Rheinisch-Westfälische Technische Hochschule Aachen
Chair of Process and Data Science
Prof. Dr. Wil van der Aalst

Process Conformance Checking in Python in SS 2021

Alignments on NFA(s) in Micropython

Group:3

Project Initiation Document

30.4.2021

supervisor: Alessandro Berti

Contents

List of Figures

1 Introduction

The goal of this project is to implement conformance checking in python with no dependencies. We aim for as few hardware and software requirements as possible. The project initialization is documented in this document. In Chapter 2 there are two business cases given that highlights the benefit for business, that motivate this project. Chapter 3 describes the feasibility of this project. Here the theoretical background is given and insides into the realisation like the used programming language are given as well. A more detailed plan for the project can be found as a gantt chart in chapter 4. The gantt chart shows how we will implement agile development in our work process. In chapter 5 the used tools are listed and explained. A personal description of each team member and a description of the assigned roles are given in chapter 6. Finally, in chapter 7 all the phase reviews of each team member are presented.

2 Business Case

2.1 Business Case - Bigando

Bigando is an e-commerce company which operates in end to end supply chain of manufacturing, storing and delivering fashion products across Europe. As per our initial discussion with them, they are facing challenges in meeting the customer demand in an efficient way particularly due to a major shift towards digital world because of Covid19 pandemic. The following business cases were identified where we can be of great help to improve their digital footprint, repair the weak links in their supply chain and ensure smooth and efficient business operations for them:

2.1.1 Logistics and Inventory Management

Product delivery time has always been an improvement area for them. Unexpected delays in it is an important parameter for any e-commerce business and can result in increased customer frustration and higher chances of losing customers to competition. Knowing where exactly the product is located in which warehouse is not enough. A constant check to ensure that the defined processes are being followed on the optimal level is equally important. At times due to steps that involve human intervention the actual process followed is different than what should have been the case due to which such delays are faced. Our product will gather all the data points and by applying various process mining techniques and conformance checking methods will not only ensure that the deviation from the optimal business process is minimum but will also recommend improvements in the defined business processes to make them more efficient in terms of cost and time.

2.1.2 Online Business Platform

First impression is the last is no more limited to physical world only. The first experience of interacting with the platform is what defines success for any online business. Better user experience, simple interfaces and quick problem solution helps improving the Click through rates resulting in more conversions and gaining good word of mouth from customers, the importance of which is self-explanatory in the below mentioned stats [5]:

- Nielsen report that 92% of consumers believe suggestions from friends and family more than advertising.

- Beyond friends and family, 88% of people trust online reviews written by other consumers as much as they trust recommendations from personal contacts.

- 74% of consumers identify word of mouth as a key influencer in their purchasing decisions.

- When specific case studies were analyzed, researchers found a 10% increase in word-of-mouth (off and online) translated into a sales lifts between 0.2 – 1.5%.

Gaining customer loyalty is the key to success in this digital world. More satisfied customers automatically mean more product ambassadors which results in an increased Customer Base that translates in increased revenues for the business which is what our product aims for. It will collect the data from the entire supply chain, consolidate it for applying process mining and conformance checking techniques to improve the customer purchase funnel by recommending improvements in existing business processes and defining new processes wherever and whenever required. As per the high level working on our initial discussions with the client, integrating our product will result in a significant uptake in both; the customer base and Bigando's revenues, which can be seen below:

