**Project Charter**

**Digitizing a Half-Century of Financial Data**

**University of Colorado Boulder**

**Computer Science Senior Capstone Team**

**Leeds School of Business**

**Prepared by Leeds Capstone Team Students**

**October 2019**

**Contents**

1. Introduction. 3

Executive Summary. 3

Business Problems/Opportunities. 3

2. Objectives and Scope. 3

Business Objectives. 3

High Level Requirements. 3

Project Scope. 4

3. Project Approach. 4

General Approach – Solution Delivery Process. 4

Assumptions. 5

Project Risks and Issues. 5

Project Changes. 5

4. Project Plan. 5

Key Deliverables. 5

Timeline. 6

Preliminary Cost Estimates. 6

5. Key Stakeholder Roles & Responsibilities. 6

Stakeholder Role/Responsibility. 6

# 

# 

# 

# 

# 

# 

# **1. Introduction**

## **Executive Summary**

The goal of this project is to digitize and gather data from the Industrial Moody’s manuals from 1920-1969. These manuals, which are in the public domain prior, contain accounting, governance and other business information for listed corporations. The main goal of this project is to fully automate the data gathering process. This data will then be made publicly available via a searchable database on the University of Colorado Boulder website.

## **Business Problems/Opportunities**

* With varying levels of familiarity each member of our team possesses in relation to programming skills, we will have to work to bring everyone up to speed.
* With previous work having been done on this project, we must familiarize ourselves with what has been done so that we may build upon it.
* Although we may focus our efforts on a specific range of years for the manuals, we hope our code can be applied to all of the manuals, even after we finish our part of this project.
* Due to the poor quality of some pages in the manuals, it may be difficult to extract all information of interest. In addition, the quality and large volume of the pages will make it difficult to check the accuracy of our results. This problem will be lessened as there are a few”gold standards” that we can compare some of the results to.
* An exciting opportunity will be presented to learn about the history of the financial world in the early 20th century in America.
* We also hope for an opportunity to improve the OCR using machine learning and other programming techniques.

# **2. Objectives and Scope**

## **Business Objectives**

* Create high quality post processed digital images from each page of the Moody's manuals.
* The resulting digital images will be converted to searchable text using a pre-process algorithm we have developed and existing OCR software.
* We will extract information on the management and directors, credit ratings, geographical and other relevant business and financial information from the searchable text.
* This information will be made available in a publicly searchable database.

A project of this scope should allow us to improve each stage of the algorithmic process and provide a clear template that future researchers can use for extracting data from other historical sources without relying on manual hand-collection that has been the standard up to this point.

## **Project Scope**

**In Scope:**

* Extraction of Data
  + Create high quality post processed digital images from each page of the Moody's manuals.
  + Create a pre-process algorithm to get searchable text from our digital images.
  + Extract information on the management and directors, credit ratings, geographical and other relevant business and financial information from the searchable text.
* Database
  + Establish a database of firm-level business information from 1909-1952 from the Moody's Manuals
  + Implement a public UI.
* OCR
  + Tune the OCR to be optimal for the Moody Manuals.

**Out of Scope:**

* Any other business areas beyond those mentioned above
* Our database will not be searchable by images or contain images.
* We will not maintain the project after the Project/School Year is over.

# **3. Project Approach**

## **General Approach – Solution Delivery Process**

* **Vision and Scope:** The project will be divided into several assignments which the sponsor will post throughout the semester. These assignments will relate to various ML, NLP, OCR, and database tasks.
* **Analysis and Design:** For each assignment, we will self-organize into smaller teams of 2-3 people to complete these assignments, in parallel. The members of the team that are interested in NLP/ML will work to improve processes and quality of our data, the ones that are interested in databases will work on setting up and organizing a DB with test data, and those that are interested in OCR will work on methods of converting the images to searchable text.
* **Build and Test :**We actively monitor and report on progress during the build and test phase. We will schedule 1-2 weekly meetings with the sponsor, depending on each sub-team’s availability, to discuss new assignments and/or progress on current assignments.
* **Finalisation:** Since sponsor has idea about how the result will look like, we will do user acceptance testing and work with sponsor in close collaboration, to achieve the common goal.

## **Project Risks and Issues**

* **Time**: We will need to be aware of deadlines to ensure deliverable dates are met on time so that the project progresses at a rate that allows completion of the project by the end of the spring 2020 semester.
* **Accuracy**: Although there are a few gold standards to compare results to, we cannot say for sure if these results will translate to all of the Moody manuals.
* **Programming Background**: No one member of our team posses all the necessary skills to complete this project. We will have to make sure we integrate our knowledge in an effective way to compile a final deliverable that meets the requirements outlined in this document.
* **Information**: Due to the poor quality of some pages in the manuals, it may be difficult to extract all information of interest and ensure accuracy of the content, thus the project might not be successfully implemented and would not be able to establish a database.

## **Project Changes**

* Changes to project scope and requirements will be reviewed by the students with guidance of the sponsor. We will assess the potential impact of the change on schedule and resources, then provide a decision for approval/disapproval. All scope/requirements changes will be documented in the **Project Log**.
* Whenever a change in scope is proposed we will discuss it as a team. If we agree with the change, we will get approval from the sponsor, then add it to the Project Charter and notify the professor and sponsor.

## **Project Log**

Here is where we will note any changes or other events that occur during the course of this project.

## **Timeline**

Deliverable Due Date

Project Charter September 27, 2019

Requirements Definition October 8, 2019

First assignments October 18, 2019

Second round of assignments November 1, 2019

Last round of assignments November 29, 2019

First deliverable December 17, 2019

**Cost**

This project will only require the use of university resources that have already been accounted for in other departments budgets. The biggest resource we will be using is the Summit supercomputer which we have all been provision access on.

**5. Key Stakeholder Roles & Responsibilities**

|  |  |
| --- | --- |
| **Stakeholder** | **Role/Responsibility** |
| **Diego Garcia** | **Project Sponsor** |
| **Lakshya Sharma** | **Project TA** |
| **Jade Wibbels** | **Project Mentor** |
|  |  |
| **Remy Dahlke** | **Student** |
| **Joshua Khoo** | **Student** |
| **Zijun Liu** | **Student** |
| **Kevin Eastman** | **Student - Team Lead** |
| **Will Shanks** | **Student** |
| **Qinglu Sun** | **Student** |
| **Jie Wang** | **Student** |
| **Zuodong Wang** | **Student** |

**BY ADDING YOU NAME BELOW YOU ACKNOWLEDGE YOUR AGREEMENT TO THE CONTENT OF THIS DOCUMENT.**

X: Kevin Eastman

X: Zijun Liu

X: Joshua Khoo

X: Zuodong Wang  
X: Jie Wang  
X: Qinglu Sun  
X: Will Shanks  
X: Remy Dahlke

X: Diego Garcia