Project

Initial Consultation with the Client

**Team: TalkTech Tinkerers**

**Group Members:**

**Rodgers Okeyo Ochieng**

**Jihee Wang**

**Divyanshi Singh**

**Chandan Patel**



ISM 6564: Text Analytics

Dr: Tim Smith

Date: October 1st 2023

Table of Contents

[1) Executive Summary 3](#_Toc147086039)

[ Objective: 3](#_Toc147086040)

[ Findings: 3](#_Toc147086041)

[ Recommendations 3](#_Toc147086042)

[2) Business Process Analysis 4](#_Toc147086043)

[a) Detailing the existing MSBAIS business process landscape 4](#_Toc147086044)

[b) Identifying Bottlenecks 4](#_Toc147086045)

[c) Exploring areas where a ChatGPT based Application like Q&A could add value 5](#_Toc147086046)

[3) Data Assessment 5](#_Toc147086047)

[ Available Data Sources: 5](#_Toc147086048)

[ Data Quality 6](#_Toc147086049)

[ Knowledge Base Assembly 6](#_Toc147086050)

[4) Scope of Use 6](#_Toc147086051)

[5) Appendix 7](#_Toc147086052)

[a) The Q&A application Flow: 7](#_Toc147086053)

[b) Interview Notes 8](#_Toc147086054)

# Executive Summary

* + Objective: Our project's objective is to create a ChatGPT-based interface tailored to enhance MSBAIS business operations at the University of South Florida (USF). The focus is on assisting prospective and new students, streamlining their initial engagement with USF, from choosing a program to the application and onboarding processes.

## Findings:

* + 1. **System in Use:** The primary tool for handling, tracking, and resolving student inquiries within the MSBAIS program at USF is the Jira system.
    2. **Assistance Approach:** USF's MS BAIS program assists students via its information request page on the graduate course website. While a comprehensive amount of data is available on the online portal, students often reach out directly for clarifications. In response, the USF IT department leverages the JIRA system to systematically address and manage these inquiries by raising tickets.
    3. **Admission Process:** The enrollment process at USF is methodical, involving an initial scrutiny of documents, data gathering, and a concluding decision based on the information and applicant's qualifications.
    4. **Student Interaction Tools:** Notably, the MS BAIS program website currently lacks any interactive interface or chatbot system, relying entirely on conventional methods to resolve student queries.
    5. **Primary Information Source::** The program’s FAQ page has been identified as a pivotal resource, capturing, and addressing the most recurrent student questions and concerns
  + Recommendations**:**
    1. **Streamlining Data Collection:** For bolstered digital support, it's essential to gather specific data concerning frequently asked MS BAIS program questions. These data points can guide automated responses once the user provides initial context. Historical data, including past JIRA tickets related to MS BAIS programs and previous email interactions, can be pivotal in creating a comprehensive corpus for any automated tool. Additionally, web scraping can be utilized to extract generic program information from the official website. The program's detailed FAQ section can also serve as an invaluable resource to train the Q&A application, ensuring it remains aligned with students' needs.
    2. **Q&A Application to be integrated with MS BAIS website:** Given the existing reliance on manual processes and the availability of the above-crafted corpus/data, introducing a Q&A application within an MS BAIS website can significantly optimize student interactions. Once operational, this application can address FAQs, guide prospective students through the application process, and reduce the workload on the administrative teams.

# Business Process Analysis

1. Detailing the existing MSBAIS business process landscape:
   * 1. **System Utilization**: At its core, USF employs the Jira system, a renowned tool for handling, tracking, and addressing student inquiries related to the MSBAIS program.
     2. **Assistance Dynamics**: While there is a robust presence of the MS BAIS program on the USF graduate course website, which provides extensive information, students often find the need to approach directly for specific clarifications. This is typically handled by the IT department of USF, which, in turn, makes use of the JIRA system to manage these queries.
     3. **Enrollment Paradigm**: The admission procedure is well-defined. It starts with a preliminary review of application documents, followed by a systematic gathering of necessary information, culminating in a final decision based on the comprehensive data and the applicant's credentials.
2. Identifying Bottlenecks**:**
   * 1. **Reliance on Manual Channels**: Despite the availability of comprehensive information online, students find themselves reaching out directly, which indicates potential clarity gaps on the portal.
     2. **Absence of Automated Support**: There's a noticeable lack of instant digital assistance tools, like chatbots, which can significantly delay response times and increase the manual workload.
     3. **Over-reliance on JIRA**: While JIRA is efficient, using it as a primary tool for all student queries may lead to potential backlogs, especially during peak admission seasons.
3. Exploring areas where a ChatGPT based Application like Q&A could add value**:**
   * 1. **Immediate Assistance**: A ChatGPT based Q&A application can provide real-time assistance to students, helping answer frequent queries instantly and reducing the time taken to provide responses.
     2. **Reducing Workload**: Automated responses to common queries can significantly decrease the manual workload on the administrative and IT teams.
     3. **Optimizing Admission Funnel**: By guiding prospective students through the application process interactively, the Q&A application can enhance user experience, potentially increasing successful applications.
     4. **Bridging Information Gaps**: A dynamic chat interface could adapt to user questions, filling the information voids that the current static web pages might not address explicitly.

# Data Assessment

* + Available Data Sources: Primary datasets include Jira ticket details provided by Professor Tim and the website content.