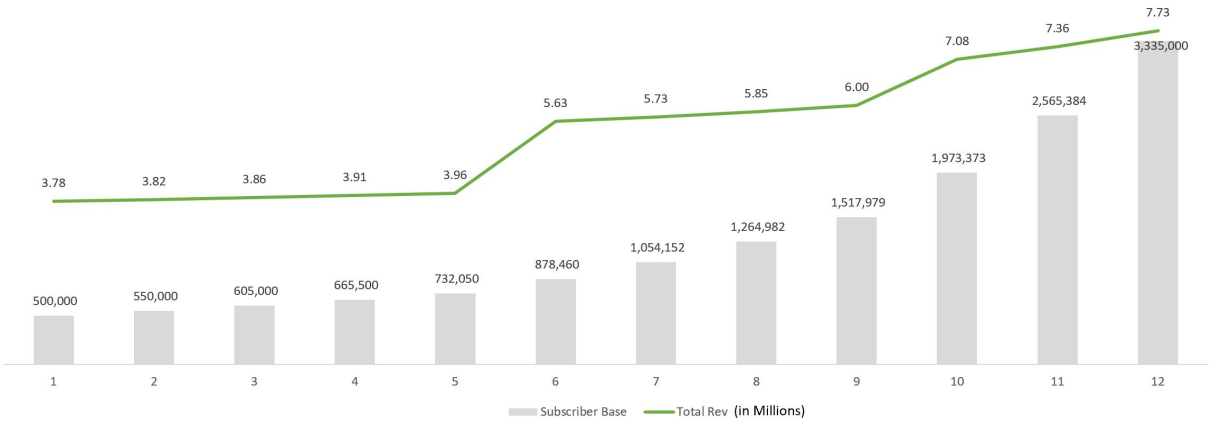

Figure 1: Subscribers and Revenues Prediction for the first year

Detailed Business case working can be seen here.

2.2 Business Case - Consulting in the field of Manufactoring

2.2.1 Client

We choose one business cases of our client company AAA GmbH to demonstrate how our low dependency conformance checking project improves their business by increasing existing revenue streams and also creates opportunities for new revenue streams. Company AAA GmbH is a consulting company that offers business process improvements for its customers using process mining. They are familiar with working with PM4PY to deliver their services. OURPRODUCT enables them to bring their loved conformance checking features known from PM4Py to low dependency microcontroller systems. In the following, a business case from the field of manufacturing is presented.

2.2.2 Manufacturing

BBB GmbH is a client of the consulting company AAA GmbH. Like many manufacturing companies, they have multiple workstations with machines that create events. At those workstations, products are assembled. The overall process starting from ordering a product to shipping the product has already been analyzed by AAA GmbH using expensive hardware that can run PM4Py and its dependencies. This leads to a faster delivery time and fewer errors in the overall process. Company BBB GmbH wants those same benefits for their assembly process directly at each workstation. The assembly process can be complex and every error made here costs a lot of money because of the loss of time and in the worst case the loss of material when the product has to be produced again because of an error in the assembly process. AAA GmbH is happy to be able to serve those needs as a lucrative follow-up sales using OURPRODUCT. As OURPODUCT offers conformance checking with no dependencies and low hardware requirements, AAA GmbH is able to apply conformance checking on a microcontroller directly to each workstation to monitor the assembly process. Therefore assembly errors can be detected way earlier which saves a lot of money because the later an error is detected the more money is already wasted. So BBB GmbH decides to buy the service offered by AAA GmbH using OURPRODUCT because

- The cheap hardware requirements allow for usage at scale
- They save money because of fewer and earlier detection of errors in the assembly process

Company AAA GmbH decides to use OURPRODUCT to serve customer BBB GmbH because

- low hardware and software dependencies allow them to generate follow-up sales that are cheap in marketing and therefore highly profitable
- they can reuse their knowledge and have a low learning curve which makes the development process cheap

3 Feasibility Study

3.1 Theoretical and Technical Feasibility

Conformance checking is an approach which is used for performance analysis. It either performs a token based replay over the model and mark the deviation or compute alignments to calculate the performance of the model [1].We in our project aim to implement conformance checking using NFA's as discussed in [2].Process trees are defined using regular expressions which can be transformed into NFA's. Regular expression is a technique developed in formal language theory where we usually search for a pattern specified by a sequence of characters. Many programming languages such as Perl, PCRE, Python, Ruby, Java use recursive backtracking for the implementation of regular expressions.

3.1.1 Performance

The recursive backtracking approach used by many programming languages for regular expression implementation is very simple but can be extremely slow. To outperform these

previous approaches we will focus on implementing regular expressions using NFA(Non Deterministic Finite Automata) with Python as the programming language. NFAs are guaranteed to work dramatically fast and with more consistent speed as explained in [3]

3.1.2 Quality

The feasibility of working with NFA has been already verified in [2] and keeping in mind that it will outperform the existing technique in python for the creation of the model using the regular expression[3] we can expect fruitful results with conformance checking as well.

