Industry Code Analyzer – NAICS Code Discovery For Startups in R

Applied Project Final Report

By

Yin (Fien) Xu

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

Table of Contents

Table of Tables

Table of Figures

Declaration

I, Yin Xu, declare that this project report submitted by me to the 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 by 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 the 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 for his contribution as a sponsor of this project and as a mentor during this project. I also want to thank all the instructors in the Management and Systems program I have taken courses with and learned a great deal.

Abstract

NAICS Code is a classification adopted by the North American Industry Classification System. Federal Statistical Agencies use the code to establish a North American standard on collecting and analyzing statistical data related to the U.S. Economy. However, NAICS is a self-assigned system. Business owners or users have to select the code that best describes their primary business activities. It is time-consuming and inefficient for business owners or users to manually use keyword search provided by the NAICS Code system, bounced back and forth among search pages and homepage. Therefore, this project aims to help startups or users of the NAICS Code system find correct and corresponding industry codes based on their business. The consultant has earlier work in building a NAICS industry code search tool in Python. The project expanded on existing preliminary work done in Python. The consultant of this project developed a tool in R to search the NAICS industry code database more intelligently. Given a business description as a text file, the industry code analyzer tool searches the NAICS industry code database to identify the industry classification corresponding to the users' uploaded business descriptions. The industry code analyzer tool uses TF-IDF text similarity scoring, returns a ranked list of industry codes, and presents the top 5 codes and descriptions to the user for selection and download. This project carried on additional experiments to define the tool capabilities and produced a user interface in Shiny for easier use of the tool. The shiny app regarding industry code analyzer served as an efficient search tool to find the correct industry code and benefit both professional and academic uses. The accuracy rate of this industry code tool reaches 79%, compared to the result generated from preliminary work in Python and code manually found by business owners. Further development regarding TF-IDF decomposition dimensionality reduction is suggested to adopt in the next phase to enhance accuracy and reducing process time in text analysis.

Keywords: NAICS Code System, Text Analysis, TF-IDF, Shiny

Abbreviations and Definitions

- *Cosine Similarity:* Cosine similarity is a commonly implemented metric in information retrieval. This metric transformed a text as a vector of terms, and cosine value between two texts' term vectors refers to the similarity between two texts.
- *NAICS Code:* NAICS code is a classification adopted in 1997 by the North American Industry Classification System to establish a North American standard on collecting and analyzing statistical data related to the U.S. industry and economy.
- *R:* R is a programming language and free software environment. R can be run on open-source tools or a set of integrated development environments such as RStudio, IntelliJ IDEA, Rattle, and so forth.
- *Text Analysis:* Text analysis is the process of parsing and translating large volumes of unstructured texts to extract machine-readable facts, insights, and patterns.
- *TF-IDF:* TF-IDF refers to the term frequency-inverse document frequency. TF-IDF is one of the widely used term weighting schemes in retrieving information. TF-IDF reflects how important a word is to a document in a corpus.
- *Shiny:* Shiny is a package of R that enables the developer to quickly build interactive web apps straight from R. Developer can host standalone apps on a webpage or embed codes in R Markdown documents or build dashboards. Moreover, a developer can extend the Shiny interface with CSS, HTML widgets, and JavaScript.

Introduction

Background information

NACIS Code, a classification developed for use by Federal Statistical Agencies, is widely adopted by various government agencies, trade associations, regulatory boards, and companies for data collection and analysis. NAICS Code utilizes a hierarchical structure, which is the relationship of one item to its particular category. The NAICS Code comprises a 2-digit sector code, 3-digit subsector code, 4-digit industry group code, 5-digit NAICS industry code, and 6-digit national industry code. The NAICS code is designed based on a production-oriented concept, meaning that NAICS Code classifies industries based on similarities in the processes used to produce goods and services. The goal of the NAICS Code classification is to have a unified industry definition for North American countries, including Canada, Mexico, and the United States. Statistical agencies refer to the NAICS Code as input to manage and analyze industry performance, productivity, unit labor costs, and employment. NAICS Code is initially designed for statistical purposes. Moreover, NAICS Code is widely adopted and applied in various administrative, regulatory, contracting, taxation, and other non-statistical purposes by government agencies, trade associations, regulatory boards, and companies. According to companies in similar or identical industries, business owners or administrators rely on the NAICS Code to classify their customers based on belonged industries and strategize their targeted marketing effort. In this way, business owners or administrators who refer to NAICS information can obtain a more profound understanding of their targeted customers, competitors, and industries by providing satisfying goods and services.

Company Name

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

New York University (NYU) is a private research university based in New York City. 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.

NYU School of Professional Studies is located at 7 East 12Th Street, NY, NY

Sponsor Information

- Name and Title: Dr. Andres Fortino: Clinical Associate Professor and MASY ACP Leader, New York University.
- Role within the organization: Dr. Andres Fortino is a clinical associate professor at NYU School of Professional Studies, where he teaches Business Analytics, Data Mining, Data Visualization, and Innovation. Dr. Andres Fortino is responsible for Applied Project – MASY – GC4100, in which students will undertake and deliver a real-world project for active practitioners in the field.
- Role on the project: Dr. Andres Fortino is the project sponsor. He will help me understand the needs of NYU – MASY and the status of preliminary work done in Python. Besides, Dr. Andres Fortino will help me as much as with the resources I seek.
- Dr. Andres Fortino (agf249@nyu.edu) can be reached over virtual conference calls as per project requirements.

Problem Description/Opportunity

Importance of the project

NAICS is a self-assigned system. Business owners or users have to select the code that best describes their primary business activities by themselves. Typically, Business owners or users have to cross-validate from a list of primary business activates containing keywords and the corresponding NAICS Codes and select the one that most closely corresponds to their primary business activities. Business owners or users sometimes have to refine their search multiple times to obtain more accurate codes. It is time-consuming and inefficient for business owners or users to manually use keyword search provided by the NAICS Code system, bounced back and forth among search pages and homepage.

To solve this problem, NYU MASY is suggested to develop a tool to find the industry code for startups. The project expands on existing preliminary work that was finished in Python. However, users with Windows without installing Python could not run the script directly. It is challenging to build an interactive website in a short time to visualize real-time analysis using Python. Therefore, the consultant of this project developed a tool in R to search the NAICS industry code database more intelligently. Given a business description as a text file, the tool searches the NAICS industry code database to identify the industry classification corresponding to the business description. It carries on additional experiments to define the tool capabilities and produce a user interface for easier use of the tool. The industry code analyzer tool uses TF-IDF text similarity scoring, returns a ranked list of industry codes, and presents the top 5 codes and descriptions to the user for selection and download. The industry code analyzer tool helps the

entrepreneur identify the industry for the startup more efficiently and perform a market analysis more quickly.

The shiny app regarding industry code analyzer also serves as an efficient search tool to find the correct industry code and benefit both professional and academic uses. The higher education industry has been growing and changing, driven by increasing student enrollment and increasing internationalization in the education sector. To better attract students and maintain competitive ranking, higher education institution strives for differentiating from various aspects, including student mix and outcomes, faculty resources, research capabilities, facilities, and community impact (André Dua & Jonathan, 2020). Higher education attaches more importance to expand community impact than before to real-world problems solving and practical training. Moreover, higher education institutions continue to make full use of their advantages in applying data and technologies to contribute to society's growth (Spear, 2020). Therefore, to strengthen community impact and improve researcher experiences, NYU MASY developed an industry code search tool in Python to find the industry code for startups. Indisputably, a computer-based tool with a user-friendly interface in R Shiny could dramatically optimize users' experiences and potentially increases NYU-MASY's positive influences on the community.

Alternate Solutions Evaluated

There are several alternative solutions to help startups to identify the correct NAICS code. Developing the industry code analyzer tool in R is adopted because of the algorithm's efficiency and accuracy after careful evaluation.

The first alternative method is to continue using the industry code analyzer tool developed in Python. This solution gives a similar and relevant result with a high accuracy rate compared to the code found by the business owner. The advantages and disadvantages of this solution are shown below.

Advantages:

- Free and Open-Source: Python comes under the open-source license, making it free to use and distribute. Developer can download, modify, and distribute the source code under the local version of Python.
- Extensive Libraries Support: Python community consists of over 200,000 packages, supporting the developer to build almost any function customized for the task. Moreover, developers can combine various libraries with building the functions and algorithms in one project, providing developers more flexibility to change packages based on project requirements.
- Interpreted Language: Python is an interpreted language that executes the code line by line. The developer can stop further execution and debug immediately when an error occurs.

Disadvantages:

- Weak in Browsers: Python is rarely used on the client-side. Python web development might consume high memory, which slows down execution. Because of the limited memory and execution speed, Python could not integrate with web browsers to display the real-time analysis on the client side.
- Possible Runtime Error: Since Python is a dynamically typed language, the data type of a variable can change anytime. A variable containing an integer or number may change to a string during execution, which increases execution time and the risk of runtime error.

The second alternative solution is to utilize Bag-of-Words (BoW) to conduct text analysis in R. BoW creates a vocabulary to extract features from the text. A bag-of-word is arguably one of the widely used approaches to represent a document with a high dimension sparse vector. Each document is encoded as a dimension with a feature vector, where the number of dimensions refers to the vocabulary size. Each dimension in the feature vector consists of the occurrence of a word in a specific document. There are several advantages and disadvantages of this solution.

Advantage:

- Simple to Implement: The boW model is a common way of representing text data as inputs to a machine learning model. Vast packages are available for developers to choose from when building the BoW model.

Disadvantage:

- High Dimension Feature Vector: The boW model typically leads to high dimensional features vector due to the large size predefined vocabulary. The distance between data points with nearest and farthest neighbors can become equidistant in high-dimensional data, potentially increasing the risk of inaccurate analysis.