**a-1) Jira ticket details:** This file contains 140 variables but we will use mostly “Summary”, “Status”, ”Created (date)” and ”Description”. We will break down “Created (date)” to grasp on the stage and categorize the Jira dataset into thematic areas (e.g., application procedures, academic queries and so on). We can further analyze the “Description” variables using text analysis and do topic modeling to help with Chat Bot Categorization. Based on Jira ticket details, we can set up Natural Language Processing (NLP). Integrating NLP to allow our chatbot to understand user intent, even if their phrasing varies. When we use the Jira ticket’s description, we can train the chatbot with various utterances for a particular intent.

**“Description” Example:**

Hii This is ~+~+~+~+~+~ UID-U#######. In process of submitting DS160. I need to send address and phone number of point of contact(school official).So can you please send address and phone number of ~+~+~+~+~+~(School official to contact uon arrival).Can i also know the first and last name of ~+~+~+~+~+~.

**a-2) Canned Responses Integration:** We found that this dataset can give us the specific categorization (i.e., interest, prerequisite review, registration for first semester, etc). This file can be used to set the first conversation flow. This file is useful to draft possible user queries and determine how the Q&A application should respond based on categorizing the information available in these files. It is beneficial to create a decision tree or a flowchart to visualize the conversation paths.

**a-3) Website from MS in Business Analytics & Information System**  
(Link: <https://www.usf.edu/business/graduate/ms-bais/faq.aspx>)

We can use the frequently asked questions and answers. We can scrap this website and gather information for founding the Q&A application.

**a-4) Implement a Feedback Loop:** We can get more information from user’s feedback. Feedback Loop can allow users to provide feedback about their experience. This will help you continually refine and improve the chatbot’s responses and functionality.

* + Data Quality**:** Preliminary assessment suggests that while the Jira dataset is structured but requires feature engineering, website data requires scraping and feature engineering too.
  + Knowledge Base Assembly**:** Central to the Q&A application training would be the FAQs, application guidelines, canned responses data and recurrent Jira ticket queries data.

# Scope of Use

* + **MS BAIS admission and process will be addressed:**
    1. Resolving standard queries .
    2. Assisting in the application process.
  + **Potential Limitations:** An application might be restricted in addressing complex or unique queries outside its training scope. Additionally, it necessitates periodic updates to keep data current.

# Appendix

## The Q&A application Flow:

A diagram of a process flow chart

Description automatically generated

* **Knowledge Base Creation & Feature Engineering:**

The foundation of our Q&A system is the knowledge base. This is essentially a structured repository that houses all the information the system can potentially use to answer queries. To create this, we start with collecting datasets related to the MSBAIS program. These datasets undergo feature engineering, a process wherein raw data is transformed or enriched to make it more useful and appropriate for machine learning algorithms. This might include processes like data normalization, encoding, and the extraction of relevant features that make it easier for algorithms to identify patterns.

* **Chunking and Embedding:**

After feature engineering, we segment or 'chunk' the data into manageable pieces. This is especially important for larger documents where only a specific part might be relevant for answering a particular query. Each of these chunks is then processed using 'word embeddings'. Word embeddings are algorithms that convert text into numerical vectors, retaining the semantic meaning of words. For example, the words 'university' and 'college' might be closer in this numerical space due to their semantic similarity.

* **Vector Database Storage:**

Once we've obtained the vectors from word embeddings, these are stored in a specialized database. This vector database is optimized for high-speed retrieval, especially when we're looking for vectors that are similar to a given input vector.

* **Query Vectorization and Similarity Search:**

When a user poses a query, that query undergoes a similar transformation process. It's first vectorized using the same word embedding method as our knowledge source text. The resulting query vector represents the essence of the user's question. To find the most relevant answer, the system then searches the vector database for the most similar vectors (or chunks of knowledge). This similarity is often measured using metrics like cosine similarity, which determines the cosine of the angle between two vectors - the closer the cosine value is to 1, the more similar they are.

* **Text Summarization:**

Once a relevant chunk or segment of text from our knowledge base is identified, it might still be too lengthy or verbose to be a direct answer. So, the system employs summarization techniques, condensing the retrieved information to a more user-friendly size, while retaining the crux of the content.

* **Contextual Augmentation:**

This summarized text isn't just directly relayed back. Instead, it's used to enhance the original query by adding valuable context. By incorporating this context, the system can bridge the gaps between what the user asked and what the system knows.

* **Learning with the LLM:**

The term 'LLM' refers to a Language Model. With the augmented query now rich in context, the Language Model steps in. It processes this enhanced query, 'learning' from the added context. Even if the LLM didn't have a direct answer before, this context helps it generate a more informed, relevant, and qualitative response.

In essence, this flow ensures the Q&A system isn't just providing static, pre-defined answers. It dynamically processes queries, searches for relevant data, enhances the context, and uses a sophisticated language model to generate tailored responses.

## Interview Notes

Most things are done through the ticketing system "Jira"

Serves both current and prospective students

Automated, automatically assigns based on input

All tickets about 5000 of them sent to Dr. Smith

Workflow

applies then archives office of admission looks at it to see if it meets requirements then send to department

Department has data warehouse and chooses the newly referred students

The Referred students are sent an email requesting for some responses

Once responses are out, Dr. Barjaji then makes final decision

Then sent to Emmy who sends admit or deny.

In the admit email, the school lists all the requirements including visa

Register them for course (One problem is that the school never asks the student to make a commitment that they will come)

Want to waive prerequisite based on profiles

Some Data maybe confidential and we may have a challenge accessing this data

Most of the data is unstructured.