Occupation Analyzer

Applied Project Final Report

By

Yuchao Wu

Spring, 2021

A paper submitted in partial fulfillment of the requirements for the degree of
Master of Science in Management and Systems
at the
Division of Programs in Business
School of Professional Studies
New York University

Contents

Declaration

I, Yuchao Wu, declare that this project report submitted by me to School of Professional Studies, New York University in partial fulfillment of the requirement for the award of the degree of Master of Science in Management and Systems is a record of project work carried out be me under the guidance of Dr. Andres Fortino, NYU Clinical Assistant Professor of Management and Systems. I grant powers of discretion to the Division of Programs in Business, School of Professional Studies, and New York University to allow this report to be copied in part or in full without further reference to me. The permission covers only copies made for study purposes or for inclusion in Division of Programs in Business, School of Professional Studies, and New York University research publications, subject to normal conditions of acknowledgment. I further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Acknowledgments

I sincerely thank Dr. Andres Fortino and Dr. Hui Soo Chae for their contribution as sponsors of this project and as mentors during this project. I also want to thank all the instructors in the Management and Systems program who I have taken courses with and learned a great deal.

Abstract

The purpose of this project is to develop a text analytical tool in R that can help student job seekers match their resumes against jobs by a mutual similarity scoring to standard occupations as defined by the BLS O*NET database. Part of the terms of job evaluation for job seekers is to discover their educational preparation for a particular desirable job. The tool scores the resume against a group of jobs and presents the user with the top-scoring jobs to select a target. The resulting top target job table is scored against the occupations. A cosine similarity score is computed between the top job/occupation scores and the resume/occupation. The ranked jobs are then presented to the user as best matched to their resume. Users can interact with this tool on a website that relied on R Shiny. The website contains an occupation database and requires users to upload their resumes and a list of jobs as inputs. The outputs include detailed information about the top 15 jobs and the bottom five jobs. The project expands on existing preliminary work done in Python. The data preprocessing part includes making all text lower cases, removing punctuations, special characters, numbers, English common stop words, extra white spaces, and stemming. It uses the term frequency-inverse document frequency (TF-IDF) algorithm and cosine similarity to measure the similarity among the resume, jobs, and occupations. The final results meet sponsors' expectations and show significant differences between previous work done in Python. Using n-grams, sliding windows, and named-entity extraction are possible methods to improve the tool's performance. I will conduct further research this summer to enhance the accuracy. All project files store in a GitHub repository.

Keywords: Occupation Analyzer, Text Analytic, R

Abbreviations and Definitions

R: R is a programming language and free software environment.

TF-IDF: TF-IDF refers to the term frequency-inverse document frequency. TF-IDF is commonly used to reflect the importance of a word in a document in a collection or corpus.

Cosine Similarity: Cosine similarity is commonly used to evaluate the similarity between two texts' term vectors.

Shiny: Shiny is an R package that allows developers to build interactive websites from R.

Introduction

Background information

NYU CAES and MASY wish to develop additional tools for student job seekers to match their resumes against jobs by a mutual similarity scoring to a standard occupation defined by the BLS O*NET database.

Company Name

New York University (NYU) School of Professional Studies and the Management and Systems program (MASY) is the sponsor company. NYU is a private research university based in 7 East 12Th Street, New York City, NY. The MASY degree is based on a unique curriculum that provides students with experiential learning opportunities to develop strong management and leadership skills and gain a comprehensive knowledge of current information technologies.

Sponsor Information

Dr. Hui Soo Chae is the Executive Director of the Center for Academic Excellence and Support (CAES) at NYU SPS.

Dr. Andres Fortino is the Clinical Associate Professor and MASY ACP Leader at NYU.

Problem Description/Opportunity

For graduate students in the United States, finding a job is an inevitable and important task. For universities, regardless of other factors, whichever one can better help students find jobs will attract more students to attend. When it comes to finding a job, students often struggle with the fact that the skills they learn in school differ somewhat from the needs of various companies in the market. For example, NYU MASY students typically graduate with strong database-related knowledge, but to get a data analyst job, they may also need to have relevant industry knowledge and competency in data visualization. Submitting the same resume to dozens of companies every day is one of the common tactics used by students to find jobs, but each company has different requirements for candidates. Theoretically, the better the match between a person's resume and the job requirements, the more likely they will be invited for an interview. To make their resumes more relevant to the job requirements, students need to spend more time customizing their resumes. But this would conflict with the strategy they commonly employ above. So students need a tool to help them achieve a balance between the two strategies. Our project will develop a text analysis tool that will help students identify jobs with the highest similarity and lowerest similarity to their resumes from various jobs. Based on analysis results, students can adopt different strategies to submit their resumes. We expect that this project will not only help students find jobs faster but will also help NYU attract more candidates.

Importance of the project

According to IBIS World's report, the university industry's revenue is expected to increase at an annualized rate of 1.1% to $580.7 billion over the five years to 2021, including a forecast increase of 0.3% in 2021 alone (Le, 2021). As for the profit margin, it is expected to average 11.1% from 2016 to 2021.

In the next five years, since the job market is expected to strengthen, higher education demand may slow down. In addition, colleges and universities are expected to endure increased competition from the massive open online course offering platforms like Academy of Mine, Udemy, and so on, which could pull students away from traditional universities by providing low-cost education. Besides, the number of students graduating from high schools directly correlates with the growth of the number of college freshmen. The high school retention rate is expected to decrease in 2021, posing a potential threat to the industry marginally (Le, 2021). Therefore, how to attract freshmen to maintain revenue is a vital issue for industry operators. Also, this urgent industry issue justifies the need for this project (Occupation Analyzer). Developing tools to help students find jobs could be a selling point for NYU.

From students' perspectives, this project will help them find the right job faster, thus enabling them to have better career development. And they will be grateful to the university that helped them. That's why this project is so important to both NYU and students.

Alternate Solutions Evaluated

There are several alternative solutions for NYU MASY to help students find desired jobs easier. However, developing a text analytical tool is the most cost-effective choice for various stakeholders.

The first alternative option could be to cooperate with companies that provide similar web-based text analytical tools.

Pros:

- Tools are available immediately.
- Don't require human resources and budgets for future maintenance.

Cons:

- Can't customize the user interface, input files, and output files.
- It might have data security problems.
- NYU can't get access to the source codes and algorithms they used. The codes and algorithms might have flaws.
- It will cost more money than developing the tool by NYU itself.

The next alternative option could be to hire career coaches to guide students in job searching, writing resumes, and preparing for interviews.

Pros :

- Career coaches can provide individualized guidance based on each student's past experience and specialties.

Career coaches can summarize and provide more detailed guidance on common student problems.

Cons:

- Career coaches don’t have enough time to guide each student and provide detailed guidance on each question if time is limited.
- Compared to the Occupation Analyzer project costs, this one requires much more money to support this solution.
- Hard to evaluate career coaches’ mentoring results.

Solution Evaluation Criteria

Cost

- Compare to other options, what’s the cost level of this choice.
- Cost should be evaluated from money, time, human resources perspectives.

Efficiency

- Compare to other options, what’s the efficiency level of this choice.
- Efficiency should be evaluated from how many students can use this option at the same time.

Quality

- Compare to other options, what’s the quality level of this choice.
- Quality should be evaluated from how easier NYU can evaluate the option’s quality and prediction about how this option could help students find jobs.

Continuous improvement

- Compare to other options, what’s the continuous improvement level of this choice.
- Continuous improvement should be evaluated from how NYU can improve this option in the future and the room for improvement.

Selection Rationale

	Cost	Efficiency	Quality	Continuous Improvement	Total Scores
Occupation Analyzer	2	1	2	1	6
Use other companies' text analytical tools	2	1	2	3	8
Career coaches	3	3	1	2	9

Table 1Selction Rationale

Each criterion will be scored from one to three. One means this option performs the best compare to other options. Three means option performs the worst compare to other options. Thus, the option that obtains the lowerest scores should be the first choice for NYU.

The result indicates that the occupation analyst gets the lowerest scores, which means it is the first choice for NYU. For the cost, occupation analyzer requires human resources and time costs, but the money cost is low. Other tools don't require human resources and time costs, but the money cost is high. Hiring coaches need support from all perspectives. For efficiency, both occupation analyzer and other tools can be used by an unlimited number of students as long as the server supports, but each only can guide one student simultaneously. For quality, both occupation analyzer and other tools can give students similarity scores between resumes and jobs, but coaches can provide more detailed guidance. For continuous improvement, the occupation analyzer has unlimited potential since NYU owns the source code and has many talents to improve this tool. Other companies will not upgrade their tools based on NYU's requirements and will not share the source code. Whether coaches will improve themselves depend on their intrinsic motivation for improvement.

In conclusion, the occupation analyzer is a rationable choice for NYU.

Approach and Methodology

This project follows the 4-step project life cycle to meet sponsors' requirements.

Initiation:

1. Discussed with clients and identify the project objectives, deliverables, risks, constraints, and priorities.
2. Based on clients' requirements, deadlines, and available resources, I developed a project proposal for clients to sign off.
3. Clients reviewed and signed the project proposal.

Planning:

1. Researched relevant papers and algorithms.
2. I generated the work breakdown schedule, risk management plan, change management plan, project sponsor acceptance document, and sponsor agreement based on the proposal.
3. Clients reviewed and signed the project sponsor acceptance document and sponsor agreement.

Execution:

1. Followed the work breakdown schedule to complete each week's tasks.
2. Held weekly Zoom meetings with clients to report project progress and adjust weekly tasks based on clients' feedback.
3. Wrote monthly status reports for clients to track the project.
4. Clients reviewed and signed monthly status reports.

Closeout:

1. Organized and uploaded all relevant files to the GitHub repository.
2. Generated project completion acceptance document.
3. I was prepared for the project final report and presentation.
4. Clients reviewed and signed the project completion acceptance document.

Project Objectives and Metrics

Goal of the project

Matching resumes against jobs are important for students to find jobs. Many NYU students have difficulties efficiently matching their resume against desired jobs. To solve this problem, NYU MASY wants to create a computer-based tool that matches a resume to jobs via discovering the mutual similarity between resume and jobs mediated by standard occupation descriptions. A similar tool based on Python was developed before. Thus, this project aims to develop the previous tool further using R.

Project Deliverables and Metrics

Project Objective 1

Compile a functional specification document that describes the tool's functions and use cases.

Metric:

Deliver the document on Feb 12 with clients' satisfaction.

Project Objective 2

Compile a database of occupation descriptions.

Metric:

1. Occupation descriptions should be the same as the descriptions in the BLS O*NET database.
2. The occupation database should be delivered on May 5.

Project Objective 3.1

Develop a text analytics tool in R that accepts the text of a person's resume and scores the resume against a list of jobs (to be input by the user) by a TF-IDF text similarity scoring algorithm.

Metric:

1. The results obtained by this tool should be the same as those obtained by the previous Python tool.
2. Deliver the functionality on May 5 with clients' satisfaction.

Project Objective 3.2

Present user with the top 25 (user input) jobs and allows the user to select their desired set of target occupations (user input).

Metric:

1. The results obtained by this tool should be the same as those obtained by the previous Python tool.
2. Deliver the functionality on May 5 with clients' satisfaction.

Project Objective 3.3

Develop the text analytics tool in R that accepts the text of a person's resume and scores the resume against occupation by a TF-IDF text similarity scoring algorithm. A TF-IDF text similarity score also scores the target set of identified jobs against O*NET occupations. The jobs are then ranked by the result of a cosine similarity analysis of the resume vector and the job vectors against the occupations.

Metric:

1. The results obtained by this tool should be the same as those obtained by the previous Python tool and Python algorithms.
2. Deliver the tool on May 5 with clients' satisfaction.

Project Objective 4

Develop an interactive website using R Shiny to present the user with the smart scoring of the target jobs to the resume via occupation-matching.

Metric:

1. Deliver the website on May 5 with clients' satisfaction.

Project Objective 5

Create and populate a GitHub repository to store all project files.

Metric:

1. Deliver the GitHub repository on May 5 with clients' satisfaction.

Risk Analysis

Risk analysis is a vital part of a project. There are several potential risks identified below.

Number	Risk	Probability Score (1,2 or 3)	Impact Score (1,2 or 3)
1	I can’t complete the project on time.	2	3
2	I can’t complete the project with high quality.	2	2
3	The project requires an extra budget.	2	1
4	Clients abandon the project.	1	3
5	The tool is not user-friendly enough.	3	1

Table 2Potential Risks

	RISK (exposure)			
Probability (of occurrence)		1.Slight	2. Moderate	3. High
	1. Very Unlikely			4
	2. Possible	3	2	1
	3. Expected	5		

Table 3Risk Matrix

In case these risks do happen, a contingency plan was prepared.

Risk	Description	Probability (1-3)	Exposure (1-3)	Contingency Plan
1	I can’t complete the project on time.	2	3	Explain to the client in advance why the project will not be completed on time and seek understanding. Ask others for help in completing projects on time.
2	I complete the project with medium or low quality.	2	2	
3	Shiny App. io may require an extra budget for the server.	2	1	
4	Clients abandon the project.	1	3	
5	Users are not 100% satisfied with the interface.	3	1	

Table 4Contingency Plan

Issues Encountered

While working on the project, the team encountered some issues. All of the issues the team faced are minor issues that do not have a major impact on the project. All issues were solved immediately once indicated so that the project was able to finish on time with high quality. Here is all type of issues project team faced in the duration of the project.

The first issue the team faced was coding errors. When coding for the occupation analyzer, countless errors occurred every day. The best helper to solve these errors is Google, and most of the useful answers come from the Stackoverflow Forum.

The second issue the team faced was clients' dissatisfaction with the user interface and output results. Most parts of the project were completed two weeks ahead of schedule. Thus, the team held several Zoom meetings to discuss with clients how to beautify the user interface and output results. After further modification, clients were satisfied with the project.

Project Chronology and Critique

The tasks and duration of the project are shown below.

Level	WBS Code	Element Name	Due By	Deliver Date
1	1	Widget Management System		
2	1.1	Initiation	March 25, 2021	March 25, 2021
3	1.1.1	Evaluation & Recommendations	January 18, 2021	January 18, 2021
3	1.1.2	Develop Project Proposal	February 10, 2021	February 10, 2021
3	1.1.3	Draft Literature Review Research	February 24, 2021	February 24, 2021
3	1.1.4	Conduct Situation Analysis & Cost-Benefit Analysis	March 3, 2021	March 3, 2021
3	1.1.5	Create Work Break Down Schedule	March 3, 2021	March 3, 2021
3	1.1.6	Develop Project Charter	March 12, 2021	March 12, 2021
3	1.1.7.	*Deliverable:* Submit Project Charter	March 17, 2021	March 17, 2021
3	1.1.8.	Project Sponsor Reviews Project Charter	March 20, 2021	March 20, 2021

3	1.1.9.	Project Charter Signed/Approved	March 25, 2021	March 25, 2021
2	1.2	Planning	March 25, 2021	March 25, 2021
3	1.2.1	Compile Functional Specification Document	February 10, 2021	February 10, 2021
3	1.2.2	Develop Project Sponsor Agreement	March 10, 2021	March 10, 2021
3	1.2.3	Submit Project Sponsor Agreement	March 10, 2021	March 10, 2021
3	1.2.4	Project Sponsor Acceptance	March 10, 2021	March 10, 2021
2	1.3	Execution	April 16, 2021	April 16, 2021
2	1.3.1	Verify Functional Specification Documents	February 12, 2021	February 12, 2021
3	1.3.2	*Deliverables 1:* Functional Specification Document Approval	February 12, 2021	February 12, 2021
3	1.3.3	Write TF-IDF Similarity Code in R	March 11, 2021	March 11, 2021

3	1.3.4	Conduct Unit Test and Peer Review for TF-IDF Similarity Code	March 12, 2021	March 12, 2021
3	1.3.5	*Deliverable 2:* Performs the TF-IDF similarity scores of resume and jobs, resume and occupations, and jobs and occupations	March 19, 2021	March 19, 2021
3	1.3.6	Write Cosine Similarity Scores Code	March 25, 2021	March 25, 2021
3	1.3.7	Conduct Unit Test and Peer Review for Cosine Similarity Scores Code	March 26, 2021	March 26, 2021
3	1.3.8	*Deliverable 3:* Performs the Cosine Similarity Scores between resume vector and jobs' vectors	April 2, 2021	April 2, 2021
3	1.3.9	Develop a Shiny App Interface	April 8, 2021	April 8, 2021

3	1.3.10	*Deliverable 4:* The Shiny App with Function to Accept the Resume and Job Files	April 9, 2021	April 9, 2021
3	1.3.11	Create an Accessible GitHub Repository for Recording All the Project Files and Supporting Documentation	April 13, 2021	April 13, 2021
3	1.3.12	*Deliverable 5:* A GitHub Repository with All Relevant Documents	April 16, 2021	April 16, 2021
2	1.4	Control	April 14, 2021	April 14, 2021
3	1.4.1	Develop Change Management Plan	April 14, 2021	April 14, 2021
3	1.4.2	Develop Risk Management Plan	April 14, 2021	April 14, 2021
2	1.5	Closeout	May 5, 2021	May 5, 2021
3	1.5.1	Draft Final Project Report	April 20, 2021	April 20, 2021
3	1.5.2	Final meeting with the client to go over the lessons learned	April 23, 2021	April 23, 2021

3	1.5.3	Compile Project Sponsor Acceptance – Project Completion Signoff	April 28, 2021	April 28, 2021
3	1.5.4	Gain Formal Acceptance	April 28, 2021	April 28, 2021
3	1.5.5	Present the project	May 5, 2021	May 5, 2021
3	1.5.6	Submit Final Project Report with Final Deliverables in GitHub Repository	May 5, 2021	May 5, 2021

Table 5Project Chronology

The literature review part could be improved. The current literature review focuses on technical papers about conducting natural language processing and improving the result accuracy. The literature review part should also include some papers about the importance of the tool and why developing it is necessary.

Lessons Learned

The whole project was able to deliver as planned with expected quality and in time, and this could not have been done without contribution and help from all team members and sponsors.

During the whole project implementation, team members have learned how to manage the scope, design the project roadmap, write the work breakdown schedule, mitigate risks, manage change requests, communicate with clients, and evaluate the project progress.

Besides, team members have acquired text mining and web development skills in R. In the text mining part, we mastered the TF-IDF and cosine similarity algorithms and how to do data preprocessing for natural language processing projects. We also mastered R packages like tm, SnowballC, and Shiny.

Conclusion and Summary

This project utilized R to develop a text analytic tool that can help students match their resumes against a group of jobs and return fifteen jobs with the highest similarity scores and five jobs with the lowest similarity scores. An interactive website using R Shiny was developed for student use. This project was based on previous work that Felix Hui did in Python. The final work reproduced similar results in R and further improved the original algorithms to increase result accuracy. Clients were satisfied with the result.

All the team members sincerely hope this tool can be further improved in the future and help all NYU students find desired jobs faster and easier.

Occupation Analyzer: https://nyuprof.shinyapps.io/OccupationAnalyzerKasper/

GitHub: https://github.com/kasper3144/Occupation_Analyzer

Limitations, Recommendations and Scope for Future Work

Even this project was able to deliver as expected. There are still some limitations within this project, and some of the limitations may be improved in future similar projects in NYU MASY.

In the data preprocessing part, n-grams, named-entity extraction (NER), and the sliding window approach might be considered. Currently, the project used unigram as input for the document term matrix, but using bigram or trigram might improve accuracy. NER aims to overcome a common problem in separating words by only using whitespace characters between the words. For example, “the Microsoft Corporation” has three tokens. “The” is a stop word and should be removed, and “Microsoft Corporation” should be treated as one token. Using NER, we can identify a set or a group of words that have a single meaning and combine them into a single token. This technique most commonly applies to the names of people or organizations. While NER usually relies on built-in word lists or capitalization of entity tokens, there are other words that consist of one or more word forms. For example, “computer science” is a phrase that frequently occurs together and has a single meaning. To identify these phrases, we can use the sliding window approach.

Besides, the user interface of the occupation analyzer can be more user-friendly, and the processing time might be further shortened.

Literature Survey

Introduction

This literature review was organized by introduction of Term Frequency and Inverse Document Frequency (TF-IDF), data preprocessing of text mining, and cosine similarity.

TF-IDF stands for Term Frequency and Inverse Document Frequency. TF is used to measure that how many times a term is present in a document (Qaiser & Ali, 2018). The inverse document frequency assigns a lower weight to frequent words and assigns a greater weight for the infrequent words (Gong, 2019). The greater or higher occurrence of a word in documents will give higher term frequency, and the less occurrence of a word in documents will yield higher importance (IDF) for that keyword searched in a particular document. TF-IDF is the multiplication of term frequency (TF) and inverse document frequency (IDF) (Silge & Robinson, 2020).

To conduct Term Frequency and Inverse Document Frequency in text mining, necessary data preprocessing steps, including removing stop words, stemming, named-entity extraction (NER), and n-grams should be done to increase the result accuracy (Vijayarani et al., 2015). The motive that stop-words should be removed from a text is to make the text look heavier and less important for analysts (Vijayarani et al., 2015). Removing stop words reduces the dimensionality of term space. Stop words are words such as “the”, “of”, “and”, etc. and usually do not contain any meaningful information for identifying document topics or similarities. Stemming is used to identify the root/stem of a word. This method aims to remove various suffixes, reduce the number of words, have accurately matching stems, and save time and memory space (Qaiser & Ali, 2018). NER aims to overcome a common problem in separating words by only using whitespace characters between the words. For example, “the Microsoft Corporation” has three tokens. “The” is a stop word and should be removed, and "Microsoft Corporation" should really be treated as one token. N-grams aims to collect word compounds in the document (Silge & Robinson, 2020). For example, "computer science", "beauty pageant", or "student athlete compensation" are all phrases that frequently occur together and have a single meaning.

Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them (Gomaa & Fahmy, 2013). Lahitani, Permanasari, and Setiawan (2016) implements the TF-IDF method and cosine similarity approach to measure the similarity level from the Indonesian essay assessment. Huang (2008) compares and analyzes the effectiveness of Euclidean Distance, Cosine Similarity, Jaccard Coefficient, Pearson Correlation Coefficient, and Averaged Kullback-Leibler Divergence in partitional clustering for text document datasets. The result shows that Pearson Correlation Coefficient and Averaged Kullback-Leibler Divergence outperform other measures.

Conclusion

Previous research shows that TF-IDF is ideal for finding important words among documents and comparing documents' similarities. But, there is no research about conducting TF-IDF analysis among resume, job, and occupation files.

The existing researches around text mining indicate that removing stop words, stemming, named-entity extraction (NER), and n-grams should be done in the resume file, job file, and occupation file before calculating their TF-IDF similarity scores.

