**ExSeedアウトプット講義**

**〜設計課題〜**

**氏名：髙橋彰仁**

**企画・目標の設定背景**

取引先の保険企業であるA社は、通常の業務でよく発生する質問や問題に対処するために、RAG機能がついたチャットボットを導入したいと考えている。ただし、A社は保険を取り扱う以上、扱う情報の正確性を重視したいと考えている。また、既にLLMにGPT-3.5を使用するという方針が定まっている。

**＜通常の業務でよく発生する質問や問題の対処の例＞**

RAG Chatbotを

1. **社内ナレッジベースとして使用：**

**社内手続きの案内：**

従業員が人事や総務に関する手続きを行う際、必要な書類やプロセスを簡単に検索できる。

**技術サポート：**

IT関連の問題やシステムの使い方について、リアルタイムで解決策を提供することができる。

**プロジェクト管理：**

プロジェクトに関連するドキュメントや進捗状況を素早く把握するために使用される。

**新入社員のオンボーディング：**

会社のポリシーや文化、業務フローについて学ぶための情報源として活用される。

1. **カスタマーサポートの強化として使用：**

**カスタマーサポートの自動化：**

FAQやよくある質問などの既存の知識を活用し、チャットボットが回答することで、人間のスタッフがより高度な対応に集中できる。

1. **業務効率化を目的として使用：**

**見積もりの自動化：**

お客様の年齢や性別、保険の種類や内容などの質問にチャットで回答してもらうことで、簡易的な見積もりを自動で出すことができる。

**マーケティング分析：**

マーケティング担当者が市場の動向や競合の戦略などを分析する際、RAG Chatbotが関連するレポートやデータを検索し、要約やインサイトを生成する。

**文書作成/要約：**

文書作成担当者がレポートやプレゼン資料などを作成する際、RAG Chatbotが参考になる文献や事例を検索し、要約して伝える、作文や原稿の添削を行う。

**企画・目標の設定**

**ユーザーゴール**

1.　必要な情報を迅速に取得し、業務を効率的に進める。

2.　顧客の問い合わせに対して素速くかつ正確な回答を提供する。

3.　業務プロセスを自動化し、時間と労力を節約する。

**ビジネスゴール**

1,3.従業員の業務プロセスを効率化させる。そして、企業のパフォーマンス向上を果たす。

2.　顧客満足度を向上させ、サポートスタッフの負荷を軽減させる。

**要件定義**

**機能要件**

・任意の文字列（RAG Chatbotへの指示文）を入力できる。

・社内データの検索結果を使用したGPT-3.5（RAG Chatbot）の生成結果が出力される。

・RAG Chatbotが応答に使用した情報を確認できるようにする。

・多言語に対応した入力と出力を行う。

・正確にベクトル検索を行う。

・社内データを管理する際の強いセキュリティ。

・ユーザー数やデータ量の増加に対応できるスケーラビリティ。

**非機能要件**

・システムが常に利用可能である（可用性）。

・大量のユーザーからの同時アクセスに対応できる高いパフォーマンス。

・直感的で使いやすいUI（ブラウザ版のChatGPTのようなUI）。

**内部設計　〜ユーザー視点の挙動/ベクトルデータサービスの選定〜**

**機能設計**

* **任意の文字列（RAGChatbotへの指示文）を入力できる。**
* ブラウザ版のChatGPTと同様に、指示文を書く入力欄、送信ボタンを用意する。送信した際は、自身のアイコンとその入力文をチャット形式で表示。
* 入力欄が空欄の場合は、送信できないようにする。
* Microsoft Copilotと同様に、入力文字数のカウント表示と制限を行う。
* **社内データの検索結果を使用したGPT-3.5の生成結果が出力される。**

RAGChatbotの生成については実装方法で述べる。出力文の表示については下記の通り。

* ブラウザ版のChatGPTと同様に、RAGChatbotのアイコンとその出力文をチャット形式で表示。
* 自身の入力文、RAGChatbotの出力文を一部選択コピー/一括コピーできる。
* RAGChatbotに再出力をリクエスト（最大2回）できる。
* **RAGChatbotが応答に使用した情報を確認できるようにする。**
* ボタンを用意する。そのボタンをクリックしたら、テキストボックスに使用した情報が表示される。

**ベクトルデータサービスの選定**

**以下はAzure AI Searchを使用することで実現できる。**

* 多言語に対応した入力と出力を行う。
* 正確にベクトル検索を行う。
* 社内データを管理する際の強いセキュリティ。
* ユーザー数やデータ量の増加に対応できるスケーラビリティ。