- Assumed Independence: BoW does not leverage co-occurrence statistics between words, meaning that the BoW model assumes all words are independent of each other.

Solution Evaluation Criteria

The industry code analyzer tool is offers recommended list of NAICS code based on users' uploaded business descriptions. The tool is expected to be accessible and adopted by business owners and administrators for searching NAICS code. In addition to the application in the business field, this tool is supposed to be used by students and professors on academic projects. Therefore, the following are the criteria to measure the solution.

- Sustainability: NYU MASY desires to instantly contribute to the community by supporting startups or entrepreneurs in their business. Moreover, NYU MASY offers students resourceful support and help in initiating and operate the business. The entrepreneurship capstone is one of the popular options for graduate students. The core courses, such as strategic marketing, also covers lecture on how to conduct industry analysis. Based on the broad application of the industry code analyzer tool in NYU MASY, the tool is supposed to be sustainable for all students and professors in NYU MASY to use for at least two years or even longer as soon as the tool is available to the public.
- Cost: The job to develop an industry analyzer tool for NYU MASY is unpaid. So, there are no benefits and costs that can be quantified. The primary cost of the solution should be human resource cost spending on maintaining the tool. Therefore, the total solution cost is supposed to under the budget of NYU MASY.
- Easy to implement: The solution is supposed to easy to implement by developers or users. For developers, the solution is expected to easy to maintain and improve. For users, the

tool is supposed to easy to access and use the functions without prior experiences or hardware setting. Users can quickly generate the result based on the brief guideline.

- Efficiency: Most importantly, the solution is expected to give accurate results based on users' needs. Moreover, the solution should quickly offer analysis to users.
- Optimized Users' Experience: The solution is expected to operate in a user-friendly interface with simple functions and readable output.

Selection Rationale

This project aims to develop a search tool in R in addition to the preliminary work to search the NAICS code more intelligently and efficiently. The five criteria listed above account for various proportions according to the importance of each criterion affect the project's overall performance. Assumed the total percentage is 100%, and each solution is expected to be calculated the weighted average based on each criterion. The original score of each criterion is 5. The solution with the maximum total score is suggested to choose. The sustainability accounts for 20% since the solution must be stable to be operated by startups, all students, and professors in NYU MASY. The cost accounts for 15% since the budget of NYU MASY is sufficient. The criterion of easy to implement accounts for 25% since it is crucial for either users or developers to implement the solutions. The efficiency accounts for 25% because the solution is expected to provide more accurate results than the manual finding. The criterion of optimized users' experience weights 15%.

	Sustainability (20%)	Cost (15%)	Easy to Implement (25%)	Efficiency (25%)	Optimized Users' Experience (15%)	Total Score
Solution1: Continue analyzer tool developed in Python	3*20%	5*15%	3*25%	3*25%	1*15%	3
Solution 2: Utilize Bag-of-Words (BoW) in R	3*20%	5*15%	4*25%	2*25%	2*15%	3.15
Solution 3: Utilize TF-IDF in R	4*20%	5*15%	5*25%	3*25%	4*15%	4.15

Table 1: Solution Selection Matrix

Approach and Methodology

This project follows the project life cycle, consisting of initiation, planning, execution, and closeout, to meet the project sponsor's requirements and deliverables.

Initiation:

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

Planning:

1. The project manager conducted a literature review on relevant papers and algorithms.
2. The project manager generated the functional specification document, work breakdown schedule, risk management plan, change management plan, project sponsor acceptance document, and sponsor agreement based on the proposal.
3. The client review and signed the project sponsor acceptance document and sponsor agreement.

Execution:

1. Followed the work breakdown schedule to complete each week's assigned tasks.
2. Meet with the client weekly on Zoom to report project progress and adjusted weekly tasks based on clients' feedback.
3. Wrote monthly status reports for the client to track the project.

4. The client reviewed and signed monthly status reports.

Closeout:

1. The project manager archived and uploaded all relevant files to the accessible GitHub repository.
2. The project manager generated a project completion acceptance document.
3. The project manager prepared the project's final report and delivered the presentation.
4. The client reviewed and signed the project completion acceptance document.

The project task outline is presented in the order from initiation, planning, execution, control, and closeout. The work breakdown structure displayed all the work and deliverables required to complete the project, and more details are shown in the project chronology.

Project Objectives and Metrics

Goal of the project

This project aims to develop a tool to find the industry code for startups. The consultant has earlier work in building a NAICS industry code search tool in Python. NYU MASY now wishes to develop a search tool in R to search the NAICS industry code database more intelligently. Therefore, this project aims to create a computer-based tool in R to identify NAICS industry codes for startups with a robust user interface.

Project Deliverables and Metrics

- ***Objective 1*** **-** Compile a functional specification document to describe a detailed step-by-step outline of each item's functionality and user flow for different user roles in a Word document by February 12.

  Metric: The document will be presented and accepted by the client through an online meeting by February 12.

- ***Objective 2*** **–** Perform and deliver a TF-IDF similarity scoring of the text description of a business against the text descriptions of the NAICS industry codes using R language. Test the algorithm's validity using SME-identified industry codes for a known set of startups. Also, test the algorithm's validity by processing the description of known established businesses from the SEC database of public companies using R language by April 22.

  Metric: The TF-IDF similarity scoring will return a ranked list of industry codes and present the top 5 codes and descriptions to the user for selection. The test will utilize the confusion matrix to show the results and improve accuracy and precision to at least 75%.

- ***Objective 3 -*** Develop a shiny app interface by April 29 in R.

  Metric: The shiny app is expected to accept the text file of the startup's business description and deliver a ranked list of the top 5 possible NAICS industry codes.

- **Objective 4 –** Create an accessible GitHub repository for recording all the project files and supporting documentation by May 5.

  Metric: All the documents are accepted by the client and can be viewed on the GitHub repository.

Risk Analysis

The risk management plan is essential for the whole project. Based on a thorough risk management plan, the project team can identify, evaluate, mitigate risks that might occur through the entire project life cycle. The project is expected to be finished on time, within budget, within the scope, and with high quality. The primary part of this project is to develop a TF-IDF algorithm and cosine similarity to analyze users' uploaded business descriptions against NAICS code with detailed business definitions. The project team identified the following risks:

Number	**Risk**	**Probability Score (1,2 or 3)**	**Impact Score (1,2 or 3)**
1	The client stops the project.	1	3
2	Functions do not work on Shiny App. io	2	3
3	Improper website design.	1	2
4	User (startups) does not want to use the tool.	2	2
5	My R skills are not sufficient to debug the tools.	3	2

Table 2: Possible Risks with Probability Score and Impact Score

To optimize the project manager's effort on controlling the risk, the project manager developed the risk matrix to visualize the exposure and probability of each risk. The project manager conducted contingency plans to the risk with high exposure and a high probability of occurrence.

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

Figure 1: Risk Matrix

The project manager conducted contingency plans to the risk with high exposure and a high probability of occurrence based on the risk matrix.

Risk	Description	Probability (1-3)	Exposure (1-3)	Contingency Plan
1	The client stops the project.	1	3	
2	Functions do not work on Shiny App. io	2	3	Run the shiny app using a personal account to ensure every function works successfully before uploading to a public account.
3	Improper website design	1	2	
4	User (startups) does not want to use the tool	2	2	

5	My R skills are not sufficient to debug the tools.	3	2	Communicate with other colleagues who have similar projects and find more learning materials online.

Table 3: Contingency Plan

Issues Encountered

While working on the project, the project team encountered some technical issues. The first issue the team faced was to develop the TF-IDF algorithm and cosine similarity. The first analysis result was different from the result generated by preliminary work done in Python. The reason was data preprocessing. The issue was solved by adopting new packages in R to remove stop words, remove special characters, and lemmatize the sentence. The second issue was that the project ran out of instance memory in Shiny.io. The project had to reload every few seconds. To solve the issue, the project manager reduced the variables in scripts to minimize the instance memory and increase operational efficiency. Moreover, the project upgraded the Shiny.io account plan to a premium one, which releases more instance memory to ensure the app operating smoothly.

Project Chronology and Critique

The following table shows the task, definition, duration, and due date.

