July 15th, 2018

**Karim Hersi**

1350 Ogilvie Road

Ottawa ON, K1J 9J6

613 716 1516

karimhersi@cmail.carleton.ca

**Howard Schwartz**

Department of Systems and Computer Engineering

Carleton University

1125 Colonel By Drive K1S 5B6

Dear Mr. Schwartz,

The following report titled "Software Engineering Work Term Report" will discuss my experience at Morgan Stanley in Montreal, Quebec. This was my fourth co-op work term for the course SYSC-3999-A. The work term took place during Summer 2018. The placement was supervised by Kenichi Sakamaki in the CFT division. My Position title was “Technology Summer Analyst”.

In this report, I will provide my initial objectives before the work term. I will discuss the purpose of the report, followed by a brief overview of the organizational context. I will also discuss my experiences, accomplishments, and challenges faced throughout the course of the term.

I acknowledge that this report adheres to the guidelines set for work term reports. I also acknowledge that the contents of this report are my own work. Thank you for taking the time to read this report.

Sincerely,

Karim Hersi

100865348

## Acknowledgements

I would like to thank all my coworkers at Morgan Stanley for providing such an enjoyable and rewarding co-op work term experience. I would like to give a special thank you to my supervisor and team lead, Kenichi Sakamaki for his guidance and mentorship. Finally, I would like to thank all my team members who ensured that I had all the required understandings for my responsibilities.

## Executive Summary

I completed my fourth co-op work term at Morgan Stanley. I worked in the office located in Montreal, Quebec. My team was responsible for automating business processes that required Machine Learning techniques. My role was to assist with developing features and improve the accuracy of named entity recognition (NER). I developed NER models, implemented feature requirements and contributed to the design process. My initial objectives were to gain exposure in Machine Learning while improving my skills as a developer.

**Morgan Stanley**

**SYSC 3999**

**Software Engineering Work Term Report**

Abdikarim Hersi,100865348

**Work Term Report #:** 2 **Email:** karimhersi@cmail.carleton.ca **Supervisor:** Kenichi Sakamaki, K.S.

**Department:** CFT

**Term:** Summer, 2018

## Table of Contents

Contents

[Acknowledgements 2](#_Toc520049094)

[Executive Summary 3](#_Toc520049095)

[Introduction 6](#_Toc520049097)

[Organizational Context 6](#_Toc520049098)

[Objectives 6](#_Toc520049099)

[The Nature of the Work 6](#_Toc520049100)

[Challenges and Solutions 8](#_Toc520049101)

[Contributions 9](#_Toc520049102)

[Relations to Academic Studies 9](#_Toc520049103)

[Career Development 9](#_Toc520049104)

### Introduction

The purpose of this report is to convey the academic, professional and interpersonal aspects of my co-op work term experience. This report will discuss my learnings, accomplishments, acquired experience and challenges encountered. The report is divided into three sections. The first section will detail the organizational context. In this section, a brief overview of my employer, department, and product will be provided. The next section will detail my work term experience. This section will discuss my initial objectives and assigned duties at Morgan Stanley. I will also include my accomplishments, challenges faced, and solutions to these challenges. The concluding section will be a reflection about my work term. I will reflect on my contributions and how my work placement connects with my academic studies.

### Organizational Context

Morgan Stanley is an investment banking and financial services company. Their services include sales, trading, wealth management and investment management. Morgan Stanley is headquartered in Manhattan, New York City. The company has offices in more than 42 countries worldwide. The Morgan Stanley office in Montreal is focused on technology and employs over 1200 technologists. It provides business solutions using technology in various areas including cloud engineering, cybersecurity, artificial intelligence, and electronic trading.

### Objectives

The objectives that I had set out for myself at the start of the term were to improve my skills as a developer and gain exposure with Machine Learning for named entity recognition. Artificial Intelligence is a rapidly growing field within technology and Machine Learning has proven to be an effective solution to a variety of common problems. It is becoming increasingly important for software developers to be familiar with Machine Learning techniques. An additional objective was to exercise my verbal communication skills. Effective communication is pivotal in delivering software solutions that address the user’s needs.

# The Nature of the Work

The product supported by our team is used to automate manual processes at Morgan Stanley. One of its responsibilities is to extract and interpret relevant fields from scanned pdf documents. To accomplish this, the scanned files were sent to an OCR tool which could extract the text from these documents. OCR (short for Optical Character Recognition) is the process of converting images with text into machine encoded text that is suitable for data-processing. The input for the OCR can be a scanned document or an image, and the output should be the text within the image. After the text is received from the OCR, the next step is to perform named entity recognition (NER) to extract specific fields of interest within the document. NER is the process of identifying and classifying entities in unstructured text. Example of entities include dates, locations, names, and addresses.

|  |
| --- |
|  |

#### The Work Environment

The code base that I was working on was written in the Java programming language. The company used the *git* version control system to record changes and manage different versions of the code base. The application used the Apache Camel framework to facilitate enterprise integration. Finally, the Stanford NLP Java library was used for Machine Learning and named entity recognition.