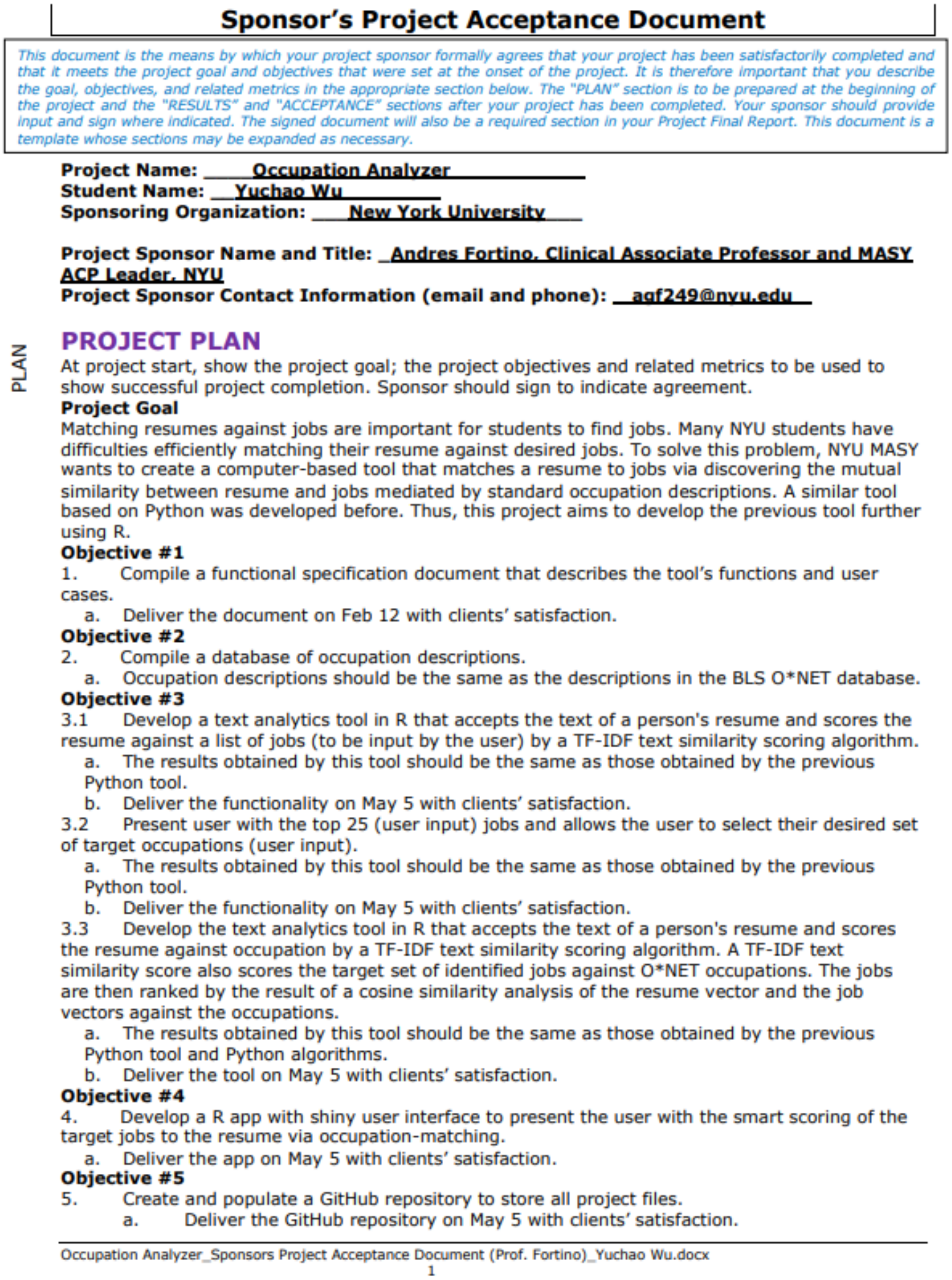
As for the method of calculating similarity scores, Pearson Correlation Coefficient and Averaged Kullback-Leibler Divergence outperforms Cosine Similarity in Huang's (2019) research. We may consider changing a calculation method.

References

1. Gomaa, W. H., & Fahmy, A. A. (2013). A survey of text similarity approaches. *International Journal of Computer Applications, 68*(13), 13-18.
2. Gong, K. (2019, October 22). *Big Data Cosine Similarity Score*. Retrieved from RPubs: https://www.rpubs.com/kgong56/542550
3. Huang, A. (2008, April). Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand* (Vol. 4, pp. 9-56).
4. Lahitani, A. R., Permanasari, A. E., & Setiawan, N. A. (2016, April). Cosine similarity to determine similarity measure: Study case in online essay assessment. In *2016 4th International Conference on Cyber and IT Service Management* (pp. 1-6). IEEE.
5. Qaiser, S., & Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications, 181*(1), 25-29.
6. Silge, J., & Robinson, D. (2020, November 10*). Analyzing word and document frequency: TF-IDF*. Retrieved from Text Mining with R: https://www.tidytextmining.com/tfidf.html#tfidf
7. Silge, J., & Robinson, D. (2020, November 10). *Relationships between words: n-grams and correlations*. Retrieved from Relationships between words: n-grams and correlations: https://www.tidytextmining.com/ngrams.html
8. Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, 5(1), 7-16.

Appendix A

Project Acceptance Document



Sponsor's Project Acceptance Document

This document is the means by which your project sponsor formally agrees that your project has been satisfactorily completed and that it meets the project goal and objectives that were set at the onset of the project. It is therefore important that you describe the goal, objectives, and related metrics in the appropriate section below. The "PLAN" section is to be prepared at the beginning of the project and the "RESULTS" and "ACCEPTANCE" sections after your project has been completed. Your sponsor should provide input and sign where indicated. The signed document will also be a required section in your Project Final Report. This document is a template whose sections may be expanded as necessary.

Project Name: Occupation Analyzer
Student Name: Yuchao Wu
Sponsoring Organization: New York University

Project Sponsor Name and Title: Andres Fortino, Clinical Associate Professor and MASY ACP Leader, NYU
Project Sponsor Contact Information (email and phone): agf249@nyu.edu

PLAN

PROJECT PLAN

At project start, show the project goal; the project objectives and related metrics to be used to show successful project completion. Sponsor should sign to indicate agreement.

Project Goal

Matching resumes against jobs are important for students to find jobs. Many NYU students have difficulties efficiently matching their resume against desired jobs. To solve this problem, NYU MASY wants to create a computer-based tool that matches a resume to jobs via discovering the mutual similarity between resume and jobs mediated by standard occupation descriptions. A similar tool based on Python was developed before. Thus, this project aims to develop the previous tool further using R.

Objective #1

1. Compile a functional specification document that describes the tool's functions and user cases.
   a. Deliver the document on Feb 12 with clients' satisfaction.

Objective #2

2. Compile a database of occupation descriptions.
   a. Occupation descriptions should be the same as the descriptions in the BLS O*NET database.

Objective #3

3.1 Develop a text analytics tool in R that accepts the text of a person's resume and scores the resume against a list of jobs (to be input by the user) by a TF-IDF text similarity scoring algorithm.
   a. The results obtained by this tool should be the same as those obtained by the previous Python tool.
   b. Deliver the functionality on May 5 with clients' satisfaction.

3.2 Present user with the top 25 (user input) jobs and allows the user to select their desired set of target occupations (user input).
   a. The results obtained by this tool should be the same as those obtained by the previous Python tool.
   b. Deliver the functionality on May 5 with clients' satisfaction.

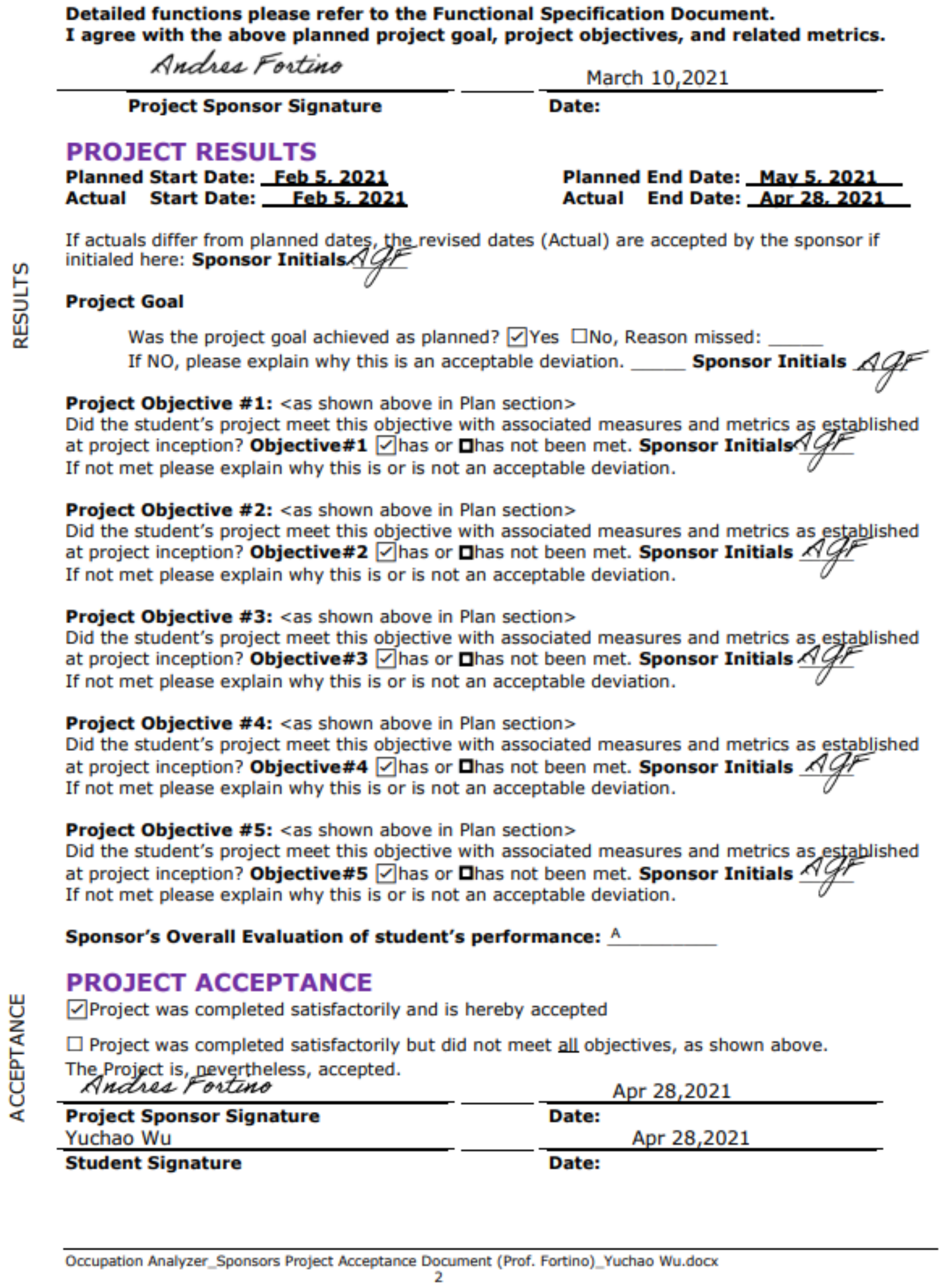
3.3 Develop the text analytics tool in R that accepts the text of a person's resume and scores the resume against occupation by a TF-IDF text similarity scoring algorithm. A TF-IDF text similarity score also scores the target set of identified jobs against O*NET occupations. The jobs are then ranked by the result of a cosine similarity analysis of the resume vector and the job vectors against the occupations.
   a. The results obtained by this tool should be the same as those obtained by the previous Python tool and Python algorithms.
   b. Deliver the tool on May 5 with clients' satisfaction.

Objective #4

4. Develop a R app with shiny user interface to present the user with the smart scoring of the target jobs to the resume via occupation-matching.
   a. Deliver the app on May 5 with clients' satisfaction.

Objective #5

5. Create and populate a GitHub repository to store all project files.
   a. Deliver the GitHub repository on May 5 with clients' satisfaction.



Detailed functions please refer to the Functional Specification Document.
I agree with the above planned project goal, project objectives, and related metrics.

Andres Fortino — March 10,2021

Project Sponsor Signature — **Date:**

RESULTS

PROJECT RESULTS

Planned Start Date: Feb 5, 2021 — **Planned End Date: May 5, 2021**
Actual Start Date: Feb 5, 2021 — **Actual End Date: Apr 28, 2021**

If actuals differ from planned dates, the revised dates (Actual) are accepted by the sponsor if initialed here: **Sponsor Initials** AGF

Project Goal

Was the project goal achieved as planned? ☑Yes □No, Reason missed: ______
If NO, please explain why this is an acceptable deviation. _____ **Sponsor Initials** AGF

Project Objective #1: <as shown above in Plan section>
Did the student's project meet this objective with associated measures and metrics as established at project inception? **Objective#1** ☑has or □has not been met. **Sponsor Initials** AGF
If not met please explain why this is or is not an acceptable deviation.

Project Objective #2: <as shown above in Plan section>
Did the student's project meet this objective with associated measures and metrics as established at project inception? **Objective#2** ☑has or □has not been met. **Sponsor Initials** AGF
If not met please explain why this is or is not an acceptable deviation.

Project Objective #3: <as shown above in Plan section>
Did the student's project meet this objective with associated measures and metrics as established at project inception? **Objective#3** ☑has or □has not been met. **Sponsor Initials** AGF
If not met please explain why this is or is not an acceptable deviation.

Project Objective #4: <as shown above in Plan section>
Did the student's project meet this objective with associated measures and metrics as established at project inception? **Objective#4** ☑has or □has not been met. **Sponsor Initials** AGF
If not met please explain why this is or is not an acceptable deviation.

Project Objective #5: <as shown above in Plan section>
Did the student's project meet this objective with associated measures and metrics as established at project inception? **Objective#5** ☑has or □has not been met. **Sponsor Initials** AGF
If not met please explain why this is or is not an acceptable deviation.

Sponsor's Overall Evaluation of student's performance: A

ACCEPTANCE

PROJECT ACCEPTANCE

☑Project was completed satisfactorily and is hereby accepted

□ Project was completed satisfactorily but did not meet all objectives, as shown above.
The Project is, nevertheless, accepted.

Andres Fortino — Apr 28,2021

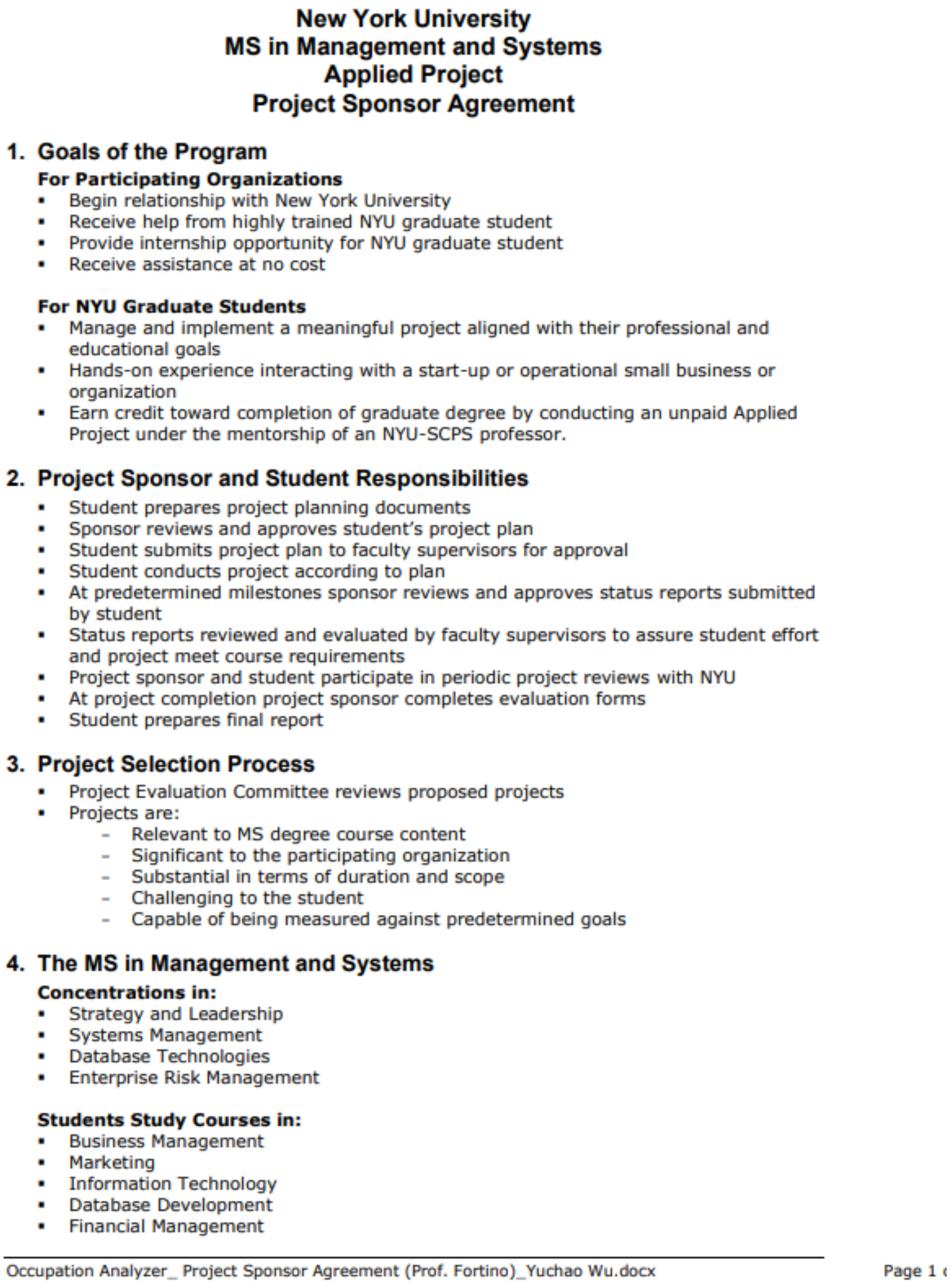
Project Sponsor Signature — **Date:**

Yuchao Wu — Apr 28,2021

Student Signature — **Date:**

Appendix B

Project Sponsor Agreement



New York University
MS in Management and Systems
Applied Project
Project Sponsor Agreement

1. **Goals of the Program**

For Participating Organizations
- Begin relationship with New York University
- Receive help from highly trained NYU graduate student
- Provide internship opportunity for NYU graduate student
- Receive assistance at no cost

For NYU Graduate Students
- Manage and implement a meaningful project aligned with their professional and educational goals
- Hands-on experience interacting with a start-up or operational small business or organization
- Earn credit toward completion of graduate degree by conducting an unpaid Applied Project under the mentorship of an NYU-SCPS professor.

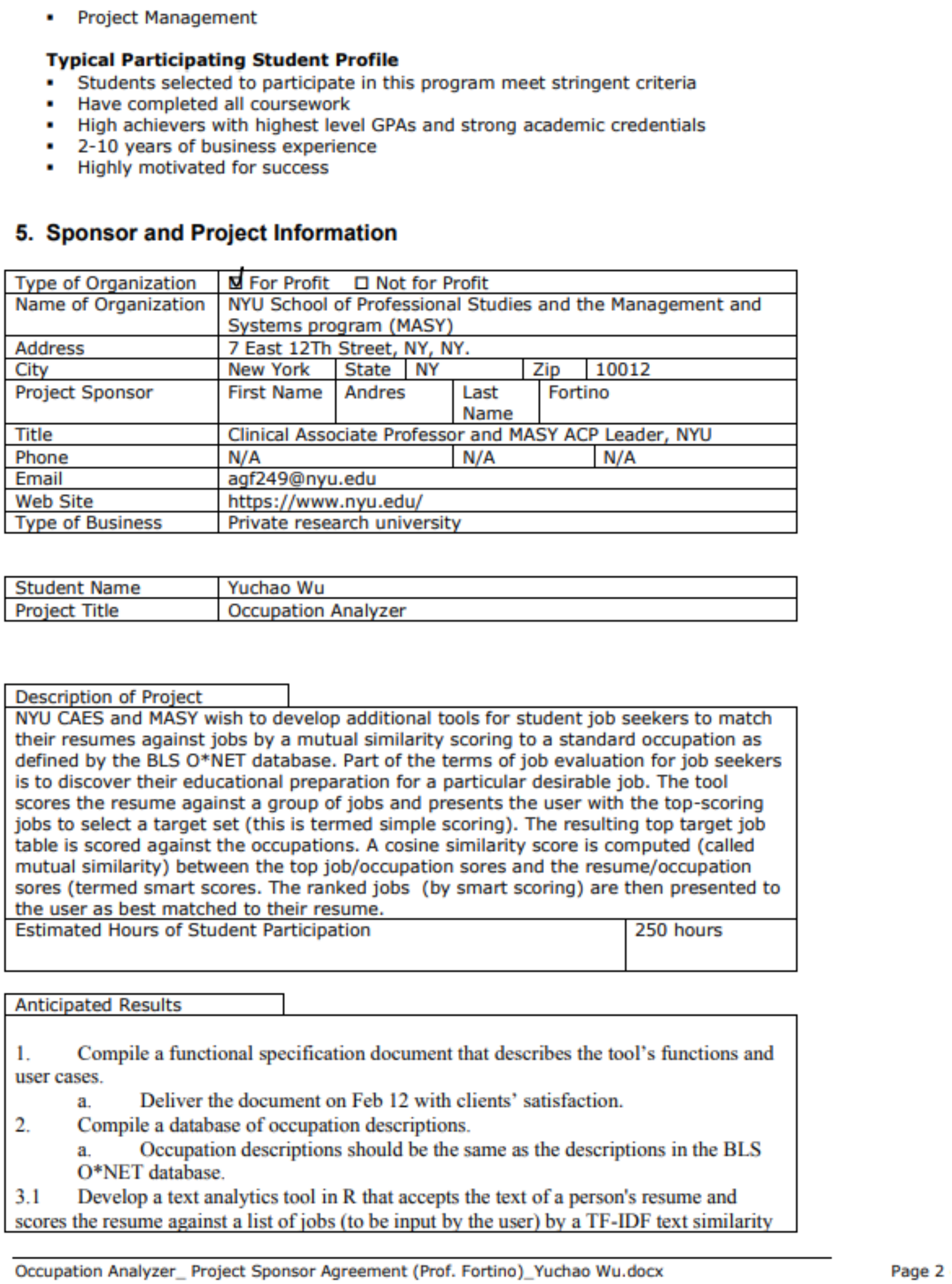
2. **Project Sponsor and Student Responsibilities**
- Student prepares project planning documents
- Sponsor reviews and approves student's project plan
- Student submits project plan to faculty supervisors for approval
- Student conducts project according to plan
- At predetermined milestones sponsor reviews and approves status reports submitted by student
- Status reports reviewed and evaluated by faculty supervisors to assure student effort and project meet course requirements
- Project sponsor and student participate in periodic project reviews with NYU
- At project completion project sponsor completes evaluation forms
- Student prepares final report