3.1.3 Implementation

We already have Regular Expression implementation using NFAs in C/Java language which can be helpful while implementing them in Python. An approach has already been proposed in [2] to calculate precision using NFAs which can be kept in mind while implementing the conformance checking technique using alignments.

3.2 Use Cases

3.2.1 Programming Language and Platform

In this era of technology, Computers have revolutionized the fields of research and Science. Many problems deemed impossible a few decades back have been solved now. The advent of Programming languages has enabled Domain experts to use the exceptional computational and memory resources of the computer and bring their research to life. However, this is not that simple as it seems, to achieve this one must be an expert of his domain as well as he must have a convincing Programming knowledge. As it is not that common, the solution to this problem requires two experts of their relevant fields. It is also not viable as two experts are collaborating for a single outcome.
To overcome this, many user-friendly and easily interpret-able programming languages have been introduced. Python overcomes the dependency of a domain expert on a Computer Scientist with easy scripting and powerful libraries for almost every scientific Discipline. Similarly, MicroPython is a software implementation of a programming language compatible with Python, developed in C, thus giving more efficient solutions for Microcontrollers. It is primarily used in systems where we have limited processing and memory resources. We do not have any such Resource limitations; hence we have opted to pursue with Python.
The advantages of using python are listed as follows,

- Simplicity
  Easy and meaningful translation of problems into the programming instructions without prior knowledge in computer science.

- Availability
  Vast range of well-documented, domain-specific libraries.

- Versatility
  Python can be used for almost everything, could it be a complex mathematical Algorithm or interactive user interface application.

3.2.2 Event Log Analysis

We will be working on an event log from a case study of The Seoul National University Bundang Hospital [4]. The hospital is one of the biggest and state of the art hospitals of Korea, having capacity of more than 1300 beds. This hospital deals with 4000+ patients on daily basis. These numbers result in a very complex process model and it could be a perfect example of checking whether the patient follow the standard process model of the hospital or not. The standard model was developed at the time when the number of patients being treated was very low as compared to the recent times. To find out if the standard model was still up to date, the Ulsan National Institute of Science and Technology (UNIST) applied methods of conformance checking to make a comparison between the two models.
Following are the key properties of the event log.

- 120,000 cases (Patients handled)
- 700,000 events (Total activities performed for these patients)
- 15 Different tasks

Together with the medical professionals, the following questions were posed:

- Does the standard model explain actual patients' movements in the hospital?
- How much of increase in patients is allowed?
- Are the process patterns different depending on the patient types?

The results were as follows:

- The comparison of Event Log and the process model resulted to 89%, which indicates that the process was being followed to a good extent.
- The analysis resulted in a simulation model, which showed that by increasing the patient count by 10%, the consultation time got increased pretty much, therefore it was indicated to not to increase patient count to that extent.
- The analysis showed the differences in process patterns between new and returning patients. It was observed that new patients stayed longer than returning patients. A smart Health care application was developed from those pattern Analysis, which Patients are able use via smartphone to find their route that is recommended by the result of pattern analysis.

After seeing the results, the medical professionals were impressed and showed a deep interest in the study.

4 Project Charter

The Gantt Chart of the our project is greatly influenced by the semester dates and the phases that have been prescribed beforehand. Each phase is encompassed in working towards a milestone of the project represented as a black bracket over the phase. Within the sprints, more specific programming milestones are included that have to do with the

