Capstone Project Proposal

Antonina Savka

Business Goals

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML/AI in solving this task? Be as specific as you can when describing how ML/AI can provide value. For example, if you're labeling images, how will this help the business?	**What industry problem are we solving?** This project solves the problem of PDF layers/structure detection. Solving this problem allows us to address **two** industry problems: 1. Helping the users to convert the PDF document to the Office document format efficient and with high accuracy, supporting a valid document structure in order to remain the result document accessible, and as result to reduce or eliminate editing and remediation time. 2. Make document automatically accessible so remediator (the person who edits the document structure to achieve compliance with accessibility standards) spends less time on manual work. **What business problem are we solving?** 1. High competitive level on a market due to continuously improving quality. 2. Higher customer satisfaction rate due to short time to address the customer's pain points. 3. Solidifying the product as experts in the domain via dealing with real-world documents. 4. Positive impact on product growth due to use of modern technologies. **How does AI add value?** 1. **Customer satisfaction**: AI can detect the document structure (e.g., text, image, table, list etc.), thus making the PDF document conversion to MS Office document format more accurate. 2. **Keep competitive level on market**: To achieve a **continues improvement** of the results with AI is

	easier than fixing the hardcoded heuristic. This allows us to remain on a high competitive level. 3. **Product growth:** Getting to the new market level by consuming modern technologies: product becomes more competitive. 4. **Reduced cost:** An AI model re-learning is also less error-prone and causes less regression than fixing the hardcoded heuristic. This reflects on maintenance and continues support time and budget.
Business Case Why is this an important problem to solve? Make a case for building this product in terms of its impact on recurring revenue, market share, customer happiness and/or other drivers of business success.	**Business case – Document conversion** With the COVID-19 situation, the digital transformation naturally sped up. As result, the demand for documents processing increased. Document conversion is one of the many problems that document processing solves. **Driver 1 – Customers demand**: The engagement with the customers uncovers that 70% of large enterprise users have a need to convert PDF to Word on daily bases as part of their document's workflow. The numbers for medium and small enterprise users are lower – about 50% or on unregular bases. **Driver 2 – Revenue**: As PDF conversion to MS Office document is a part of the workflow for the large enterprise customers, it allows us to win more big deals. Absents of such a feature leads to significant revenue loss. **Driver 3 – Market fit**: Conversion is a "must-have" feature of PDF editor and became an unspoken standard. **Business case – Accessibility** **Driver 1 – Cover market gap**: All our competitors support the ability to create an accessible document. Absents of such a feature means that particular market segments are closed for us, such as government, education etc. **Driver 2 – Customers demand**: To have an accessible PDF document became low. Many organisations such as educational, government, healthcare etc. must ensure that any public document is accessible for the people with special needs. Moreover, with COVID-19 more and more educational program moved online and this increased demand for accessible PDF. **Driver 3 – Customers happiness:** Remediate PDF document is not a trivial task. The remediation specialist

	spends many hours making a few pages of PDF document accessible. The task is time-consuming and requires experience and focus. Automatically generated document structure with high accuracy saves many hours of remediator's work.
Application of ML/AI What precise task will you use ML/AI to accomplish? What business outcome or objective will you achieve?	Using AI allows us to detect the PDF document structure with high performance and accuracy. PDF document as a format has nothing close to represent as a document structure (paragraph, heading, table, list etc.) – it is just a set of rules on how to draw the content when rendering a PDF. A human can easily detect where e.g. the table is in the rendered PDF document; a code to detect the same table in the PDF format is a difficult calculation. By using AI we target to achieve "human's eyes" detection where is which element of the document. **Business outcomes**: 1. **Increased user experience** through improved accuracy, reliability and performance (reduced execution time) 2. **Improved customer satisfaction** through reducing end-user time on the result document editing (e.g., fixing accessibility and formatting of the document after conversion) 3. **Saving costs** through removing a need for expensive fixing of hardcoded heuristic logic that often leads to regression.