I.D.	Element Name	Definition	Duration	Due Date
1	NAICS code analyzer tool	All work to implement a new NAICS code analyzer tool	45	April 22, 2021
2	Initiation	The work to initiate the project.	10	March 25, 2021
3	Evaluation & Recommendations	The project manager works with the client to evaluate solution sets and make recommendations.	10	January 18, 2021
4	Develop Project Proposal	The project manager develops the project proposal.	3	February 10, 2021
5	Draft Literature Review Research	The project manager finds 10+ pieces of literature related to the project and conducts a literature review.	5	February 24, 2021
6	Conduct Situation Analysis & Cost-Benefit Analysis	The project manager conducts the situation analysis and cost analysis of the project.	5	March 3, 2021

7	Create Work Break Down Schedule	The project manager creates the work breakdown schedule with detailed project procedures and a timeline.	3	March 3, 2021
8	Develop Project Charter	The project manager develops the Project Charter.	5	March 12, 2021
9	*Deliverable:* Submit Project Charter	Project Charter is delivered to the Project Sponsor.	5	March 17, 2021
10	Project Sponsor Reviews Project Charter	The project sponsor reviews the Project Charter.	5	March 20, 2021
11	Project Charter Signed/Approved	The project sponsor signs the Project Charter and authorizes the project manager to move to the Planning Process.	2	March 25, 2021

12	Planning	The work is for the planning process for the project.	10	March 25, 2021
13	Compile Functional Specification Document	The project manager creates a functional specification document	5	February 10, 2021
14	Develop Project Sponsor Agreement	The project manager develops a project sponsor agreement with a detailed description of deliverables.	3	March 10, 2021
15	Submit Project Sponsor Agreement	The project sponsor agreement is submitted to the project sponsor	3	March 10, 2021
16	*Milestone:* Project Sponsor Acceptance	Under the project sponsor's confirmation and approval, the project manager begins to move on execution step.	3	March 10, 2021

17	Execution	Work involved executing the project.	10	April 16, 2021
18	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.	5	February 12, 2021
19	*Deliverables 1:* Functional Specification Document Approval	The project plan is approved, and the project manager has permission to execute the project according to the project plan.	5	February 12, 2021
20	Perform and Delivery TF-IDF Similarity Code with Samples Using R Language	The project implementer should finish the sample exercises in TF-IDF.	10	March 1, 2021

21	Conduct Unit Test of Sample Code for Each Section in TF-IDF Code	The project implementer should conduct unit tests of each section in the TF-IDF algorithm.	10	March 5, 2021
22	*Deliverable 2:* Perform and deliver a TF-IDF similarity scoring of a business's text description against the text descriptions of the NAICS industry codes using R language.	The TF-IDF similarity scoring will return a ranked list of industry codes and present the top 5 codes and descriptions to the user for selection.	2	March 19, 2021
23	Conduct Unit Test of Code of Business Description Against NACIS Code	The test will utilize the confusion matrix to show the results and improve accuracy and precision to at least 85%.	5	March 26, 2021

24	Develop a Shiny App Interface	The project implementer develops a shiny app using R.	5	April 8, 2021
25	*Deliverable 3:* The Shiny App with Function to Accept the Text File of Business Description and Deliver a Ranked List of the Top 5 Possible NACIS Industry Codes	The shiny app is expected to accept the text file of the startup's business description and deliver a ranked list of the top 5 possible NAICS industry codes.	3	April 14, 2021
26	Create an Accessible GitHub Repository for Recording All the Project Files and Supporting Documentation	All files and records are updated to GitHub Repository.	5	April 13, 2021

27	*Deliverable 4:* A GitHub Repository with All the Documents that Viewed and Accepted by the Clients	All the documents are accepted by the client and can be viewed on the GitHub repository.	3	April 21, 2021
28	Control	The work involved the control process of the project.	5	April 14, 2021
29	Develop Change Management Plan	The project manager conducts a change management plan with possible solutions regarding changes.	5	April 14, 2021
30	Develop Risk Management Plan	The project manager conducts a risk management plan with possible contingency plans.	5	April 14, 2021
31	Closeout	The work to close out the project.	5	May 5, 2021

32	Draft Final Project Report	The project manager drafts the final project report.	5	April 20, 2021
33	Final meeting with the client to go over the lessons learned	The project manager, along with the project sponsor, performs a lesson learned meeting.	2	April 23, 2021
34	Compile Project Sponsor Acceptance – Project Completion Signoff	The project sponsor signs the final project completion document.	2	April 28, 2021
35	Gain Formal Acceptance	The Project Sponsor formally accepts the project by signing the acceptance document included in the project plan.	2	April 28, 2021
36	Present the project	The project manager presents the project to the client using dynamic PowerPoint.	1	May 5, 2021

37	Submit Final Project Report with Final Deliverables in GitHub Repository	All project-related files and documents are formally archived and submitted.	1	May 5, 2021

Table 4: Project Chronology Table

Lessons Learned

The whole project was able to deliver planned with expected quality and in time, and this could not have been done without contribution and help from the sponsor. During the project, I learned a lot from the following perspectives:

- Project Management Skills and Experience: My project management skills were dramatically enhanced after this project. I accumulated valuable experience in managing a project from start to end.
- Programming Skills: I learned a new language and code the program in R. My programming skills enhanced remarkably, strengthening my competence to land a professional job in the future.
- App Development Experience: I obtained valuable experience in developing a shiny app using the R language. I learned how the shiny server interacts with the interface in the browser.
- Git Hub Knowledge: I understood how Git Hub enables projects to collaborate simultaneously in different servers. I opened my Git Hub account and managed my projects.
- Problem-solving and Communication Skills: I communicated with my client frequently to ensure the project is on the right track. Besides, I maintained a positive attitude to solve multiple tasks in the project with a data-driven mindset.

Conclusion and Summary

This project performs med TF-IDF algorithm and cosine similarity in R to analyze users' uploaded business descriptions against the NAICS codes definitions. The project achieved the goal and objectives to return the top 5 and top 10 industry codes for users' selection and allow users to download the complete list of recommended industry codes based on analysis. After visualizing the analysis result in Shiny, the similarity score refers to how similar the business descriptions against the NAICS code industry definitions. Compared to the result generated by owners, the result provided by the industry code analyzer tool in R has a 79% accuracy rate, indicating that the industry code analyzer tool in R can be an effective search tool to generate correct NAICS code for startups. Compared to the previous work done in Python, the industry code analyzer tool in R offered an interactive and user-friendly interface for users to view the real-time analysis result.

Shiny App Link: https://ranalyzer.shinyapps.io/Industrycode/

Git Hub Repository: https://github.com/yinxufientsui/R_NAICS-Code-Analyzer.git

Limitations, Recommendations, and Scope for Future Work

Even this project was able to deliver as excepted, there are still some limitations within this project, and some of the limitations may be improved in the future similar projects in NYU MASY. First of all, limited by the knowledge on coding in R, the text analysis of this project is supposed to conduct dimension reduction to compare the accuracy rate. Singular value decomposition is suggested to use in future projects to generate a more accurate result. In addition to cosine similarity, more advanced algorithms, such as K-NN, K-means, is suggested to use in the future project on accuracy improvement.

Secondly, limited time shortened the test period for the final project. Insufficient tests might cause the developer to neglect the problems, which might affect the results. Therefore, the project team suggests that future projects leave more time on testing and cross-validation to ensure the optimized result.

The following scope for future work might consider increasing more functions on the Shiny app to allow users to customize the download versions and refine search to generate more accurate results. Moreover, further development regarding TF-IDF decomposition dimensionality reduction is suggested to adopt in the next phase to enhance accuracy and reducing process time in text analysis.

Literature Survey

This literature review is organized by the background of the North American Industrial Classification System (NAICS) code, the challenge of startups to identify correct industry code, the introduction of Term Frequency - Inverse Document Frequency (TF-IDF) on cosine similarity, and application of R programming languages in the shiny app.

NAICS code, known as industry classification, is adopted by the North American Industry Classification system in 1997 to replace the Standard Industrial Classification (SIC) system code. O'Connor, L. (2000) introduced that the NAICS code is designed with a production-based framework with a hierarchical six-digit structure while SIC code is product-based. Kile and Phillips (2009) pointed out that the NAICS code system covers a broader range of emerging industries and technologies. The NAICS code system was initially designed for statistical analysis. Pierce and Schott (2012) indicated that United States statistical data was generated by various agencies in different formats and standards. Therefore, a standardized industry classification was used to ensure all the economic data is gathered based on a consistent benchmark. NAICS code is helpful for companies to classify their industry, business scope, and customers. More specifically, startups can use NAICS code to identify their belonged industries, determine their target market, and generate consistent data on other companies in similar or identical industries. However, the NAICS code is a self-assigned system, which means that users need to find the code based on the system's standardized description versus their business descriptions. Therefore, it is very challenging for startups to identify the correct NAICS industry code at the first time when access to the system.

Term Frequency - Inverse Document Frequency (TF-IDF) is a technique to compute each word's relative weight that occurred in a specific document compared to the inverse proportion of word over the entire corpus (Ramos, 2003). Bafna, Pramod, and Vaidya (2016) indicated that TF-IDF is one of the most commonly used approaches to turn strings into vectors and extract the most relevant terms from the corpus. Besides, Wu, Luk, Wong, and Kwok (2008) applied the binary independence model, logistic regression model, vector space model, and extended Boolean model as examples to prove that TF-IDF is one of the efficient models of relevant information retrieval.

Cosine similarity is a vector-based algorithm to measure the similarity between two vectors (Gunawan, Sembiring, & Budiman, 2018). Tata and Patel (2007) described that cosine similarity transforms each term and string into a vector with specific dimensions. The angle between two vectors signifies how similar those strings are. Bafna, Pramod, and Vaidya (2016) indicated that among similarity measures commonly used, such as Euclidean, Person correlation, Cosine, and Extended Jaccard, cosine similarity has the advantage of simplifying the comparison process and accelerating evaluation speed by calculating the angle of two non-zero vectors.

The shiny framework is an accompanying graphical user interface (GUI) that establishes an interactive web app directly from R. The shiny framework allows users to visualize the result in a real-time interface (Crisan, Munzner, & Gardy, 2019). Potter, Wong, Alcaraz, and Chi (2016) showed the advantages of shiny apps over the traditional presentation methods with dynamic reporting and easy access online via a web browser, which dramatically improve users' experiences.

Most of the selected research concluded that it is significant for startups to utilize the NAICS code to identify the correct industry and gather insights from consumers and competitors in similar or identical industries. Meanwhile, the authors pointed out that it is time-consuming and challenging for small businesses to find the correct industry code. TF-IDF is suggested to use in

information retrieval and text mining. Most authors admitted that cosine similarity's efficiency in scoring similarity between vectors and identify relevance between terms. However, a lack of research applied TF-IDF to explore the similarity between NAICS code and accurate business description in R-based application. Therefore, this project is expected to fill the gap between existing research and practice about using R-based application and real-time interface to generate the similarity comparison between NAICS code and startups business description and visualize the result using an interactive interface.