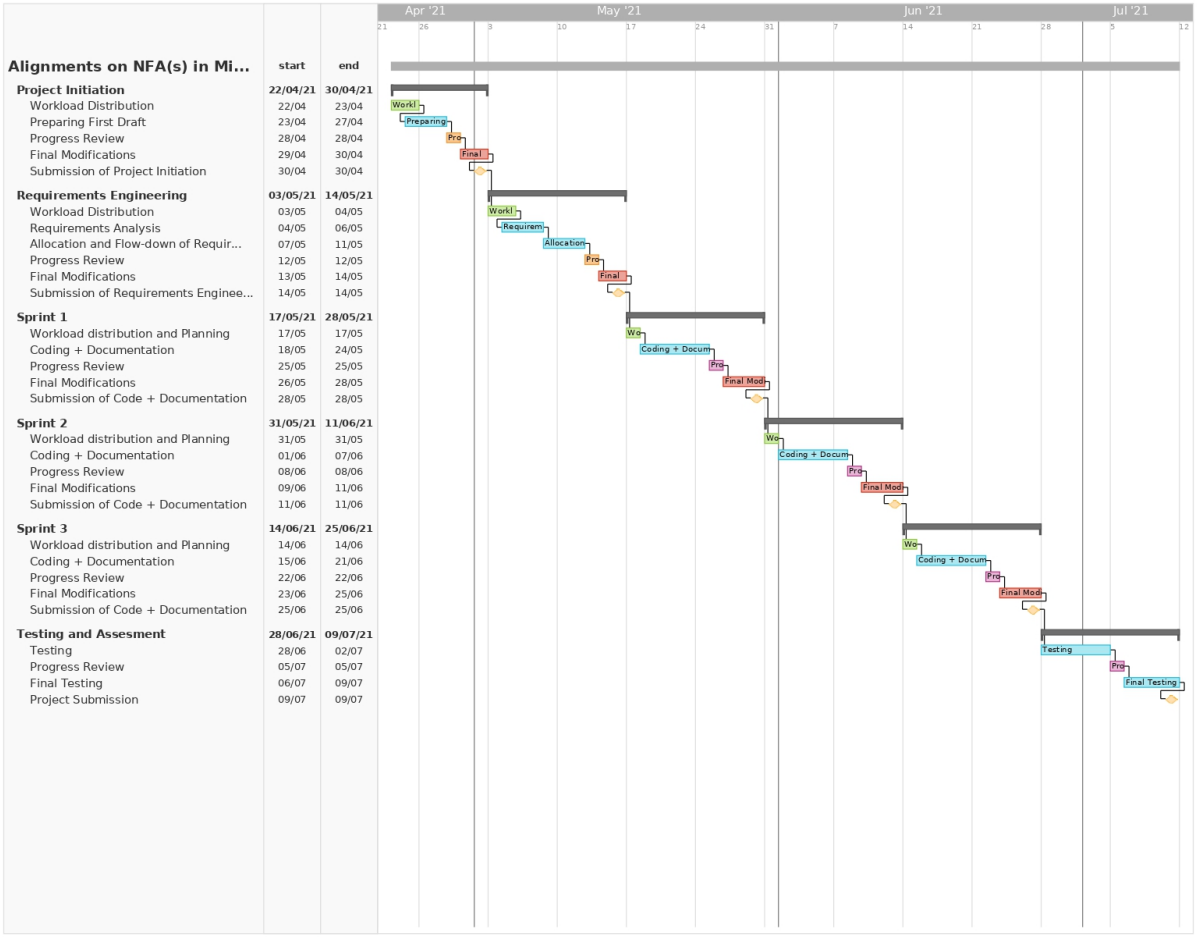

Figure 2: Gantt Chart of the project with milestone phase dates.(TeamGannt)

actual product. Each phase is structured in a similar way to make it easier for all members to get into a regular and known schedule. Typically, at the start of a phase,the team will gather and distribute the overall tasks and have a discussion of the overall plan, internal deadlines and final deliverable of the phase. After distribution, each team member will have time to work on their tasks until a progress review is done on the work done. The progress review's purpose is to see at what stage every member is, to get feedback on the already completed work, to bring up problems that need solving and to keep up to date on the overall progress. Of course, in addition to the big progress review, the team members will conduct smaller scrum meetings via different communication means to keep each other updated.After the last of the bigger progress review before the end of a phase, the final modification of the phase have to be made if there are any comments to finalize the deliverable. The deadline for the final modification is a few hours before submission to make use of time before submission. On the final day of submission, a final check of the phase will take place to make sure that any smaller errors are caught before submission. It is important to stress that big changes will not be made on the final day of each phase as this may cause more problems.

