York University MLC - CSML1030
Project proposal
Project category : NLP
July 2021 - September 2021
Queenie Tsang, Crystal Zhu, Gouri Kulkarni
Project hosted by: Athennian

1.Abstract

Rule based coding approaches to analysing PDFs can provide insights and reports, but are not scalable. In this paper, we have explored some state of the art Natural Language Processing (NLP) approaches to generalizing the task of extracting key information and building relationships to develop workflows. The best approach would be the one which provides accurate results and satisfies key metrics.Through using Pre Trained models and Fine Tuning, we can achieve results aligned to the use case and domain. Specifically in this project we applied NLP tasks available in Spacy Natural Language Processing, Rosette Deep Learning Neural Network, and ELMO. We also explored Optical Character Recognition (OCR) and the use of available public datasets in the business/legal domain.

2.Keywords

NLP, Named-entity recognition, text extraction, information retrieval, spaCy, OCR, ELMO, Rosette, public datasets

3. Introduction

As part of the machine learning capstone course, we had the opportunity to address a real business problem faced by Athennian, turn it into a scalable machine learning task and research possible solutions. Athennian provides entity management products and services to law firms. Entity minute books have historically been stored by law firms on behalf of their clients in binders following traditional filing practices. Athennian's clients are law firms that have migrated from legacy systems to their SaaS solution. The cloud-based work space enables streamlining of day-to-day processes and keep transactions organized. Majority of the work done at law firms is based off reading, assimilating and analyzing information from documents. An entity is a client whose books are managed by the law firm. The size of data handled per entity varies. So does the size of the law firm. The number of entities managed by a law firm can range from 50 to up to 25,000. Contract review is the process of thoroughly reading a contract to understand the content. It is one of the most repetitive and most tedious jobs performed by law firm associates and is also an expensive and inefficient use of skills.The same data points are to be identified in every document. A list of data points is shown below. Automation involves developing new workflows and using data driven approaches to read, analyse and interpret documents. Generalizing the process to work on any input document cannot be handled with traditional coding.
Law firms that are on the journey to becoming paperless already have scanned PDF

copies of minute books. Manually summarizing and extracting key information from PDFs involves a lot of time and effort.

Variable Category	Variable Name
Basic Entity Data	Current Entity Name
Basic Entity Data	Entity Names forming Amalgamation (optional)
Basic Entity Data	Previous Entity Name - (optional)
Basic Entity Data	Entity Type
Basic Entity Data	Corporation Number
Basic Entity Data	Jurisdiction
Basic Entity Data	Entity Formation Date
Basic Entity Data	Registered Office
Share Classes	Share Class Name
Share Classes	Voting Rights (True/False)
Share Classes	Votes per Share
Share Classes	Status
Share Classes	Authorized Amount Unlimited (True/Fase)
Share Classes	Authorized Amount
Share Classes	Share Class Creation Date
Share Classes	Certificate Code
Share Classes	Share Class Termination Date
Articles	Min Directors
Articles	Max Directors
Articles	Unlimited Directors (True/False)
Articles	Transfer Restrictions Texts
Articles	Other Provisions
Articles	Restrictions