3. **Project Selection Process**
- Project Evaluation Committee reviews proposed projects
- Projects are:
    - Relevant to MS degree course content
    - Significant to the participating organization
    - Substantial in terms of duration and scope
    - Challenging to the student
    - Capable of being measured against predetermined goals

4. **The MS in Management and Systems**

Concentrations in:
- Strategy and Leadership
- Systems Management
- Database Technologies
- Enterprise Risk Management

Students Study Courses in:
- Business Management
- Marketing
- Information Technology
- Database Development
- Financial Management



- Project Management

Typical Participating Student Profile

- Students selected to participate in this program meet stringent criteria
- Have completed all coursework
- High achievers with highest level GPAs and strong academic credentials
- 2-10 years of business experience
- Highly motivated for success

5. Sponsor and Project Information

Type of Organization	☑ For Profit □ Not for Profit				
Name of Organization	NYU School of Professional Studies and the Management and Systems program (MASY)				
Address	7 East 12Th Street, NY, NY.				
City	New York	State	NY	Zip	10012
Project Sponsor	First Name	Andres	Last Name	Fortino	
Title	Clinical Associate Professor and MASY ACP Leader, NYU				
Phone	N/A		N/A	N/A	
Email	agf249@nyu.edu				
Web Site	https://www.nyu.edu/				
Type of Business	Private research university				

Student Name	Yuchao Wu
Project Title	Occupation Analyzer

Description of Project

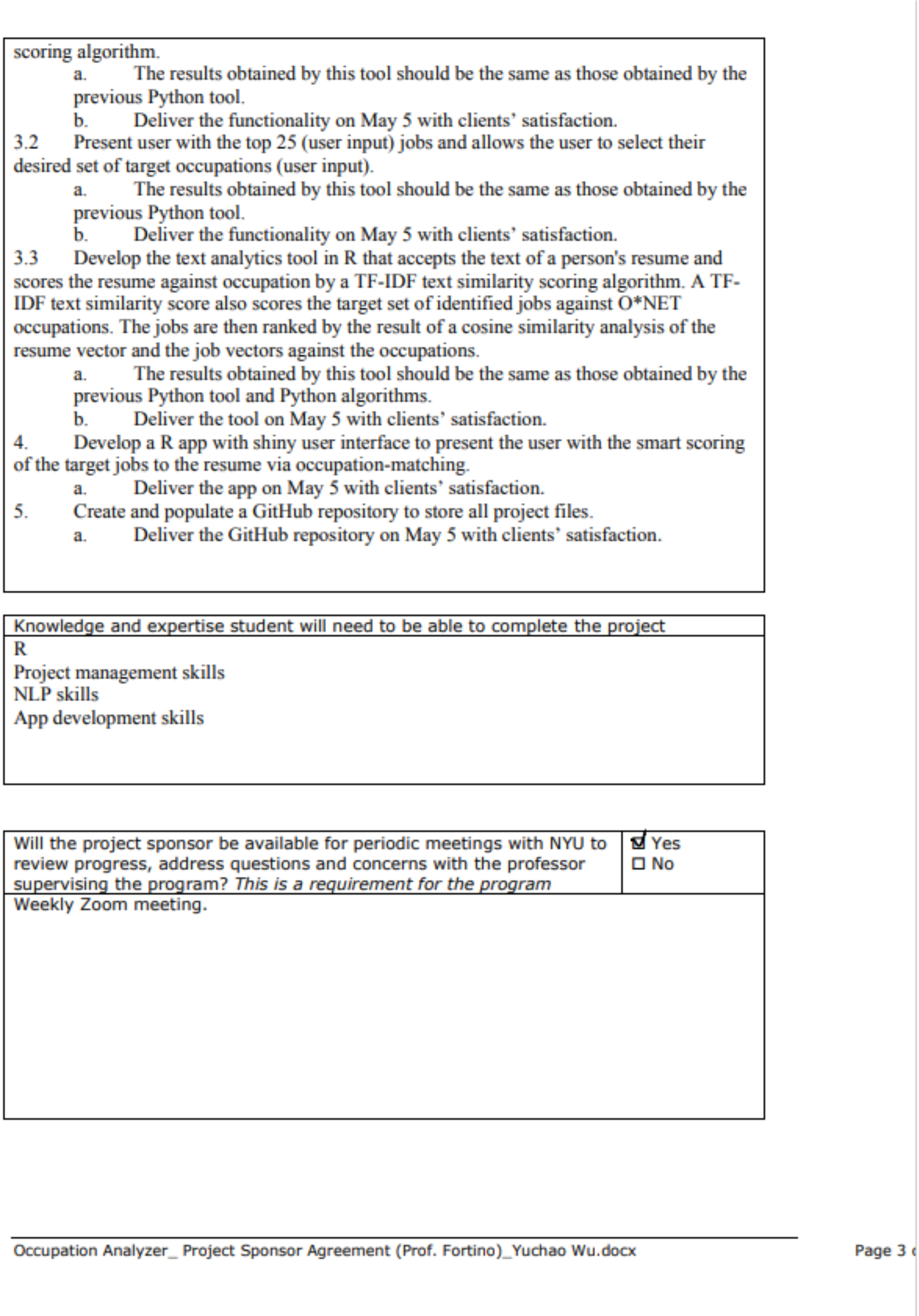
NYU CAES and MASY wish to develop additional tools for student job seekers to match their resumes against jobs by a mutual similarity scoring to a standard occupation as defined by the BLS O*NET database. Part of the terms of job evaluation for job seekers is to discover their educational preparation for a particular desirable job. The tool scores the resume against a group of jobs and presents the user with the top-scoring jobs to select a target set (this is termed simple scoring). The resulting top target job table is scored against the occupations. A cosine similarity score is computed (called mutual similarity) between the top job/occupation sores and the resume/occupation sores (termed smart scores. The ranked jobs (by smart scoring) are then presented to the user as best matched to their resume.

Estimated Hours of Student Participation	250 hours

Anticipated Results

1. Compile a functional specification document that describes the tool's functions and user cases.
   a. Deliver the document on Feb 12 with clients' satisfaction.
2. Compile a database of occupation descriptions.
   a. Occupation descriptions should be the same as the descriptions in the BLS O*NET database.

3.1 Develop a text analytics tool in R that accepts the text of a person's resume and scores the resume against a list of jobs (to be input by the user) by a TF-IDF text similarity



scoring algorithm.

a. The results obtained by this tool should be the same as those obtained by the previous Python tool.

b. Deliver the functionality on May 5 with clients' satisfaction.

3.2 Present user with the top 25 (user input) jobs and allows the user to select their desired set of target occupations (user input).

a. The results obtained by this tool should be the same as those obtained by the previous Python tool.

b. Deliver the functionality on May 5 with clients' satisfaction.

3.3 Develop the text analytics tool in R that accepts the text of a person's resume and scores the resume against occupation by a TF-IDF text similarity scoring algorithm. A TF-IDF text similarity score also scores the target set of identified jobs against O*NET occupations. The jobs are then ranked by the result of a cosine similarity analysis of the resume vector and the job vectors against the occupations.

a. The results obtained by this tool should be the same as those obtained by the previous Python tool and Python algorithms.

b. Deliver the tool on May 5 with clients' satisfaction.

4. Develop a R app with shiny user interface to present the user with the smart scoring of the target jobs to the resume via occupation-matching.

a. Deliver the app on May 5 with clients' satisfaction.

5. Create and populate a GitHub repository to store all project files.

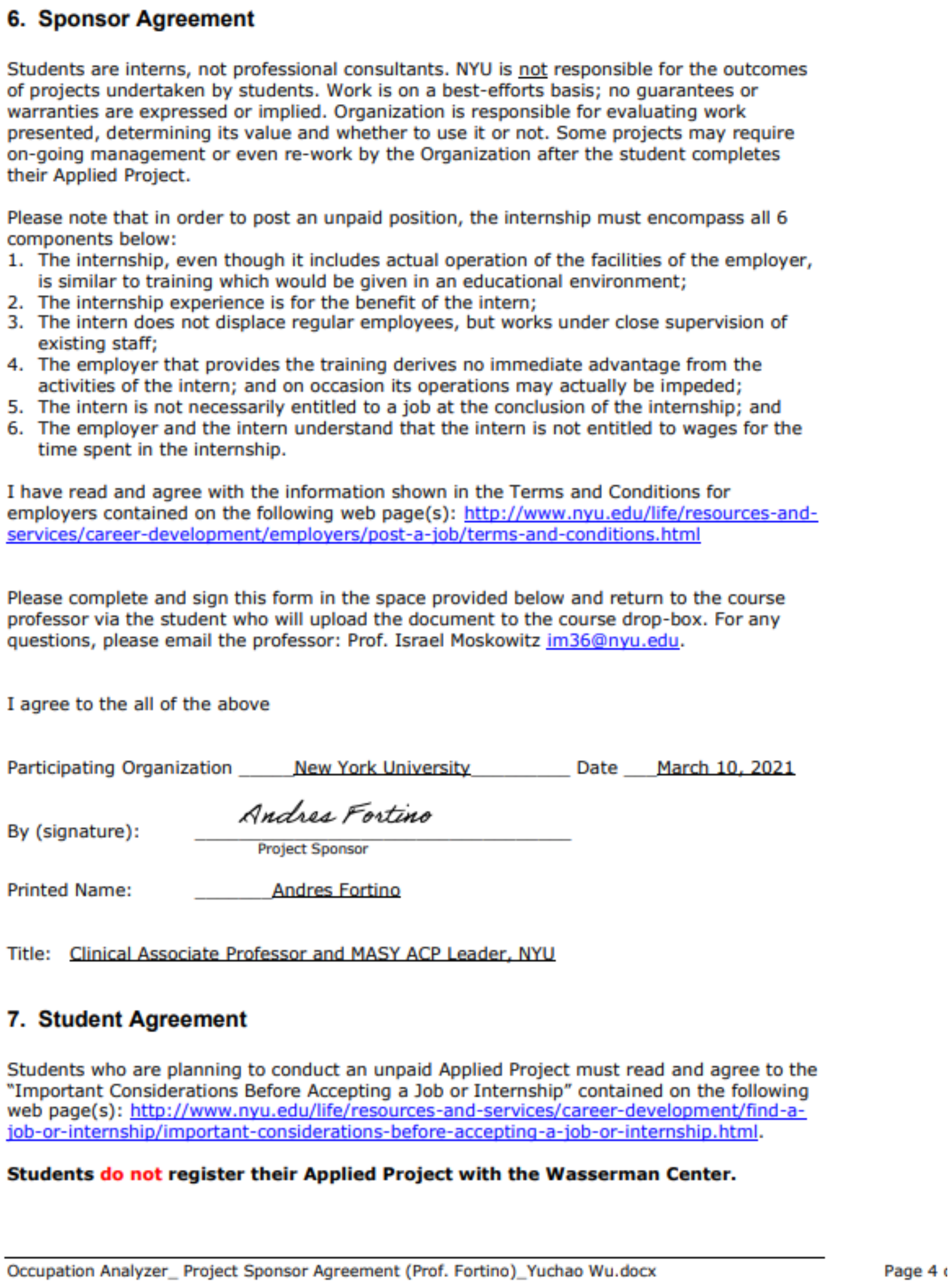
a. Deliver the GitHub repository on May 5 with clients' satisfaction.

Knowledge and expertise student will need to be able to complete the project
R Project management skills NLP skills App development skills

Will the project sponsor be available for periodic meetings with NYU to review progress, address questions and concerns with the professor supervising the program? *This is a requirement for the program*	☑ Yes ☐ No
Weekly Zoom meeting.	

Occupation Analyzer_ Project Sponsor Agreement (Prof. Fortino)_Yuchao Wu.docx

Page 3



6. Sponsor Agreement

Students are interns, not professional consultants. NYU is not responsible for the outcomes of projects undertaken by students. Work is on a best-efforts basis; no guarantees or warranties are expressed or implied. Organization is responsible for evaluating work presented, determining its value and whether to use it or not. Some projects may require on-going management or even re-work by the Organization after the student completes their Applied Project.

Please note that in order to post an unpaid position, the internship must encompass all 6 components below:

1. The internship, even though it includes actual operation of the facilities of the employer, is similar to training which would be given in an educational environment;
2. The internship experience is for the benefit of the intern;
3. The intern does not displace regular employees, but works under close supervision of existing staff;
4. The employer that provides the training derives no immediate advantage from the activities of the intern; and on occasion its operations may actually be impeded;
5. The intern is not necessarily entitled to a job at the conclusion of the internship; and
6. The employer and the intern understand that the intern is not entitled to wages for the time spent in the internship.

I have read and agree with the information shown in the Terms and Conditions for employers contained on the following web page(s): http://www.nyu.edu/life/resources-and-services/career-development/employers/post-a-job/terms-and-conditions.html

Please complete and sign this form in the space provided below and return to the course professor via the student who will upload the document to the course drop-box. For any questions, please email the professor: Prof. Israel Moskowitz im36@nyu.edu.

I agree to the all of the above

Participating Organization ____New York University________ Date ___March 10, 2021

By (signature): Andres Fortino
Project Sponsor

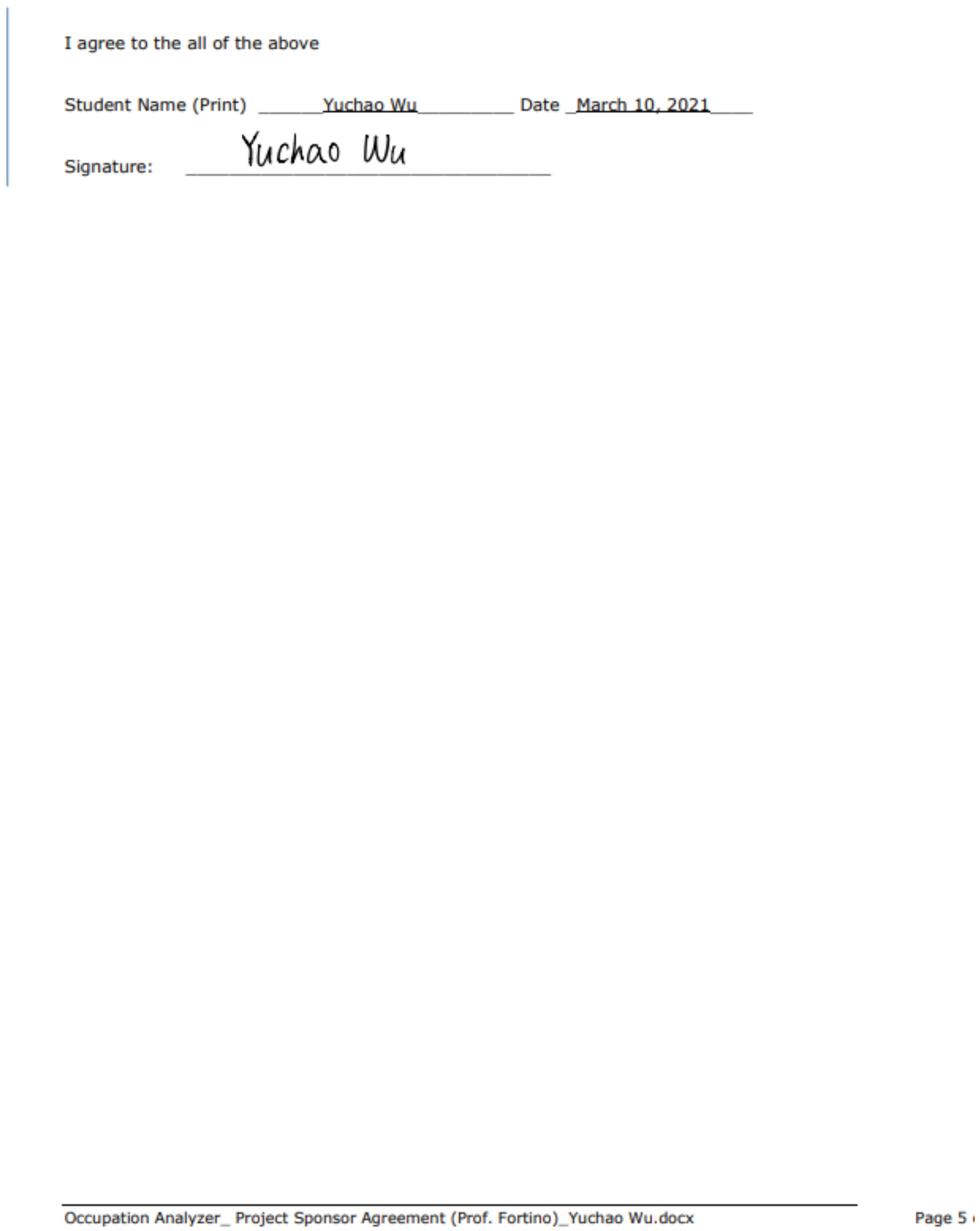
Printed Name: ________Andres Fortino

Title: Clinical Associate Professor and MASY ACP Leader, NYU

7. Student Agreement

Students who are planning to conduct an unpaid Applied Project must read and agree to the "Important Considerations Before Accepting a Job or Internship" contained on the following web page(s): http://www.nyu.edu/life/resources-and-services/career-development/find-a-job-or-internship/important-considerations-before-accepting-a-job-or-internship.html.

Students do not register their Applied Project with the Wasserman Center.



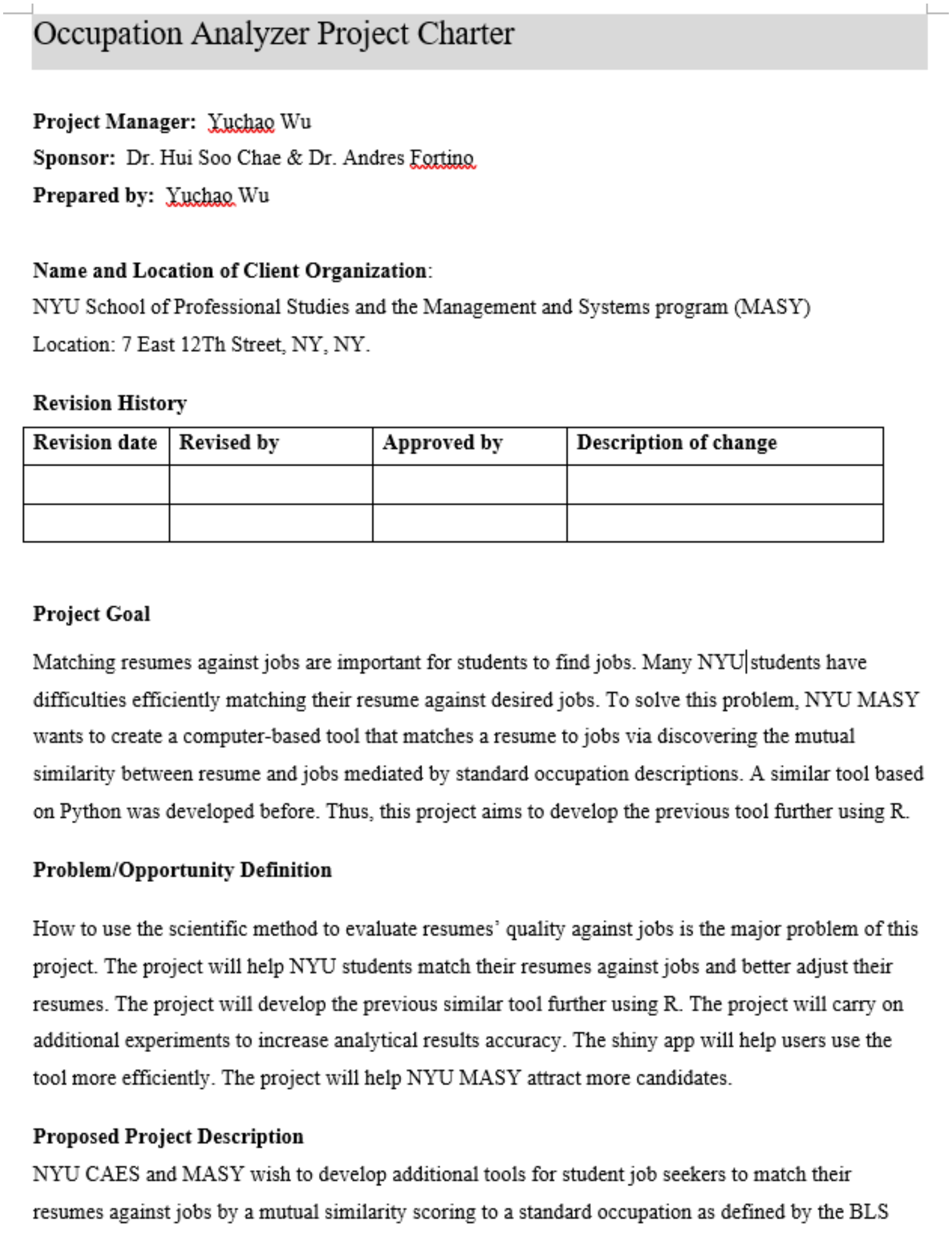
I agree to the all of the above

Student Name (Print) ______Yuchao Wu_________ Date _March 10, 2021____

Signature: Yuchao Wu

Appendix C

Project Charter



Occupation Analyzer Project Charter

Project Manager: Yuchao Wu

Sponsor: Dr. Hui Soo Chae & Dr. Andres Fortino

Prepared by: Yuchao Wu

Name and Location of Client Organization:

NYU School of Professional Studies and the Management and Systems program (MASY)

Location: 7 East 12Th Street, NY, NY.

Revision History

Revision date	Revised by	Approved by	Description of change

Project Goal

Matching resumes against jobs are important for students to find jobs. Many NYU students have difficulties efficiently matching their resume against desired jobs. To solve this problem, NYU MASY wants to create a computer-based tool that matches a resume to jobs via discovering the mutual similarity between resume and jobs mediated by standard occupation descriptions. A similar tool based on Python was developed before. Thus, this project aims to develop the previous tool further using R.

Problem/Opportunity Definition

How to use the scientific method to evaluate resumes' quality against jobs is the major problem of this project. The project will help NYU students match their resumes against jobs and better adjust their resumes. The project will develop the previous similar tool further using R. The project will carry on additional experiments to increase analytical results accuracy. The shiny app will help users use the tool more efficiently. The project will help NYU MASY attract more candidates.

Proposed Project Description

NYU CAES and MASY wish to develop additional tools for student job seekers to match their resumes against jobs by a mutual similarity scoring to a standard occupation as defined by the BLS

O*NET database. Part of the terms of job evaluation for job seekers is to discover their educational preparation for a particular desirable job. The tool scores the resume against a group of jobs and presents the user with the top-scoring jobs to select a target set (this is termed simple scoring). The resulting top target job table is scored against the occupations. A cosine similarity score is computed (called mutual similarity) between the top job/occupation sores and the resume/occupation sores (termed smart scores. The ranked jobs (by smart scoring) are then presented to the user as best matched to their resume.