5 Tools

5.1 Trello

Trello is a project management tool, which supports essential project management activities. At its core, it is organized as a project board that documents the current work that has been done or needs to be done in a given time frame as a note. With the possibility of dividing it into several sub notes. This allows for a dedicated distribution of work that supports our agile approach in delivering a high-quality product.
Link to the Trello Board can be found here.

5.2 Git&GitHub

Git is an open-source version control system, which is de facto the industry standard for distributed version control system that is already included in many IDE´s. GitHub is a version control using Git. It offers the distributed version control and source code management (SCM) functionality of Git, plus its own features. It provides access control and several collaboration features such as bug tracking, feature requests, task management, continuous integration and wikis for every project.GitHub launched a new program called the GitHub Student Developer Pack to give students free access to popular development tools and services. Therefore, it will be used in this project in combination with the online service GitHub to manage our software with multiple people.
Link to the Git Repository can be found here.

5.3 IDE-PyCharm

PyCharm is an integrated development environment (IDE) used in computer programming, specifically for the Python language.It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems (VCSes), and supports web development with Django as well as data science with Anaconda.Also PyCharm as it is a cross-platform, with Windows, macOS and Linux versions.So PyCharm is very important for the success of our project

5.4 PM4Py

PM4Py is a python library that supports (state-of-the-art) process mining algorithms in python. It is completely open source and intended to be used in both academia and industry projects. PM4Py is a product of the Fraunhofer Institute for Applied Information Technology.This library provides a simplified interface for conformance checking, which has two important techniques: token-based replay and alignments. Therefore, PM4Pyis fundamental to complete our project.

5.5 Overleaf

Overleaf is a collaborative cloud-based LaTeX editor used for writing, editing and publishing scientific documents. It partners with a wide range of scientific publishers to provide official journal LaTeX templates, and direct submission links. We used Overleaf to work collaboratively on our project documentation and produce a well organized documentation.
Link to the document can be found here.

6 Team

6.1 Personal Description

6.1.1 Lukas Liß

I am a computer science master student at RWTH Aachen University. I gained practical experience in developing and managing software during the last 4 years because I founded a startup to improve appointment accessibility in the German healthcare system. In that context, I learned a lot about frontend and backend development with JavaScript. Moreover, it familiarized me with agile development and production environments that require robust unit testing. In addition, I have some practical experience when it comes to applied process mining in the industry, as I work part-time as a developer for a consulting company that started its first approaches to utilize process mining a year ago. This was also when I attended the lecture Business Process Intelligence that introduced me to process mining. That brings me to my theoretical background and interests. I have already attended Business Process Intelligence and Introduction to Datascience. This semester I am attending Advanced Process Mining. I hope and think that this background allows me to make valuable contributions to the project we perform as a group. This project is interesting for me because I am looking forward to working with a team to come up with a solution for the given conformance checking task. As I worked mostly with C# and JavaScript in the past I only conducted some basic knowledge in python from some hobby projects but that should be enough to enable me to transfer my knowledge from other programming languages.

6.1.2 Mahmoud Emara

Currently undertaking a master's degree in Data Science and having accomplished a bachelor's degree in Computer Science, I have had some experience in software engineering practices and data science disciplines such as process mining. My main motivation in the field is to deepen my knowledge and increase my experience creating efficient software not only in performance, but also in terms of data use. During my studies, I did pursue some of the data science related courses, I had the opportunity to demonstrate some of the following courses by theoretical and practical implementation of the most popular and reliable algorithms: Machine Learning, Algorithmic foundation of Data Science , Natural Language Processing, Process mining mainly using Python. For the two years before starting my Master study, I have been a software engineer developing iOS applications for some big companies. I lack experience in the research field, thus being assigned the role of understanding the theoretical aspect of the project will strengthen my weakness. Lastly, my interest in process mining was sparked in the lecture introduction to data science and business process intelligence, which I have taken successfully. As I have little experience in applying process mining in real industry projects I am excited to use this opportunity to grow and learn the necessary skill set to comfortably apply process mining techniques.

