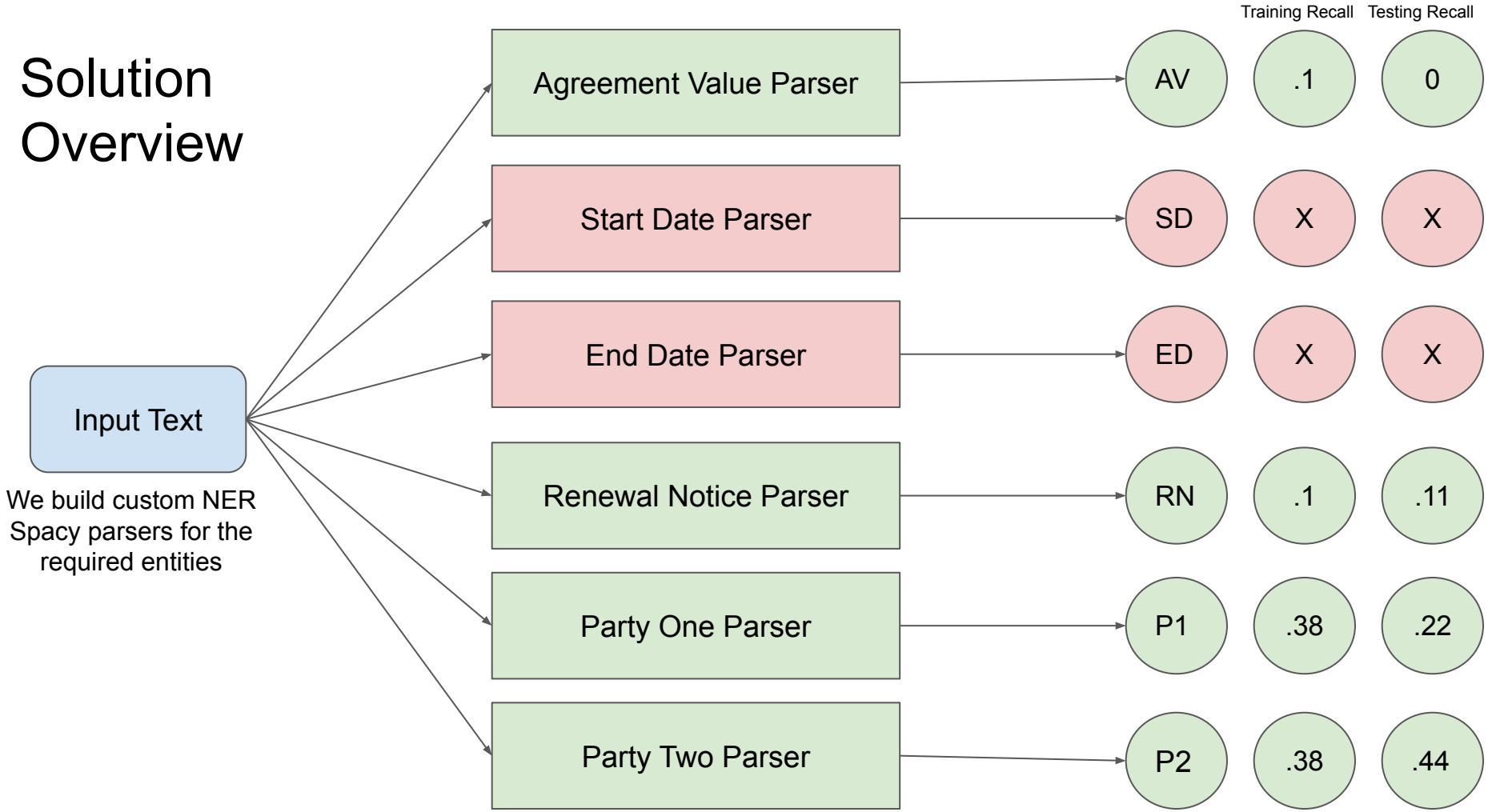
NER Model For Rent Agreement Parsing

Objective - Training a Rent Agreement Parser
that fetches the following from a *.pdf.docx* file:
- Agreement Value
- Start Date
- End Date
- Renewal Notice Period
- Party One
-Party Two

Code:
https://www.github.com/sahil3vedi

Solution Overview



1. Data Preprocessing

- We begin by cleaning the incorrectly formatted or non existent references from our dataset.
- We extract the training csv file as a pandas dataframe.
- We also fetch the training word files from their location in the dataset
- The final training dataframe has 42 records.

