imas
INVESTMENT MANAGEMENT
ASSOCIATION OF SINGAPORE

AUTOMATED DATA PARSING, EXTRACTION AND ANALYSIS TOOL FOR ASSET MANAGEMENT

CONTEXT

Schroders is an international asset and investment management company with over 200 legal entities globally, offering a wide range of asset classes in diverse geographies. With their global footprint and diversified business model, Schroders seeks to deliver long-term value for its clients and shareholders.

To grow their clients' finances and portfolio, asset managers from Schroders play a key role in making well-timed investment decisions, understanding the clients' financial position and risk, to best determine potential opportunities and investment options. To do so, the financial analyst reviews content-heavy financial and Environmental, Social and Governance (ESG) documents from their clients, conduct their due diligence checks and present key insights for internal analysis and reports. Documents can include financial statements, sustainability reports, and other 3rd party data sources; most of these data and information lack standardisation across different documents, companies and industries.

In processing such high volume of unstandardised documents at a crunch time, Schroders' financial analysts face three key challenges:

1. **Extracting Relevant Data Across Different Data Sources** – Currently, Schroders employs generic extraction tools to support financial analysts to extract the relevant data from the clients' documents. However, these data parsing solutions are often not robust enough to identify and extract the essential information from the various document formats, skewing the accuracy of the numbers and figures extracted. Such inaccuracy is a major pain during analysis.

2. **Understanding and Contextualising Documents –** Besides extracting key data and figures, the accompanying analysis in the reports are also highly subjective. Hence, while standard information such as revenue and profit can be found consistently across financial statements from different companies, detailed information such as market segmentation, ESG reporting are highly dependent on how different companies choose to report them. As such, the extraction of essential and relevant information/data is a task that still requires a certain level of subject matter expertise to better understand and interpret. Analysts need to manually comb through the materials to integrate the information for their internal analysis and reporting.

3. **Ensuring Accuracy and Compliance to Legal Regulations** – When interfacing with legal documents, such as commercial agreements and contracts, constant review against changes made by regulators is necessary, to ensure its compliance and relevancy (e.g. LIBOR transition), and to trigger renegotiations when required. These

legal documents are highly specialized in their language and terminology, which cannot be adequately served by current NLP libraries (e.g. sPacy).

While Schroders has tried various RPA and data parsing solutions, the dynamicity of the documents is still a challenge, and this remains mostly a manual data extraction and analysis process. With today's direct access to digital documents and databases, there is opportunity to develop an industry specific data parsing and extraction solution to automate the extraction and analysis of the required data accurately and reliably, so that data can be leveraged and maintained in a manner that can be easily consumed by downstream processes. This would allow the data to be more effectively leveraged for different layers of consumptions across various services and solutions, providing true value to the users in Schroders.

PROBLEM STATEMENT

How might we create a smart document parser to streamline and enhance the process of information extraction and analysis for the financial analysts?

WHAT ARE WE LOOKING FOR?

A prototype seamless and user-friendly data parsing and extraction solution specific for the asset management industry that is able to perform most of the following functions:

- **Intelligent Data Extraction:** To process and extract data and information from structured and unstructured documents that have different formats and layouts accurately, such as:
    - Identification and extraction of data from tables that span across multiple pages of the document, tables with multi-column layouts (e.g. 2-column page layout), or from graphs (image-based data extraction)
    - Differentiate image-based PDFs from digital PDFs upon upload, so that digital PDFs would not go through the OCR data extraction process and to ensure 100% data extraction accuracy, regardless of the type of PDF file
    - Flag out potential data extraction and formatting errors when processing a document with bespoke designs and layouts (edge cases), so that users at Schroders can review the documents and follow-up accordingly
- **NLP-based Financial Topic Analysis:** A financial industry-specific natural language processing engine that can process complex queries so that it can assist users in the financial industry by providing users with insights (e.g. relevance, key figures) based on the data and information available in the source documents. It would also be good if the solution could summarise complex documents into bite-sized abstracts (e.g. under 150 words).
- **Intelligent Data Contextualisation:** To contextualise relevant information and content that are extracted, while identifying and highlighting relevant edge case information for users to review and analyse. (For example, the solution is able to extract key information from ESG reports and highlight the volunteerism under CSR.) This would significantly reduce the time needed for users to look for and process relevant information and streamline the overall process of understanding the client's financial

position.

- **Centralised Document Repository:** To allow for the storage of documents (and the extracted data and information) based on clients and/or projects that allows easy and intuitive uploading and accessing of documents for all users. If new documents (within the same project and/or client) with amendments are uploaded, the NLP engine should process the amendments and highlight past/archive documents that may be affected, which would then notify the user to review and follow-up

- **Feedback Loop:** To allow users to review, edit, annotate and comment on documents NLP engine that generates insights, summaries and sentiment analysis must improve in accuracy and efficiency over time.

