Executive Summary

Management consultants generate a large number of documents when working with a client. They need to be able to rapidly access subsections of these documents that are relevant to a particular area of their consulting. One approach is to tag sections of the document with keywords for quick information retrieval. This is usually done manually, which is slow and resource expensive.

This project develops an A.I. to automate the division of human-written business documents – which include diagrams and tables – into meaningful chunks, and then assigns keywords to each chunk. The user can provide a pre-defined set of keywords in a chunk/keyword training set CSV.

The complexity of automatically reading PDFs, spreadsheets, documents, and slides requires a sophisticated understanding of language and vision in parallel. OpenAI recently released a vision/text parallel understanding system called GPT-4-Vision, utilised in this project to 'read' documents, including their tables and figures, and divide them into meaningful chunks for tagging/keyword extraction. OpenAI also has a language system, GPT-3.5-Turbo, trained in this project based on chunk/keyword examples from Discy and then used to extract keywords from the chunks generated by GPT-4-Vision. Given OpenAI's privacy statements on their API, Discy is comfortable with the usage of the API.

The chunking by GPT-4-Vision is compared manually to more traditional chunking methods and is found to generate much more meaningful chunks. It is also capable of creating text chunks that describe diagrams and tables. The GPT-3.5-Turbo keyword extraction technique scores provide an accuracy score, known as F1, of 81% on the original Discy data (120 samples). Considering there are 22 tags/keywords and only 120 samples of data, further improvement is likely possible.