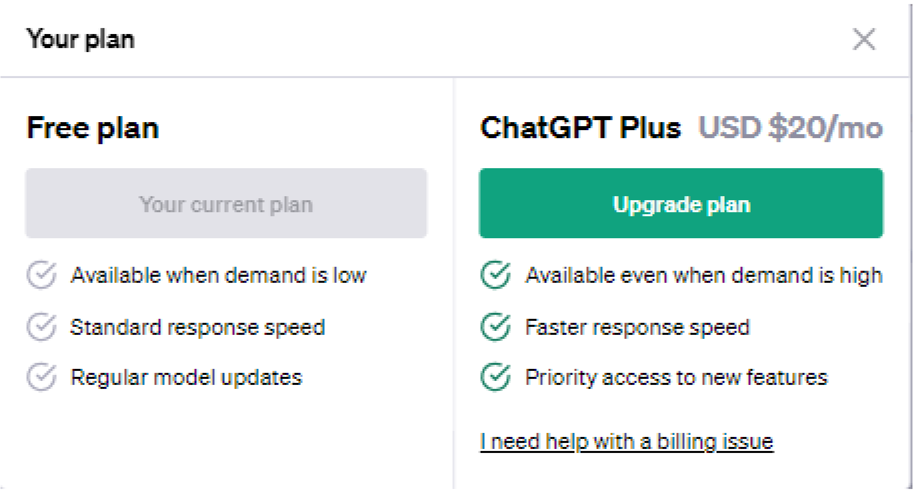
Research Document

Overview:

Our goal is to extract information from various document types such as invoices, NDAs (Non-Disclosure Agreements), and contracts. The purpose of this document is to look into how we can extract this information.

ChatGPT for Extracting Data:

As things stand ChatGPT can only take in textual inputs, what this means is that it can only ingest text-based data and not any other format such as PDFs or images. [1] Behind the scenes, ChatGPT is using the GPT-3.5 which is again a model only capable of analyzing text. However, recently a new model has been launched GPT-4, which is only available to ChatGPT Plus users (see image attached below for pricing details). However, even with GPT-4 images are only allowed to be uploaded via the User Interface (UI) and not the API. The API currently has a long waiting list (you can sign up here: https://share.hsforms.com/1u4goaXwDRKC9-x9IvKno0A4sk30). I signed up a few months ago, but am yet to hear back from them. [2]



Your plan

Free plan

Your current plan

- ✓ Available when demand is low
- ✓ Standard response speed
- ✓ Regular model updates

ChatGPT Plus USD $20/mo

Upgrade plan

- ✓ Available even when demand is high
- ✓ Faster response speed
- ✓ Priority access to new features

I need help with a billing issue

Other Alternatives:

There are many specialized services that can analyze documents such as invoices, NDAs, and contracts. The most popular services that are generally used to extract data from documents are:

1. Amazon Textract
2. Google Document AI

While there are other tools and services that can extract data from documents the above 2 services are developed by Amazon and Google respectively, as such they are the best solutions in terms of reliability, efficiency, and accuracy. In the next section, I will be diving deeper into how Amazon Textract and Google Cloud Document AI actually work under the hood.

Amazon Textract:

Amazon Textract is a cloud-based service provided by Amazon Web Services (AWS) that enables automated text and data extraction from scanned documents, PDF files, and images. It uses advanced machine learning algorithms and optical character recognition (OCR) techniques to analyze and extract structured data from unstructured documents.

The underlying algorithms and techniques used by Amazon Textract can be divided into several key steps:

1. Document Analysis: Textract begins by breaking down the input document into individual pages and identifying key structural elements such as tables, forms, and lines of text. It uses computer vision algorithms to analyze the layout and structure of the document, identifying blocks of text, tables, and other elements.

2. Optical Character Recognition (OCR): Textract applies OCR techniques to recognize and extract the text from the document images. It uses deep learning models to analyze the individual characters and words within the document. The OCR algorithms are trained on a vast amount of data to accurately recognize characters and handle variations in fonts, styles, and languages.

3. Text and Data Extraction: Once the text is extracted from the document, Textract performs additional analysis to identify and extract specific types of data. It can recognize and extract data such as form fields, key-value pairs, tables, and checkboxes. Textract uses a combination of machine learning and rule-based techniques to identify and extract structured data from unstructured documents.

4. Document Structure Analysis: Textract also analyzes the structure of the document to identify hierarchical relationships between different elements such as headings, paragraphs, and sections. This information can be useful for organizing and understanding the document's content.