Project Sponsor

- Name and Title
  - Dr. Hui Soo Chae, Executive Director of the Center for Academic Excellence and Support (CAES) at the NYU School of Professional Studies (NYU SPS)
  - Dr. Andres Fortino, Clinical Associate Professor and MASY ACP Leader, NYU
- Role within the organization
  - Dr. Hui Soo Chae develops tools, courses, and projects that can promote student development.
  - Dr. Andres Fortino is an instructor and the ACP Leader at NYU MASY.
- Role on the project
  - Dr. Hui Soo Chae and Dr. Andres Fortino will illustrate the project's details, provide necessary resources, and evaluate the project's progress.

Objectives:

Technical Objectives:

- Develop an R app with a shiny user interface to present the user with the smart scoring of the target jobs to the resume via occupation-matching before May 5, 2021.

Timing objectives

- Complete the entire project before May 5, 2021.

Resource objectives:

- Collaborate with project sponsors to finish the project before May 5, 2021.

Budget objectives

- No additional fees are required.

Budget objectives:

	Planned	**Actual**
Salaries	0	0
Documentation	0	0
Construction	0	0
Mover	0	0
Total	$ 0	$ 0

Scope objectives:

- Create a computer-based tool that matches a resume to jobs via discovering the mutual similarity between resume and jobs mediated by standard occupation descriptions using R.

Project Selection & Ranking Criteria

Project benefit category:

❑ Compliance/Regulatory ❑ Efficiency/Cost reduction ✓ Revenue increase

Portfolio fit and interdependencies

Not determined.

Project urgency

Medium.

Cost/Benefit Analysis

Tangible Benefits

Benefit: No identifiable benefits

Value & Probability: N/A

Assumptions Driving Value: N/A

Intangible Benefits

Benefit: N/A

Value & Probability: N/A

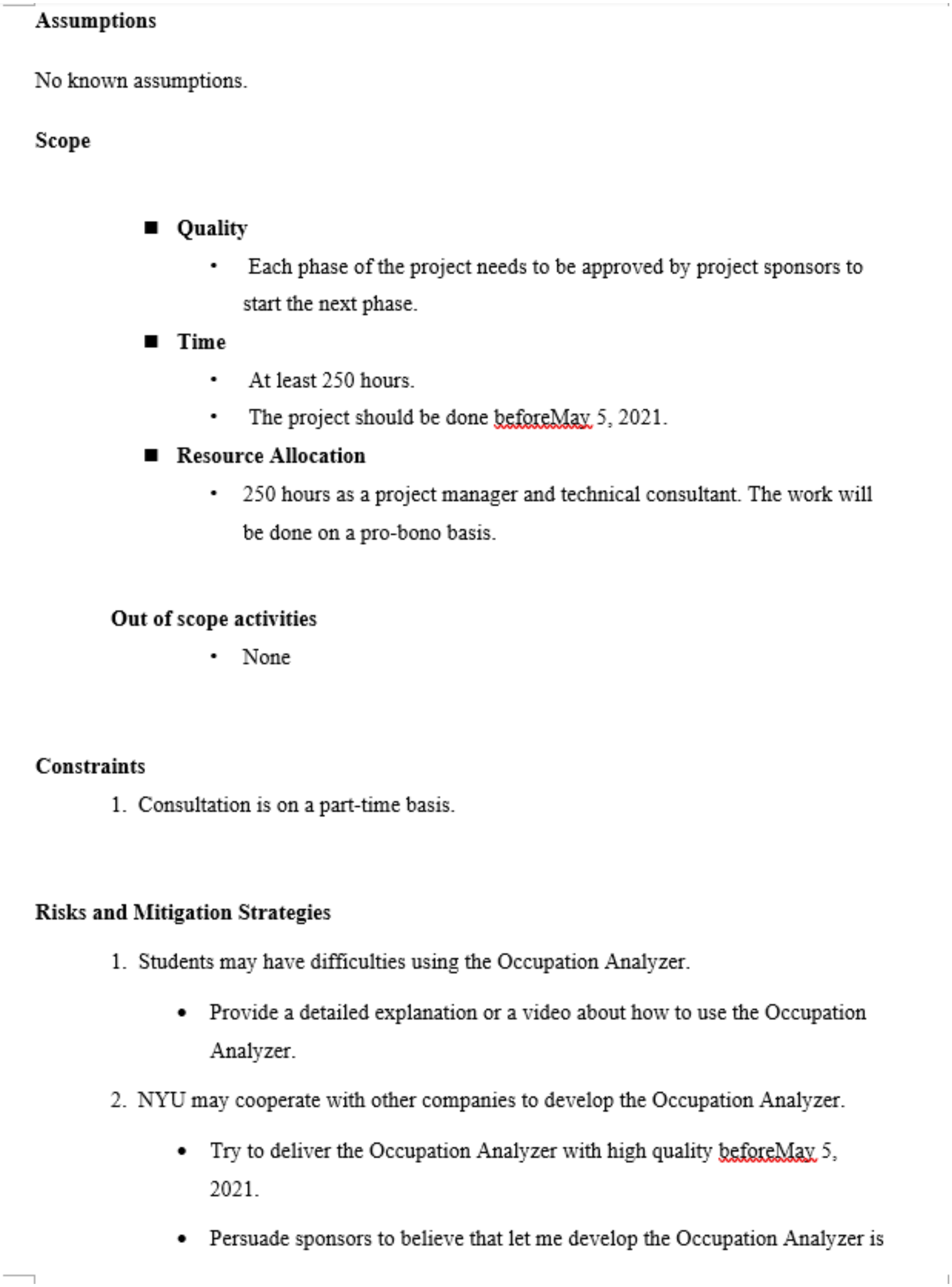
Assumptions Driving Value: N/A

Cost Categories	**Amount**
Internal Labor hours	N/A
External Costs	N/A
Labor (consultants, contract labor)	N/A
Equipment, hardware or software	N/A
List other costs such as travel & training	N/A

Financial Return

Other Business Benefits

The intangible benefit of Occupation Analyzer could be to increase NYU's reputation and attract more students. But it is hard to quantify. Based on the expected working hour and compensation rate, the project implementor's salary could be $29,000. This project costs nothing because this is a non-paid project.



Assumptions

No known assumptions.

Scope

- **Quality**
  - Each phase of the project needs to be approved by project sponsors to start the next phase.
- **Time**
  - At least 250 hours.
  - The project should be done beforeMay 5, 2021.
- **Resource Allocation**
  - 250 hours as a project manager and technical consultant. The work will be done on a pro-bono basis.

Out of scope activities

- None

Constraints

1. Consultation is on a part-time basis.

Risks and Mitigation Strategies

1. Students may have difficulties using the Occupation Analyzer.
   - Provide a detailed explanation or a video about how to use the Occupation Analyzer.
2. NYU may cooperate with other companies to develop the Occupation Analyzer.
   - Try to deliver the Occupation Analyzer with high quality beforeMay 5, 2021.
   - Persuade sponsors to believe that let me develop the Occupation Analyzer is

the cost-effective way.

Communications Plan

1. Frequency: Once a week.
2. Method: Zoom.
3. Content: Updates, progress reports, and resolution of issues.

Schedule Overview

Project Start Date: Feb 3, 2021

Estimated Project Completion Date: May 5, 2021

Major Milestones

1. Get Functional Specification Document approval before Feb 12, 2021.
2. Performs the TF-IDF similarity scores of resume and jobs, resume and occupations, and jobs and occupations before Mar 19, 2021
3. Performs the cosine similarity scores between resume vector and jobs' vectors before Apr 2, 2021.
4. Develop the shiny app with a function to accept the resume and job files before Apr 9, 2021.
5. Create a GitHub repository with all relevant documents before Apr 16, 2021

External Milestones Affecting the Project

None identified.

Impact of Late Delivery

I will get a negative performance review from project sponsors.

Resources Required

Role	Responsibilities	Duration of work	Qualifications needed
Project Manager	Manage the project	50 hours	Experience in project management
Technical Developer	Code for the project	200 hours	Experience in programming

Facilities , Software, Hardware and other Resources

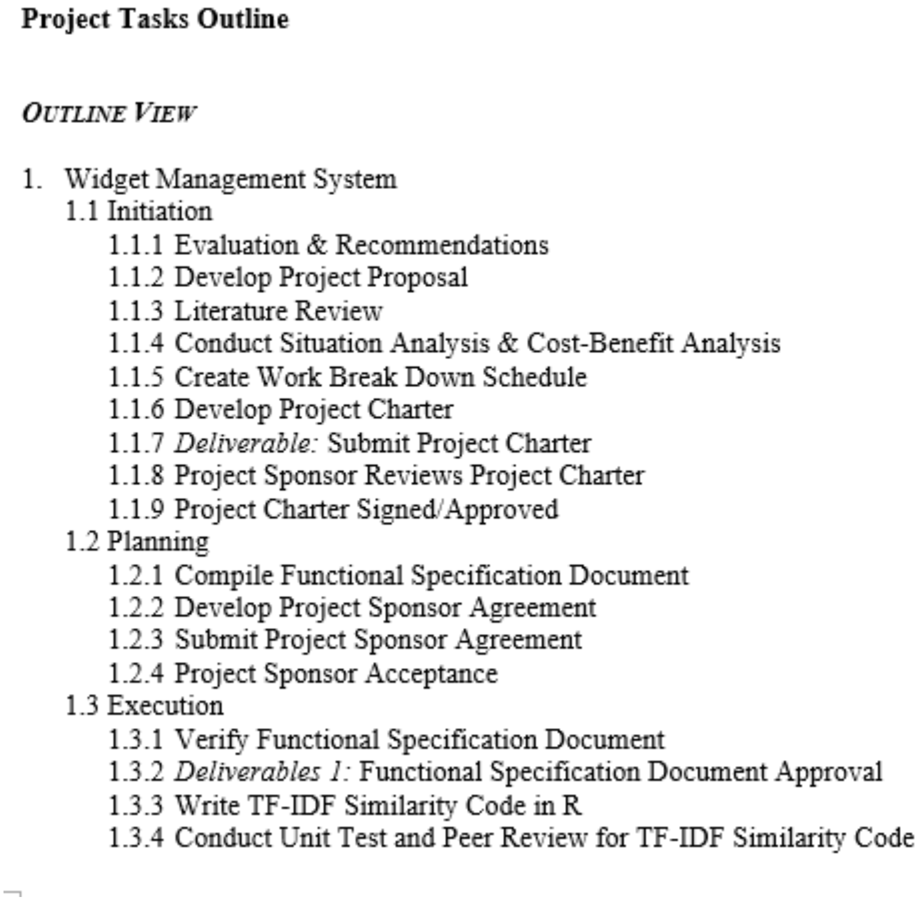
Personal computer, R Studio, and Zoom

Project Evaluation

1. **Project schedule:** refer to Work Breakdown Schedule (Table 5)
2. **Project weekly status report and dashboard:** it will be a verbal report about what I did, what I am doing, and plans for next week.
3. **Project communication plan, issues log, risk register, change:** I will meet my sponsors weekly via Zoom and report to them what I did, what I am doing, and plans for next week. Weekly progress will be documented in the Progress Report. Issues will be documented in the Issues Log. Any changes will be documented in the Change Management Plan.
4. **Project monthly status report:** refer to Appendix H.

Appendix D

Project Plan



Project Tasks Outline

Outline View

1. Widget Management System
   - 1.1 Initiation
     - 1.1.1 Evaluation & Recommendations
     - 1.1.2 Develop Project Proposal
     - 1.1.3 Literature Review
     - 1.1.4 Conduct Situation Analysis & Cost-Benefit Analysis
     - 1.1.5 Create Work Break Down Schedule
     - 1.1.6 Develop Project Charter
     - 1.1.7 *Deliverable:* Submit Project Charter
     - 1.1.8 Project Sponsor Reviews Project Charter
     - 1.1.9 Project Charter Signed/Approved
   - 1.2 Planning
     - 1.2.1 Compile Functional Specification Document
     - 1.2.2 Develop Project Sponsor Agreement
     - 1.2.3 Submit Project Sponsor Agreement
     - 1.2.4 Project Sponsor Acceptance
   - 1.3 Execution
     - 1.3.1 Verify Functional Specification Document
     - 1.3.2 *Deliverables 1:* Functional Specification Document Approval
     - 1.3.3 Write TF-IDF Similarity Code in R
     - 1.3.4 Conduct Unit Test and Peer Review for TF-IDF Similarity Code

- 1.3.5 *Deliverable 2:* Performs the TF-IDF similarity scores of resume and jobs, resume and occupations, and jobs and occupations
- 1.3.6 Write Cosine Similarity Scores Code
- 1.3.7 Conduct Unit Test and Peer Review for Cosine Similarity Scores Code
- 1.3.8 *Deliverable 3:* Performs the Cosine Similarity Scores between resume vector and jobs' vectors
- 1.3.9 Develop a Shiny App Interface
- 1.3.10 *Deliverable 4:* The Shiny App with Function to Accept the Resume and Job Files
- 1.3.11 Create an Accessible GitHub Repository for Recording All the Project Files and Supporting Documentation
- 1.3.12 *Deliverable 5:* A GitHub Repository with All Relevant Documents

1.4 Control

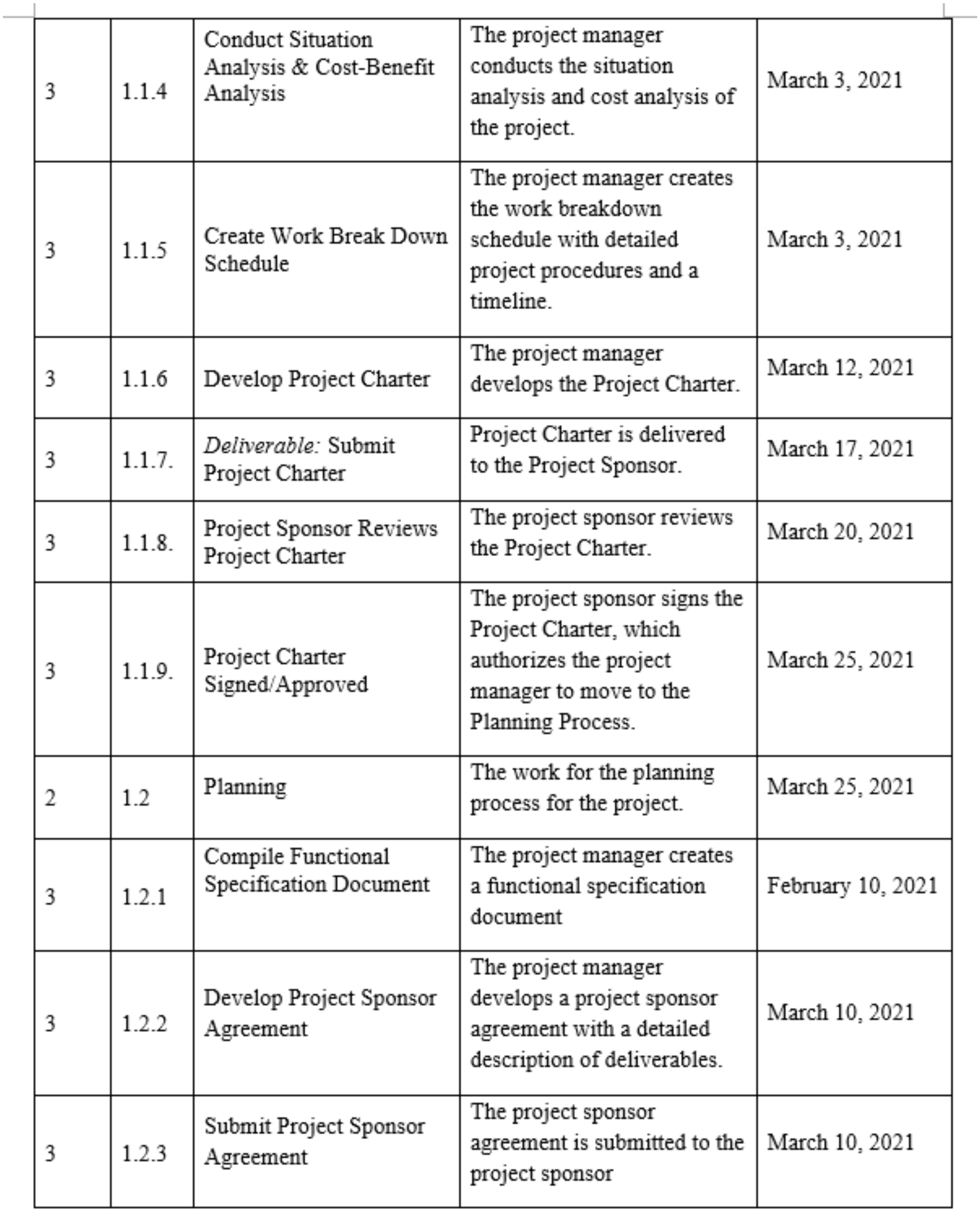
- 1.4.1 Develop Change Management Plan
- 1.4.2 Develop Risk Management Plan

1.5 Closeout

- 1.5.1 Draft Final Project Report
- 1.5.2 Final meeting with the client
- 1.5.3 Compile Project Sponsor Acceptance – Project Completion Signoff
- 1.5.4 Gain Formal Acceptance
- 1.5.5 Present the project
- 1.5.6 Submit Final Project Report with Final Deliverables in GitHub Repository

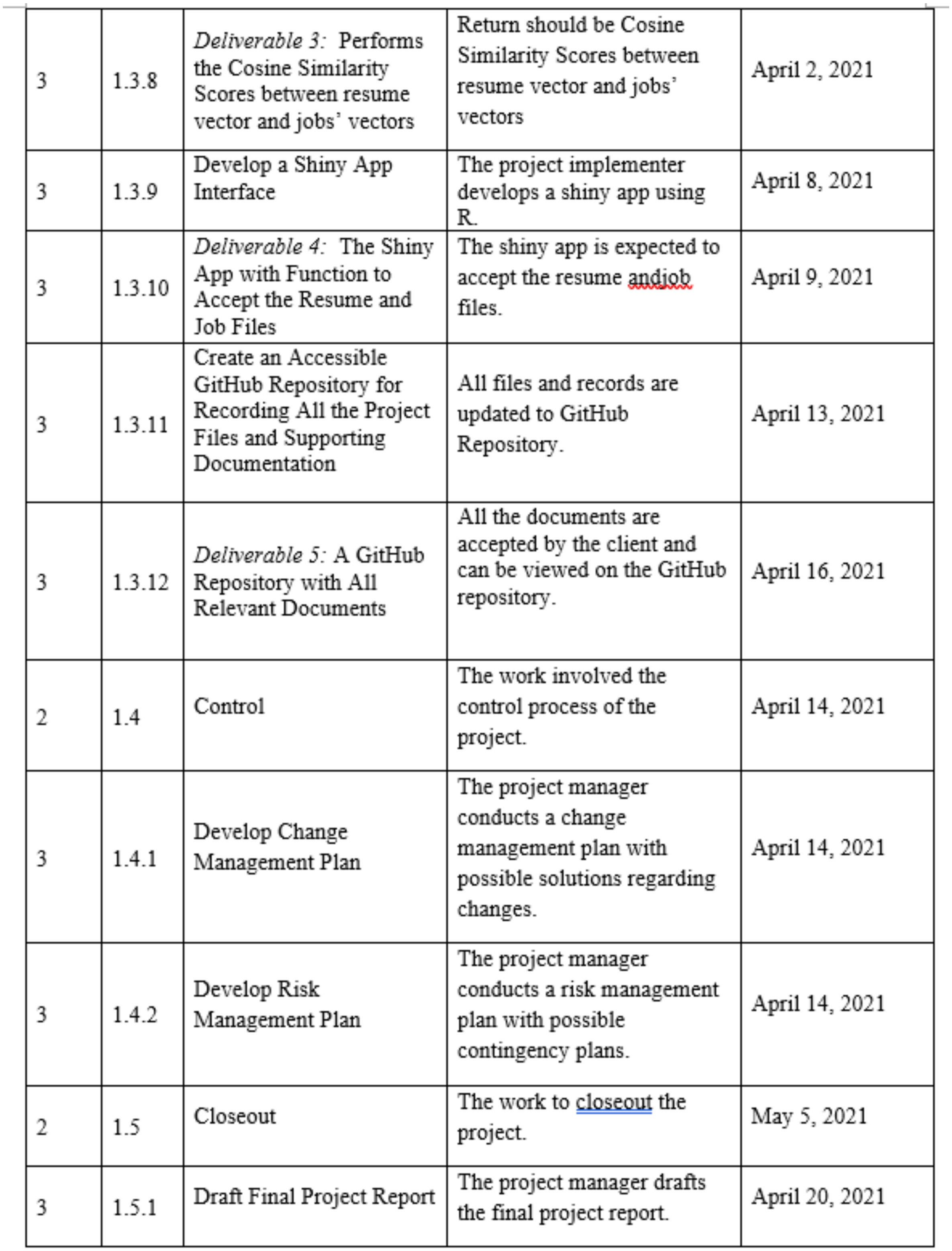
Work Breakdown Task Definition and Schedule

Level	WBS Code	Element Name	Definition	Due By
1	1	Widget Management System	All work to implement a new widget management system.	
2	1.1	Initiation	The work to initiate the project.	March 25, 2021
3	1.1.1	Evaluation & Recommendations	The project manager works with the client to evaluate solution sets and make recommendations.	January 18, 2021
3	1.1.2	Develop Project Proposal	The project manager develops the project proposal.	February 10, 2021
3	1.1.3	Draft Literature Review Research	The project manager finds 10+ pieces of literature related to the project and conducts a literature review.	February 24, 2021