References

Bafna, P., Pramod, D., & Vaidya, A. (2016). Document clustering: TF-IDF approach. *International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), 16*(3), 61-66.

Crisan, A., Munzner, T., & Gardy, J. L. (2019). Adjutant: An R-based tool to support topic discovery for systematic and literature reviews. *Bioinformatics*, *35*(6), 1070-1072.

Gunawan, D., Sembiring, C. A., & Budiman, M. A. (2018). The implementation of cosine similarity to calculate text relevance between two documents. *Journal of Physics: Conference Series*, *978*(1), 121-127.

Hashemzadeh, B., & Abdolrazzagh-Nezhad, M. (2020). Improving keyword extraction in multilingual texts. *International Journal of Electrical & Computer Engineering, 10*(6), 5909-5916.

Kile, C. O., & Phillips, M. E. (2009). Using industry classification codes to sample high-technology firms: Analysis and recommendations. *Journal of Accounting, Auditing & Finance*, *24*(1), 35-58.

O'Connor, L. (2000). Approaching the challenges and costs of the North American industrial classification system (NAICS). *The Bottom Line: Managing Library Finances, 13*(2), 83-89.

Pierce, J. R., & Schott, P. K. (2012). A concordance between ten-digit U.S. Harmonized System Codes and SIC/NAICS product classes and industries. *Journal of Economic and Social Measurement*, *37*(1-2), 61-96.

Potter, G., Wong, J., Alcaraz, I., & Chi, P. (2016). Web application teaching tools for statistics using R and shiny. *Technology Innovations in Statistics Education*, *9*(1).

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. *Proceedings of the first instructional conference on machine learning, 242*(1), 29-48.

Tata, S., & Patel, J. M. (2007). Estimating the selectivity of tf-idf based cosine similarity predicates. *ACM Sigmod Record*, *36*(2), 7-12.

Wu, H. C., Luk, R. W. P., Wong, K. F., & Kwok, K. L. (2008). Interpreting tf-idf term weights as making relevant decisions. *ACM Transactions on Information Systems (TOIS)*, *26*(3), 1-37.

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: Industry Code Analyzer – NAICS Code Discovery for Startups in R
Student Name: Yin Xu (Fien)
Sponsoring Organization: New York University School of Professional Studies and the Management and Systems program (MASY)
Project Sponsor Name and Title: Dr. Andres Fortino, Clinical Associate Professor and MASY ACP Leader, New York University.
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: Consultant of this project will create a computer-based tool in R to identify NAICS industry codes for startups with a robust user interface.
Objective #1: Compile a functional specification document to describe a detailed step-by-step outline of each item's functionality and user flow for different user roles in a Word document by February 12.
Objective #2: Perform and deliver a TF-IDF similarity scoring of the text description of a business against the text descriptions of the NAICS industry codes using R language. Test the algorithm's validity using SME identified industry codes for a known set of startups. Also, test the algorithm's validity by processing the description of known established businesses from the SEC database of public companies using R language by April 22.
Objective #3: Develop a shiny app interface by April 29 in R.
Objective #4: Create an accessible GitHub repository for recording all the project files and supporting documentation by May 5.

I agree with the above planned project goal, project objectives, and related metrics.

Andres Fortino _____ 3/8/21 _____
Project Sponsor Signature **Date:**

RESULTS

PROJECT RESULTS

Planned Start Date: January 28, 2021 **Planned End Date:** May 5, 2021
Actual Start Date: January 28, 2021 **Actual End Date:** April 23, 2021

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

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** _____

Project Objective #1: Compile a functional specification document to describe a detailed step-by-step outline of each item's functionality and user flow for different user roles in a Word document by February 12.
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: Perform and deliver a TF-IDF similarity scoring of the text description of a business against the text descriptions of the NAICS industry codes using R language. Test the algorithm's validity using SME identified industry codes for a known set of startups. Also, test the algorithm's validity by processing the description of known established businesses from the SEC database of public companies using R language by April 22.
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: Develop a shiny app interface by April 29 in R.
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: Create an accessible GitHub repository for recording all the project files and supporting documentation by May 5.
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.

Sponsor's Overall Evaluation of student's performance: A <expand, as necessary>

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	4/28/21
Project Sponsor Signature	**Date:**
Yin Xu	
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	New York University School of Professional Studies and the Management and Systems Program (MASY)				
Address	7 East 12Th Street				
City	New York	State	New York	Zip	10003
Project Sponsor	First Name	Andres	Last Name	Fortino	
Title	Clinical Associate Professor and MASY ACP Leader				
Phone					
Email	agf249@nyu.edu				
Web Site	https://www.sps.nyu.edu/homepage/academics/masters-degrees/ms-in-management-and-systems.html				
Type of Business	Private Research University				

Student Name	Yin Xu (Fien)
Project Title	Industry Code Analyzer – NAICS Code Discovery for Startups in R

Description of Project	
The consultant of this project will develop a tool in R to search the NAICS industry code database more intelligently. Besides, the project will expand on existing preliminary work that was finished in Python. Given a business description as a text file, the tools will search the NAICS industry code database to identify the industry classification corresponding to the business. It will carry on additional experiments to define the tool capabilities and produce a user interface for easier use of the tool. The industry code analyzer tool will use TF-IDF text similarity scoring, return a ranked list of industry codes, and present the top 5 codes and descriptions to the user for selection.	
Estimated Hours of Student Participation	300

Anticipated Results

- **Deliverable 1** - Compile a functional specification document to describe a detailed step-by-step outline of each item's functionality and user flow for different user roles in a Word document by February 12.
  - *Metric for measurement*: The document will be presented and accepted by the client through an online meeting by February 12.
- **Deliverable 2** – Perform and deliver a TF-IDF similarity scoring of the text description of a business against the text descriptions of the NAICS industry codes using R language. Test the algorithm's validity using SME identified industry codes for a known set of startups. Also, test the algorithm's validity by

processing the description of known established businesses from the SEC database of public companies using R language by April 22.
- Metric for measurement: The TF-IDF similarity scoring will return a ranked list of industry codes and present the top 5 codes and descriptions to the user for selection. The test will utilize the confusion matrix to show the results and improve accuracy and precision to at least 85%.
- **Deliverable 3** - Develop a shiny app interface by April 29 in R.
- Metric for measurement: The shiny app is expected to accept the text file of the startup's business description and deliver a ranked list of the top 5 possible NAICS industry codes.
- **Deliverable 4** – Create an accessible GitHub repository for recording all the project files and supporting documentation by May 5.
- Metric for measurement: All the documents are accepted by the client and can be viewed on the GitHub repository.

Knowledge and expertise student will need to be able to complete the project
• Project Management: Student will apply the methods and theories to plan, execute, control, and launch the project. Besides, student will strictly follow the process of project management to finish the required documents, providing sufficient resources that support follow-up projects in the future. • Research Process & Methodology: Student will apply the skills and methods to conduct detailed research and deliver high-quality academic research papers as required. • Database Technology for Web Application: Student will apply coding skills to design a user-friendly app interface. • Coding in R Language: Student will learn the knowledge of coding in R, perform TF-IDF similarity score, and develop a shiny app interface. • GitHub Knowledge: Student will be familiar with GitHub and create a valid GitHub repository to record the final project files and supporting documents.

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
Describe the form and frequency of supervision of the student by the Project Sponsor. Student and Dr. Andres Fortino (Project Sponsor) will meet weekly via Zoom meeting.	

6. Sponsor Agreement

Students are interns, not professional consultants. NYU is <u>not</u> 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 NYU School of Professional Studies Date 3/8/21

By (signature): Andres Fortino

Printed Name: Dr. Andres Fortino

Title: _ Project Sponsor Clinical Associate Professor of Management and Systems

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) Yin Xu (Fien) Date 03/08/2021

Signature: Yin Xu

Appendix C

Project Charter

Industry Code Analyzer – NAICS Code Discovery for Startups in R Project Charter

Project Manager: Yin (Fien)Xu
Sponsor: Dr. Andres Fortino
Prepared by: Yin (Fien)Xu

Name and Location of Client Organization:
NYU School of Professional Studies and the Management and Systems program (MASY)
NYU School of Professional Studies is located at 7 East 12Th Street, NY, NY

Revision History

Revision date	Revised by	Approved by	Description of change

Project Goal
It is time-consuming for an entrepreneur or a new user to the NAICS system to find a correct and corresponding industry code for business. To solve this problem, NYU MASY needs to develop a tool to find the industry code for startups. The consultant has earlier work in building a NAICS industry code search tool in Python. NYU MASY now wishes to develop a search tool in R to search the NAICS industry code database more intelligently. Therefore, this project aims to create a computer-based tool in R to identify NAICS industry codes for startups with a robust user interface.

Problem/Opportunity Definition
Given a startup's description of the business as a text file, the industry code analyzer will search the NAICS industry code database to identify the startup's industry classification. The tool will help the entrepreneur identify the industry for the startup more efficiently and perform a market analysis more easily. The project will expand on existing preliminary work done in Python. The project will carry on additional experiments and test to increase the accuracy of analytical results. The shiny app regarding industry code analyzer will serve as an efficient search tool to find the right industry code and benefit both professional and academic uses.

Proposed Project Description
The consultant of this project will develop a tool in R to search the NAICS industry code database more intelligently. Besides, the project will expand on existing preliminary work that was finished in Python. Given a business description as a text file, the tools will search the NAICS industry code database to identify the industry classification corresponding to the business. It will carry on additional experiments to define the tool capabilities and produce a user interface for easier use of the tool. The industry code analyzer tool will use TF-IDF text similarity scoring, return a ranked list of industry codes, and present the top 5 codes and descriptions to the user for selection.

