermore, a role discovery algorithm is applied to the event log to group resources into roles. Finally, the mining results enhanced with data, performance, and organizational characteristics are integrated into one comprehensive simulation model.

The different perspectives are joined into a single model and represented as a Colored Petri net (CPN). From a theoretical point of view, our project is feasible.

3.2 Technical View

We use Python 3.8 as the programming language and our project is mainly implemented using open source python libraries/frameworks like pm4py, SimPy, and Flask and deployment using Docker containers.

The pm4py library is the leading open-source process mining platform written in Python and is jointly developed by the Process Mining group at PADS with the Fraunhofer Institute for Applied Information Technology (FIT). The library allows importing and exporting event logs stored in various data formats, implements the best-known process discovery algorithms, and allows visualizing Petri nets.

SimPy is a process-based discrete-event simulation framework based on standard Python. It allows simulations to be performed in real-time or by manually stepping through the events.

Flask is a web application framework written in Python. This makes it easier to build reliable, scalable, and maintainable web applications. We use this framework to handle HTTP requests. Docker allows developers to build lightweight and portable software containers that simplify development, testing, and deployment. Therefore, from a technical point of view, our project is feasible.

3.3 Technical Risks

Table 1: Table of identified technical risks

Risk No.	**Risk Description**	**Likelihood**	**Impact**	**Planned Actions to Reduce Risk**
1.	Unclear and unrealistic project designs and API specifications	Medium	High	Encourage project team to take up relevant knowledge sessions on designing an architecture, design patterns and API specifications for a project.
2.	Project quality	Medium	Medium	Ensure all features have test cases and are well tested.
3.	Knowledge of technologies used (eg: Flask, pm4py)	Medium	Medium	Provide knowledge transfer sessions within the team.
4.	Lack of project related resources (eg: event logs, infrastructure)	Low	High	Procure all the project related material required before planning for the project implementation.
5.	Integration risk	Medium	High	Involve people early in the process. Be clear and transparent about what is changing in the project at all stages.

3.4 Non-Technical Risks

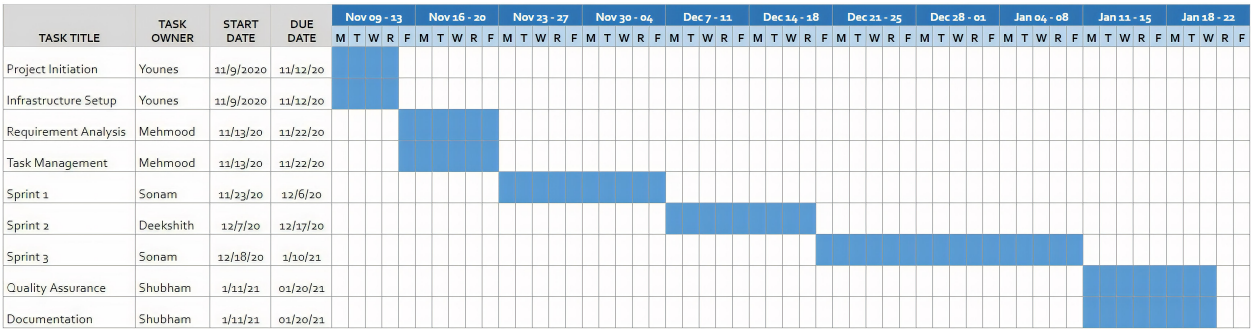
Table 2: Table of identified non-technical risks

Risk No.	**Risk Description**	**Likelihood**	**Impact**	**Planned Actions to Reduce Risk**
1.	Missing, unclear requirements of exactly what is needed	Medium	High	Perform proper planning, read relevant documents carefully, and communicate with the project supervisors frequently.
2.	Estimation and scheduling	Medium	High	Get a solid understanding of the technical prerequisites before story pointing and estimating efforts.
3.	Project schedule is not clearly defined or understood	Low	High	Organize regular meetings with the project team so that they understand the plan so that the likelihood of missing a deliverable is reduced.
4.	Employee turnover	Low	Medium	Ensure every team member gets interesting work and rotate the responsibilities, if possible.
5.	Disputes between the team members (due to different thoughts, expectations between different people)	Low	Low	Conduct meetings on a regular basis and let all the team members discuss the issues openly and a relevant solution is provided as soon as possible.

4 Project Plan

4.1 Project Activities

Figure 1: Gantt chart illustrating the project schedule and activities.



4.2 Project Deliverables

Table 3: Project deliverables and their responsible individuals

Deliverable	**Responsible Person**	**Due Date**
Project Initiation Document	Younes Müller	12.11.2020
Requirement Analysis Document	Mehmood Ahmed	22.11.2020
Sprint 1 Product & Documentation	Sonam Chugh	06.12.2020
Sprint 2 Product & Documentation	Deekshith Shetty	17.12.2021
Sprint 3 Product & Documentation	Sonam Chugh	10.01.2021
Final Product & Documentation	Shubham Balyan	20.01.2021

5 Project Team

Table 4: Project team organization

Name	**Role**
Younes Müller	Project Manager
Mehmood Ahmed	Scrum Master
Shubham Balyan	Product Owner
Sonam Chugh	System Architect
Deekshith Shetty	Quality Engineer

References

[1] A. Rozinat, R.S. Mans, M. Song, and W.M.P. van der Aalst. Discovering simulation models. In *Information Systems*, pages 305 – 327, 2009.

[2] Celonis. *Process Mining.* `https://www.celonis.com/process-mining/what-is-process-mining/#why-is-process-efficiency-so-important`.