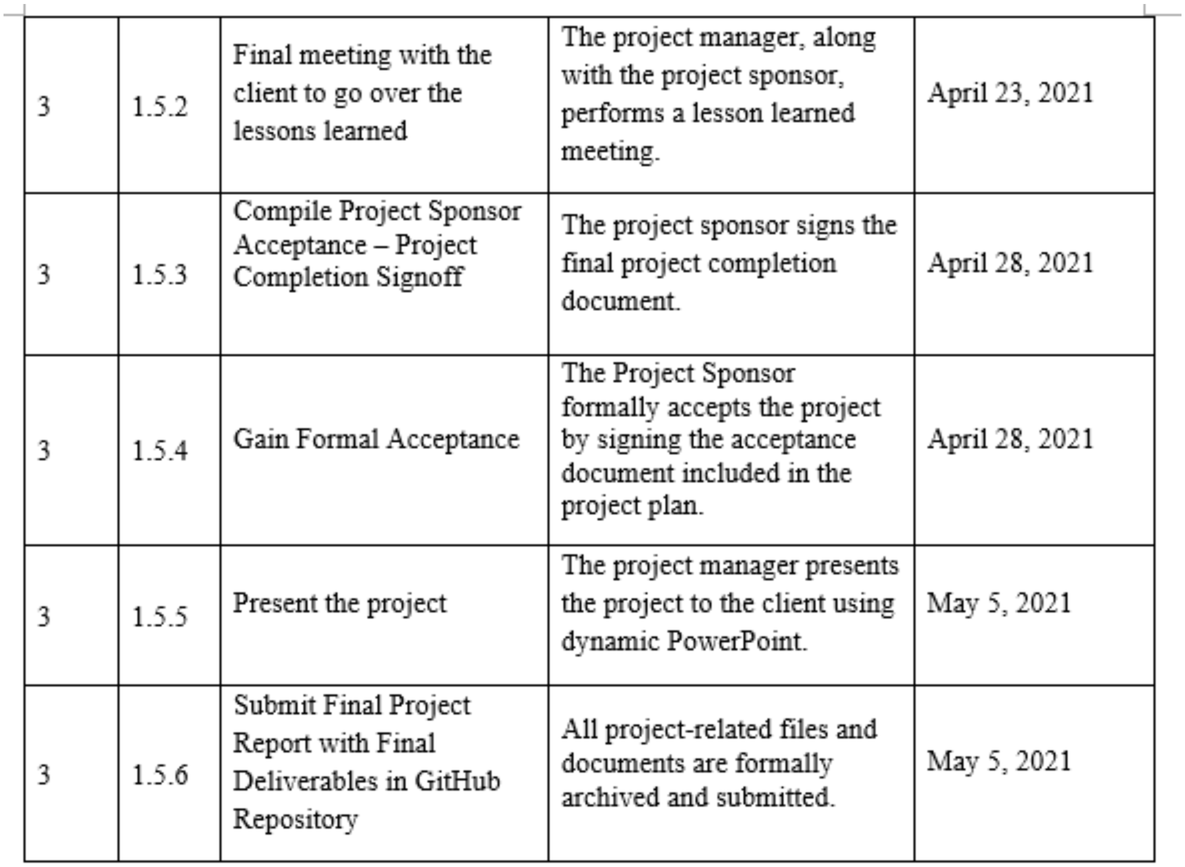


3	1.1.4	Conduct Situation Analysis & Cost-Benefit Analysis	The project manager conducts the situation analysis and cost analysis of the project.	March 3, 2021
3	1.1.5	Create Work Break Down Schedule	The project manager creates the work breakdown schedule with detailed project procedures and a timeline.	March 3, 2021
3	1.1.6	Develop Project Charter	The project manager develops the Project Charter.	March 12, 2021
3	1.1.7.	*Deliverable:* Submit Project Charter	Project Charter is delivered to the Project Sponsor.	March 17, 2021
3	1.1.8.	Project Sponsor Reviews Project Charter	The project sponsor reviews the Project Charter.	March 20, 2021
3	1.1.9.	Project Charter Signed/Approved	The project sponsor signs the Project Charter, which authorizes the project manager to move to the Planning Process.	March 25, 2021
2	1.2	Planning	The work for the planning process for the project.	March 25, 2021
3	1.2.1	Compile Functional Specification Document	The project manager creates a functional specification document	February 10, 2021
3	1.2.2	Develop Project Sponsor Agreement	The project manager develops a project sponsor agreement with a detailed description of deliverables.	March 10, 2021
3	1.2.3	Submit Project Sponsor Agreement	The project sponsor agreement is submitted to the project sponsor	March 10, 2021

3	1.2.4	Project Sponsor Acceptance	Under the project sponsor's confirmation and approval, the project manager begins to move on execution step.	March 10, 2021
2	1.3	Execution	Work involved executing the project.	April 16, 2021
2	1.3.1	Verify Functional Specification Documents	The functional specification documents described a detailed step-by-step outline of each item's functionality and user flow for different user roles.	February 12, 2021
3	1.3.2	*Deliverables 1:* Functional Specification Document Approval	The project plan is approved, and the project manager has permission to proceed to execute the project according to the project plan.	February 12, 2021
3	1.3.3	Write TF-IDF Similarity Code in R	The project implementer should finish the TF-IDF Similarity code in R.	March 11, 2021
3	1.3.4	Conduct Unit Test and Peer Review for TF-IDF Similarity Code	The project implementer should conduct unit tests and peer review TF-IDF similarity code.	March 12, 2021
3	1.3.5	*Deliverable 2:* Performs the TF-IDF similarity scores of resume and jobs, resume and occupations, and jobs and occupations	The TF-IDF similarity scoring will return TF-IDF similarity scores of resume and jobs, resume and occupations, and jobs and occupations.	March 19, 2021
3	1.3.6	Write Cosine Similarity Scores Code	The project implementer should finish the Cosine Similarity Scores code in R.	March 25, 2021
3	1.3.7	Conduct Unit Test and Peer Review for Cosine Similarity Scores Code	The project implementer should conduct unit tests and peer review Cosine Similarity Scores code.	March 26, 2021



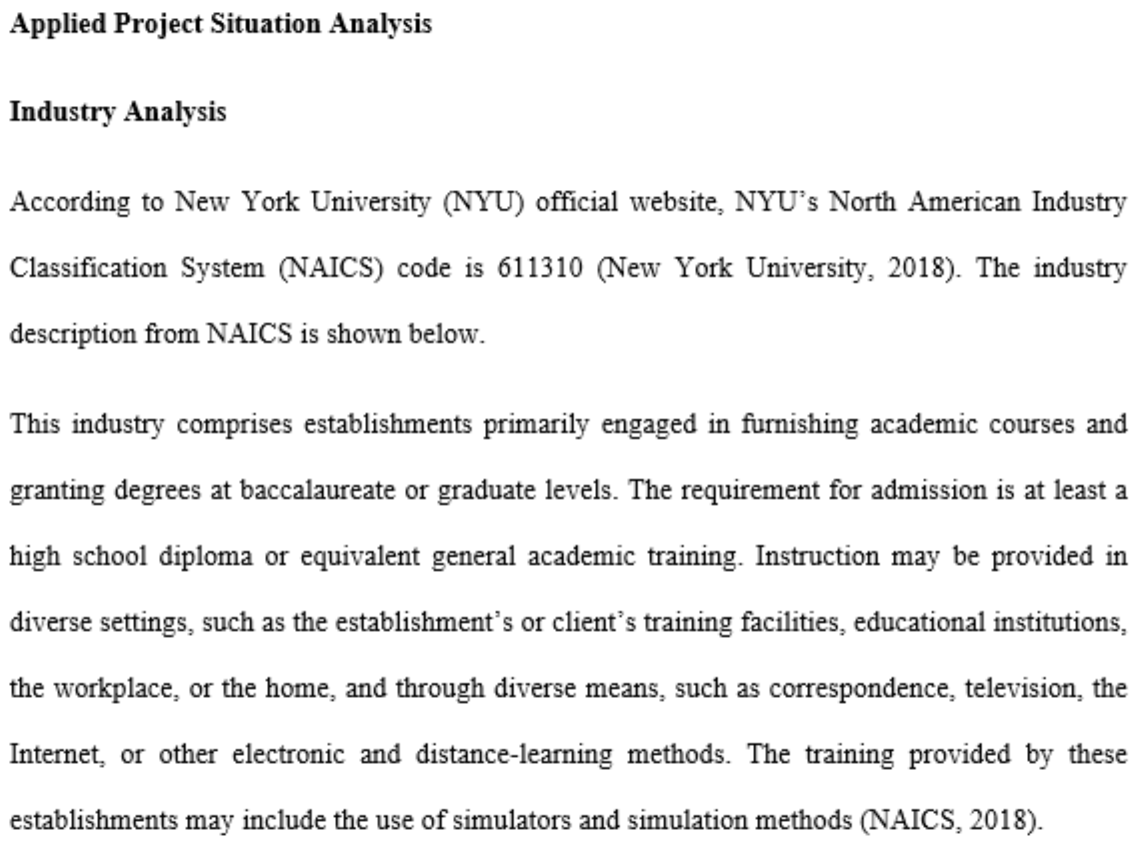
3	1.3.8	*Deliverable 3:* Performs the Cosine Similarity Scores between resume vector and jobs' vectors	Return should be Cosine Similarity Scores between resume vector and jobs' vectors	April 2, 2021
3	1.3.9	Develop a Shiny App Interface	The project implementer develops a shiny app using R.	April 8, 2021
3	1.3.10	*Deliverable 4:* The Shiny App with Function to Accept the Resume and Job Files	The shiny app is expected to accept the resume andjob files.	April 9, 2021
3	1.3.11	Create an Accessible GitHub Repository for Recording All the Project Files and Supporting Documentation	All files and records are updated to GitHub Repository.	April 13, 2021
3	1.3.12	*Deliverable 5:* A GitHub Repository with All Relevant Documents	All the documents are accepted by the client and can be viewed on the GitHub repository.	April 16, 2021
2	1.4	Control	The work involved the control process of the project.	April 14, 2021
3	1.4.1	Develop Change Management Plan	The project manager conducts a change management plan with possible solutions regarding changes.	April 14, 2021
3	1.4.2	Develop Risk Management Plan	The project manager conducts a risk management plan with possible contingency plans.	April 14, 2021
2	1.5	Closeout	The work to closeout the project.	May 5, 2021
3	1.5.1	Draft Final Project Report	The project manager drafts the final project report.	April 20, 2021



3	1.5.2	Final meeting with the client to go over the lessons learned	The project manager, along with the project sponsor, performs a lesson learned meeting.	April 23, 2021
3	1.5.3	Compile Project Sponsor Acceptance – Project Completion Signoff	The project sponsor signs the final project completion document.	April 28, 2021
3	1.5.4	Gain Formal Acceptance	The Project Sponsor formally accepts the project by signing the acceptance document included in the project plan.	April 28, 2021
3	1.5.5	Present the project	The project manager presents the project to the client using dynamic PowerPoint.	May 5, 2021
3	1.5.6	Submit Final Project Report with Final Deliverables in GitHub Repository	All project-related files and documents are formally archived and submitted.	May 5, 2021

Appendix E

Situational Analysis



Applied Project Situation Analysis

Industry Analysis

According to New York University (NYU) official website, NYU's North American Industry Classification System (NAICS) code is 611310 (New York University, 2018). The industry description from NAICS is shown below.

This industry comprises establishments primarily engaged in furnishing academic courses and granting degrees at baccalaureate or graduate levels. The requirement for admission is at least a high school diploma or equivalent general academic training. Instruction may be provided in diverse settings, such as the establishment's or client's training facilities, educational institutions, the workplace, or the home, and through diverse means, such as correspondence, television, the Internet, or other electronic and distance-learning methods. The training provided by these establishments may include the use of simulators and simulation methods (NAICS, 2018).

According to IBIS World's report, the university industry's revenue is expected to increase at an annualized rate of 1.1% to $580.7 billion over the five years to 2021, including a forecast increase of 0.3% in 2021 alone (Le, 2021). As for the profit margin, it is expected to average 11.1% from 2016 to 2021.

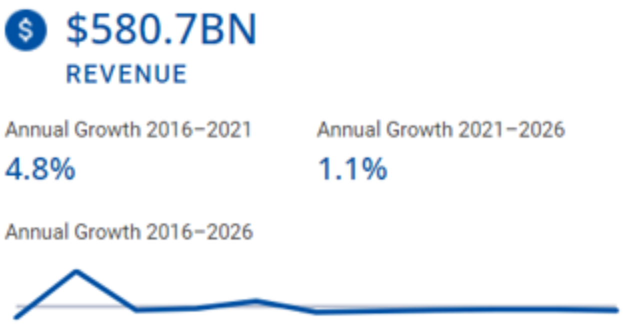


Figure 1Industry Revenue, IBIS World

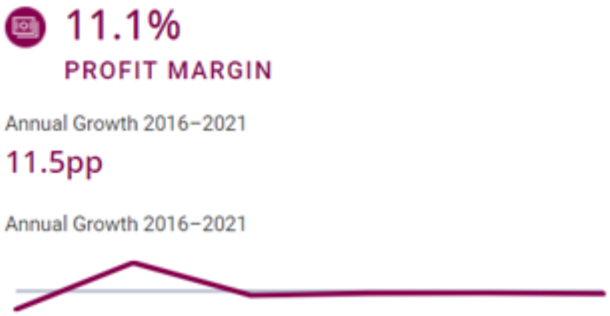


Figure 2Industry Profit Margin, IBIS World

In the next five years, since the job market is expected to strengthen, higher education demand may slow down. In addition, colleges and universities are expected to endure increased competition from the massive open online course offering platforms like Academy of Mine, Udemy, and so on, which could pull students away from traditional universities by providing low-cost education. Besides, the number of students graduating from high schools directly correlates with the growth of the number of college freshmen. The high school retention rate is expected to decrease in 2021, posing a potential threat to the industry marginally (Le, 2021). Therefore, how to attract freshmen to maintain revenue is a vital issue for industry operators. Also, this urgent industry issue justifies the need for this project (Occupation Analyzer). Developing tools to help students find jobs could be a selling point for NYU.

Competitors

Location and education quality are two important factors for higher education institutions. Considering the two factors, in New York City, NYU's competitors could be Columbus University, State University of New York, and Pace University. In the United States, competitors could be expanded to the Top 20 universities on the U.S. News University Ranking like Yale University, Standford University, and Duke University.

Stakeholders

The stakeholders of Occupation Analyzer could be project sponsors, the project manager, the project implementer, NYU students, NYU faculties, company recruiters, NYU shareholders, NYU students' families, U.S. Department of Labor, and NYU executives. The project sponsors are Dr. Andres Fortino and Dr. Hui Soo Chae. They will provide scope, data, budget, and technical support to this project. The project manager and project implementer are Yuchao Wu, who will write project plans, write relevant documents, and develop the tool. NYU students and faculties benefit from the project, and they will be users giving valuable feedback. Using Occupation Analyzer, students and faculties can easily find jobs that best match their resumes. If students find a dream job by using Occupation Analyzer, it may increase his/her families' happiness. To some extend, company recruiters will be able to shorten the time to find appropriate candidates. If this tool is very useful, it may attract more students to apply for NYU, increasing NYU's revenue and benefit NYU shareholders and executives. The occupation data is collected from the U.S. Department of Labor.

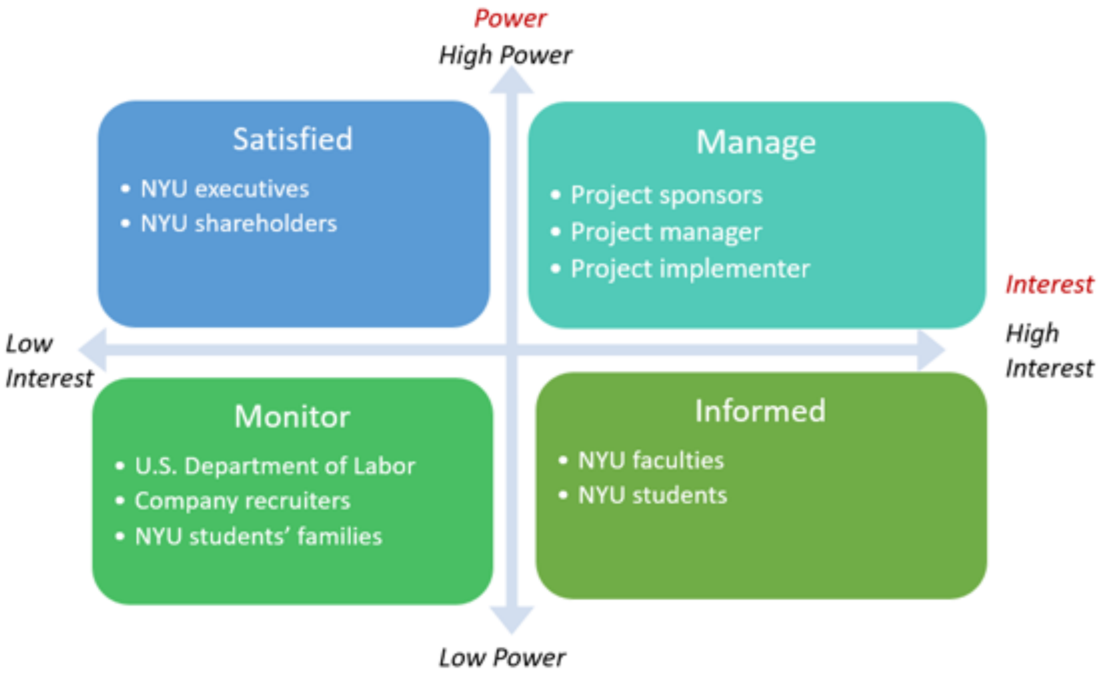


Figure 3Stakeholder Analysis

Porter's Five-Forces Model

The Threat of Potential New Entrants

The entry barrier of the university-industry is high. Key success factors of this industry include experienced professors, having a good reputation, ability to take advantage of government subsidies and other grants, ability to raise revenue from additional sources, ability to respond to students' needs, and effective cost controls (Le, 2021). Each of the above factors requires lots of money, resources, time, and human resources. No players can quickly enter this industry. Thus, the threat of potential new entrants is low.

The Threat of Substitutes

Online education platforms could be a substitute for universities. However, this kind of platform has limitations like limited course time, no research sponsors, no laboratory, no networking events, inconvenient Q&A, low recognition on the job market, and limited professional teachers. Thus, the threat of substitutes is low.

Bargaining Power of Customers

Students are customers of the university-industry. Since the high school retention rate is expected to decrease in 2021 and the job market is expected to strengthen, there will be fewer students apply for universities. Also, elite students normally will get offers from several top universities, and they have the power to choose. Thus, universities are in a passive position, which means customers' bargaining power should be medium.

Bargaining Power of Suppliers

The suppliers of the university-industry could be instructors and infrastructure providers. There are many instructors on the market, but top scholars will always be competed for by universities. Also, there are many infrastructure providers on the market, but some infrastructure like high-performance computers can only be provided by limited companies. Thus, the bargaining power of suppliers is low.

Competitive Rivalry

The competitors of a university are other universities. There are 1,400 universities around the world. Although NYU is the top university, location, tuition, professors, research power, fundings, and many other factors affect students' choice. Thus, competitive rivalry is high.

References

Le, T. (2021, February). *Colleges & Universities in the U.S.* Retrieved from IBIS World: https://my-ibisworld-com.proxy.library.nyu.edu/us/en/industry/61131a/industry-at-a-glance#key-statistics-snapshot

NAICS. (2018). *Colleges, Universities, and Professional Schools*. Retrieved from NAICS Association: https://www.naics.com/naics-code-description/?code=611310

New York University. (2018). *Ohter Information*. Retrieved from New York University: https://www.nyu.edu/research/resources-and-support-offices/sponsored-programs/proposal-development/nyu-administrative-information-for-proposal-preparation/other-information.html

Table of Figures

Cost/Benefit Analysis

Tangible Benefits

Benefit: No identifiable benefits

Value & Probability: N/A

Assumptions Driving Value: N/A

Intangible Benefits

Benefit: N/A

Value & Probability: N/A

Assumptions Driving Value: N/A

Cost Categories	**Amount**
Internal Labor hours	N/A
External Costs	N/A
Labor (consultants, contract labor)	N/A
Equipment, hardware or software	N/A
List other costs such as travel & training	N/A

Financial Return

Breakeven analysis

Other Business Benefits

The intangible benefit of Occupation Analyzer could be to increase NYU’s reputation and attract more students. But it is hard to quantify. Based on the expected working hour and compensation rate, the project implementor’s salary could be $29,000. This project costs nothing because this is a non-paid project.

Appendix F

Risk Management Plan

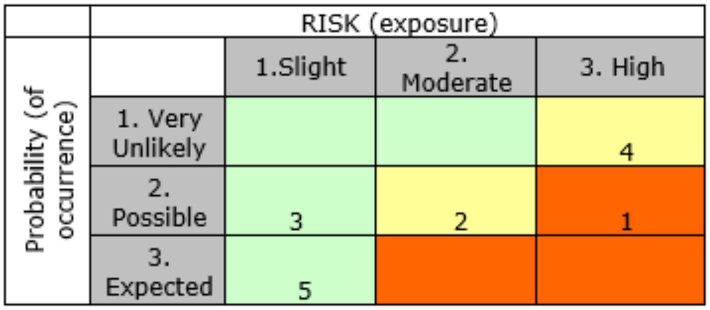
Project

Occupation Analyzer: a text analytics tool in R that accepts a user's resume and scores the resume against a list of jobs, and the return is the scoring of the target jobs to the resume via occupation-matching.