5. Advanced Features: Textract includes additional advanced features to enhance the extraction process. For example, it can handle documents with mixed content types, such as documents that contain both text and images. It can also process documents with multiple pages and extract data from tables with complex structures.

It's important to note that the exact algorithms and techniques used by Amazon Textract are proprietary to Amazon and not publicly disclosed. The service is continuously improved and updated by Amazon based on feedback and ongoing research in the field of document analysis and OCR.

Overall, Amazon Textract leverages machine learning, OCR, computer vision, and document analysis techniques to automate the extraction of text and structured data from various types of documents, enabling developers to build applications that can extract and process information from unstructured sources more efficiently. [3]

Google Cloud Document AI:

Google Cloud Document AI is a cloud-based service provided by Google Cloud that focuses on document processing and analysis. It leverages a combination of machine learning algorithms, OCR techniques, and natural language understanding to extract structured data from unstructured documents such as scanned images, PDFs, and other file formats.

Here's an overview of the algorithms and techniques used by Google Cloud Document AI:

1. Document Parsing: Document AI starts by parsing the input document into individual pages and segments, identifying various document elements such as paragraphs, tables, images, and form fields. It utilizes computer vision algorithms to analyze the document layout, structure, and textual content.

2. Optical Character Recognition (OCR): Document AI employs OCR techniques to recognize and extract text from document images. It applies deep learning models that have been trained on a vast amount of data to accurately recognize characters, words, and other textual elements. OCR enables the transformation of scanned or photographed documents into machine-readable text.

3. Natural Language Understanding (NLU): Once the text is extracted from the documents, Document AI utilizes natural language processing techniques to understand the semantic meaning of the text. It can perform tasks such as entity recognition, and language detection

4. Form Parsing and Extraction: Document AI specializes in parsing and extracting data from structured forms such as invoices, receipts, and applications. It uses machine learning models trained specifically for form understanding to identify key-value pairs, checkboxes, and other form fields. This enables the automated extraction of relevant information from documents that follow predefined templates or formats.

5. Table Extraction: Document AI includes algorithms for table extraction, enabling the identification and extraction of tabular data from documents. It can handle complex table structures and extract structured information such as rows, columns, and headers. Table extraction is particularly useful for processing documents with structured data organized in tabular form, such as financial reports or research papers.

6. Document Classification: Document AI can classify documents into predefined categories based on their content. It uses machine learning models to analyze the textual information and classify documents into relevant classes or categories. This can be helpful for automating document organization and routing based on their content.

It's important to note that the specific algorithms and techniques employed by Google Cloud Document AI are proprietary and not publicly disclosed. Google continuously improves and updates the service based on research advancements and user feedback in the domain of document processing and analysis.

In summary, Google Cloud Document AI combines machine learning, OCR, computer vision, and natural language understanding techniques to automate the extraction of structured data from unstructured documents. By leveraging these technologies, developers can build applications that perform advanced document analysis, enabling efficient data extraction and understanding from diverse document sources. [4]

Amazon Textract or Google Cloud Document AI?

Both Amazon Textract and Google Cloud Document AI offer similar capabilities for document processing and data extraction. The choice between them may depend on factors such as pricing, specific requirements, existing infrastructure, and the overall ecosystem of services offered by AWS or Google Cloud.

However, for the purpose of this project, I have decided to go with Amazon Textract for the following reasons:

- Amazon Textract has been available for a longer period and has gained significant traction in the industry, providing a more mature and battle-tested solution.
- Most organizations already utilize Amazon Web Services (AWS), so integrating Textract into the existing infrastructure is seamless.
- Textract's ability to handle various types of documents, including invoices, contracts, and forms, combined with its support for multiple languages and the extraction of data from handwritten text, makes it a versatile choice
- AWS offers a broad range of complementary services, such as storage, analytics, and machine learning, which can enhance document processing workflows.
- Most importantly, I am going with Amazon Textract because it is a tool that I am familiar with as I have used their API in several projects and I am comfortable working with it.

High-Level Code Design:

The 2 APIs of Amazon Textract that are useful for us are:

1. AnalyzeExpense: This will be used primarily for extracting data from invoices.

   The AnalyzeExpense API of Amazon Textract is designed specifically for processing expense-related documents, such as invoices, receipts, and bills. It enables the extraction of key information from these documents, such as line items, dates, totals, vendor details, and currencies. Here's a step-by-step breakdown of how the AnalyzeExpense API works:
   Input Document: You provide the API with an input document, which can be an image file (JPEG or PNG) or a PDF file. The document should be in a format typically associated with expense documents.