Success Metrics

Success Metrics What business metrics will you apply to determine the success of your product? Good metrics are clearly defined and easily measurable. Specify how you will establish a baseline value to provide a point of comparison.	**Business metrics – Document conversion:** A little bit of background for the numbers: bad quality of conversion leads to rejection of the deal on the pilot stage; if the pilot goes well but the pain points which are discovered later or are highlighted during the pilot are not addressed lately – retention rate drops. 1. **Customer retention:** gain retention of 90% of previously unhappy with the conversion quality and reliability. 2. **Customer acquisition and revenue gain:** win 2-3 big customers per year who previously declined

	our deal due to bad conversion quality. 3. **Customer support cases:** the number of customer support cases about the document conversion drops by 50%. 4. **Maintenance cost:** the cost of re-training and re-deploying the new model is the same as less than the cost of fixing hardcoded heuristic logic. 5. **Time to market:** time between the issue is reported and the solution with a re-trained model is delivered to the customer is less than the time delta for fixing the hardcoded logic approach. **Business metrics – Accessibility:** 1. **Customer acquisition**: win deal with 2 customers who previously declined our deal due to absents of accessibility feature; win deal with 5 new customers in government or educational market. 2. **Revenue**: increase the number of sold licenses by 5% to existing customers by acquisition their departments which are responsible for the document's structure remediation.

Data

Data Acquisition Where will you source your data from? What is the cost to acquire these data? Are there any personally identifying information (PII) or data sensitivity issues you will need to overcome? Will data become available on an ongoing basis, or will you acquire a large batch of data that will need to be refreshed?	**Initial data source:** 1. In-house existing source of the document, collected by many years of the company's existence. It is our testing data set. IMPORTANT: dataset needs review in terms of personal information. Dataset is relevantly small (over 1500) but is good as a start point for POC. 2. Dataset banks: e.g., SciTSR, DocBank etc. **Dataset extension:** 1. Consider building in-app consent with the users to use their documents for the model training. 2. Make sure that the documents received from the user are pre-processed by masking data to prevent personal data leak. Clearly communicate about data masking in customer-facing documentation. 3. Make consent optional (easy to reject) and clearly communicate it in customer-facing documentation.

Data Source Consider the size and source of your data; what biases are built into the data and how might the data be improved?	**Data source:** The same as the answer above. As the different data sources may give a different input format (e.g. PDF as images or PDF as a real document), the input data has to be pre-processed (standardised) before it is feed to the model. SUBJECT FOR DISCUSSION WITH AI ENGINEER: which input data format should we use: 1) PDF as images 2) PDF as a document (OCRed) or 3) PDF represented as a JSON/XML? Can be found out by running POC. **Biases**: There are at least 2 factors that have an influence on the document structure and may cause biases. **Bias 1 – document origination**. Most of the datasets are English and West Europe or US documents. This may create a bias in terms of the document structure, such as test orientation and formatting. To improve the dataset we need to consider the acquisition of datasets from East Europe, Asian and other regions. When it comes to the educational sector of the market, we need to consider the documents written in less common languages. **Bias 2 – documents purpose**. We need to ensure diversity of the document types: educational (e.g. vocabularies), books (for adults and for kids), business documents (agreements, contracts), legal documents, forms/applications, magazines, brochure etc. **Data Size:** Variety of the document content types leads to needing at least 11M documents for training to achieve decent quality. The number is taken from the result report about Microsoft LayoutML library performance – a potential candidate to be used in our model. For POC, an existing in house dataset of 1500+ documents is a good start.
Choice of Data Labels What labels did you decide to add to your data? And why did you decide on these labels versus any other option?	Choice of Data Labels is driven by PDF standard and corresponds the standard set of the PDF tags (elements of the document structure): - Paragraph - Figure - Heading