Risks

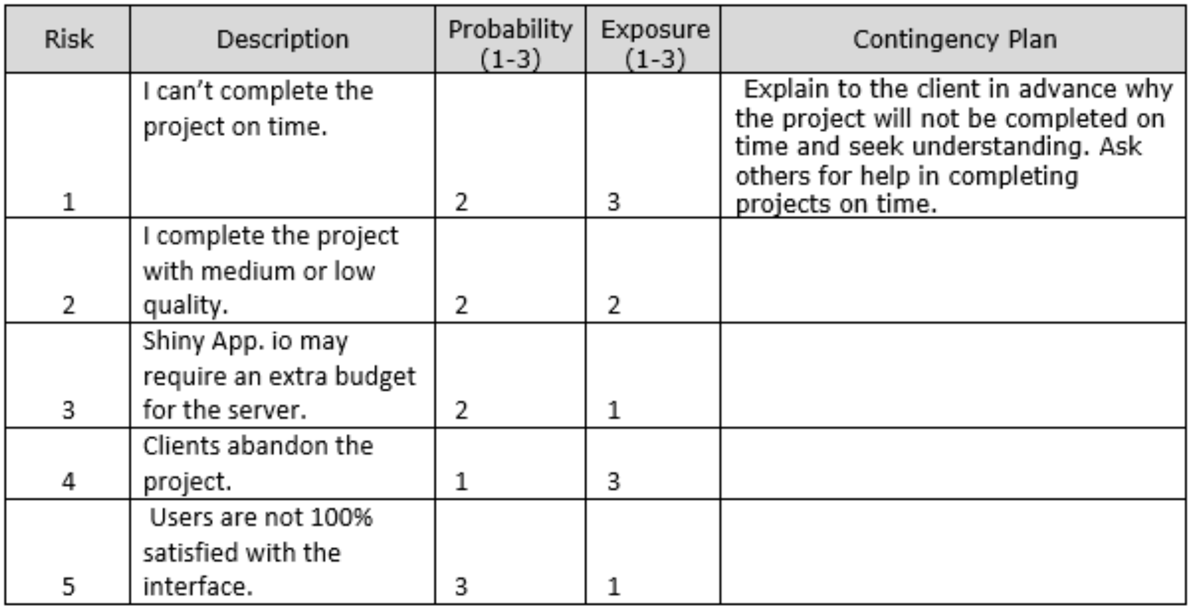
Number	Risk	Probability Score (1,2 or 3)	Impact Score (1,2 or 3)
1	I can’t complete the project on time.	2	3
2	I can’t complete the project with high quality.	2	2
3	The project requires an extra budget.	2	1
4	Clients abandon the project.	1	3
5	The tool is not user-friendly enough.	3	1

Risk Matrix



		RISK (exposure)		
Probability (of occurrence)		1.Slight	2. Moderate	3. High
	1. Very Unlikely	🟩	🟩	🟨 4
	2. Possible	🟩 3	🟨 2	🟧 1
	3. Expected	🟩 5	🟧	🟧

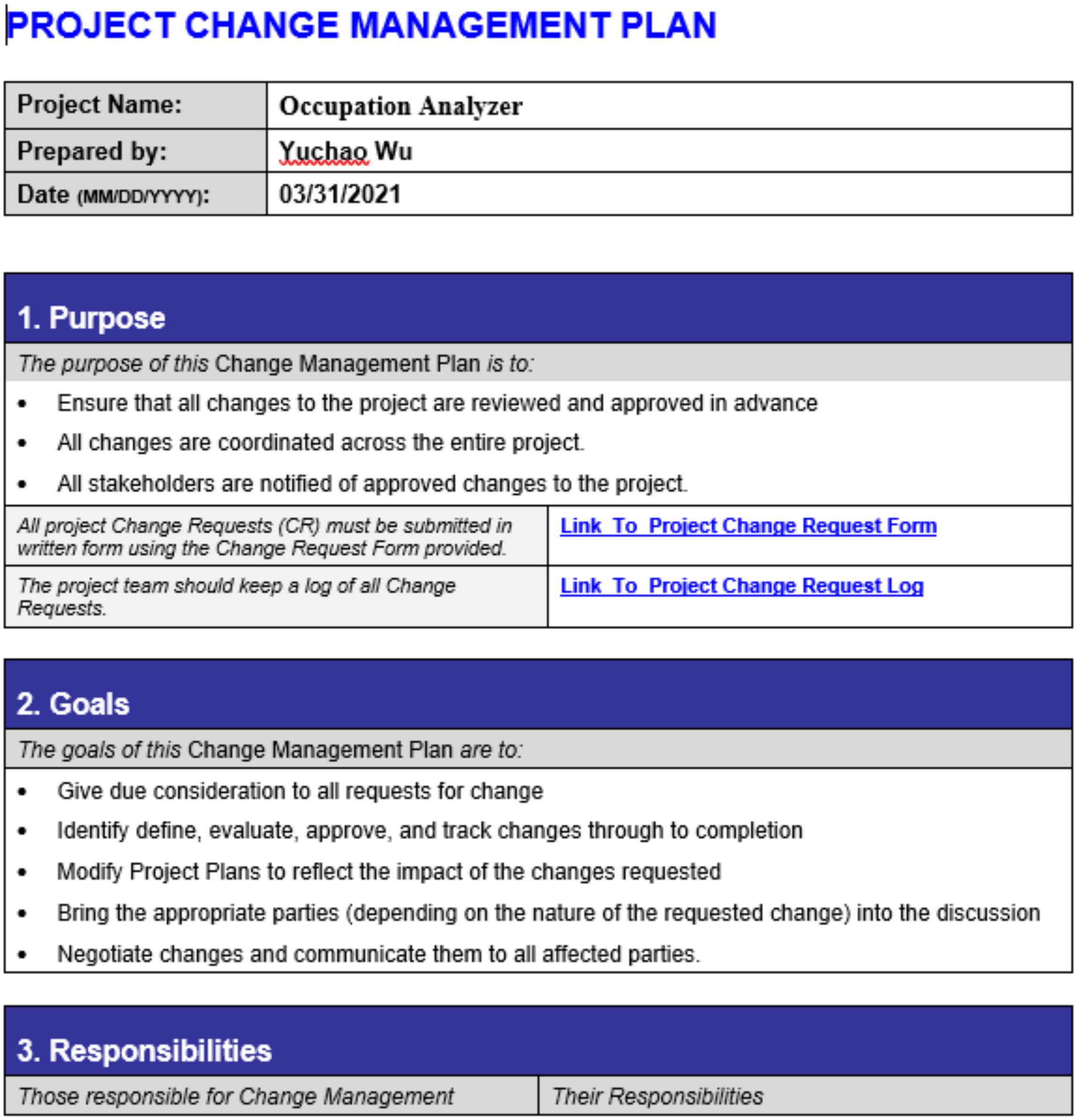
Contingency Plan



Risk	Description	Probability (1-3)	Exposure (1-3)	Contingency Plan
1	I can't complete the project on time.	2	3	Explain to the client in advance why the project will not be completed on time and seek understanding. Ask others for help in completing projects on time.
2	I complete the project with medium or low quality.	2	2	
3	Shiny App. io may require an extra budget for the server.	2	1	
4	Clients abandon the project.	1	3	
5	Users are not 100% satisfied with the interface.	3	1	

Appendix G

Change Management Plan



PROJECT CHANGE MANAGEMENT PLAN

Project Name:	**Occupation Analyzer**
Prepared by:	**Yuchao Wu**
Date (MM/DD/YYYY):	**03/31/2021**

1. Purpose

The purpose of this Change Management Plan *is to:*

- Ensure that all changes to the project are reviewed and approved in advance
- All changes are coordinated across the entire project.
- All stakeholders are notified of approved changes to the project.

All project Change Requests (CR) must be submitted in written form using the Change Request Form provided.	Link To Project Change Request Form
The project team should keep a log of all Change Requests.	Link To Project Change Request Log

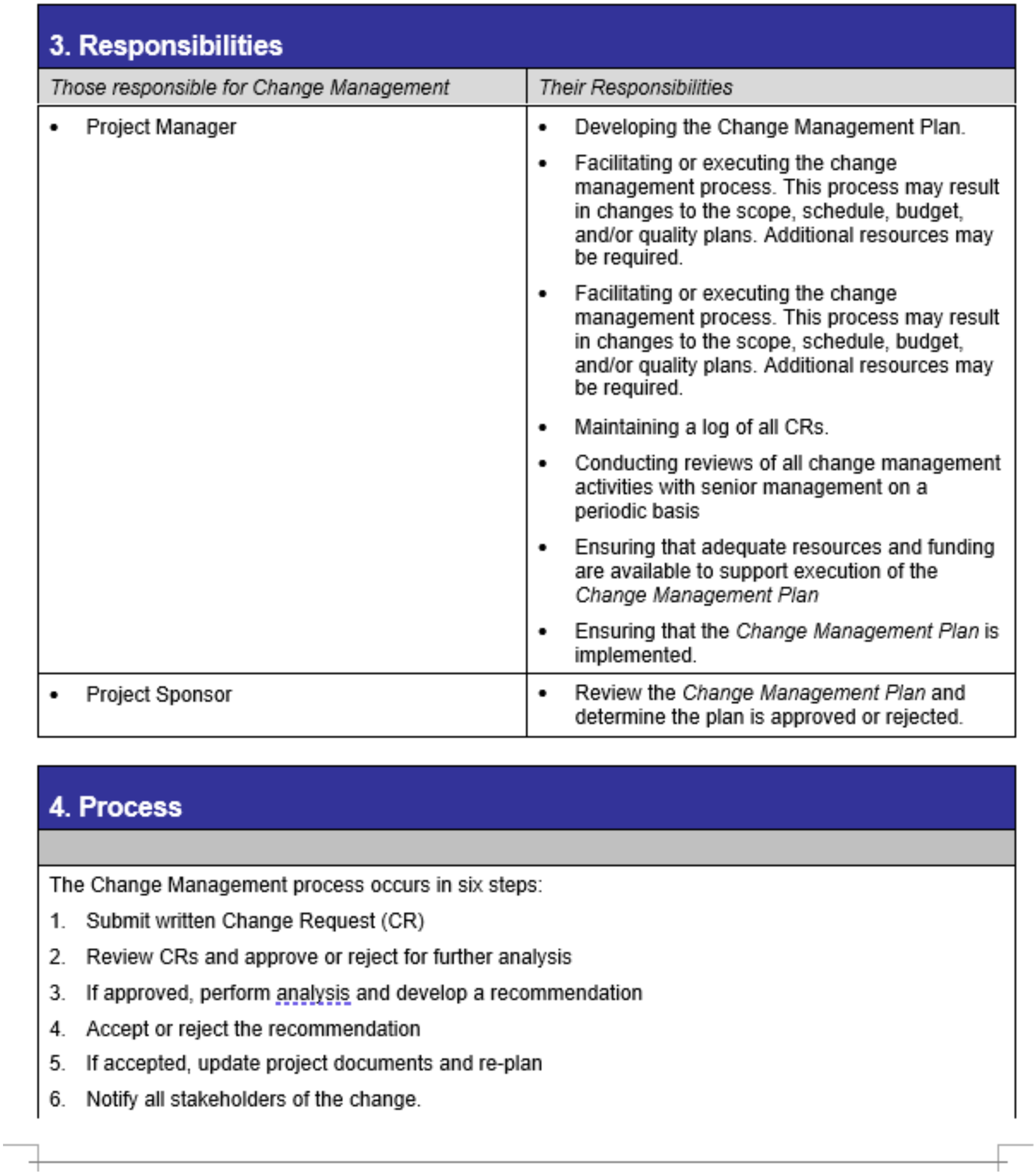
2. Goals

The goals of this Change Management Plan *are to:*

- Give due consideration to all requests for change
- Identify define, evaluate, approve, and track changes through to completion
- Modify Project Plans to reflect the impact of the changes requested
- Bring the appropriate parties (depending on the nature of the requested change) into the discussion
- Negotiate changes and communicate them to all affected parties.

3. Responsibilities

Those responsible for Change Management	*Their Responsibilities*



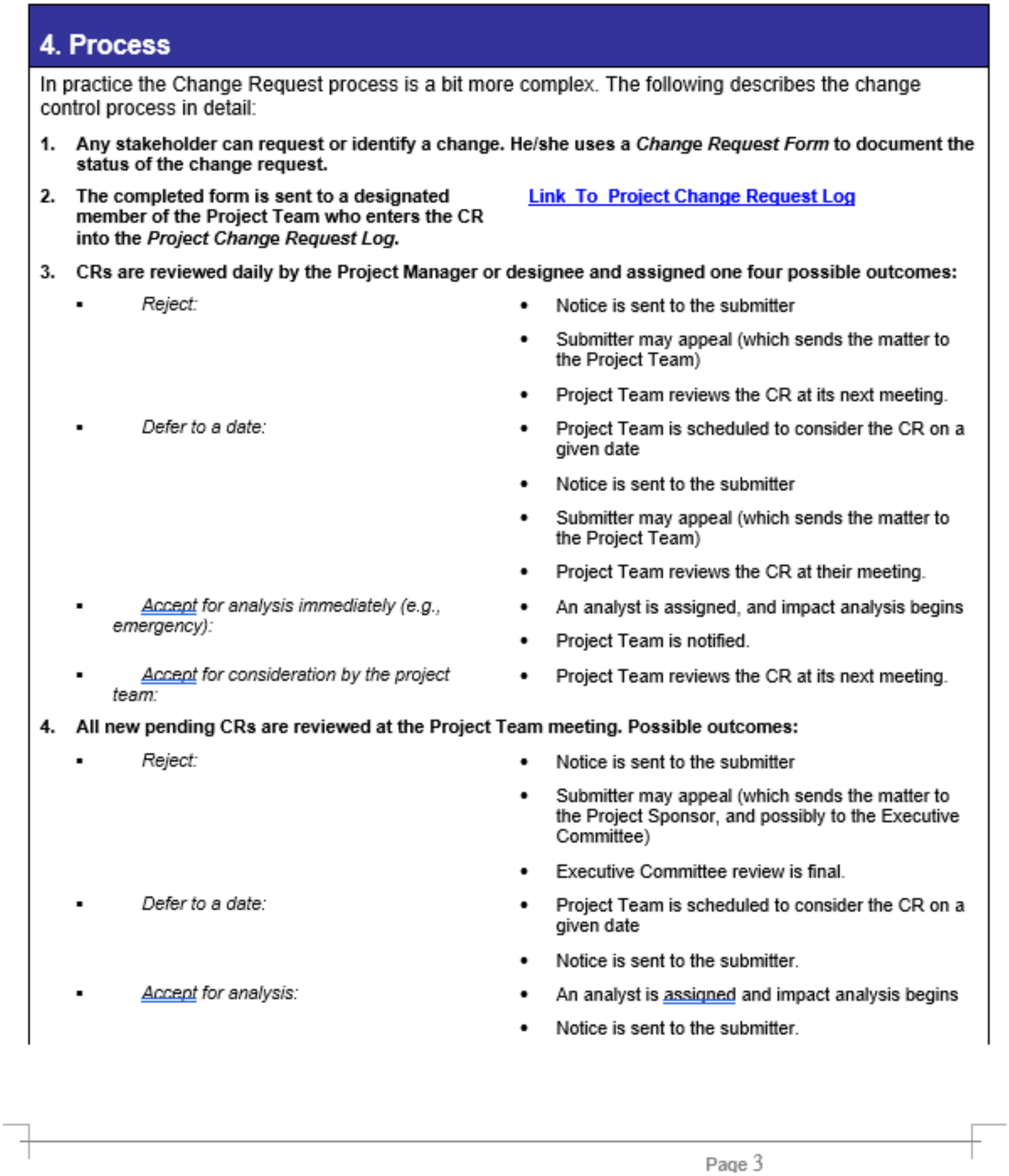
3. Responsibilities

Those responsible for Change Management	*Their Responsibilities*
• Project Manager	• Developing the Change Management Plan. • Facilitating or executing the change management process. This process may result in changes to the scope, schedule, budget, and/or quality plans. Additional resources may be required. • Facilitating or executing the change management process. This process may result in changes to the scope, schedule, budget, and/or quality plans. Additional resources may be required. • Maintaining a log of all CRs. • Conducting reviews of all change management activities with senior management on a periodic basis • Ensuring that adequate resources and funding are available to support execution of the *Change Management Plan* • Ensuring that the *Change Management Plan* is implemented.
• Project Sponsor	• Review the *Change Management Plan* and determine the plan is approved or rejected.

4. Process

The Change Management process occurs in six steps:

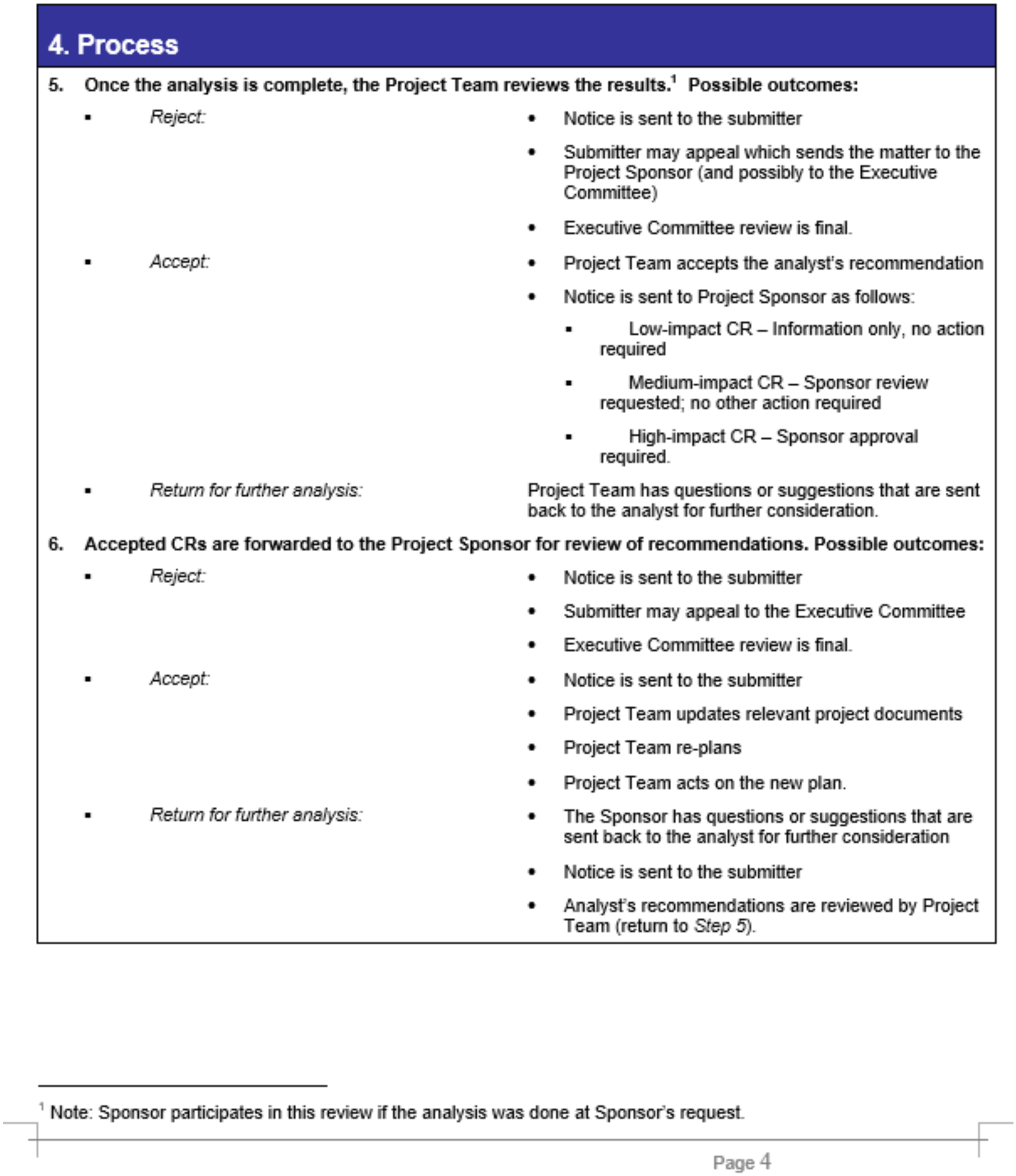
1. Submit written Change Request (CR)
2. Review CRs and approve or reject for further analysis
3. If approved, perform analysis and develop a recommendation
4. Accept or reject the recommendation
5. If accepted, update project documents and re-plan
6. Notify all stakeholders of the change.



4. Process

In practice the Change Request process is a bit more complex. The following describes the change control process in detail:

1. **Any stakeholder can request or identify a change. He/she uses a *Change Request Form* to document the status of the change request.**
2. **The completed form is sent to a designated member of the Project Team who enters the CR into the *Project Change Request Log*.** [Link To Project Change Request Log]
3. **CRs are reviewed daily by the Project Manager or designee and assigned one four possible outcomes:**
   - *Reject:*
     - Notice is sent to the submitter
     - Submitter may appeal (which sends the matter to the Project Team)
     - Project Team reviews the CR at its next meeting.
   - *Defer to a date:*
     - Project Team is scheduled to consider the CR on a given date
     - Notice is sent to the submitter
     - Submitter may appeal (which sends the matter to the Project Team)
     - Project Team reviews the CR at their meeting.
   - *Accept for analysis immediately (e.g., emergency):*
     - An analyst is assigned, and impact analysis begins
     - Project Team is notified.
   - *Accept for consideration by the project team:*
     - Project Team reviews the CR at its next meeting.
4. **All new pending CRs are reviewed at the Project Team meeting. Possible outcomes:**
   - *Reject:*
     - Notice is sent to the submitter
     - Submitter may appeal (which sends the matter to the Project Sponsor, and possibly to the Executive Committee)
     - Executive Committee review is final.
   - *Defer to a date:*
     - Project Team is scheduled to consider the CR on a given date
     - Notice is sent to the submitter.
   - *Accept for analysis:*
     - An analyst is assigned and impact analysis begins
     - Notice is sent to the submitter.



4. Process

5. Once the analysis is complete, the Project Team reviews the results.[1] Possible outcomes:

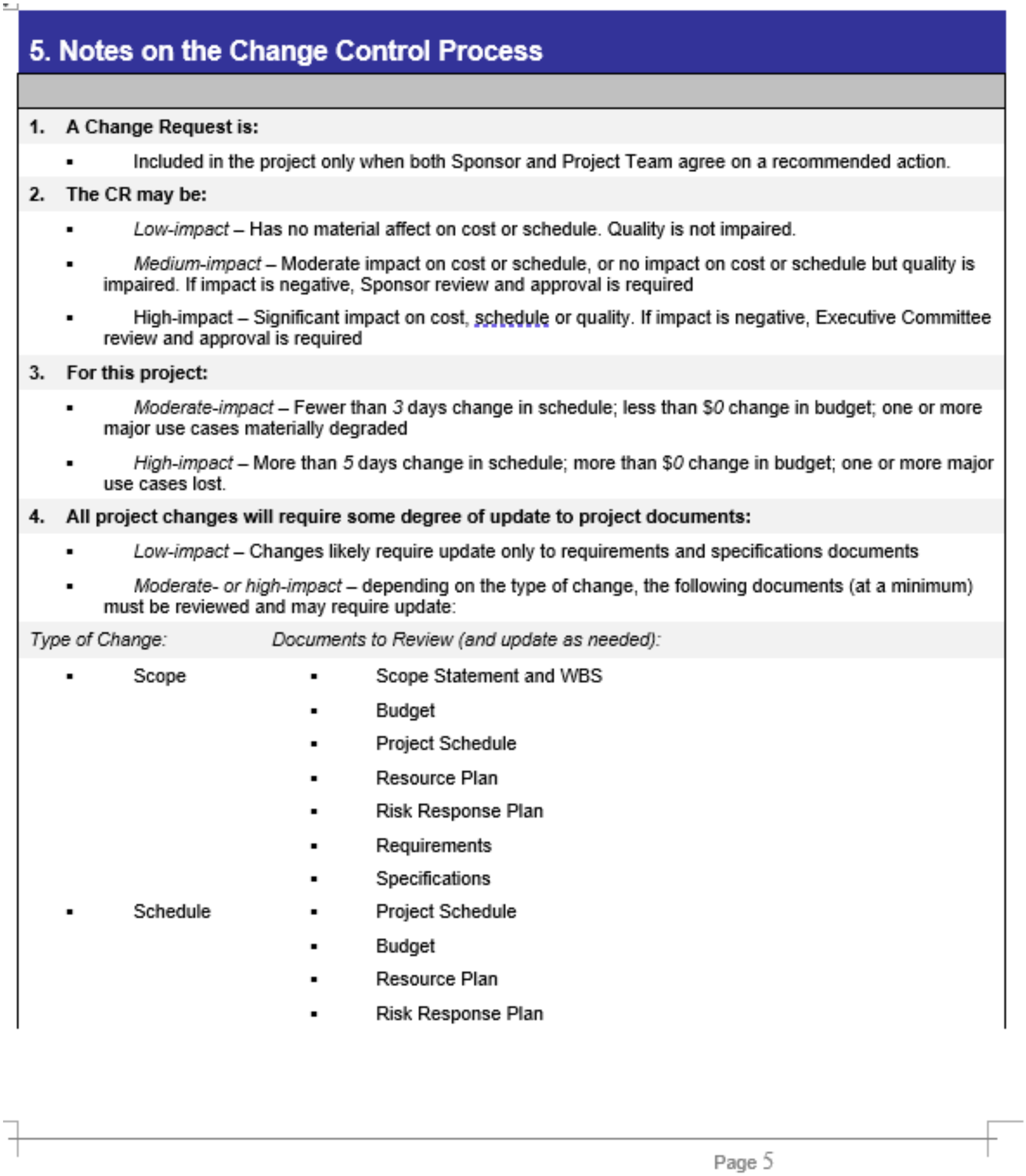
▪ *Reject:*	• Notice is sent to the submitter • Submitter may appeal which sends the matter to the Project Sponsor (and possibly to the Executive Committee) • Executive Committee review is final.
▪ *Accept:*	• Project Team accepts the analyst's recommendation • Notice is sent to Project Sponsor as follows: ▪ Low-impact CR – Information only, no action required ▪ Medium-impact CR – Sponsor review requested; no other action required ▪ High-impact CR – Sponsor approval required.
▪ *Return for further analysis:*	Project Team has questions or suggestions that are sent back to the analyst for further consideration.