- Document Processing: Amazon Textract performs an initial processing step on the input document. It uses OCR techniques to extract the text from the document, including the textual content of tables.

- Text Extraction: The API analyzes the extracted text to identify and extract relevant information. It looks for key elements such as line items, dates, vendor names, currencies, and totals within the document.

- Entity Recognition: Amazon Textract employs machine learning models to recognize and extract specific entities from the document. For example, it can identify and extract individual line items, their descriptions, quantities, unit prices, and extended prices.
- Key-Value Pair Extraction: The API also identifies key-value pairs within the document. It recognizes labels (keys) associated with their corresponding values, such as "Total" with its corresponding value.
- Structure and Formatting: Amazon Textract considers the structure and formatting of the document to enhance the extraction process. It understands the organization of information within tables, identifies headers and rows, and associates data accordingly.
- Output: The AnalyzeExpense API provides the extracted information in a structured format, such as JSON. It includes information about line items, vendor details, dates, totals, and other relevant data found within the document.
  (Read more detail on how the API works at https://docs.aws.amazon.com/textract/latest/dg/API_AnalyzeExpense.html) [5]

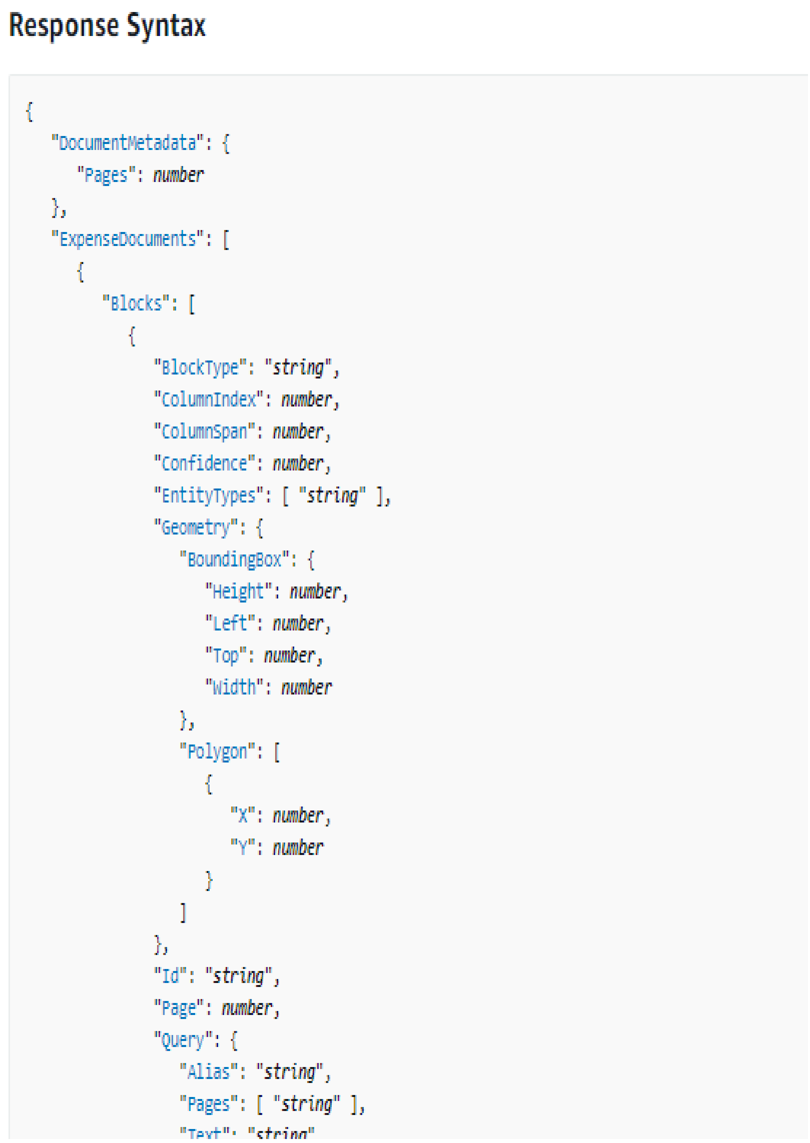
2. AnalyzeDocument: This will be used primarily for extracting data from contracts and NDAs.

