RookieCoders

Document Finder

An efficient way to screen a large number of documents

https://github.com/shaikhmq20/RookieCoders_5

Understanding the Problem Statement

Provide an efficient way of searching for content within a single or a set of documents when we have been given a large set of documents.

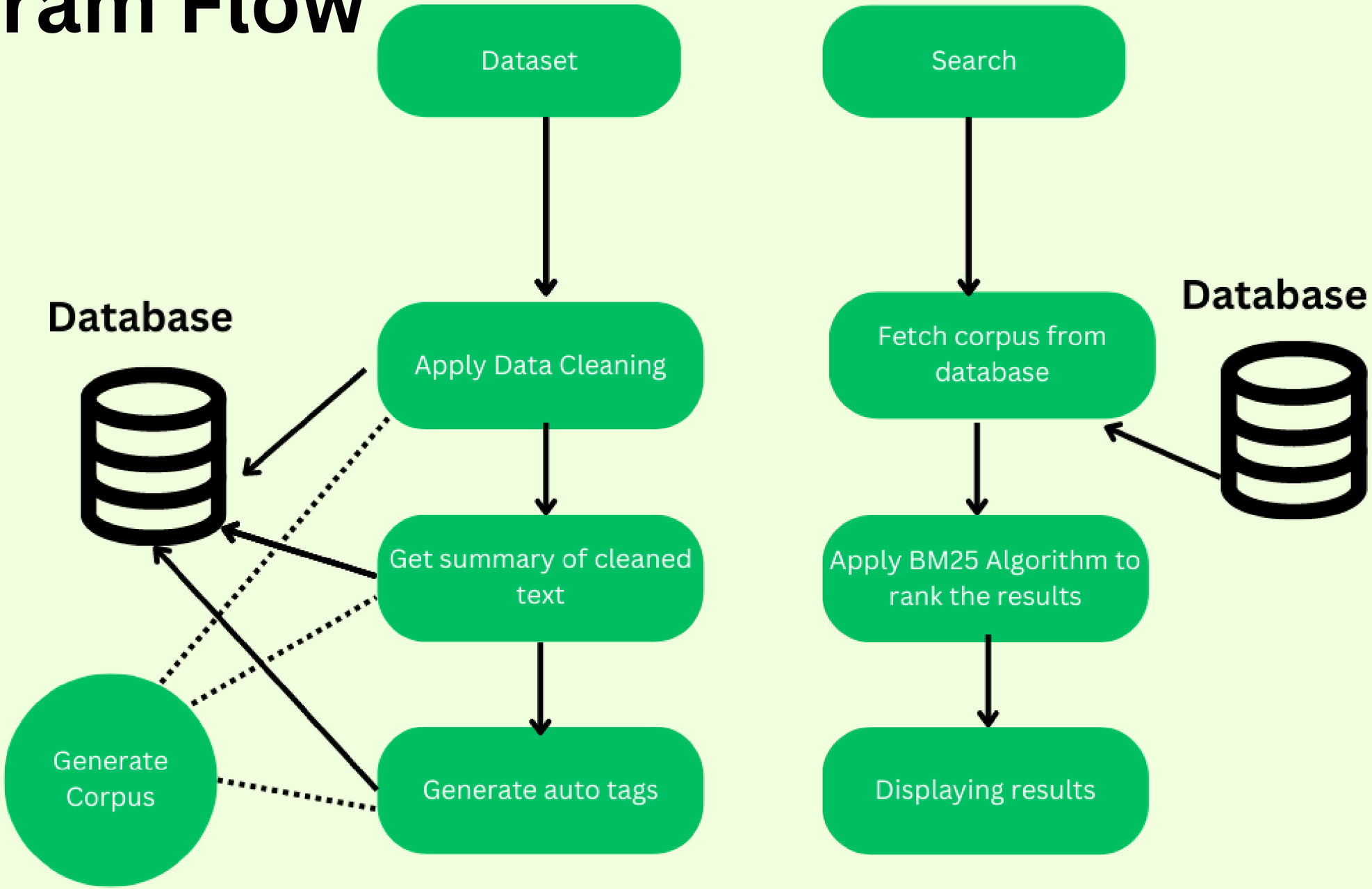
Proposed Solution

- The Solution makes use of the concept of natural language processing to understand what the end-user is trying to search for.

- It then extracts the relevant keywords and fetches for the keywords in a list of documents.

- The respective list of documents is then provided to the user.

Program Flow



POSSIBLE AREAS OF EXPANSION

The provision for uploading files.

The provision for downloading files which appear.

Show similar documents.

CHALLENGES FACED

Narrowing down the most appropriate tech stack for the implementation of the given feature.

The authentication of azure bot to create a web-chat based search engine.

Integration of firebase and Algolia library. The issue was with the authentication and accessibility of API keys .

Key Learning Outcomes

Technical

- We learnt how to create an azure account, get started with azure bot service, add a service to the resource group and deploying the service for a basic echo bot.

- Tried our hands at different document-indexing based libraries in Javascript as well as python like algolia, lucene and gensim.

Key Learning Outcomes

Non-Technical

- Patiently and meticulously going through various solutions to a given problem and choosing the most viable option.

Tech-stack used

Python	Flask
NLTK	JavaScript
SQLite	NLP Libraries

References

https://www.nltk.org/book/ch05.html

https://www.geeksforgeeks.org/nlp-gensim-tutorial-complete-guide-for-beginners/

https://medium.com/analytics-vidhya/simple-text-summarization-using-nltk-eedc36ebaaf8

https://www.youtube.com/@prettyprinted/videos