Project Sponsor

- Name and Title: Dr. Andres Fortino: Clinical Associate Professor and MASY ACP Leader, New York University.
- Role within the organization: Dr. Andres Fortino is a clinical associate professor at NYU School of Professional Studies, where he teaches Business Analytics, Data Mining, Data Visualization, and Innovation. Dr. Andres Fortino is responsible for Applied Project – MASY – GC4100, in which students will undertake and deliver a real-world project for active practitioners in the field.
- Role on the project: Dr. Andres Fortino is the project sponsor. He will help me to understand the needs of NYU – MASY and status of preliminary work done in Python. Besides, Dr. Andres Fortino will help me as much as with the resources I seek.

Objectives:

Technical Objectives:

- Perform and deliver a TF-IDF similarity scoring of a business's text description against the text descriptions of the NAICS industry codes using R language. Test the algorithm's validity using SME-identified industry codes for a known set of startups. Also, test the algorithm's validity by processing the description of known established businesses from the SEC database of public companies using R language by April 9.
- Develop a shiny app interface by April 14 in R.
- Create an accessible GitHub repository for recording all the project files and supporting documentation by April 20.

Timing objectives

- Complete the entire project before May 5, 2021.

Resource objectives:

- Compile a functional specification document to describe a detailed step-by-step outline of each item's functionality and user flow for different user roles in a Word document by February 12.
- Collaborate with the project sponsor to finish the project before May 5, 2021.

Budget objectives

- There is no cost for the entire project since the job to develop an industry analyzer tool for NYU MASY is unpaid. The intangible cost is hard to be quantified.

Budget objectives:

	Planned	**Actual**
Salaries	$0	$0
Domain name	$0	$0
Total	**$0**	$0

Scope objectives:

- Create a NAICS industry code analyzer tool that will perform TF-IDF similarity scoring of a text description of business against the NAICS 4-digit and 6-digit code and return a ranked list of top5 industry codes for users to select and download. The tool will be shown in a shiny app interface that allows users to submit the requested files on a browser web page and view the output. All the documents are accepted by the client and can be viewed on the GitHub repository by May 5, 2021.

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:

- Improve efficiency for startup owners to find the correct industry code
- Help students and professors to quickly identify the corresponding industry code for academic research
- Strengthen community impact by providing an efficient search tool for startups
- Improve users' experiences

Value & Probability: N/A

Assumptions Driving Value: N/A

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

Financial Return

Intangible Profit

Other Business Benefits

There is an intangible benefit of improving efficiency for startup owners to find the correct industry code. Moreover, it is hard to quantify the dollar of how efficient and precise that startups owners to identify the correct industry code. If count as consultant salary based on expected working hour and compensation rate, the consultant salary costs $28,000. But the job to develop an industry analyzer tool for NYU MASY is unpaid. So, there are no benefits and costs that can be quantified.

Assumptions

1.The data, including the NAICS code, business descriptions provided by NYU - MASY, are accurate.
2. The industry analyzer search tool is developing, which has room for improvement.

Scope

- **Quality**
  - The industry code search tool will return a ranked list of industry codes and present the top 5 codes and descriptions to the user for selection.
  - The shiny app is expected to accept the text file of the startup's business description and deliver a ranked list of the top 5 possible NAICS industry codes.
  - The shiny app shall allow the users to download the ranked code as a CSV file.

- **Time**
  - The industry code search tool should be finished by April 9.
  - The finalized shiny app in R should be finished by April 14.
  - The entire project requires 300 hours.
  - The project should be finished before May 05.
- **Resource Allocation**
  - Reference preliminary work is done in Python to develop a new search tool in R
  - Utilize resources and sampled data to test the TF-IDF algorithm
  - Require 300 hours as a project manager and project implementor. The work will be done on a pro-bono basis.

Out of scope activities

None

Constraints

1. NAICS code requires a long time to update. Startups might create a brand new business sector that is not categorized by the NAICS code. Therefore, new startups' business descriptions might be covered by the NAICS code system.

2. TF-IDF algorithm computes documents directly in a word-count space, which might be super slow for text with large vocabularies. Besides, TF-IDF does not compare the semantic similarities between words, which might bring bias to the result.

3. Consultant is on a part-time basis.

Risks and Mitigation Strategies

Risk 1: Startups might have difficulties using the industry code search tool at first.

Strategy 1: Provide a detailed instruction document to show users how to use the tool

Risk 2: There might be another team developing a similar project that might be adopted by NYU – MASY before May.

Strategy 2: Make sure the project can be finished in high quality before May and meet the requirement of NYU – MASY.

Communications Plan

1. Frequency: Once per week
2. Method: Affected by the epidemic, the main communication methods are emails and zoom meetings.
3. Content: Status report; Milestone notification; Project status update; Zoom video recordings.

Schedule Overview

Project Start Date: January 28, 2021

Estimated Project Completion Date: May 5, 2021

Major Milestones

Milestone 1: Project Sponsor Acceptance - March 10, 2021

Milestone 2: Submit Project Charter - March 17, 2021

Deliverable 1: Functional Specification Document Approval - February 12, 2021

Deliverable 2: Perform and deliver a TF-IDF similarity scoring of a business's text description against the text descriptions of the NAICS industry codes using R language - March 19, 2021

Deliverable 3: The Shiny App with Function to Accept the Text File of Business Description and Deliver a Ranked List of the Top 5 Possible NACIS Industry Codes - April 9, 2021

Deliverable 4: A GitHub Repository with All the Documents that Viewed and Accepted by the Clients - April 14, 2021

External Milestones Affecting the Project

Not identified.

Impact of Late Delivery

The project sponsor will deduct the students' marks if the project is delivered late.

Resources Required

Role	Responsibilities	Duration of work	Qualifications needed
Project Manager	Manage the project	50 hours	Experience in project management
Project Implementor	Code for the project	300 hours	Experience in programming

Facilities, Software, Hardware, and Other Resources

Personal computer, Microsoft Office Suite, R Studio, and Zoom for weekly meetings

Procedures/ Methodology

Refer to the Work Breakdown Structure (Appendix 1).

Project Evaluation

1. **Project schedule:** refer to Work Breakdown Schedule (Appendix 1)
2. **Project weekly status report and dashboard**: I keep the project sponsor informed by my project progress during weekly zoom meetings. The zoom meeting recordings have all my verbal progress report to the project sponsor.
3. **Project communication plan, issues log, risk register**: I will meet the project sponsor weekly via Zoom meeting and report any of my project progress. Besides, I will show the project supervisor what I plan to do in the next step. Weekly progress will be documented in weekly zoom meeting recordings. Any project change will be documented in the Change Management Plan with corresponding contingency plans.
4. **Project monthly status report:** Refer to Appendix 2.

Appendix 1

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	*Milestone:* 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	Perform and Delivery TF-IDF Similarity Code with Samples Using R Language	The project implementer should finish the sample exercises in TF-IDF.	March 1, 2021
3	1.3.4	Conduct Unit Test of Sample Code for Each Section in TF-IDF Code	The project implementer should conduct unit tests of each section in the TF-IDF algorithm.	March 5, 2021
3	1.3.5	*Deliverable 2:* Perform and deliver a TF-IDF similarity scoring of a business's text description against the text descriptions of the NAICS industry codes using R language.	The TF-IDF similarity scoring will return a ranked list of industry codes and present the top 5 codes and descriptions to the user for selection.	March 19, 2021
3	1.3.6	Conduct Unit Test of Code of Business Description Against NACIS Code	The test will utilize the confusion matrix to show the results and improve accuracy and precision to at least 85%.	March 26, 2021
3	1.3.7	Develop a Shiny App Interface	The project implementer develops a shiny app using R.	April 8, 2021
3	1.3.8	*Deliverable 3:* The Shiny App with Function to Accept the Text File of Business Description and Deliver a Ranked List of the Top 5 Possible NACIS Industry Codes	The shiny app is expected to accept the text file of the startup's business description and deliver a ranked list of the top 5 possible NAICS industry codes.	April 9, 2021
3	1.3.9	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	1.3.10	*Deliverable 4:* A GitHub Repository with All the Documents that Viewed and Accepted by the Clients	All the documents are accepted by the client and can be viewed on the GitHub repository.	April 14, 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

Appendix2

Project Status Areas:	**Execution Week <x>**		
	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
• The functional specification document was approved. • The project sponsor agreement was approved. • The sponsor project acceptance document was approved. • Finished and test the sampled code.

2 – Project Schedule	
Tasks that are not on schedule per workplan	Impact
1.	1.

3 – Project Deliverables
COMPLETED DELIVERABLES: Functional Specification Document ***UPCOMING DELIVERABLES:*** TF – IDF coding Shiny App interface GitHub Repository

4 – Issues

5 – Project Risks	
Potential Risks	**Possible Mitigation**

6– Resources and Collaboration
•

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

Comments/Actions

8 – Sponsor Signoff	
Sponsor indicates agreement with the above status report.	

Assessment Guidelines

The assessment is designated by one of the three "Traffic Light" colors utilizing the following guidelines:

Each project should establish the appropriate project slippage metrics for yellow vs red status

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.

In your filename for this document, prefix with Green-, Red-, or Yellow-. G- or R- or Y- and show the date and your name

Appendix D

Project Plan

NYU SCHOOL OF PROFESSIONAL STUDIES

MASTER OF SCIENCE IN MANAGEMENT AND SYSTEMS
Applied Project Capstone
MASY GC- 4100

MEMORANDUM

TO: Dr. Andres Fortino
FROM: Yin (Fien)Xu
DATE: March 3, 2021

RE: **Assignment 3B – Work Breakdown Structure and Schedule**

Project Tasks Outline

The project task outline is presented in the order from initiation, planning, execution, control, and closeout. The work breakdown structure displayed all the work and deliverables required to complete the project, and more details are shown below.