**選択理由**

　Chromadbは、リリースされてから日が浅いことから、サポート体制やドキュメンテーションの充実度が他2つより劣るなど、信頼性に欠ける。[[1]](#endnote-1)

Pineconeは、単体で、非ベクトルデータ（日本語、英語など）をデータベースに保存できない。[[2]](#endnote-2)

私は、検索性能、価格、実装のしやすさ、リスクの小ささ、シチュエーションに適しているかを鑑み、Azure AI Searchが、最も適したベクトルデータベースサービスだと考えた。[[3]](#endnote-3)

**Azure AI Searchの特徴**

* ベクトル検索とフルテキスト検索の両方を提供する。[[4]](#endnote-4)
* 柔軟な価格設定。[[5]](#endnote-5)

**Azure AI Searchの使い方**

[Azure AI Searchでのデータインポートとベクター化の実践ガイド #AdventCalendar2023 - Qiita](https://qiita.com/akiraokusawa/items/68cc2c6a5912ebe09013)

**画面設計**

画面は、ブラウザ版のChatGPTのUIに、Copilotの入力文字数のカウント表示と制限を加えたものにする。

テキスト, 手紙

自動的に生成された説明

四角形 が含まれている画像

自動的に生成された説明

**内部設計　〜具体的にどう実装するか〜**

**データ設計**

* **Q&Aデータベース**
* 通常の業務でよく発生する質問や問題を集め、それぞれの解答や解決策とともにデータベースを作成する。これにより、チャットボットはユーザーからの質問や問題をデータベースと照合し、適切な回答や解決策を提供することができる。
* **保険情報データベース**
* A社が取り扱う保険商品の詳細情報をデータベース化する。これにより、チャットボットは具体的な保険商品に関する質問に対して、正確な情報を提供することができる。

**モジュール設計**

* **GPT-3.5を使用するためのクラス**

GPT系モデルを使用するためのAPIキーを定義する。

* **システム設定を保存する/入力リストとログを初期化する関数（self,system\_setting:str) ->None**
* システム設定を引数として受け取り、それをクラスの内部変数に保存する。また、入力リストとログを初期化する。
* System\_settingには、モデルに対する全体的な指示文を格納する。
* **モデルの指定/ハイパーパラメータの設定/与えられた入力に対する出力を生成する関数（self,input\_text:str)->str**
* ユーザーからの入力を受け取り、指定したモデル（ここではGPT-3.5）でそれに対する応答を生成する。また、温度、最大出力トークン数、top\_pなどのハイパーパラメータなどをここで設定する。
* input\_textには、ユーザーからの入力文を格納する。
* **Azure AI Searchを使用するためのAPIキーと関数**

Azure AI SearchのエンドポイントとAPIキーを定義する。

* **Azureサービスに対する認証に使用するキー資格情報を保持する関数（api\_key:str)->AzureKeyCredential**
* **Azureサービスとの通信を行うための関数（endpoint:str, index\_name:str, credential:str)->SearchClient**
* **ユーザーからの入力を受け取り、それを検索クエリとしてAzure AI Searchで検索し、最上位の検索結果を返す関数（input\_text:str）->str**
* **GPT-3.5へのプロンプトを作成する関数(input\_text:str, search\_result:str)->str**
* ユーザーからの指示文と、最上位の検索結果をf文字列で1つのプロンプトにまとめる。

（例）f"{input\_text}　参照情報を用いて回答しなさい。　参照情報：{search\_result}”

* search\_resultには、最上位の検索結果を格納する。
* **GPT-3.5で応答を生成する関数（pronpt:str)->str**
* 作成したプロンプトを、入力に対する出力を生成する関数の引数に渡して、GPT-3.5に応答を生成させる。

1. [Releases · chroma-core/chroma (github.com)](https://github.com/chroma-core/chroma/releases) [↑](#endnote-ref-1)
2. <https://www.pinecone.io/learn/vector-database/> [↑](#endnote-ref-2)
3. [Azure AI Searchでのデータインポートとベクター化の実践ガイド #AdventCalendar2023 - Qiita](https://qiita.com/akiraokusawa/items/68cc2c6a5912ebe09013) [↑](#endnote-ref-3)
4. [Azure AI Search の概要 - Azure AI Search | Microsoft Learn](https://learn.microsoft.com/ja-jp/azure/search/search-what-is-azure-search) [↑](#endnote-ref-4)
5. [価格 - Azure AI 検索 | Microsoft Azure](https://azure.microsoft.com/ja-jp/pricing/details/search/) [↑](#endnote-ref-5)