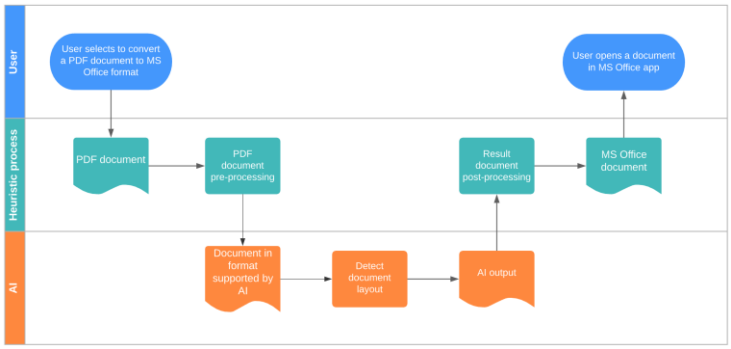
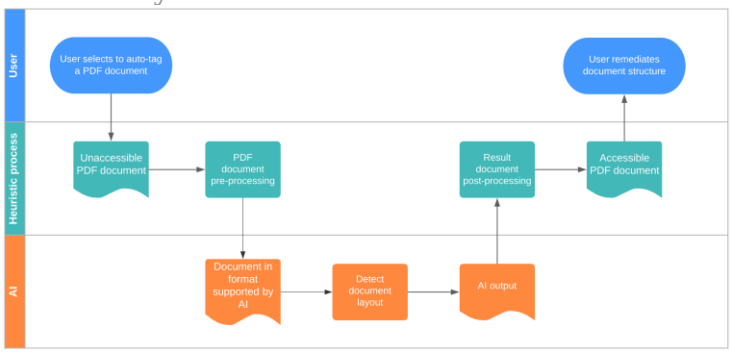
	- Table - List - Header - Footer - Table of Content - Hyperlink - Annotation First, the AI detects the objects bounding boxes in the document and then assignee the label to each found object. Here important to understand that the model contains multiple layers, each of which is responsible for detecting a specific label.

Model

Model Building How will you resource building the model that you need? Will you outsource model training and/or hosting to an external platform, or will you build the model using an in-house team, and why?	The model will be build using an in-house team (hiring additional resources will be required). The model training will be performed by: - In-house team - Existing out-sourcing team. Both teams have rich domain knowledge and already have full access to the test resources (the initial set of the training data). The build model process will be split on 2 stages: 1. **POC.** It requires an existing engineer + PM + QA resource – people with rich domain knowledge. Data hosting – inhouse until the dataset is relevantly small. 2. **Bringing the project to MVP.** Moving forward hiring an AI specialist is highly recommended: with the dataset growing and the initial model gets mature we need to address data storage, security and maintenance questions. Consider data hosting on existing in-use external resources (data warehouses). Building any product with the in-house team is a company policy.
Evaluating Results	**Performance metrics - Document conversion:** - Recall. Highly important for detecting primitive

Which model performance metrics are appropriate to measure the success of your model? What level of performance is required?	elements such as Paragraph vs. Figure - Precision. This metric is more important (compared to recall) for more complex layout such as tables and lists. - Time spent on conversion the document (target: seconds, comparing to minutes using hardcoded heuristic logic) NOTE: unfortunately we cannot measure the time spent on editing converted document format as this will require gathering metrics from MS Office application where we do not have access to. For this metric, we need to rely on the customers and test group feedback. **Performance metrics – Accessibility:** - Recall. Highly important for more complex layouts like tables, lists, references etc. - Precision. Highly important for detecting primitive elements such as Paragraph vs. Figure. - Time spent on manual remediation (less is better) The difference between precision and recall metrics for these 2 business cases are based on the work that an end user has to do after conversation/auto-tagging (making a document accessible). For conversion, it is more important to keep the document rendering as accurate as possible, while for accessibility remediator it is critical to reduce time spend on remediating complex objects, e.g. tables (to remediate complex object to simple object is easier than vice versa).