OUTLINE VIEW

1. Widget Management System
   - 1.1 Initiation
     - 1.1.1 Evaluation & Recommendations
     - 1.1.2 Develop Project Proposal
     - 1.1.3 Draft Literature Review Research
     - 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 *Milestone:* 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 Perform and Delivery TF-IDF Similarity Code with Samples Using R Language
1.3.4 Conduct Unit Test of Sample Code for Each Section in TF-IDF Code
1.3.5 *Deliverable 2:* Performs the TF-IDF Similarity Scoring the text description of a business against the text descriptions of the NAICS industry codes and returns A Ranked Lists and Displayed the Top 5 Industry Codes and Business Descriptions to Users for Selection
1.3.6 Conduct Unit Test of Code of Business Description Against NACIS Code
1.3.7 Develop a Shiny App Interface
1.3.8 *Deliverable 3:* The Shiny App with Function to Accept the Text File of Business Description and Deliver a Ranked List of the Top 5 Possible NACIS Industry Codes
1.3.9 Create an Accessible GitHub Repository for Recording All the Project Files and Supporting Documentation
1.3.10 *Deliverable 4:* A GitHub Repository with All the Documents that Viewed and Accepted by the Clients
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

Create a WBS Schedule including task definitions and due dates similar to the table below:

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	*Milestone:* 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	Perform and Delivery TF-IDF Similarity Code with Samples Using R Language	The project implementer should finish the sample exercises in TF-IDF.	March 1, 2021
3	1.3.4	Conduct Unit Test of Sample Code for Each Section in TF-IDF Code	The project implementer should conduct unit tests of each section in the TF-IDF algorithm.	March 5, 2021
3	1.3.5	*Deliverable 2:* Perform and deliver a TF-IDF similarity scoring of a business's text description against the text descriptions of the NAICS industry codes using R language.	The TF-IDF similarity scoring will return a ranked list of industry codes and present the top 5 codes and descriptions to the user for selection.	March 19, 2021

3	1.3.6	Conduct Unit Test of Code of Business Description Against NACIS Code	The test will utilize the confusion matrix to show the results and improve accuracy and precision to at least 85%.	March 26, 2021
3	1.3.7	Develop a Shiny App Interface	The project implementer develops a shiny app using R.	April 8, 2021
3	1.3.8	*Deliverable 3:* The Shiny App with Function to Accept the Text File of Business Description and Deliver a Ranked List of the Top 5 Possible NACIS Industry Codes	The shiny app is expected to accept the text file of the startup's business description and deliver a ranked list of the top 5 possible NAICS industry codes.	April 14, 2021
3	1.3.9	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	1.3.10	*Deliverable 4:* A GitHub Repository with All the Documents that Viewed and Accepted by the Clients	All the documents are accepted by the client and can be viewed on the GitHub repository.	April 21, 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

NYU SCHOOL OF PROFESSIONAL STUDIES

MASTER OF SCIENCE IN MANAGEMENT AND SYSTEMS
Applied Project Capstone
MASY GC- 4100

MEMORANDUM

TO: Dr. Andres Fortino
FROM: Yin (Fien)Xu
DATE: March 2, 2021

RE: **Assignment 3A – Situational Analysis**

Applied Project Situation Analysis

Industry Analysis

The company of the industry code analyzer is New York University (NYU). Founded in 1831, NYU is one of the world-foremost private research universities in the higher education industry. MASY is a master's program offered through NYU's Management and Technology Department within the Division of Programs in Business. The high education industry consists of students, teachers, institutions providing similar higher education, institutions offering courses or certifications that substitute traditional in-person teaching mode, and institutions offering associate degrees or certificates.

The higher education industry has been growing and changing, driven by increasing student enrollment and increasing internationalization in the education sector. To better attract students and maintain competitive ranking, higher education institution strives for differentiating from various aspects, including student mix and outcomes, faculty resources, research capabilities, facilities, and community impact (André Dua & Jonathan, 2020). Higher education attaches more

importance to expand community impact than before to real-world problems solving and practical training. Moreover, higher education institutions continue to make full use of their advantages in applying data and technologies to contribute to society's growth (Spear, 2020). Therefore, aiming to strengthen community impact and improve researcher experiences, NYU MASY developed an industry code search tool in Python to find the industry code for startups. To optimize users' experiences, NYU MASY aims to create a computer-based tool with a user-friendly interface in R to identify industry codes based on startup business descriptions.

Competitors

Location is a significant competitive advantage of higher education institutions, and higher education institutions that have similar locations are considered main competitors. Other private higher education institutions in New York City, such as Columbus University, Fordham University, Yeshiva University, are considered a competitor of NYU.

Stakeholders

In the industry analyzer project, the shareholders include the project sponsor, project implementer, project manager, faculties, students, researchers, officers of NYU administrative department, NYU shareholders, startups, officers of NACIS Association, LLC, and U.S. Census Bureau. The industry code will serve as an efficient search tool to help students, teachers, researchers, and startups identify the correct industry code for professional and academic uses. Moreover, NYU will be benefited from this project to expand cooperation and influence among the community by supporting startups. NAICS Association will obtain higher brand awareness as an industry code provider when more users utilize the industry analyzer. U.S. Census Bureau is responsible for providing data to NAICS Association. It will play a monitor role as stakeholders to ensure the data credibility and quality of industry description and classification.

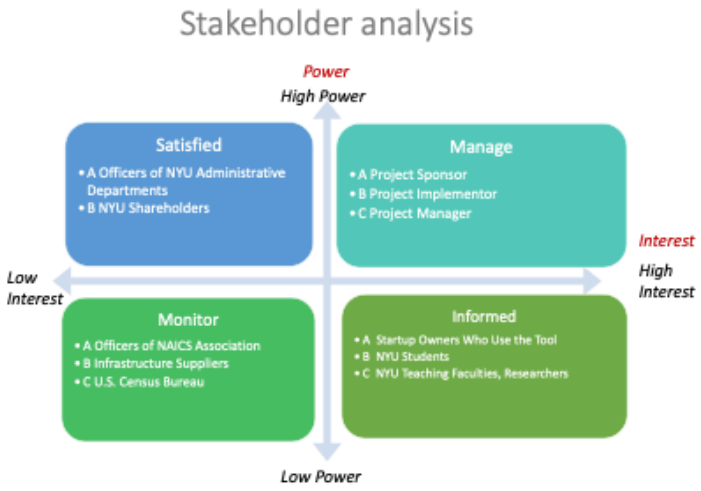

Figure 1: Shareholder Analysis

Porter's Five-Forces Model

The Threat of Potential New Entrants

Professional training certificates are becoming prevalent since more and more companies provide their employees with necessary job skills training. Those companies, such as IBM, Bloomberg, AWS, possess enterprise-level technology and infrastructures. They are able to collaborate with educational institutions to provide training certificates for their employees with privilege or any person who is willing to pay full price to receive certificates. The certification is official issued by the companies or certified institutions, but certifications offered by widely recognized higher education institutions have higher credibility. Therefore, the threat of potential new entrants is relatively low.

The Threat of Substitutes

One of the advantages of private higher education research institutions is elite-oriented teaching in small classes. As the Internet popularize, online courses will be the substitute for people to have higher education. More and more higher education institutions offer online graduate programs for students worldwide to enroll in the class without location restrictions. Therefore, the level of threat of substitutes is medium since online programs have advantages in flexible course time and location while they lack a close-knit community to students.

Bargaining Power of Customers

Students are the primary consumers of higher education institutions. Tuition is one of the major deciding factors when students a university. Typically, the tuition of private research universities is far higher than the one of public universities. In addition to paying tuition, students need to meet the universities' enrollment criteria, and students have less bargaining power over deciding the final tuition even though they have a scholarship or student loan.

Bargaining Power of Suppliers

Instructors and faculties are primary suppliers of higher education. Instructors and faculties are paid based on their academic research or teaching experiences, and they have more bargaining power in salary negotiation than those who teach in a public education institution. In addition to faculties, infrastructure providers are also the suppliers of schools. Higher education institutions require a lot of infrastructures to ramp up labs, classrooms, offices, and those infrastructures typically cost considerable expenses. Schools can choose any level of infrastructure based on the budget; therefore, the power of suppliers is medium or low.

Competitive Rivalry

The competitors of NYU are the similar private research universities that provide higher educations. The National Center for Educational Statistics showed that the number of four-year postsecondary public and private institutions has increased at a 54 percent rate from 1,957 in 1981 to 3,026 in 2013, while the rise in the overall number of students is only 16.8% between 2016 and 2025, showing that more schools might compete intensely for fewer students (National Center for Education Statistics, 2017). Therefore, the competitive rivalry is high.

References

André Dua, & Jonathan, L. (2020). *As education leaders consider their options in the age of the COVID-19 crisis, they must rethink conventional wisdom.* Retrieved from McKinsey & Company: https://www.mckinsey.com/industries/public-and-social-sector/our-insights/reimagining-higher-education-in-the-united-states

National Center for Education Statistics. (2017). *Educational Institutions*. Retrieved from How many educational institutions exist in the United States: https://nces.ed.gov/fastfacts/display.asp?id=84

Spear, E. (2020). *10 Trends in Higher Education to Watch in 2020*. Retrieved from Precision Campus: https://precisioncampus.com/blog/trends-higher-education/

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 (if appropriate)

Other Business Benefits

N/A

There is an intangible benefit of improving efficiency for startup owners to find the correct industry code. Moreover, it is hard to quantify the dollar of how efficient and precise that startups owners

to identify the correct industry code. If count as consultant salary based on expected working hour and compensation rate, the consultant salary costs $28,000. But the job to develop an industry analyzer tool for NYU MASY is unpaid. So, there are no benefits and costs that can be quantified.