- **Smart Search Capabilities:** To assist users with accurately and conveniently searching for key information across multiple documents

Overall, the solution must reduce manpower, time and resources needed for the data extraction and data review processes, enhance the productivity of users through improved human-machine collaboration, and improve the quality of reports produced by financial analysts at Schroders.

Schroders will work closely with the selected solution provider to develop a financial industry-specific NLP engine that can manage customised and complex financial industry- specific queries, and identify to identify the data fields and parameters that need to be identified and extracted from documents

There are no restrictions on the geographical location of the problem solvers who may choose to apply to this challenge. However, the prototype must be demonstrated in Singapore.

POSSIBLE USE CASES

1. **Data Extraction**: Adam, a Financial Analyst at Schroders, needs to review several documents for Company X to create bespoke customisations and develop financial models for Company X. As such documents often varies in formats, Adam usually has to manually comb through the documents to extract the relevant data which often takes up to hours. With the new data parsing solution, he is able to instantly upload the various documents received form the client. The solution is able to process the documents with different formats and layout, and automatically identifies and extracts the relevant information and data with high accuracy. All the documents that were uploaded by Adam, and the extracted data and information, can be found in a document repository that is easily accessible and searched. This allows Adam's colleagues to search for the relevant information with ease during other downstream processes

2. **Documentation Understanding**: Rachel, a Financial Analyst at Schroders, is currently reviewing the ESG and sustainability reports from a client. As these documents are very content-heavy and the accompanying analysis subjective, she usually takes 1 to 2 hours just to process 1 document. However, with the new financial industry-specific data extraction solution that has NLP capabilities, she can simply upload these documents, and the solution would process the content found in the documents to highlight key information that would prove useful for Rachel's analysis. In addition, the solution allows Rachel to submit complex search queries, and allows her to accurately and convenient search for and extract key

information across multiple documents, thus streamlining her workflows and increasing her efficiency.

WHAT'S IN IT FOR YOU

- SGD 20,000 of prize money for each winner of this challenge (see Award Model)
- Gain access to IMDA's Technology resources and facility for prototyping
- Collaborate with IMAS (Investment Management Association of Singapore) to reach out to the greater community for exposure, refinement and deployment
- Opportunity to pitch to industry audience in IMAS Digital Events
  (For more information, visit www.imasdap.com)

EVALUATION CRITERIA

The Applicants shall be evaluated in accordance with the evaluation criteria set out below.

Solution Fit	• To what extent does the proposed solution address the problem statement effectively?
Solution Readiness	• How ready is the proposed solution to go to the market? • Is there any evidence to suggest capacity to scale?
Solution Advantage	• Is the solution cost effective and truly innovative? • Does it make use of new technologies in the market, and can it potentially generate new IP? • What are the top 3 key benefits will the solution bring to the Asset Management industry? Optional • To share estimated cost for pilot trial, deployment and software support.
Company Profile	• Does the product have user and revenue traction? • Do the team members possess strong scientific/technical background and when is the company founded? • Does the team have relevant clients/ use cases? • Does the team have plans to grow and propagate the solution in Singapore?

AWARD MODEL

30% of the prize money will be awarded to each selected finalist at the start of the POC/prototype development process, with the remainder 70% to be awarded after completion of the POC/prototype solution, based on milestones agreed between Problem Owner(s) and the solver. Prize money will be inclusive of any applicable taxes and duties that any of the parties may incur.

Note that a finalist who is selected to undertake the prototype development process will be required to:

- Enter into an agreement with Problem Owner(s) that will include more detailed conditions pertaining to the prototype development;
- Complete an application form with IMDA that will require more financial and other related documents for the co-funding support.

Teams with public research performers are required to seek an endorsement from their respective innovation and enterprise office, and submit the attached IEO form together with the proposal.

DEADLINE

All submissions must be made by **19th March 2021, 1600 hours (SGT/GMT +8)**. Problem Owner(s) and IMDA may extend the deadline of the submission at their discretion. Late submissions on the OIP, or submissions via GeBIZ, will not be considered.

FAQ

Question: Could you talk about Schroders track record of working with startups? E.g., I know that last year Schroders was involved in the OIP as well. Did that exercise a success?

Answer: Yes, the project was a success. The project is near completion and they have recently presented their project in IMAS Digital Summit 2021.

Question: Could you please name the aspects of spaCy that made you look for a different solution? Is that the special types of entities you need to extract, terminology customisation, or something else?

