CS 410 Project Proposal

Created	@October 26, 2023 5:43 PM

A Different Approach to Notion Search

Team Members

- Michael Sun (ms48@illinois.edu) (Captain)
- Maxwell Wang (mlwang3@illinois.edu)

Free Topic Description

We are aiming to create a different approach to Notion Search that will hopefully better serve our needs. This project was motivated by our own dissatisfaction with the default implementation of Notion search, which uses direct keyword matching and frequently yields irrelevant search results with low recall. It's not just us either, Notion search is one of the top complaints for users, and with millions of users using search everyday, even a slightly improved search function would save massive amounts of time and frustration for users.

We would like to approach this problem by implementing the VSM similar to the one discussed in lectures, but with Notion specific weighting functions. We believe that Notion is perfectly suited for this because text data is conveniently split into cells of various types, and we can exploit how users typically use these cells to improve the ranking function of our documents. For example, we may give greater weight to terms that appear in the "callout" cell shown below:

> ❗ Terms in this cell are likely to be more representative of the content in a given page when compared to regular text.

Our tentative plan to implement this search tool is:

1. Interface with Notion API to retrieve pages and turn them into vectors

2. Create an inverted index of documents and words
3. Use BM25 and IDF to weigh words and rank documents
4. Discuss, implement, and test Notion-specific weighting heuristics

We plan to use **Python** to implement this, and we believe that the implementation of this task will take at least 40 hours since we will need to interface with an API that is foreign to both of us, and we will need to combine all the features in the state of the art retrieval functions discussed in class with Notion specific features.

The expected outcome of this project is a search tool that will in a set of Notion pages and a query, and return a ranking of relevant documents according to our VSM model. Since our data set is limited to the set of documents in our personal Notion boards, the evaluation will be mostly our subjective opinion of relevance of retrieved documents.