Appendix F

Risk Management Plan

NYU SCHOOL OF PROFESSIONAL STUDIES

MEMORANDUM

TO: Dr. Andres Fortino
FROM: Yin (Fien)Xu
DATE: April 14, 2021

RE: **Assignment 7 – Risk Management Plan**

Project

Industry Code Analyzer - NAICS Code Discovery for Startups in R

Given a business description as a text file, the tools will search the NAICS industry code to identify the industry classification corresponding to the business. The industry code analyzer tool will use TF-IDF text similarity scoring, return a ranked list of industry codes, and present the top 5 codes and descriptions to the user for selection.

Risks

Number	Risk	Probability Score (1,2 or 3)	Impact Score (1,2 or 3)
1	The client stops the project.	1	3
2	Functions do not work on Shiny App. io	2	3
3	Improper website design.	1	2
4	User (startups) does not want to use the tool.	2	2
5	My R skills are not sufficient to debug the tools.	3	2

Risk Matrix

	RISK (exposure)

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

Contingency Plan

Risk	Description	Probability (1-3)	Exposure (1-3)	Contingency Plan
1	The client stops the project.	1	3	
2	Functions do not work on Shiny App. io	2	3	Run the shiny app using a personal account to ensure every function works successfully before uploading to a public account.
3	Improper website design	1	2	
4	User (startups) does not want to use the tool	2	2	
5	My R skills are not sufficient to debug the tools.	3	2	Communicate with other colleagues who have similar projects and find more learning materials online.

Appendix G

Change Management Plan

PROJECT CHANGE MANAGEMENT PLAN

Project Name:	**Industry Code Analyzer - NAICS Code Discovery for Startups in R**
Prepared by:	**Yin (Fien) Xu**
Date (MM/DD/YYYY):	**03/30/2021**

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*
• 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. • 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 the 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.

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.
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

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

	• 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*).

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
• 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 Class

- 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:

[_] Other:

- The following individuals have edit access to project documents:

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

•	•
•	•

6. Project Change Management Plan / Signatures

Project Name:	Industry Code Analyzer – NAICS Code Discovery for Startups in R
Project Manager:	Yin (Fien) Xu

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

Name	Role	Signature	Date (MM/DD/YYYY)
Yin (Fien) Xu	Project Manager; Project Implementor	Yin (Fien) Xu	03/30/2021

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

Project Status Report

<Industry Code Analyzer - NAICS Code Discovery for Startups in R > Status Report – March <March 2021>

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

From: Yin (Fien) Xu

Date: March 30, 2021

YOUR ANTICIPATED COMPLETION DATE: May 5, 2021_____

COMPLETION SEMESTER: Spring, 2021_____**(e.g. Summer, 2025)**

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
• Deliverable 1 Finished: The functional specification document was approved. • The project sponsor agreement was approved. • The sponsor project acceptance document was approved. • The project charter was approved. • Deliverable 2 Finished: Performed TF-IDF similarity scoring. • Finished the unit test using sampled data. • Finished shiny app learning courses

2 – Project Schedule	
Tasks that are not on schedule per workplan	Impact

3 – Project Deliverables
COMPLETED DELIVERABLES: Deliverable 1: Functional specification document approval - February 12, 2021 Deliverable 2: Perform and deliver a TF-IDF similarity scoring of a business's text description against the text descriptions of the NAICS industry codes using R language - March 19, 2021 ***UPCOMING DELIVERABLES:*** Deliverable 3: The Shiny App with function to accept the text file of business description and deliver a ranked list of the Top 5 possible NACIS Industry Codes - April 9, 2021 Deliverable 4: A GitHub repository with all the documents viewed and accepted by the clients - April 14, 2021

4 – Issues

5 – Project Risks	
Potential Risks	**Possible Mitigation**

6– Resources and Collaboration
• Data: NAICS Industry code, Business Description • Personal Computer • Microsoft Office Suite • R Studio • Zoom for weekly meetings

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

Comments/Actions

8 – Sponsor Signoff	
Sponsor indicates agreement with the above status report.	
AJF	3/31/21

Assessment Guidelines

The assessment is designated by one of the three "Traffic Light" colors utilizing the following guidelines:

Each project should establish the appropriate project slippage metrics for yellow vs red status

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.

<Industry Code Analyzer - NAICS Code Discovery for Startups in R > Status Report – April < April 2021>

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

From: Yin (Fien) Xu

Date: April 14, 2021

YOUR ANTICIPATED COMPLETION DATE: May 5, 2021_____

COMPLETION SEMESTER: Spring, 2021_____ (e.g. Summer, 2025)

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
• Deliverable 1 Finished: The functional specification document was approved. • The project sponsor agreement was approved. • The sponsor project acceptance document was approved. • The project charter was approved. • Deliverable 2 Finished: Performed TF-IDF similarity scoring. • Finished the unit test using sampled data. • Finished shiny app learning courses • Deliverable 3 Finished: Created a shiny app interface • Finished unit test of shiny app functions

2 – Project Schedule	
Tasks that are not on schedule per workplan	Impact

3 – Project Deliverables
COMPLETED DELIVERABLES: Deliverable 1: Functional specification document approval - February 12, 2021 Deliverable 2: Perform and deliver a TF-IDF similarity scoring of a business's text description against the text descriptions of the NAICS industry codes using R language - March 19, 2021 Deliverable 3: The Shiny App with function to accept the text file of business description and deliver a ranked list of the Top 5 possible NACIS Industry Codes - April 9, 2021 ***UPCOMING DELIVERABLES:*** Deliverable 4: A GitHub repository with all the documents viewed and accepted by the clients - April 14, 2021

4 – Issues

5 – Project Risks	
Potential Risks	**Possible Mitigation**

6– Resources and Collaboration
• Data: NAICS Industry code, Business Description • Personal Computer • Microsoft Office Suite • R Studio • Zoom for weekly meetings • Shinyapps.io

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

The assessment is designated by one of the three "Traffic Light" colors utilizing the following guidelines:

Each project should establish the appropriate project slippage metrics for yellow vs red status

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

NYU SCHOOL OF PROFESSIONAL STUDIES

MASTER OF SCIENCE IN MANAGEMENT AND SYSTEMS

Applied Project Capstone

MASY GC- 4100

MEMORANDUM

TO: Dr. Andres Fortino

FROM: Yin (Fien)Xu

DATE: 02/21/2021

RE: **Assignment 2A – Ten to Fifteen References**

References

1. Anastasia, C. (2015). Exploring definitions of small business and why it is so difficult. *Journal of Management Policy and Practice*, *16*(4), 88.

   Abstract: Small businesses dot the American landscape and have stimulated the economy for the past two centuries. The definitions for small businesses have changed consistently over the course of the last five decades. While some businesses are considered small if they have fewer than 500 employees, other businesses seem to get lost in the process

simply because they fall into a microenterprise category, including the Mom and Pop companies, family-owned companies, and those individuals that are considered self-employed. The Small Business Administration (SBA) has been the authority on defining what is or is not a small business. With this in mind, this study used a mixed methodology to explore the definitions of small businesses and microenterprises by surveying 388 MBAs and CPAs in the United States to create a better understanding of what constitutes a small business and a microenterprise. When all was said and done, the definitions of small businesses and micro-enterprises still remain a mystery. However, a broader understanding of the difficulties associated with defining the two is brought to the forefront.

This article explained why small business administration (SBA) utilizes industry code for small businesses to identify their industries. It is important for startups to reference industry codes to define business and market scope. This article will be used in the literature review to show the importance of industry code to small businesses. Besides, this article shows vague definitions of small business and microenterprise place primary challenges for small business to identify the corresponding industry code.

2. Bafna, P., Pramod, D., & Vaidya, A. (2016). Document clustering: TF-IDF approach. *International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), 16*(3), 61-66.

Abstract: Recent advances in computer and technology resulted in an ever-increasing set of documents. The need is to classify the set of documents according to the type. Laying related documents together is expedient for decision-making. Researchers who perform

interdisciplinary research acquire repositories on different topics. Classifying the repositories according to the topic is a real need to analyze the research papers. Experiments are tried on different real and artificial datasets such as NEWS 20, Reuters, emails, research papers on various topics. Term Frequency-Inverse Document Frequency algorithm is used along with fuzzy K-means and hierarchical algorithm. Initially, an experiment is being carried out on a small dataset and performed cluster analysis. The best algorithm is applied to the extended dataset. Along with different clusters of the related documents, the resulted silhouette coefficient, entropy, and F-measure trend are presented to show algorithm behavior for each data set.

The authors utilized the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm in R programming tools to examine the relevance of words to documents. Differently, the authors applied fuzzy K-means and hierarchical algorithm to find similar words. This article will be used in the literature review to serve as a comparison of cosine similarity.

3. Crisan, A., Munzner, T., & Gardy, J. L. (2019). Adjutant: An R-based tool to support topic discovery for systematic and literature reviews. *Bioinformatics*, *35*(6), 1070-1072.

Abstract: Adjutant is an open-source, interactive, and R-based application to support mining PubMed for literature reviews. Given a PubMed-compatible search query, Adjutant downloads the relevant articles and allows the user to perform an unsupervised clustering analysis to identify data-driven topic clusters. Following clustering, users can also sample documents using different strategies to obtain a more manageable dataset for further analysis. Adjutant makes explicit tradeoffs between speed and accuracy, which are modifiable by the user, such that a complete analysis of several thousand

documents can take a few minutes. All analytic datasets generated by Adjutant are saved, allowing users to easily conduct other downstream analyses that Adjutant does not explicitly support.