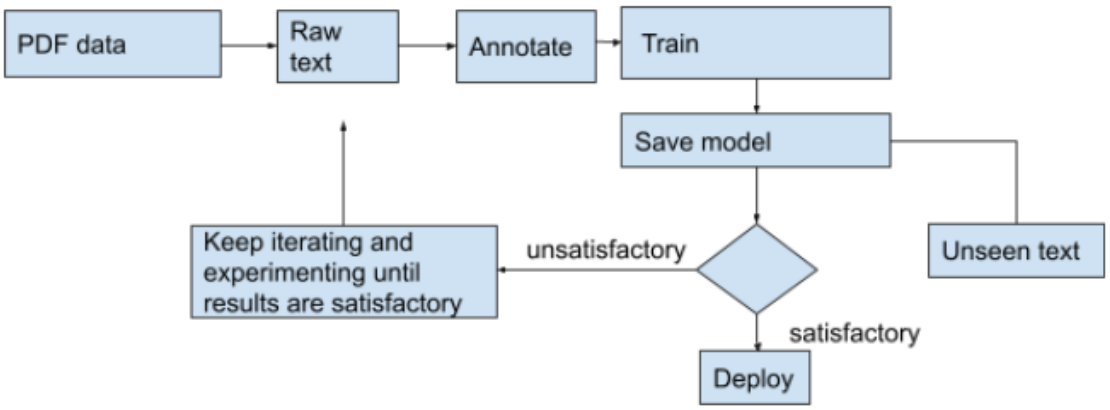
Key variables of interest

4.Business problem

The business problem involves assisting a document processing company in text extraction of specific variables from scanned pdfs of legal documents for use in downstream tasks. The proposed solution involves applying Natural Language Processing techniques to extract and process the unstructured text data from the pdf files, and train models for named entity extraction and relationship development.

York University MLC - CSML1030
Project proposal
Project category : NLP
July 2021 - September 2021
Queenie Tsang, Crystal Zhu, Gouri Kulkarni
Project hosted by: Athennian

5.Process



Data collection: The input provided was in the form of 12 scanned minute book files in PDF format shared with the group by the company Athennian via a shared Google Doc folder.

Text pre-processing : Text that forms the corpus to be used in NLP tasks was extracted from the pdf files using different libraries including PyPDF2, textract, tika, fitz(PyMuPDF), pdfplumber, PDFminer. Text was also extracted using OCR libraries such as pytesseract.

Text mining : The unstructured text does not have a predefined format. Text mining NLP techniques were used to do initial EDA, text summarization, NER for further analysis.

Annotation : Transforming the available data into a format baseline or proposed models which can be used to make predictions, annotation is the process of generating labelled data for supervised learning.The format of annotations required by spaCy, Rosette, ELMO , or any Pre Trained model differs (BILOU, CoNLL, IOB, json, spacy formats).

Text embedding/ feature engineering : With word embeddings from the selected downloaded pretrained models and transfer learning, we extracted features of use for downstream models.

Entity extraction: The small pretrained model from spaCy, pre-trained deep neural networks provided in Rosette and ELMO word embeddings were used to extract named entities such as organization, location and number, contained in the pdf files. Next, variables of interest, such as current entity name, entity type and corporation number, could be further extracted by applying rule-based conditions to the named entities. The links between variables that form the relationships can be annotated in the training data.

Training/refining: For every proposed method,evaluation of the performance of the saved model based on key metrics is an iterative process. If the results of one evaluation are not satisfactory, the annotations to the training data may need to be changed or datasets may have to be switched until satisfactory outputs are obtained.

6. Proposed methods

- ## spaCy for named entity recognition and annotation of raw text

  We attempted to use spaCy to generate training data from the input PDFs. spaCy is designed specifically for production use and helps build applications that process and understand large volumes of text using several pretrained models and it also has transformer capability. It can be used to build information extraction or natural language understanding systems. It has components for named entity recognition.The spaCy API has pipelines and matchers that can be used to develop annotated training data suited to the downstream tasks.
  Using text as input, spaCy 2.2.4 and the pre-trained small english model, the task of extracting key variables was converted into a NLP NER problem.
  The minute book text is sparse and does not consist of complete sentences, so POS tagging was not done.
  As the default NER (PERSON, ORG, GPE, MONEY etc.) labels that come with the model may or may not suit our task, we attempted to create a blank model, add a new NER pipeline and custom labels to the pipeline. By creating patterns with the PhraseMatcher , we generated training data by using the spans in the NLP document. A NER model was trained to extract input entities. This approach can be extended to extract multiple entities from the text and generate multiple labels.
  It was also found that text extractors such as PyPDF2 , fitz, textract do not extract the desired text from forms. Text extractors are generally accurate, but the PDFs being noisy, none were able to give good results. Thus, the NER model was unable to extract certain variables from the text. Also, the small pretrained model

relies on CPU. Training the model to extract one variable from text extracted from 200 page PDFs was very slow. Testing the model in unseen data worked.
This approach can be used to create annotated training data in spaCy format for use in further entity/variable extraction. spaCy can also be used to develop applications and automate the generation of annotations. Due to CPU restrictions and poor quality of extracted text, this approach was not pursued.