   The AnalyzeDocument API of Amazon Textract is a more generalized service that can be used for analyzing various types of documents. It provides comprehensive document understanding capabilities, allowing you to extract structured data and perform advanced analysis. Here's a detailed explanation of how the AnalyzeDocument API works:
   Input Document: You provide the API with an input document, which can be an image file (JPEG or PNG) or a PDF file. The document can contain various types of content, including text, images, tables, and forms.

- Preprocessing: Amazon Textract preprocesses the document by extracting the text and analyzing the layout and structure. It recognizes key elements such as paragraphs, lines, tables, and form fields.
- OCR and Text Extraction: The API utilizes OCR techniques to recognize and extract the textual content from the document. It applies deep learning models to accurately recognize characters, words, and other textual elements. The extracted text serves as the foundation for subsequent analysis.
- Document Structure Analysis: Amazon Textract analyzes the structure of the document, identifying hierarchical relationships between different elements such as headings, paragraphs, and sections. This information helps in understanding the organization and layout of the document.
- Key-Value Pair Extraction: The API identifies key-value pairs within the document, recognizing labels (keys) associated with their corresponding values. It can extract data such as field labels and their corresponding values, making it useful for processing forms and structured documents.
- Table Extraction: Amazon Textract employs table extraction algorithms to identify and extract tabular data from documents. It recognizes rows, columns, and headers, allowing for the extraction of structured information organized in tabular form.

- Advanced Analysis: The AnalyzeDocument API also provides additional analysis capabilities, including entity recognition, and content classification. It enables the identification of entities such as names, addresses, and dates associated with specific portions of the document.
- Output: The extracted information is provided in a structured format, typically JSON. It includes data such as text, tables, key-value pairs, and other relevant elements found within the document.
  (Read more detail on how the API works at https://docs.aws.amazon.com/textract/latest/dg/API_AnalyzeDocument.html) [6]

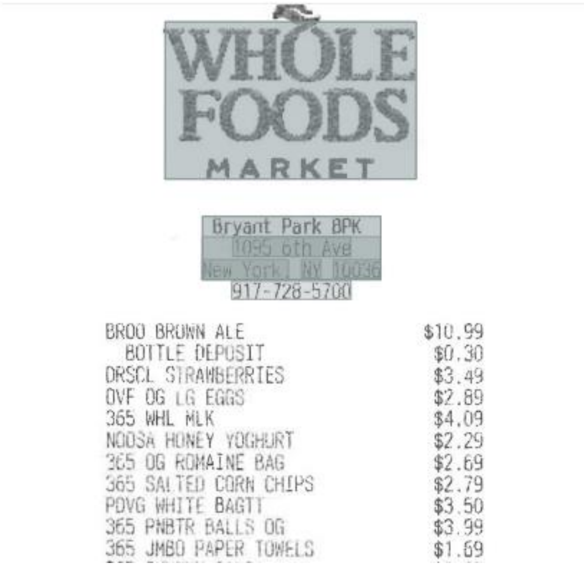
When accessing through the API this is what the response object (extracted data) would look like:



Response Syntax

```
{
   "DocumentMetadata": {
      "Pages": number
   },
   "ExpenseDocuments": [
      {
         "Blocks": [
            {
               "BlockType": "string",
               "ColumnIndex": number,
               "ColumnSpan": number,
               "Confidence": number,
               "EntityTypes": [ "string" ],
               "Geometry": {
                  "BoundingBox": {
                     "Height": number,
                     "Left": number,
                     "Top": number,
                     "Width": number
                  },
                  "Polygon": [
                     {
                        "X": number,
                        "Y": number
                     }
                  ]
               },
               "Id": "string",
               "Page": number,
               "Query": {
                  "Alias": "string",
                  "Pages": [ "string" ],
                  "Text": "string"
```