Minimum Viable Product (MVP)

Design What does your minimum viable product look like? Include sketches of your product.	**Document conversion**  **Accessibility**  **Model characteristics:** 1. AI layer is a complex model that detects bounding box for each element of the document and then decides whether the elements forms into Paragraph, Figure, Table, List, Heading etc. 2. Before AI model can process the input, a PDF document has to be pre-processed (standardised). Depends on what library will be used for AI, the input documents have to be represented as a "PDF as image" or "PDF as JSON/XML". **Subject for discussion with AI engineer.** 3. Then the output result from the AI is post-processed and converted to the output document (MS Office document or Accessible PDF).
Use Cases What persona are you designing for? Can you describe the major epic-level use cases your product	**Use case – Document conversion** A Udacity student needs to download and fill the Capstone project template but for some technical reason, the docx link is broken. The user downloads a PDF template instead, opens it in our PDF editor, selects to

addresses? How will users access this product?	convert it to MS Word document. Now, without a need to fix the result document formatting, they need just fill the template with the answers. **Use case – Accessibility** A teacher in Udacity needs to include to the course resources a Cheat Sheet for students about LinkedIn profile best practices. The teacher got PDF document from the designers but the document is not accessible – the screen reader read the document in a chaotic order, e.g. a little bit of the text from column one and a little bit of the text from column two. The employee needs to make a document accessible to meet online educational standards. The Udacity teacher opens the PDF document in our PDF editor and selects to auto-tag the document (create a document structure). Now, with minimum effort of tuning the document structure the document (remediation) becomes accessible. The AI model is a replacement for the existing error-prone hardcoded logic in the product. As result, the user experience will not dramatically change after the project is done. However, the waiting time and the quality of the result document become significantly better. User experience improves in terms of editing/remediation of the result document.
Roll-out How will this be adopted? What does the go-to-market plan look like?	**Approach 1:** Initial enrolment will assume A/B testing. We leave 80% of our users with existing hardcoded heuristic logic implementation and 20% of the users will work with the product that uses the AI model instead. **Approach 2:** Run an “early access” program: advertise the new “engine” for conversion/accessibility among existing customers and advert signing to “early access” schema for evaluation. These customers will get the product with the AI model instead of hardcoded heuristic logic as well as they accept consent on sharing their documents for further model training (with masking data). When the model proves its quality and performance after a few iterations, all customers will be upgraded to the version with the AI model.

Post-MVP-Deployment

Designing for Longevity How might you improve your product in the long-term? How might real-world data be different from the training data? How will your product learn from new data? How might you employ A/B testing to improve your product?	The main approach to improve the product is through learning from the remediating work. By gathering metrics from accessibility remediation we can spot the weak areas of AI in detecting the document structure. By engaging the customers in the "early access" schema or by getting their consent on using their data (e.g. during A/B testing), we can get more real world documents to improve the model performance. **IMPORTANT**: we need to keep an eye on data balance to prevent introducing biases by adding new training documents from the customers. The real world data will be different from the training data to some degree: depending on the document purpose and region of origin we will deal with a little bit different set of the layouts. The companies who follow the accessibility standards are more accurate with the document formatting as they usually keep accessibility in mind when creating a PDF document. However, other organisations and retail users may pay less attention to such details and end up with the document with messy formatting and heavy content layouts (e.g. PDF created from presentations).
Monitor Bias How do you plan to monitor or mitigate unwanted bias in your model?	**Monitor bias:** 1. Use data from remediators to detect whether the bad model performance was caused by bias. E.g. if the most common remediation is about replacing the list with the paragraph element in the structure, there is a clear bias toward list detection. 2. Targeted testing. By including a diverse set of documents in the testing set may help detect the bias. The documents test set must be selected by a diverse group to prevent human bias transferred to the model. 3. Random testing. Pick random output documents

	to evaluate their accuracy. Use a diverse group of testers. **Mitigate bias:** 1. Keep a balance of training documents in terms of document layouts and the origin of the documents. 2. Have a diverse QA group during the model training. 3. Keep a balance between the documents when the real document comes from a customer for training purpose. If the customer willing to help have single-type documents (e.g. all documents are bank sheets with fee rates) then such a benefit will become a bias.