6. Accepted CRs are forwarded to the Project Sponsor for review of recommendations. Possible outcomes:

▪ *Reject:*	• Notice is sent to the submitter • Submitter may appeal to the Executive Committee • Executive Committee review is final.
▪ *Accept:*	• Notice is sent to the submitter • Project Team updates relevant project documents • Project Team re-plans • Project Team acts on the new plan.
▪ *Return for further analysis:*	• The Sponsor has questions or suggestions that are sent back to the analyst for further consideration • Notice is sent to the submitter • Analyst's recommendations are reviewed by Project Team (return to *Step 5*).

[1] Note: Sponsor participates in this review if the analysis was done at Sponsor's request.

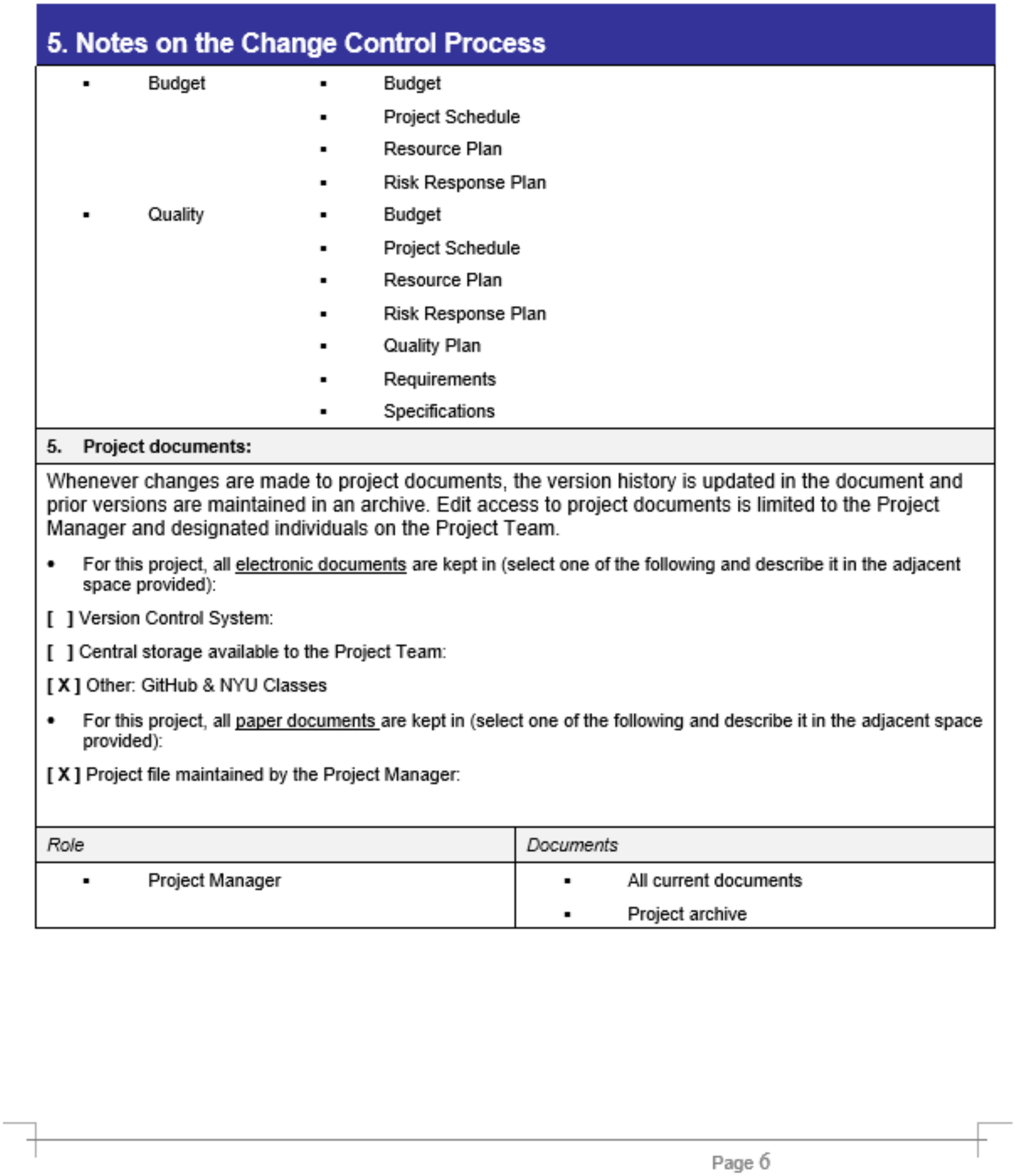
Page 4



5. Notes on the Change Control Process

1. **A Change Request is:**
   - Included in the project only when both Sponsor and Project Team agree on a recommended action.
2. **The CR may be:**
   - *Low-impact* – Has no material affect on cost or schedule. Quality is not impaired.
   - *Medium-impact* – Moderate impact on cost or schedule, or no impact on cost or schedule but quality is impaired. If impact is negative, Sponsor review and approval is required
   - High-impact – Significant impact on cost, schedule or quality. If impact is negative, Executive Committee review and approval is required
3. **For this project:**
   - *Moderate-impact* – Fewer than *3* days change in schedule; less than *$0* change in budget; one or more major use cases materially degraded
   - *High-impact* – More than *5* days change in schedule; more than $*0* change in budget; one or more major use cases lost.
4. **All project changes will require some degree of update to project documents:**
   - *Low-impact* – Changes likely require update only to requirements and specifications documents
   - *Moderate- or high-impact* – depending on the type of change, the following documents (at a minimum) must be reviewed and may require update:

Type of Change:	*Documents to Review (and update as needed):*
▪ Scope	▪ Scope Statement and WBS ▪ Budget ▪ Project Schedule ▪ Resource Plan ▪ Risk Response Plan ▪ Requirements ▪ Specifications
▪ Schedule	▪ Project Schedule ▪ Budget ▪ Resource Plan ▪ Risk Response Plan



5. Notes on the Change Control Process

▪ Budget	▪ Budget ▪ Project Schedule ▪ Resource Plan ▪ Risk Response Plan
▪ Quality	▪ Budget ▪ Project Schedule ▪ Resource Plan ▪ Risk Response Plan ▪ Quality Plan ▪ Requirements ▪ Specifications

5. Project documents:

Whenever changes are made to project documents, the version history is updated in the document and prior versions are maintained in an archive. Edit access to project documents is limited to the Project Manager and designated individuals on the Project Team.

- For this project, all electronic documents are kept in (select one of the following and describe it in the adjacent space provided):

[ ] Version Control System:

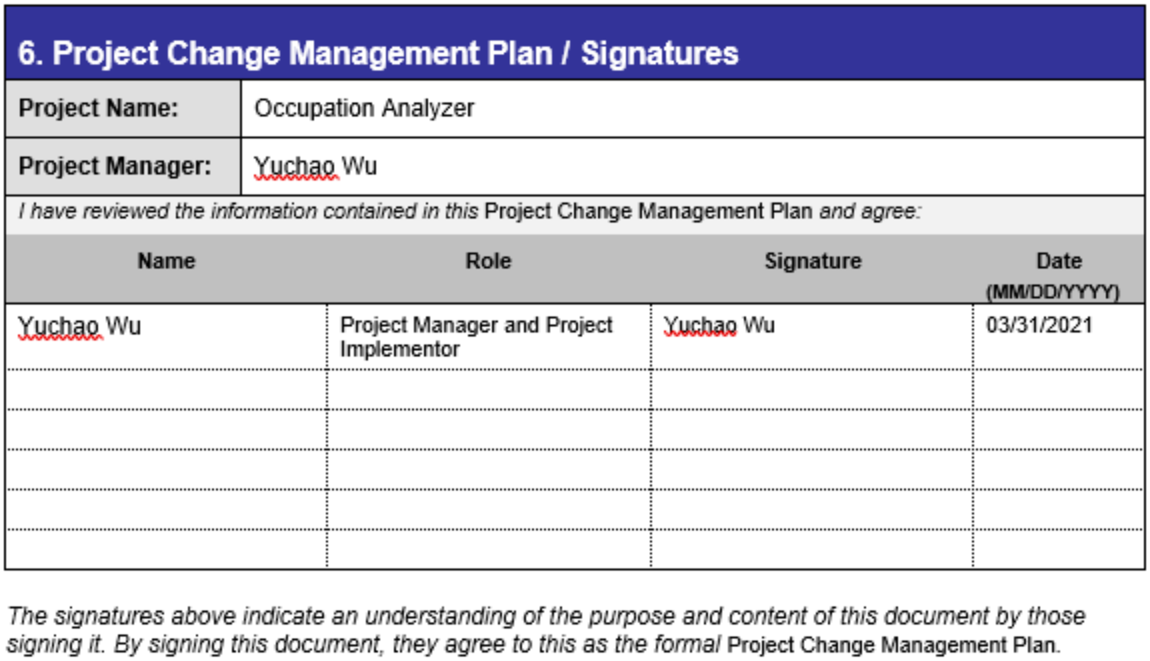
[ ] Central storage available to the Project Team:

[X] Other: GitHub & NYU Classes

- For this project, all paper documents are kept in (select one of the following and describe it in the adjacent space provided):

[X] Project file maintained by the Project Manager:

Role	*Documents*
▪ Project Manager	▪ All current documents ▪ Project archive



6. Project Change Management Plan / Signatures

Project Name:	Occupation Analyzer
Project Manager:	Yuchao Wu

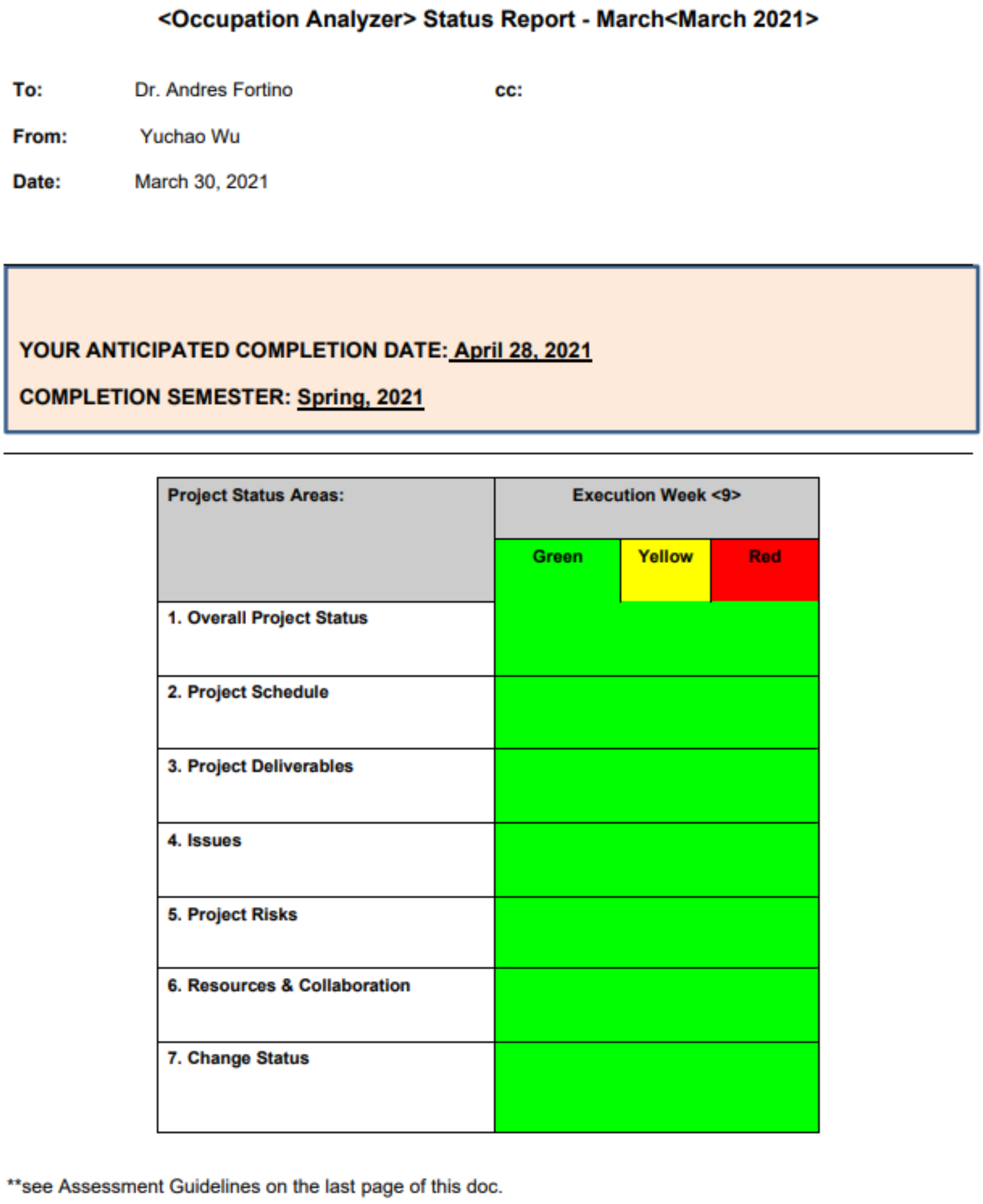
I have reviewed the information contained in this Project Change Management Plan *and agree:*

Name	Role	Signature	Date (MM/DD/YYYY)
Yuchao Wu	Project Manager and Project Implementor	Yuchao Wu	03/31/2021

The signatures above indicate an understanding of the purpose and content of this document by those signing it. By signing this document, they agree to this as the formal Project Change Management Plan.

Appendix H

Status Reports



<Occupation Analyzer> Status Report - March<March 2021>

To: Dr. Andres Fortino **cc:**

From: Yuchao Wu

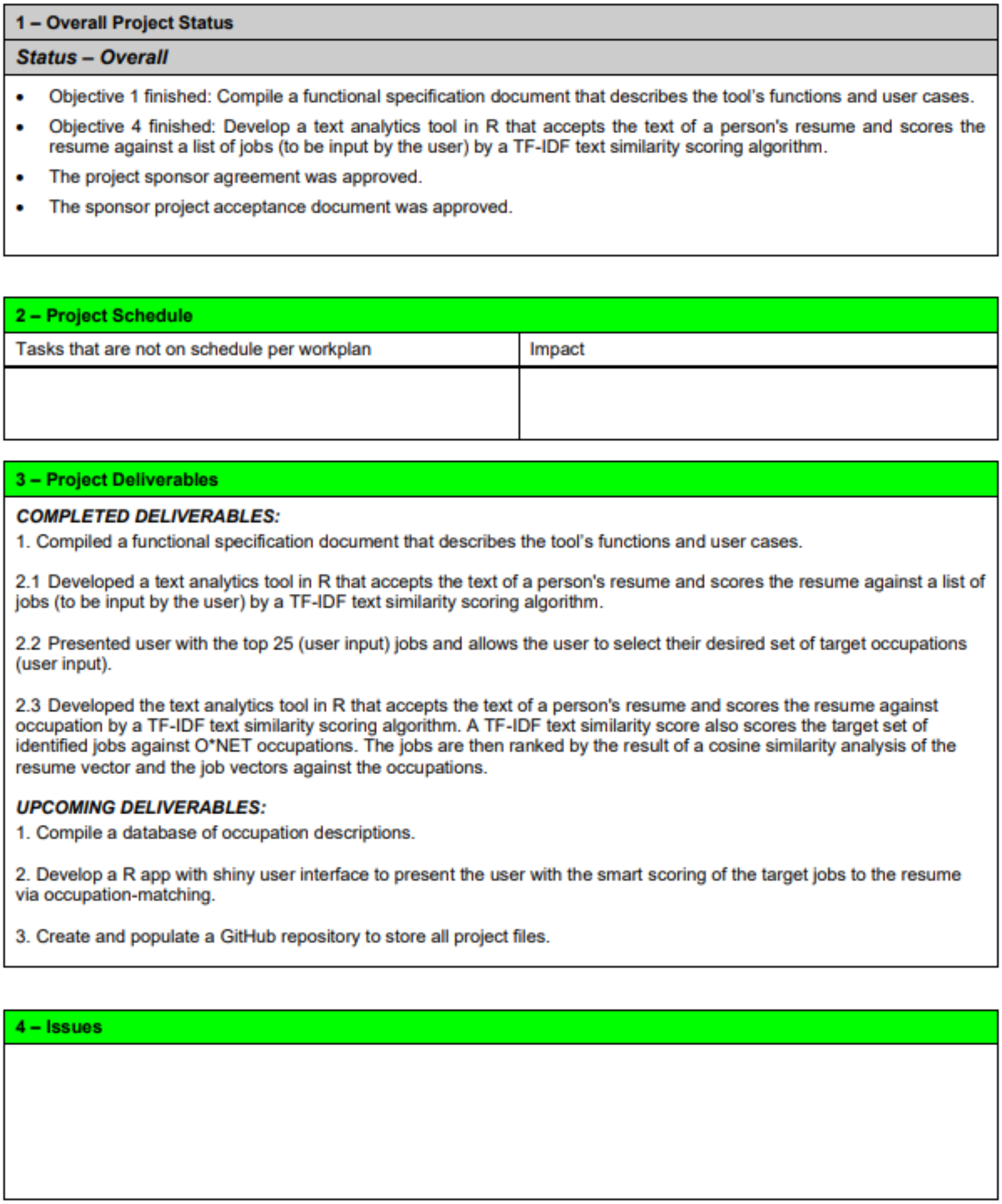
Date: March 30, 2021

YOUR ANTICIPATED COMPLETION DATE: **April 28, 2021**

COMPLETION SEMESTER: **Spring, 2021**

Project Status Areas:	**Execution Week <9>**		
	Green 🟩	**Yellow** 🟨	**Red** 🟥
1. Overall Project Status	🟩		
2. Project Schedule	🟩		
3. Project Deliverables	🟩		
4. Issues	🟩		
5. Project Risks	🟩		
6. Resources & Collaboration	🟩		
7. Change Status	🟩		

**see Assessment Guidelines on the last page of this doc.



1 – Overall Project Status

Status – Overall

- Objective 1 finished: Compile a functional specification document that describes the tool's functions and user cases.
- Objective 4 finished: Develop a text analytics tool in R that accepts the text of a person's resume and scores the resume against a list of jobs (to be input by the user) by a TF-IDF text similarity scoring algorithm.
- The project sponsor agreement was approved.
- The sponsor project acceptance document was approved.

2 – Project Schedule

Tasks that are not on schedule per workplan	Impact

3 – Project Deliverables

COMPLETED DELIVERABLES:

1. Compiled a functional specification document that describes the tool's functions and user cases.

2.1 Developed a text analytics tool in R that accepts the text of a person's resume and scores the resume against a list of jobs (to be input by the user) by a TF-IDF text similarity scoring algorithm.

2.2 Presented user with the top 25 (user input) jobs and allows the user to select their desired set of target occupations (user input).

2.3 Developed the text analytics tool in R that accepts the text of a person's resume and scores the resume against occupation by a TF-IDF text similarity scoring algorithm. A TF-IDF text similarity score also scores the target set of identified jobs against O*NET occupations. The jobs are then ranked by the result of a cosine similarity analysis of the resume vector and the job vectors against the occupations.

UPCOMING DELIVERABLES:

1. Compile a database of occupation descriptions.

2. Develop a R app with shiny user interface to present the user with the smart scoring of the target jobs to the resume via occupation-matching.

