**ArmorDoc Project Plan**

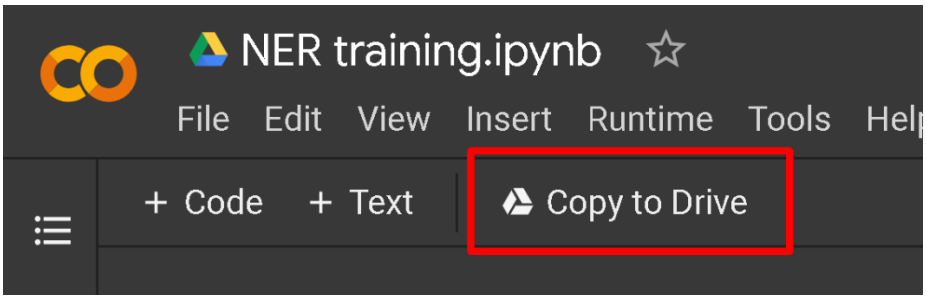
**Goal:**

To gain insight into the approaches, techniques and execution a Data Scientist would use to solve for Use Cases specific to ArmorDoc's service offerings. To find out if this is work that is found to be interesting.

**Timing:** *Start Date*: 5/24/21; *End Date*: 9pm est on 5/28/21

**Material Provided:**

1. PDF file with 149 pages consisting of a mix of 1st and 2nd Notes (a type of loan document.)
2. Collab notebook for training an NER model with Spacy (this is a hyperlink)




**Scope of Work:**

1. PDF with embedded bookmark: leverage parts of the document which distinguish it from a 1st Note compared to a 2nd Note.

Differences between the two documents can be found in the title of the document in which the word "Second" is included for 2nd Notes. They can also be found under section 3. PROMISE TO PAY SECURED.

Build a ML model for classification.

In addition to the ML model, you can optionally add hard coded rules to improve accuracy.

OCR will be needed to obtain the text for modeling.

2. Extract the following data points with a high level of precision using OCR:

a. Lender (found in section 1. DEFINITIONS)
    a. "Lender" means….
    b. Text field
    **c. Training an NER model using Spacy 2**
b. Maximum Principal Amount (found in section 2. BORROWER'S PROMISE TO PAY; INTEREST)
    a. "maximum principal amount of…"
    b. Currency field
    **c. Using Regular Expression**
c. Maturity Date (found in section 2. BORROWER'S PROMISE TO PAY; INTEREST)
    a. "are due and payable on…."
    b. Date field
    **c. Using Regular Expression**

**Deliverables:**

1. PDF with embedded bookmarks to distinguish between 1st and 2nd Notes.

2. Data extract with Page Number, Document Label (1st or 2nd Note), Lender, Maximum Principal Amount and Maturity Date in a csv file.

3. .ipynb or .py file(s) for all codes.

**Additional Notes:**

- You may need to perform some research to assist in execution
- OCR stands for Optical Character Recognition (Use Tesseract for python)
- Since the bookmarks will only be 1st or 2nd Note, you can just add a number at the end of bookmark in the pdf to prevent repeating bookmarks like "1st Note_0"
- NER stands for Name Entity Recognition (We're using Spacy 2; please do so as well)
- Refer to the Collab notebook for how to train and use NER with Spacy2, and the data format needed for training
- There are tools you can find online for NER annotation (Label Studio)