Answer: I mentioned spacy, it was in used in our LIBOR use case because spacy is NLP and we used it to help us leverage the lemmatization. For example, when we are working with our legal colleagues and they gave us the search terms, it included terms like act of God, act of Gods – singular and plurals. So, we leverage libraries like spaCy to handle the variations or lemmatizations of the word rather than, having to type them all out. But spaCy is for NLP – for the 4 documents types that I have shown, the challenge is identifying the relationship between the figures and the roles itself. Having an accurate extraction is actually the challenge, it’s not so much on NLP side.

Question: What languages do you need to support?

Answer: Let us stick to English, that is where all the information is coming from, that is where it is being processed. It covers 80-90 plus percent.

Question: Is there an internal repository with source documents to be assumed tagged for each client/account? How standardized (document file types like pdf.) are these?

Answer: The likes of the data providers from the Bloomberg's, FactSets, Morning Star, Refinitiv,, they do have financial information and data but where we see the challenge is actually in terms of emerging market or growth markets as well – that's where the data are sparser, the coverage is lesser but that's where a lot of activities are because our teams do need to analyse and understand the performance or even the modelling of the data from such countries. By saying tagging from it, where I think you are trying to relate to, would be more towards to machine learning – to train the models against a specific set. Where I see this is that it is going to be challenging in terms of finding out all the source to really tag it out – because our team, while we have covered Singapore, Indonesian, London, and Zurich, we are a very small team. One key thing which is why we leverage the OIP and IMAS partnership is that we do not have the bandwidth to research and build up the product on our own. That is why we are doing very selected business use cases to solve for the business challenge based on whichever client that send us more documents – it is a bigger load for the business, that is when we will go in and solve it. There is a benefit analysis that we work with in order to identify which processes and which are the things we should solve for. Where I have heard of, or at least discussions with several vendors as well – not going to name them, I think if you go down the ML route, it is possible but it's going to be a massive effort as well. I think I have a certain reservation for it because it is a niche, we are looking for a much more narrow or dedicated AI solutions if we really want to say in that way.

Question: Are you open to proprietary taxonomy / term management solutions?

Answer: That is fine. I came from a vendor space previously. Our team, our main focus is delivery the solution for the business. It does not have to be something that we build, we have leverage (inaudible), new algorithm, or research results that are public from Github etc. What I would go is always how well does the solution work, can it fit our business use case – that means if I send you a document, we can give you certain parameters, can it extract the information, is it accurate, and from the information extracted, can we then use it for downstream processes? That is why I would rather draw a blue box, or a black box would work as well, but I think everyone should understand it.

Question: What are the current sustainability standards Schroders are compiling with? (e.g., GRI, SASB)?

Answer: I need some time to check on this but let us not restrict the participants based on this.

Question: In terms of IP that could be developed to answer this challenge, would Schroders be willing to own the algorithms being developed or would the remain the propriety of the Start-up?

Answer: This will be subject to further discussions.

Question: Will you provide samples to train models for text extraction?

Answer: Yes. A selected set of samples and directions can be provided as guidelines

Question: Since you mentioned LIBOR as a use case, wondering if you will also be looking to address the challenge related to extraction of LIBOR fallback information too?

Answer: No. The LIBOR use case has been completed. But if there is an existing solution that can be shared we would be open to hear about it

Question: The documents are electronic PDF or scanned PDF? (since electronic PDF we don't need to do OCR)

Answer: Let us focus on electronic PDF for now.

Question: Are you open to proprietary taxonomy / term management solutions?

Answer: Yes.

Question: Are you looking at an app or web platform for your solution?

Answer: No restrictions as the data and information would be used in further downstream processes.

Question: Will you provide samples to train models for text extraction?

Answer: Sample documents are available but they are not tagged as BAU activity is focused on extraction rather than tagging of the document

Question: May I know who will be the main user of your platform? Sustainability compliance team or operation team?

Answer: The Operations Innovation team would front it to feed it to downstream processes

Question: Is the focus to just extract and store in database? Not in scope: a. Information retrieval: index, search sort, etc; Inference: b. Summarisation, paraphrasing models in scope

Answer: The storage is for ease of reference. If there are additional processing to enrich the information retrieval that would depend on its relevance to the information

Question: How many different formats for PDFs needs to be supported?

Answer: Let us focus on the general PDF standards. But if its externally produced PDFs the format might change (e.g. PDF/A, PDF/E) but that can be kept for later stages.

Question: What are the main document types to be process? Is it solely on financial statement type of data?

Answer: The financial document varies but financial statement is a key example in this case.