- ## Rosette pre-trained deep learning neural network model

  We attempted to use the pre-trained deep neural networks provided in Rosette to extract the variables from the scanned pdf files. Rosette is a platform that uses natural language processing, statistical modeling, and machine learning to analyze unstructured and semi-structured text. The Rosette entity extraction endpoint uses statistical or deep neural network based models, patterns, and exact matching to identify entities in documents.
  The statistical models are based on computational linguistics and human-annotated training documents. The patterns are regular expressions that identify entities such as dates, times, and geographical coordinates. The exact matcher uses lists of entities to match words exactly in one or more languages. Rosette also has a deep neural network model (currently in BETA) that can be used in place of the statistical model for selected languages. The deep neural networks involve Multilayer Bidirectional Long Short Term Memory (LSTM) structure, combined with character encoding represented for each word. Multiple languages are supported by the deep learning models trained in Rosette but accuracy differs among languages.
  It was found that deep learning neural networks are able to detect more entity types than statistical models, and the entity extraction is more accurate when applied to clean text files.

- ## Named Entity Recognition and Word embeddings using ELMO (Embeddings from Language Models)

  ELMO is a NLP framework developed by AllenNLP which uses a bi-directional Long Short Term Memory (LSTM) neural network. ELMo is trained on predicting the next word in a sequence.We attempted to use ELMO to extract features from the raw text. The pretrained ELMO model is trained on a large dataset. Following several tutorials, we attempted to pre-process the input text data similar to get an

annotated corpus for Named Entity Recognition. Tutorial input texts were already annotated to be BILOU tagged or already annotated using Groningen Meaning Bank, where different tags such as sentence number, word, part of speech (POS) and Tag were already labeled.

It was used to extract word embeddings from our input text. Every input sentence was thus converted into an ELMO vector. Saving the extracted embeddings, a model was trained to predict named entities.

- Pretrained LayoutLM model fine tuned on form, receipt and image datasets

  We attempted to fine tune the pre-trained LayoutLM model on FUNSD, a dataset of annotated forms, receipts and business memos. The dataset consists of PDFs converted into PNG images. Every PNG file has a corresponding json file containing annotations in question - answer format. The LayoutLM model fine tuned on the FUNSD dataset, having learned the annotations, was used to make inferences on PNG images of unseen PDF minute books. Inference required loading an image, running OCR to get the bounding boxes of the text and then running LauoytLM on the individual tokens and visualizing the results. Through this exercise, we concluded that in order to get the desired output using this method, (extraction of variables), we need to fine tune the LayoutLM model on a dataset of annotated minute books.

  .

Benefits and Challenges for Each method

Approach	**Pros**	**Cons**
Spacy Natural Language Processing	• Availability of a variety of pretrained models, transformer models in various languages and domains • Open source, free • Flexible pipeline	• Training the model on text can be slow depending on the pretrained model chosen • spaCy's annotation tool Prodigy is not free • Annotations in spaCy format can be generated

	development through reusable components suited to the task	but require extensive coding and exploration of the available text
Rosette pre-trained deep learning neural network model	• No need to spend much time training a deep NER model from scratch • Huge amount of training text data is not needed • 29 entity types, such as LOCATION, ORGANIZATION, PERSON, PRODUCT, TITLE, NATIONALITY, etc., and 450+ subtypes are available out-of-the-box • Use the Wikidata knowledge base to link Person, Location, and Organization entities • Highlights the entities that are most relevant to the content of a document with a salience score • Cloud or enterprise deployments	• Not a free source (a free 30-day trial is available) • Relatively clean txt files are needed to apply the models • Language models that may be suited for the task may be slower to load and train • Language models that are close to the domain (contracts, agreements, English, tabular, images) don’t exist
NER and Word embeddings using ELMO	• Can be used on unstructured text • Different words retain their different	• Difficult to implement due to having to use pre-processed data in a specific format