6.1.3 Syed Faizan Hassan

I'm a Computer Science graduate and currently doing my Masters in Data Science at RWTH Aachen University. I have more than two years of professional experience performing data analysis for the industrial sector. I served as an Analyst Software Engineer

at an off-shore setup of a US-based firm: Afiniti.An average day at Afiniti, for me, involved data extraction and data analysis via established SQL platforms. My work on statistical analysis allowed me to expand my knowledge base and contribute to the company's decisions for call modelling. During my masters I undertook the courses related to Process Mining which directed my interest to Process Discovery and Conformance Checking. Then, I also participated in a seminar on 'Machine Learning Applications in Process Mining'. My main area of interest is finding out the bottlenecks in a process and to find out ways of improving processes. I possess a lot of theoretical knowledge about process mining and now aim to learn more about the practical applications of those methods,I am proficient in programming with Python and also well aware of the Process Mining tools such as Disco, Prom, Rapid Miner.As the work in the lab involves working with the pm4py library and Python as the programming language it would serve as a great platform for me to not only deepen my knowledge, but also equip myself with more skills and experience in the field of process mining.

6.1.4 Asad Tariq

I am a Computer science graduate and currently doing my master's in data science at RWTH. I have worked for more than 3 years in Software Development industry and thus I am quite familiar with the Software Development cycle. I am highly skilled in programming and fully aware of the development process. I have worked in development teams under sprint model therefore I am familiar with Agile Software Development Life Cycle. At I9, under Prof. Dr. Wil van der Aalst, I have studied courses related to process mining (Introduction to data science, Business Process Intelligence and Advance Process mining). I am fully aware of the process discovery algorithms (Alpha miner, Inductive miner, Heuristic miner, DFG, EST-miner) and have them implemented in different tools (Jupyter notebook using python, Prom, Disco, Celonis, pm4py). I have also taken Seminar in process mining "Machine Learning Applications in Process Mining (SE)" which has made my knowledge vaster related to process mining. I believe that this lab course is an excellent opportunity for me to bring my theoretical knowledge to Practice. I am expecting to enhance my Python skills and mainly the knowledges about process mining through working in a group to solve practical issues.

6.1.5 Mina Khalid

I am a Data Science Masters student at RWTH Aachen university. With my Bachelor's in Computer Science and an MBA in Marketing, I have an industrial experience of around 1 year as a Software Developer and more than 3 years as a Product and Business Analyst. Working on the entire chain of product delivery pipeline has helped me experience various challenges at each step and how to think from a 360-degree perspective. The paradigm shift towards automation, predictive modelling and role of Data in gaining insights for coming up with Data driven solutions compelled me to pursue a degree in this field. The course Introduction to Data Science that I took last semester ignited my interest in the domain of Process mining and I am really looking forward to learning more about it by a hands-on learning approach in this lab.

6.2 Roles

6.2.1 Product Owner: Mina Khalid

As a Product Owner, Mina will be working with the Client and other stakeholders to define and prioritize user stories in the Team Backlog to streamline the execution of program priorities while maintaining the conceptual integrity of the Features.

6.2.2 Scrum Master: Asad Tariq

As a scrum master, he is responsible for the organizational environments of his team to create a high-value product. He will be planning the Sprints along with the team members and plan about the Product releases. He must manage the Trello Board and GitLab of the project, assign tasks to each member and help the team to approach their work without difficulty. He is responsible for the smooth development of the product, its deployment, and its Release.

6.2.3 Tools Administrator: Lukas Liß

Lukas is handling the role of the Tool Administrator. He has to know how to utilize the used tools like trello and git such that they support the workflow of the team. In addition he has to familiarize everyone that has not used the tools before with the tools. He is the one to talk to whenever there is the need for a tools that is not yet used.

6.2.4 Theory Expert: Asad Tariq and Syed Faizan Hassan

Asad and Faizan would be handling the role of Domain expert in this project. They must have knowledge regarding the existing algorithms being used in this domain. They should be able to help the team understanding the algorithm and related limitations.