#### The Work Experience

The work term began with the task of setting up my work environment. This included configuring desktop settings, installing software and many more set up procedures. It also included getting access to the project repository and forking the project using Git. Once the environment was set up correctly, the second stage was to gain a broad understanding of the software’s business logic. Fortunately, there was a lot of well written documentation to assist with this stage. I had no trouble understanding the business logic and requirements after reading up on all the documentation. I also had to read and gain exposure with the Apache Camel framework. I spent a few days watching video tutorials, reading the documentation, and talking with team members to get a broad understanding of the framework.

The first task was to ensure that all input files were unsecured before they were sent to the OCR. Our OCR tool does not extract text from secured pdf files. To handle this issue, a program was created that could take a screenshot of a pdf image. This screenshot is then formatted to have the same dimensions as the original scan. The formatted image is sent to the OCR tool instead of the original secured scan. The security removal module was successfully integrated with the rest of the product.

The second task during the work term was to improve the accuracy of the OCR results. Unfortunately, there is no OCR tool today that works at a high-level accuracy. Common OCR errors include the character ‘B’ being interpreted as an ‘8’, and the character ‘O’ being interpreted as a ‘0’. Most OCR tools have an error rate higher than 10%. My task was to use post-processing to reduce the error rate. Post-processing techniques allow the OCR to make errors and then use business logic to detect and correct the errors. For instance, since the entity ‘first name’ is not an alpha-numeric field, we can deduce that the value “Bo8” contains an OCR error and should be “Bob”.

The last portion of the work term was used to create a program that can extract key fields in a company-specific class of documents. A machine learning model that would perform at an accuracy of at least 90% was required. The first step was to generate annotated data that would be used to help train the model. Stanford NLP’s Java library was used to assist with the NER. The library is built internally using linear chain Conditional Random Field (CRF) sequence models. These models predict the probability that a piece of text maps to an entity. The words in the text are classified based on these probabilities. The Stanford NLP library also includes functionality that assists with testing a model. If the model is not performing with an acceptable accuracy, more data can be added, the algorithm can be tuned by modifying parameters, and outlier values can be removed from the training set.

|  |
| --- |
|  |

# Challenges and Solutions

There were a few challenges faced throughout the co-op work term. The first challenge was decrypting pdf files without compromising on image quality. Taking a screenshot of a secured file results in the screen shot to have less pixels compared to the original. The lower the image quality, the less accurate the text from the OCR will be. To minimize image quality loss, a Java library that specialized in image enhancement was used.

Another challenge was getting the program to effectively classify dates. There are many different commonly used notations for dates. In addition, most documents had multiple dates and each date needed to be classified correctly (i.e. date of birth versus current date). The solution to this problem was providing a diverse and thorough training set and using a feedback loop system that would allow the machine to learn from past mistakes.

# Contributions

I was able to implement and integrate a module that could ensure that files entering the OCR were not secured. I was able to introduce post-processing which detected OCR errors from files and ensured that the documents were correctly interpreted. I also developed a machine learning model that can recognize entities of interest. Finally, I contributed to the process of testing and deciding which OCR tool was the most suitable for our company’s needs.

# Relations to Academic Studies

The tasks at my co-op work term were directly related to what I learned at school. There were applications of Artificial Intelligence, Object Oriented Programming, Software Design Project and Algorithms and Data Structures. Strong understanding of these courses was essential for doing the tasks that were assigned. There were uses of Database Management Systems as well. Although I was previously exposed to these materials in school, it took some time being able to effectively apply them in a work environment.

# Career Development

The work term helped strengthen many of my technical and non-technical work skills. From a technical standpoint, I was able to gain exposure working with commonly used Machine Learning technologies. I was introduced to various industry design patterns for managing information technology systems. From a non-technical standpoint, I was able to collaborate daily with others, many of which worked in entirely different fields. These skills are very difficult to gain in an academic environment.

#### Summary

I achieved the objectives that I had set out at the start of the work term. I received some exposure to industry standard software development. I learned a few tools and technologies that are very relevant to this field. I gained a lot of insight on what technical and non-technical skills I need to work on to be successful in a software work environment.

**Appendix A - References:**

1. “Efficient Training Methods for Conditional Random Fields”. [Online]. Available: http://homepages.inf.ed.ac.uk/csutton/publications/sutton-thesis.pdf. [Accessed: 12 - July - 2018].

1. "An Introduction to Conditional Random Fields for Relational Learning ". [Online]. Available: https://people.cs.umass.edu/~mccallum/papers/crf-tutorial.pdf. [Accessed: 12 - July - 2018].

**Appendix B - Glossary:**

**Machine Learning:** A software or program used for the management of changes to computer programs.

**Stanford NLP:** A software or program used for the management of changes to computer programs.

**Code Base:** A collection of programming source code used to develop software.

**Classification:** A software or program used for the management of changes to computer programs.

**Annotated Data:** A software or program used for the management of changes to computer programs.