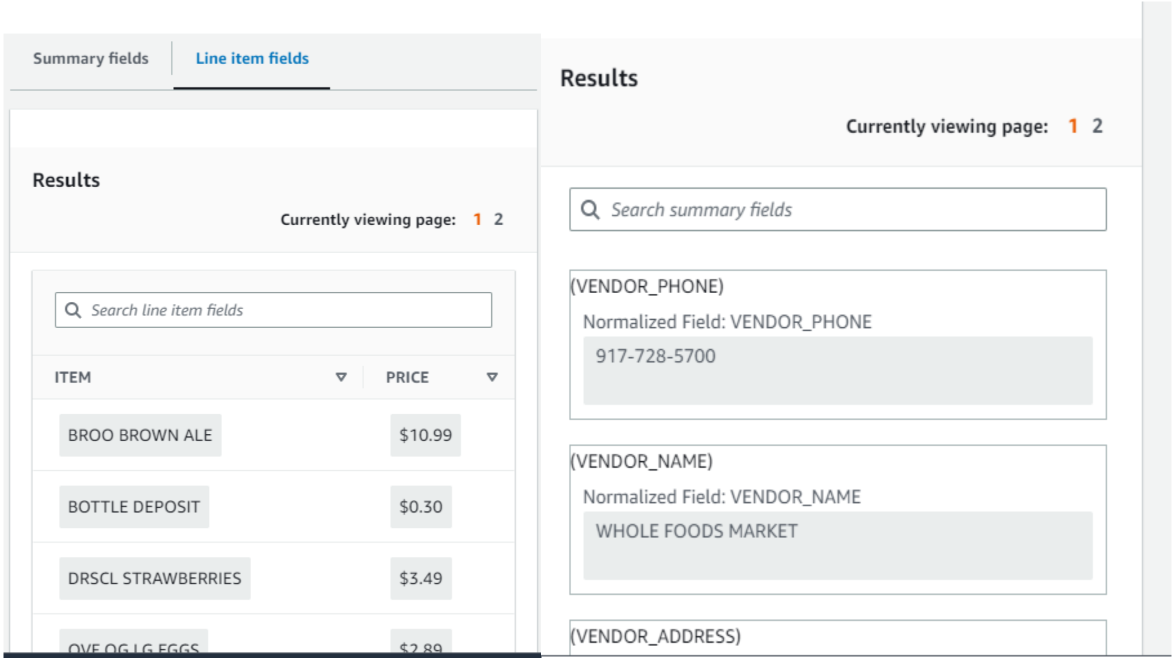
This is a sample receipt:


WHOLE FOODS MARKET

Bryant Park BPK
1095 6th Ave
New York, NY 10036
917-728-5700

BROO BROWN ALE	$10.99
BOTTLE DEPOSIT	$0.30
DRSCL STRAWBERRIES	$3.49
OVF OG LG EGGS	$2.89
365 WHL MLK	$4.09
NOOSA HONEY YOGHURT	$2.29
365 OG ROMAINE BAG	$2.69
365 SALTED CORN CHIPS	$2.79
PDVG WHITE BAGTT	$3.50
365 PNBTR BALLS OG	$3.99
365 JMBO PAPER TOWELS	$1.69

And this is the information extracted from this receipt through Amazon Textract:



Results

Currently viewing page: 1 2

Search line item fields

ITEM	PRICE
BROO BROWN ALE	$10.99
BOTTLE DEPOSIT	$0.30
DRSCL STRAWBERRIES	$3.49
OVF OG LG EGGS	$2.89

Results

Currently viewing page: 1 2

Search summary fields

(VENDOR_PHONE)
Normalized Field: VENDOR_PHONE
917-728-5700

(VENDOR_NAME)
Normalized Field: VENDOR_NAME
WHOLE FOODS MARKET

(VENDOR_ADDRESS)

[7]

It's important to note that while the high-level steps and functionalities are outlined here, the specific algorithms and techniques used by Amazon Textract are proprietary and not publicly disclosed.

Where does ChatGPT fit into this?

While ChatGPT cannot be used to extract data from documents due to the limitations described above, we can certainly use it to answer questions about the extracted data. For example, once the data has been extracted we can pass it to ChatGPT via OpenAI's APIs and ask it questions related to that data.

Bibliography

1. GPT-4 Technical Report
   https://arxiv.org/abs/2303.08774
2. How to access GPT4?
   https://help.openai.com/en/articles/7102672-how-can-i-access-gpt-4
3. AWS Textract Documentation (This is the home page for all the documentation)
   https://aws.amazon.com/textract/
4. Google Cloud Document AI Documentation (This is the home page for all the documentation)
   https://cloud.google.com/document-ai#section-5
5. AnalyzeExpense API Documentation
   https://docs.aws.amazon.com/textract/latest/dg/API_AnalyzeExpense.html
6. AnalyzeDocument API Documentation
   https://docs.aws.amazon.com/textract/latest/dg/API_AnalyzeDocument.html
7. Code Demo (will not be accessible, associated with my account)
   https://us-west-1.console.aws.amazon.com/textract/home?region=us-west-1#/analyzeexpense