6.2.5 Software Expert: Mahmoud Emara

The Software Lead is responsible for being an expert behind the software requirements of the project and making sure that the project follows it. This means that the Software Lead should be able to answer any relevant technical question behind the practical implementation of the software, take responsibility for ensuring application of best software practices and finally making sure the documentation is sufficient for each part.

6.2.6 Quality and Test Management: Syed Faizan Hassan

As the Quality manager, he is responsible for the organization of tests and assurance of quality.

7 Phase Review

7.1 Asad Tariq

I found the whole team very energetic and ambitious. I am very much impressed by the teamwork and honesty everyone has shown. We have met many times on online meetings and discussed the deliverable thoroughly. Everyone showed a great level of responsibility and have done their respective tasks in time. Now that every team member has found their

positions in this project, and they got used to cooperation tools like Trello and whatsapp, I am expecting more efficient teamwork in upcoming phases of this project.

7.2 Mahmoud Emara

This first stage was our first challenge. I think we handled it well. We were able to lay the groundwork for our project by assigning a specific job to each of us and discussing the difficulties that lie ahead and how we should tackle the project.Overall it wasn't that hard to come up with ideas for the project initiation as we had a reference document that could be used as inspiration for the content and also there exist many resources online.The way I see things is that if we stick to this path, we will be able to deliver an efficient bug-free software project.

7.3 Syed Faizan Hassan

I am very happy and satisfied with the way every person owned their task and completed it according to deadlines we decided in our internal meetings. Even though its just the first deliverable we had the internal meeting thrice and I am very impressed that everyone took out the time to join all the meetings despite the fact that the exams from winter semester are still going on. The meetings were very helpful as each one of us tried to pitch in their ideas and helped clear the misunderstandings regarding the requirements and task we are supposed to perform. Good thing was that we did not only focus on our part but we as a Team reviewed everyone's part so that we can help incase anyone is facing difficulties in completing their respective part. Another good thing was that everyone worked actively on their task and maintained the dashboard in parallel. It was not needed to remind anyone to update their work on the Trello dashboard or on the overleaf document. I am very excited to work with this team in the upcoming phases of the project and hope that we will keep on working this way.

7.4 Lukas Liß

We used the first sprint to get to know each other and set up the project. I really enjoy working with the team. Everyone delivered all artifacts on time and I think we worked well together when reviewing each other's work. We have different background knowledge that we have combined well, especially in the zoom meetings we had. I enjoyed the way we all shared our knowledge to kick off this project. The way we instantiated the project allows us to be productive in the following sprints, which is something I am very much looking forward to. I know that there will most likely be tough problems ahead of us, but I am confident that we can solve them, by keeping up the effort and work we spend on this sprint.

7.5 Mina Khalid

The project initiation phase served as an ice breaker for the entire team. We got to know each other better and I am really impressed by how fast everyone has jelled in. I find the group mates quite supportive and open in terms of discussions and welcoming ideas. Different backgrounds and expertise have added more value in each discussion we had so far. Everyone tries to propose better alternate ways that add an entire new perspective

for doing things in a better way. If we follow the same pace and attitude, I am certain we would be able to pull off a successful project delivery.

References

[1] Wil van der Aalst. Data Science in Action. Springer Berlin Heidelberg, 2016.

[2] Leemans, S.J.J., Fahland, D. van der Aalst, W.M.P. Scalable process discovery and conformance checking. Softw Syst Model 17, 599–631 (2018). https://doi.org/10.1007/s10270-016-0545-x

[3] https://swtch.com/ rsc/regexp/regexp1.html

[4] Ulsan National Institute of Science and Technology. Process Mining at Seoul National University Bundang Hospital, 2014. https://www.tf-pm.org/resources/casestudy/process-mining-at-seoul-national-university-bundang-hospital.

[5] Reference for Word of Mouth Marketing Statistics: https://www.bigcommerce.com/blog/word-of-mouth-marketing/what-is-word-of-mouth-marketing