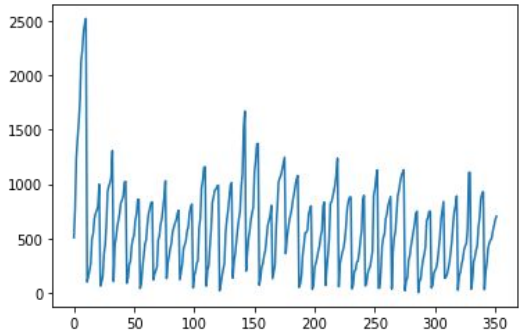
2. Using Spacy to Label Training Data

- We need multiple classifiers for different labels.

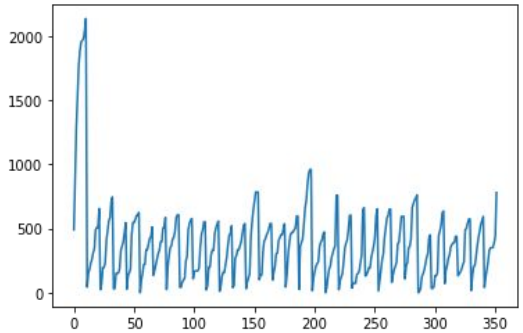
- We also use one additional general purpose multilabel classifier.

- We use Spacy to form multiple training sets for total 5 classifiers:
  - General Purpose Multilabel Classifier (GP)
  - Agreement Value Classifier (AV)
  - Renewal Notice Classifier (RN)
  - Party One Classifier (P1)
  - Party Two Classifier (P2)

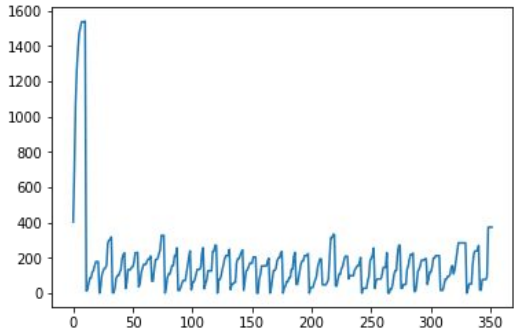
3. Training Loss Overview



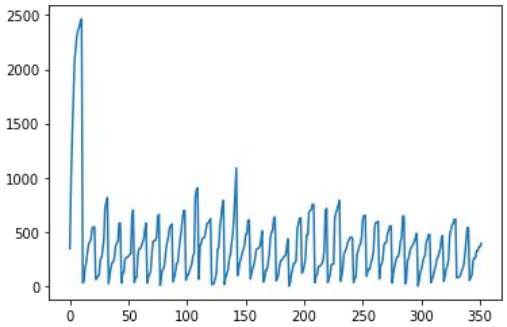
GP



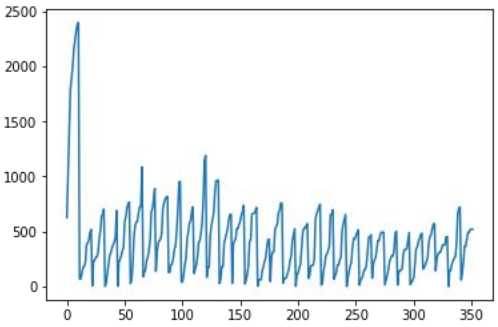
AV



RN



P1



P2

3. Validation Recall Overview

- We validate the 5 classifiers on the validation set.

- Recall Values:
  - AV: 0 / 9
  - RN: 1 / 9
  - P1: 2 / 9
  - P2: 4 / 9

4. Potential Improvements to the Model

- More training data / epochs
- Hyperparameter tuning in classifiers.
- Converting Dates and Money to Common Format using tokenizers.
- Optimising Text Length
- Using SoTA techniques like Transformers / LSTMs