3. Create and populate a GitHub repository to store all project files.

4 – Issues

5 – Project Risks

Potential Risks	**Possible Mitigation**

6– Resources and Collaboration

- R Studio
- Zoom
- Personal Computer

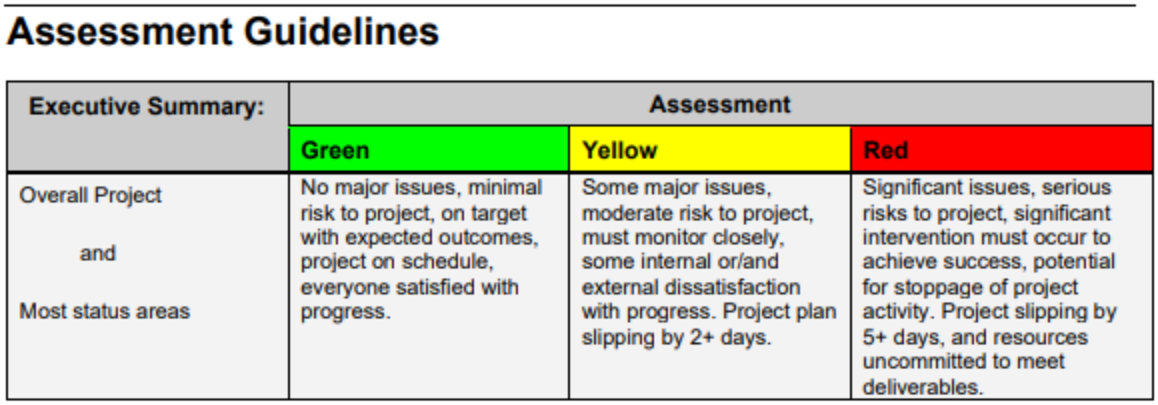
7 – Change Status

Scope Changes	**Status** (Requested | Approved | Completed)

Comments/Actions

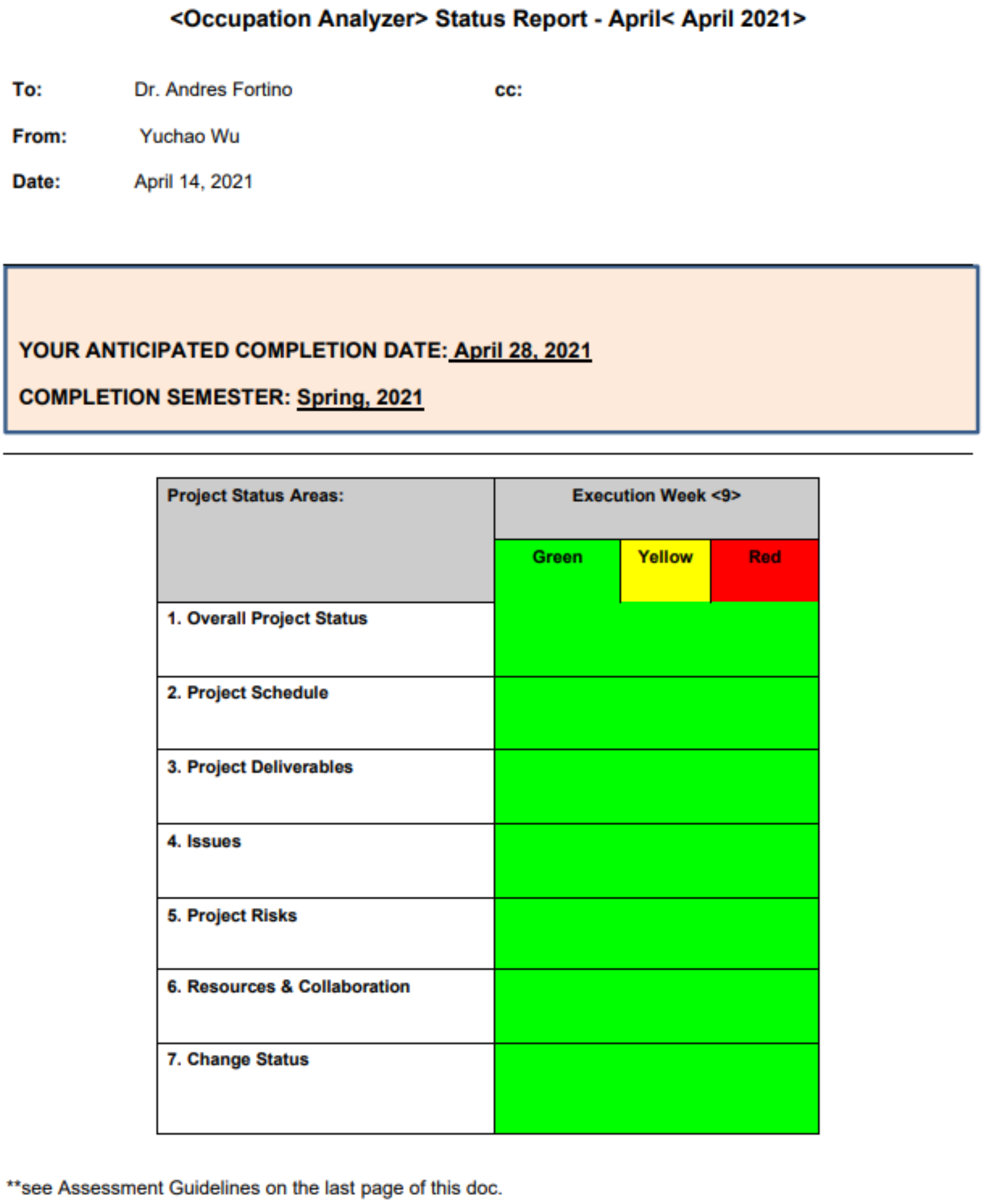
8 – Sponsor Signoff

Sponsor indicates agreement with the above status report.	
AGF	3/31/21



Assessment Guidelines

Executive Summary:	Assessment		
	Green	Yellow	Red
Overall Project and Most status areas	No major issues, minimal risk to project, on target with expected outcomes, project on schedule, everyone satisfied with progress.	Some major issues, moderate risk to project, must monitor closely, some internal or/and external dissatisfaction with progress. Project plan slipping by 2+ days.	Significant issues, serious risks to project, significant intervention must occur to achieve success, potential for stoppage of project activity. Project slipping by 5+ days, and resources uncommitted to meet deliverables.



<Occupation Analyzer> Status Report - April< April 2021>

To: Dr. Andres Fortino **cc:**

From: Yuchao Wu

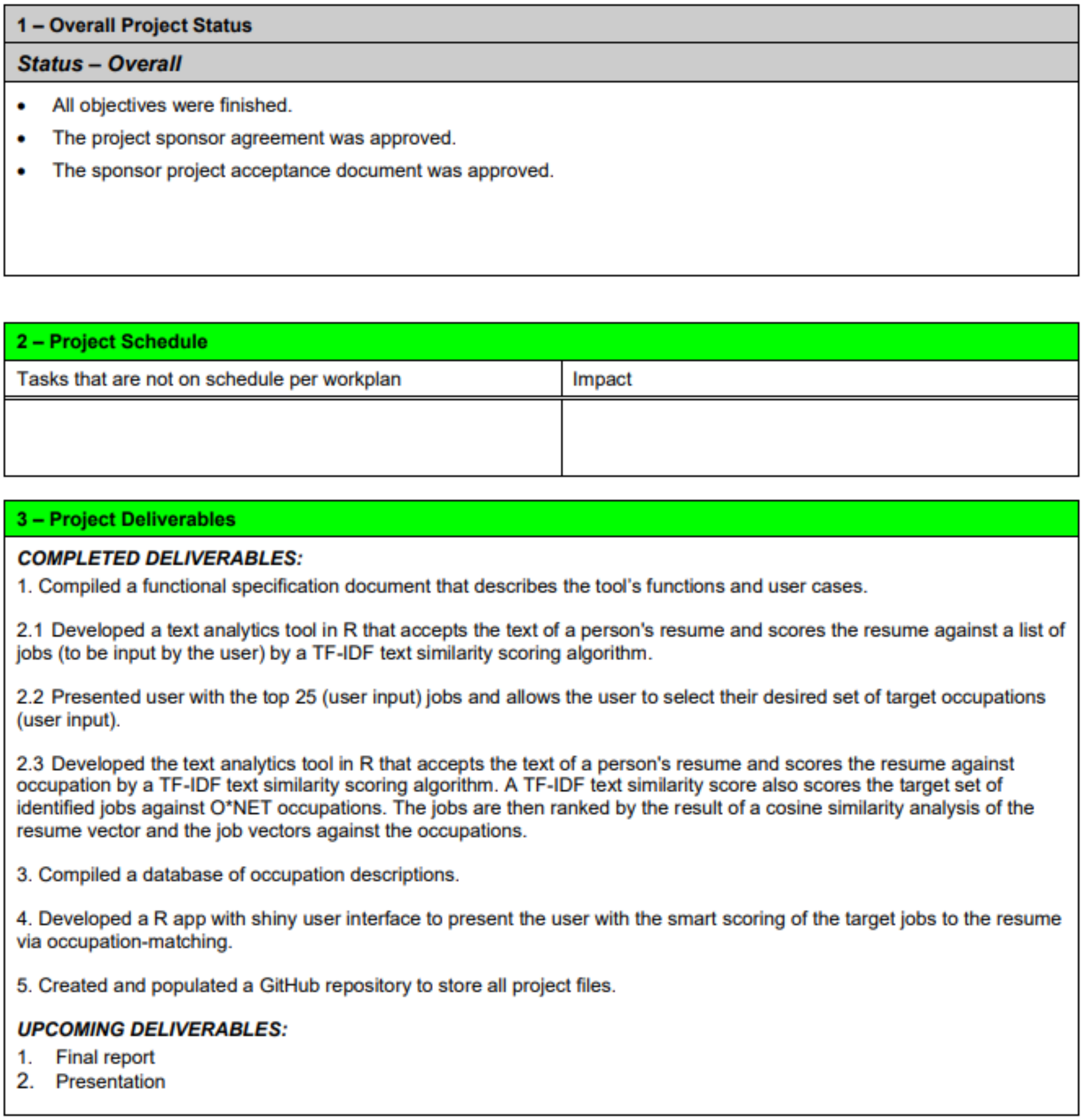
Date: April 14, 2021

YOUR ANTICIPATED COMPLETION DATE: **April 28, 2021**

COMPLETION SEMESTER: **Spring, 2021**

Project Status Areas:	**Execution Week <9>**		
	Green	**Yellow**	**Red**
1. Overall Project Status	🟩		
2. Project Schedule	🟩		
3. Project Deliverables	🟩		
4. Issues	🟩		
5. Project Risks	🟩		
6. Resources & Collaboration	🟩		
7. Change Status	🟩		

**see Assessment Guidelines on the last page of this doc.



1 – Overall Project Status

Status – Overall

- All objectives were finished.
- The project sponsor agreement was approved.
- The sponsor project acceptance document was approved.

2 – Project Schedule

Tasks that are not on schedule per workplan	Impact

3 – Project Deliverables

COMPLETED DELIVERABLES:

1. Compiled a functional specification document that describes the tool's functions and user cases.

2.1 Developed a text analytics tool in R that accepts the text of a person's resume and scores the resume against a list of jobs (to be input by the user) by a TF-IDF text similarity scoring algorithm.

2.2 Presented user with the top 25 (user input) jobs and allows the user to select their desired set of target occupations (user input).

2.3 Developed the text analytics tool in R that accepts the text of a person's resume and scores the resume against occupation by a TF-IDF text similarity scoring algorithm. A TF-IDF text similarity score also scores the target set of identified jobs against O*NET occupations. The jobs are then ranked by the result of a cosine similarity analysis of the resume vector and the job vectors against the occupations.

3. Compiled a database of occupation descriptions.

4. Developed a R app with shiny user interface to present the user with the smart scoring of the target jobs to the resume via occupation-matching.

5. Created and populated a GitHub repository to store all project files.

UPCOMING DELIVERABLES:

1. Final report
2. Presentation

4 – Issues

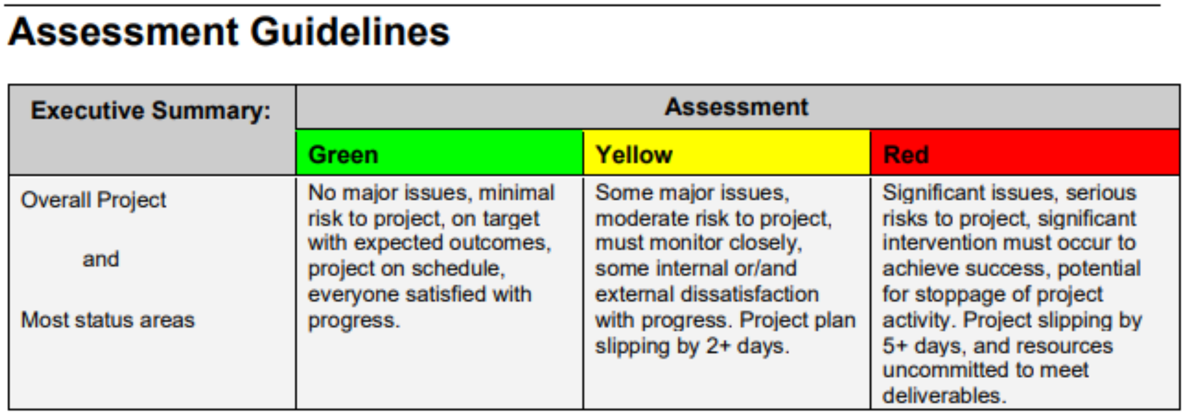
5 – Project Risks	
Potential Risks	Possible Mitigation

6– Resources and Collaboration
• R Studio • Zoom • Personal Computer

7 – Change Status	
Scope Changes	Status (Requested | Approved | Completed)

Comments/Actions

8 – Sponsor Signoff	
Sponsor indicates agreement with the above status report.	
Andres Fortino	4/14/21

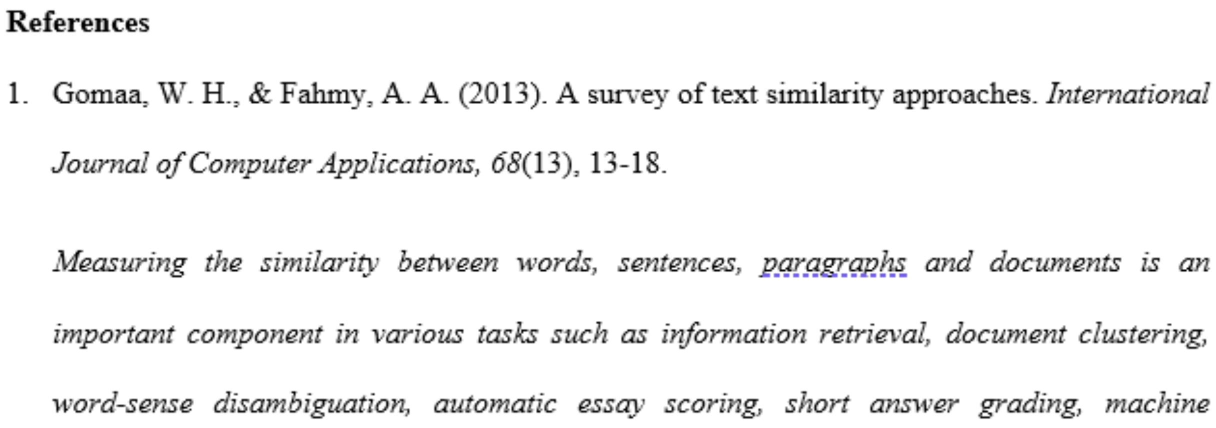


Assessment Guidelines

Executive Summary:	Assessment		
	🟩 Green	🟨 Yellow	🟥 Red
Overall Project and Most status areas	No major issues, minimal risk to project, on target with expected outcomes, project on schedule, everyone satisfied with progress.	Some major issues, moderate risk to project, must monitor closely, some internal or/and external dissatisfaction with progress. Project plan slipping by 2+ days.	Significant issues, serious risks to project, significant intervention must occur to achieve success, potential for stoppage of project activity. Project slipping by 5+ days, and resources uncommitted to meet deliverables.

Appendix I

Annotated Bibliography



References

1. Gomaa, W. H., & Fahmy, A. A. (2013). A survey of text similarity approaches. *International Journal of Computer Applications, 68*(13), 13-18.

   Measuring the similarity between words, sentences, paragraphs and documents is an important component in various tasks such as information retrieval, document clustering, word-sense disambiguation, automatic essay scoring, short answer grading, machine *translation and text summarization. This survey discusses the existing works on text similarity through partitioning them into three approaches; String-based, Corpus-based and Knowledgebased similarities. Furthermore, samples of combination between these similarities are presented.*

This paper introduces existing text similarity approaches. Among them, cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. This research can help me have a better understanding of cosine similarity.

2. Gong, K. (2019, October 22). *Big Data Cosine Similarity Score*. Retrieved from RPubs: https://www.rpubs.com/kgong56/542550

   This website provides a detailed explanation about how to calculate cosine similarity between two documents using R.

   This website provides R codes about how to calculate cosine similarity between two documents. These codes can guide me in coding the cosine similarity part.

3. Huang, A. (2008, April). Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand* (Vol. 4, pp. 9-56).

   Clustering is a useful technique that organizes a large quantity of unordered text documents into a small number of meaningful and coherent clusters, thereby providing a basis for intuitive and informative navigation and browsing mechanisms. Partitional clustering algorithms have been recognized to be more suitable as opposed to the hierarchical clustering schemes for processing large datasets. A wide variety of distance functions and *similarity measures have been used for clustering, such as squared Euclidean distance, cosine similarity, and relative entropy.*

In this paper, we compare and analyze the effectiveness of these measures in partitional clustering for text document datasets. Our experiments utilize the standard Kmeans algorithm and we report results on seven text document datasets and five distance/similarity measures that have been most commonly used in text clustering.

This article compares and analyzes the effectiveness of Euclidean Distance, Cosine Similarity, Jaccard Coefficient, Pearson Correlation Coefficient, and Averaged Kullback-Leibler Divergence in partitional clustering for text document datasets. The result shows that Pearson Correlation Coefficient and Averaged Kullback-Leibler Divergence outperform other measures. This helps me to decide which method I should use to compare vectors' similarities.

Lahitani, A. R., Permanasari, A. E., & Setiawan, N. A. (2016, April). Cosine similarity to determine similarity measure: Study case in online essay assessment. In *2016 4th International Conference on Cyber and IT Service Management* (pp. 1-6). IEEE.

Abstract- Development of technology in educational field brings the easier ways through the variety of facilitation for learning process, sharing files, giving assignment and assessment. Automated Essay Scoring (AES) is one of the development systems for determining a score automatically from text document source to facilitate the correction and scoring by utilizing applications that run on the computer. AES process is used to help the lecturers to score efficiently and effectively. Besides it can reduce the subjectivity scoring problem. However, implementation of AES depends on many factors and cases, such as language and *mechanism of scoring process especially for essay scoring. A number of methods implemented for weighting the terms from document and reaching the solutions for handling comparative level between documents answer and expert's document still defined. In this research, we implemented the weighting of Term Frequency – Inverse Document Frequency (TF-IDF) method and Cosine Similarity with the measuring degree concept of similarity terms in a document. Tests carried out on a number of Indonesian text-based documents that have gone through the stage of preprocessing for data extraction purposes. This process results is in a ranking of the document weight that have closesness match level with expert's document.*

This paper presents the implementation of the TF-IDF method and cosine similarity approach to measure the similarity level from the Indonesian essay assessment. This research can guide me in conducting TF-IDF cosine similarity analysis.

5. Leskovec, J., Rajaraman, A., & Ullman, J. D. (2020). *Mining of massive data sets*. Cambridge university press.

At the highest level of description, this book is about data mining. However, it focuses on data mining of very large amounts of data, that is, data so large it does not fit in main memory. Because of the emphasis on size, many of our examples are about the Web or data derived from the Web. Further, the book takes an algorithmic point of view: data mining is about applying algorithms to data, rather than using data to "train" a machine-learning engine of some sort.

This book introduces the theory of dimensionality reduction approaches. Among those approaches, the singular-value decomposition approach is often used in natural

language processing projects to keep important text features and decrease processing time. Leskovec provides a detailed explanation about how to use it. This book can guide me in decreasing program processing time.

6. McKinnon, C., Baazeem, I., & Angus, D. (2015, December). How few is too few? Determining the minimum acceptable number of LSA dimensions to visualise text cohesion with Lex. In *Proceedings of the Australasian Language Technology Association Workshop 2015* (pp. 75-83).

   Building comprehensive language models using latent semantic analysis (LSA) requires substantial processing power. At the ideal parameters suggested in the literature (for an overview, see Bradford, 2008) it can take up to several hours, or even days, to complete. For linguistic researchers, this extensive processing time is inconvenient but tolerated— but when LSA is deployed in commercial software targeted at non-specialists, these processing times become untenable. One way to reduce processing time is to reduce the number of dimensions used to build the model. While the existing research has found that the model's reliability starts to degrade as dimensions are reduced, the point at which reliability becomes unacceptably poor varies greatly depending on the application. Therefore, in this paper, we set out to determine the lowest number of LSA dimensions that can still produce an acceptably reliable language model for our particular application: Lex, a visual cohesion analysis tool. We found that, across all three texts that we analysed, the cohesion-relevant visual motifs created by Lex start to become apparent and consistent at 50 retained dimensions.

This article indicates that keeping 300 features is appropriate for natural language processing projects. It tells me that I should keep 300 features if I use singular value decomposition to deduct dimensions in my research.

7. Qaiser, S., & Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications, 181*(1), 25-29.

   In this paper, the use of TF-IDF stands for (term frequency-inverse document frequency) is discussed in examining the relevance of key-words to documents in corpus. The study is focused on how the algorithm can be applied on number of documents. First, the working principle and steps which should be followed for implementation of TF-IDF are elaborated. Secondly, in order to verify the findings from executing the algorithm, results are presented, then strengths and weaknesses of TD-IDF algorithm are compared. This paper also talked about how such weaknesses can be tackled. Finally, the work is summarized and the future research directions are discussed.

   This paper illustrates the background of term frequency-inverse document frequency, its algorithms, necessary data preprocessing procedure, and its limitations. In the data preprocessing part, all stop words should be removed. In the limitation part, Qaiser mentions that the stemming process can be used to improve TF-IDF. This research will be used to explain TF-IDF algorithm and improve data preprocessing.

8. Silge, J., & Robinson, D. (2020, November 10*). Analyzing word and document frequency: TF-IDF*. Retrieved from Text Mining with R: https://www.tidytextmining.com/tfidf.html#tfidf

A central question in text mining and natural language processing is how to quantify what a document is about. Can we do this by looking at the words that make up the document? One measure of how important a word may be is its term frequency (tf), how frequently a word occurs in a document, as we examined in Chapter 1. There are words in a document, however, that occur many times but may not be important; in English, these are probably words like "the", "is", "of", and so forth. We might take the approach of adding words like these to a list of stop words and removing them before analysis, but it is possible that some of these words might be more important in some documents than others. A list of stop words is not a very sophisticated approach to adjusting term frequency for commonly used words.

Another approach is to look at a term's inverse document frequency (idf), which decreases the weight for commonly used words and increases the weight for words that are not used very much in a collection of documents. This can be combined with term frequency to calculate a term's TF-IDF (the two quantities multiplied together), the frequency of a term adjusted for how rarely it is used.

This website provides R codes about how to analyzing words and TF-IDF. These codes can guide me in coding the TF-IDF part.

9. Silge, J., & Robinson, D. (2020, November 10). *Relationships between words: n-grams and correlations*. Retrieved from Relationships between words: n-grams and correlations: https://www.tidytextmining.com/ngrams.html

So far we've considered words as individual units, and considered their relationships to sentiments or to documents. However, many interesting text analyses are based on the *relationships between words, whether examining which words tend to follow others immediately, or that tend to co-occur within the same documents.*

In this chapter, we'll explore some of the methods tidytext offers for calculating and visualizing relationships between words in your text dataset. This includes the token = "ngrams" argument, which tokenizes by pairs of adjacent words rather than by individual ones. We'll also introduce two new packages: ggraph, which extends ggplot2 to construct network plots, and widyr, which calculates pairwise correlations and distances within a tidy data frame. Together these expand our toolbox for exploring text within the tidy data framework.

This website provides R codes about how to extract and analyze n-grams. These codes can guide me in coding the n-grams analysis part.

10. Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, 5(1), 7-16.

Data mining is used for finding the useful information from the large amount of data. Data mining techniques are used to implement and solve different types of research problems. The research related areas in data mining are text mining, web mining, image mining, sequential pattern mining, spatial mining, medical mining, multimedia mining, structure mining and graph mining. This paper discussed about the text mining and its preprocessing techniques. Text mining is the process of mining the useful information from the text documents. It is also called knowledge discovery in text (KDT) or knowledge of intelligent text analysis. Text mining is a technique which extracts information from both structured and unstructured data *and also finding patterns. Text mining techniques are used in various types of research domains like natural language processing, information retrieval, text classification and text clustering.*

This paper introduces the preprocessing techniques for text-mining projects. The techniques include removing stop words, stemming, and named-entity extraction. This research can be a reference for my data preprocessing part.