The authors explained the detailed R's shiny framework and introduced several packages to support the R-based applications. Besides, the article showed the data visualization using a shiny app. The article will be adopted in the literature review to provide background information and the interface of a shiny app.

4. Gunawan, D., Sembiring, C. A., & Budiman, M. A. (2018). The implementation of cosine similarity to calculate text relevance between two documents. *Journal of Physics: Conference Series*, *978*(1), 121-127.

Abstract: Rapidly increasing the number of web pages or documents leads to topic-specific filtering in order to find web pages or documents efficiently. This is preliminary research that uses cosine similarity to implement text relevance in order to find a topic-specific document. This research is divided into three parts. The first part is text-preprocessing. In this part, the punctuation in a document will be removed, then convert the document to lower case, implement stop word removal, and then extracting the root word by using the Porter Stemming algorithm. The second part is keywords weighting. Keyword weighting will be used in the next part, the text relevance calculation. Text relevance calculation will result from the value between 0 and 1. The closer value to 1, then both documents are more related, vice versa.

The authors showed the detailed process of implementing Term Frequency- Inverse Document Frequency based on cosine similarity to calculate the relevance between two text documents from text processing to the result visualization stage. This

article will be adopted in the literature review as the reference of industry analyzer design.

5. Hashemzadeh, B., & Abdolrazzagh-Nezhad, M. (2020). Improving keyword extraction in multilingual texts. *International Journal of Electrical & Computer Engineering, 10*(6), 5909-5916.

Abstract: The accuracy of keyword extraction is a leading factor in information retrieval systems and marketing. In the real world, the text is produced in various languages, and the ability to extract keywords based on information from different languages improves the accuracy of keyword extraction. In this paper, the available information of all languages is applied to enhance a traditional keyword extraction algorithm from a multilingual text. The proposed network extraction procedure is an unsupervised algorithm and designed based on selecting a word as a keyword of a given text, if in addition to that, language holds a high rank based on the keywords criteria in other languages, as well. To achieve this aim, the average TF-IDF of the candidate words was calculated for the same and the other languages. Then the words with the higher averages TF-IDF were chosen as the extracted keywords. The obtained results indicate that the algorithms' accuracies of the multilingual texts in terms of frequency-inverse document frequency (TF-IDF) algorithm, graph-based algorithm, and the improved proposed algorithm are 80, 60.65, and 91.3%, respectively.

This article pointed out the significant application of text mining and keyword extraction in real-world scenarios. Besides, the author compares the matching result of documents generated by Term Frequency – Inverse Document Frequency (TF-IDF) with a graph-based algorithm and the improved proposed algorithm. The

article will be used in the literature review to show that TF-IDF is an acceptable algorithm used in keyword matching among various documents.

6. Kile, C. O., & Phillips, M. E. (2009). Using industry classification codes to sample high-technology firms: Analysis and recommendations. *Journal of Accounting, Auditing & Finance*, *24*(1), 35-58.

Abstract: This study develops procedures for selecting and partitioning samples of high-technology firms for North American Industry Classification System (NAICS), Global Industry Classification System (GICS), and Standardized Industry Classification (SIC) codes. This study also assesses whether NAICS or GICS codes offer an improvement over SIC codes as a methodology to identify high-technology firms, including Internet firms. SIC codes are predominantly used in accounting, information systems, business, and economics studies to select high-technology firms for analyses. However, recent studies such as Clarke (1989), Fan and Lang (2001), Ramnath (2001), and Bhojraj, Lee, and Oler (2003) have called into question the accuracy of using SIC codes for firm classification. Our study constructs a benchmark industry classification, constructed from analyses of detailed descriptions of firms' businesses from Description of Business and footnote disclosures (required in all 10-Ksand registration statements). Based on a comparison with our benchmark industry classification, we developed combinations of codes from each system that we recommend being used to construct samples of high-technology firms or specifically targeted high-technology industries. We conclude from our analyses that GICS codes offer an improvement over SIC and NAICS codes for targeting technology firms.

The author conducts a comprehensive comparison among the NAICS code, GICS code, and SIC code. This article will be used to provide background information on the NAICS code and illustrate why the NAICS code is beneficial for startups.

7. O'Connor, L. (2000). Approaching the challenges and costs of the North American industrial classification system (NAICS). *The Bottom Line*.

Abstract: The transition from the standard industrial classification (SIC) system to the North American industrial classification system (NAICS) will not be rapid, but its effects will be profound for business researchers and information professionals. Most government agencies are already in the midst of this six-year transition and, although private information producers are not compelled to switch from SIC to NAICS, most are planning to do so. The far-reaching impact of NAICS on business information will affect libraries of all types. This article describes the challenges and costs associated with this change and makes recommendations for materials and training.

This article introduces the transition from the Standard Industrial Classification (SIC) system to the North American Industrial Classification System (NAICS). Besides, this article shows the profound effect of the NAICS code on entrepreneurs. This article will be used to further proves that it is meaningful to create an industry analyzer to help startups to find the correct industry code.

8. Pierce, J. R., & Schott, P. K. (2012). A concordance between ten-digit US Harmonized System Codes and SIC/NAICS product classes and industries. *Journal of Economic and Social Measurement*, *37*(1-2), 61-96.

Abstract: This paper provides and describes concordances between the ten-digit Harmonized System (HS) categories used to classify products in US international trade and the four-digit SIC and six-digit NAICS industries that cover the years 1989 to 2006. We also provide concordances between ten-digit HS codes and the five-digit SIC and seven-digit NAICS product classes used to classify US manufacturing production. Finally, we briefly describe how these concordances might be applied in current empirical international trade research.

This article pointed out the concordances between Harmonized System (HS) categories adopted in products classification in the US international trade and System Industrial Classification (SIC) and North American Industrial Classification System (NAICS) that are used to industry classification. This article will be used to explain metrics that the NAICS code uses to locate the specific industry which has a connection to specialized groups of products.

9. Ramos, J. (2003). Using TF-IDF to determine word relevance in document queries. *Proceedings of the first instructional conference on machine learning, 242*(1), 29-48.

Abstract: In this paper, we examine the results of applying Term Frequency Inverse Document Frequency (TF-IDF) to determine what words in a corpus of documents might be more favorable to use in a query. As the term implies, TF-IDF calculates values for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in. Words with

high TF-IDF numbers imply a strong relationship with the document they appear in, suggesting that if that word were to appear in a query, the document could be of interest to the user. We provide evidence that this simple algorithm efficiently categorizes relevant words that can enhance query retrieval.

This article explained the algorithm of Term Frequency Inverse Document Frequency (TF-IDF) and showed how TF-IDF could be used in text mining. The author indicated that words with high TF-IDF numbers show a strong relationship with the document they appear in. This article will be used in the literature review as a reference to prove that TF-IDF is one of the commonly used ways to find similarities among word documents.

10. Tata, S., & Patel, J. M. (2007). Estimating the selectivity of TF-IDF-based cosine similarity predicates. *ACM Sigmod Record*, *36*(2), 7-12.

Abstract: An increasing number of database applications today require sophisticated approximate string-matching capabilities. Examples of such application areas include data integration and data cleaning. Cosine similarity has proven to be a robust metric for scoring the similarity between two strings, and it is increasingly being used in complex queries. An immediate challenge faced by current database optimizers is to find accurate and efficient methods for estimating the selectivity of cosine similarity predicates. To the best of our knowledge, there are no known methods for this problem. In this paper, we present the first approach for estimating the selectivity of TF-IDF-based cosine similarity predicates. We evaluate our approach on three different real datasets and show that our method often produces estimates that are within 40% of the actual selectivity.

The authors utilized the TF-IDF algorithm based on cosine similarity. The authors explained detailed metrics of cosine similarity for scoring the similarity between vectors. Even though the report pointed that the selectivity of cosine similarity is not solved, the report proved that cosine similarity still an efficient metric to identify the similarity between terms. This article will be used as a reference in the literature review to illustrate the function, application, and significance of cosine similarity.

11. Wu, H. C., Luk, R. W. P., Wong, K. F., & Kwok, K. L. (2008). Interpreting TF-IDF term weights as making relevant decisions. *ACM Transactions on Information Systems (TOIS)*, *26*(3), 1-37.

Abstract: A novel probabilistic retrieval model is presented. It forms a basis to interpret the TF-IDF term weights as making relevant decisions. It stimulates the local relevance decision-making for every location of a document and combines all of these "local" relevance decisions as to the "document-wide" relevance decision for the document. The significance of interpreting TF-IDF in this way is the potential to (1) establish a unifying perspective about information retrieval as relevant decision-making and (2) develop advanced TF-IDF-related term weights for future elaborate retrieval models. Our novel retrieval model is simplified to a basic ranking formula that directly corresponds to the TF-IDF term weights. In general, we show that the term-frequency factor of the ranking formula can be rendered into different term-frequency factors of existing retrieval systems. In the basic ranking formula, the remaining quantity - log p(r¯|t ∈ d) is interpreted as the probability of randomly picking a nonrelevant usage (denoted by r¯) of term t. Mathematically, we show that this quantity can be approximated by the inverse

document-frequency (IDF). Empirically, we show that this quantity is related to IDF, using four references TREC ad hoc retrieval data collections.

This article introduced the metrics and applications of Term Frequency Inverse Document Frequency (TF-IDF) in text mining and document-wide matching. Besides, the authors showed and developed an advanced TF-IDF-related term weight for future elaborate text similarity retrieval models. This article will be used to provide a theoretical explanation of TF-IDF and offer potential solutions to improve model performance.