	meanings depending on the context of the text • ELMO looks at the entire sentence before assigning each word an embedding • ELMO LSTM can be trained on a massive dataset to create a customized language model, and then can be used in other Natural Language Understanding models	(ie. BILOU tagged, CoNLL) • Requires setting up Spark NLP
Pretrained LayoutLM model fine tuned on form, receipt and image datasets	• Uses layout and style information that is vital for document image understanding and which is missed when using pure text for embeddings. • Text and layout are jointly used in a single framework for document-level pre-training which suits our input data. • As the text dataset has been turned into an image dataset, image preprocessing techniques such as augmentation can be used to achieve better results. On noisy datasets.	• Results depend on the dataset used for fine tuning which can be hard to obtain depending on the domain. • Results depend on the annotation strategy used which needs to be set up according to the chosen NLP task. • OCR annotation formats can get complicated when the text is dense.

8. Proposal for future work

1. Explore pdf text converter libraries that produce cleaner text
2. Explore pre-trained neural network models trained in a similar domain
3. Explore annotated text in a similar domain
4. Explore using pre trained embeddings as input vs using pre trained embeddings as trainable layers
5. Explore applying transfer learning to annotate domain specific text to fine-tune pre-trained models

9. Conclusion

In this project, we started with scanned pdf files, and then applied several text extraction tools such as PyPDF2, textract, tika, fitz(PyMuPDF), pdfplumber, PDFminer to convert pdf files to text files. We then explored several state-of-the-art NLP techniques, including SpaCy, Rosette's deep neural network and ELMO, to extract named entities and also explored extracting text using pre-trained fine tuned models using OCR for text detection.

Through using pre-trained models, we can improve the generalizability of tasks if we have small training datasets or wish to speed up training on large datasets.

10.References

Basis Technology Corp, n.d., *Entity Extraction and Linking,* Rosette, accessed 10 September 2021, <https://developer.rosette.com/features-and-functions#entity-extraction-and-linking>

Kfir Bar (2019), Named Entity Recognition, <https://www.youtube.com/watch?v=TUXbXwu17KE&ab_channel=PyData>

Devlin, J., Chang, M., Lee, K. and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv: 1810.04805*

Intuition Engineering (2018), Deep learning for specific information extraction from unstructured texts, <https://towardsdatascience.com/deep-learning-for-specific-information-extraction-from-unstructured-texts-12c5b9dceada>

York University MLC - CSML1030
Project proposal
Project category : NLP
July 2021 - September 2021
Queenie Tsang, Crystal Zhu, Gouri Kulkarni
Project hosted by: Athennian

PyPDF2 https://pypi.org/project/PyPDF2/

ELMo

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Mattmann, C. A., Wilson, B. D., Zhao, D., Durri, K., Palsulich, T., Germuska, J., Shvedov, V., Vieira, D., Ahmadia, A., Singh, K., Nasyrov, R., Brooking, J., Tanna, Y., Tokarev, I., Parker, I., K. A. Didier, Elosua, J., de Oliveira Antunes, C., Tika-Python. Github Repository. Accessed from https://github.com/chrismattmann/tika-python

Overview: Extracting and serving feature embeddings for machine learning

<https://cloud.google.com/architecture/overview-extracting-and-serving-feature-embeddings-for-machine-learning >

LayoutLM: Pre-training of Text and Layout for Document Image Understanding

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou (2020)
arXiv:1912.13318v5

CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review

Dan Hendrycks, Collin Burns, Anya Chen, Spencer Ball (2021)
arXiv:2103.06268

York University MLC - CSML1030
Project proposal
Project category : NLP
July 2021 - September 2021
Queenie Tsang, Crystal Zhu, Gouri Kulkarni
Project hosted by: Athennian