Provectus

Synthetic Invoice Dataset Generator

Abstract

In this paper, we present the implementation of an invoice generator, with random filling of data and various template formats. This type of data can meet the needs of a machine learning system, to automate and improve the flow of complex document processing. Using a synthetic dataset solves the problem of sensitivity of the training data for such models. In addition, this invoice generator solves the problem of data annotation by automating the generation process. The quality of the generated dataset was checked using a custom model for information extraction. PDF or PNG formats, and output annotations to JSON or XML formats. The final dataset consists of invoices with varied ground truth and layout.

Introduction

A synthetic dataset is a type of data that is generated by a program, and not collected from real life experience. Its main goal is to be flexible and rich enough to help conduct research with machine learning models.

In the process of building a machine learning model for data recognition and extraction from invoices, it is necessary to have a sufficiently large annotated dataset for the training process. Due to the sensitivity of information, such datasets are not publicly available, so we decided to build a document generator similar in format and content to examples of real-life invoices. This approach can be used not only for invoices, but also for other types of documents.

We propose the approach of building HTML invoice templates, along with further extraction of bounding box coordinates and labels during rendering. Our generator allows output documents to be converted to PDF or PNG formats, and output annotations to JSON or XML formats. The final dataset consists of invoices with varied ground truth and layout.

Invoice Generation

After examining examples of invoices in real life, a typical structure for this type of document was formed, which is in three parts: the Head contains company and client information, invoice number and date information; the Central part contains tables with product information; and the Footer contains information about payment, total cost and taxes. Other information blocks are optional and can have different positions in the document layout.

To create an instance of an invoice, the information blocks of interest were first determined, for the extraction of values for which the model would be trained. For the field content of information blocks, random data were used for text fields and mathematical functions were used to calculate some numerical values. To generate the layout, random behavior of information blocks were organized in predetermined sections of the invoice structure.

There are several steps involved in creating invoices: creating HTML, extracting bounding box coordinates, extracting label data, and converting HTML files to PDF or PNG formats.

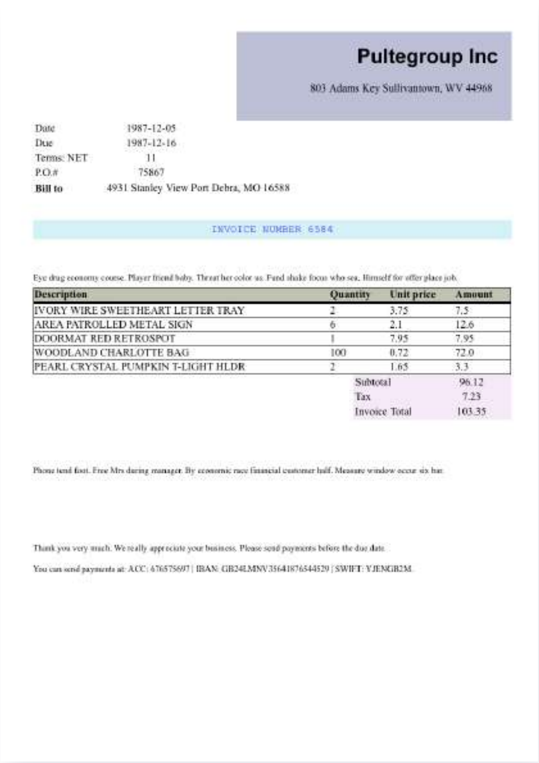
Data Generation

To generate random text data we used the Python package Faker. Two prepared synthetic datasets were used for company names (Provectus Contract Generator) and product tables. Some of the fields were determined through functional dependencies, for example, the calculation of total cost and due date values.

Based on the analysis of real-life documents of this type, groups of possible names of target fields were collected to account for diversity in training. For example, "Invoice Date," "Issue Date," "Date," "Date of Issue," and "Issued on," can all represent the same target field in different documents.

Layout Generation

The document layout was built with Jinja as a templating engine for Python, which allows for dynamic creation of HTML documents. Information blocks were defined for each part of the template (Head, Central, Footer). The rendering of all information blocks was randomized using specific css properties (background color, font size, border, etc.). Mixed render layouts were achieved in this way (Figure 1).



Pultegroup Inc

803 Adams Key Sullivantown, WV 44968

Date	1987-12-05
Due	1987-12-16
Terms: NET	11
P.O.#	75867
Bill to	4931 Stanley View Port Debra, MO 16588

INVOICE NUMBER 6584

Eye drug economy course. Player friend baby. Threat her color us. Fund shake focus who sea. Himself for offer place job.

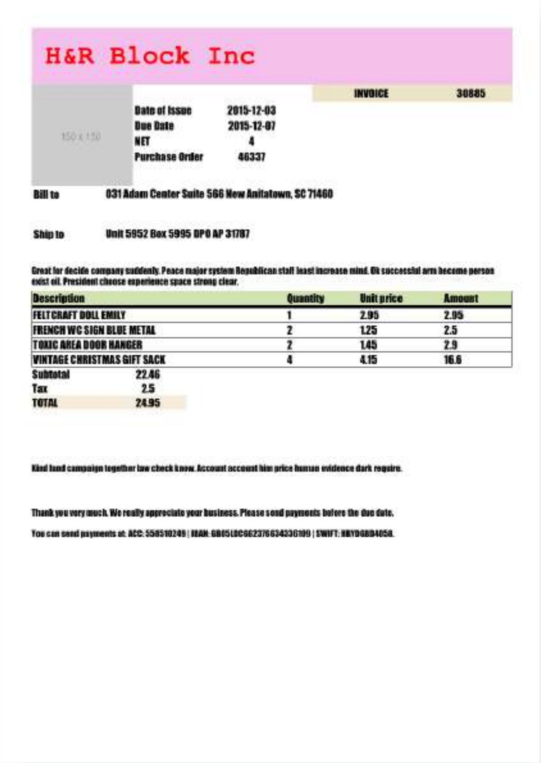
Description	Quantity	Unit price	Amount
IVORY WIRE SWEETHEART LETTER TRAY	2	3.75	7.5
AREA PATROLLED METAL SIGN	6	2.1	12.6
DOORMAT RED RETROSPOT	1	7.95	7.95
WOODLAND CHARLOTTE BAG	100	0.72	72.0
PEARL CRYSTAL PUMPKIN T-LIGHT HLDR	2	1.65	3.3

Subtotal	96.12
Tax	7.23
Invoice Total	103.35

Phone tend foot. Free Mrs during manager. By economic race financial customer half. Measure window occur six bar.

Thank you very much. We really appreciate your business. Please send payments before the due date.

You can send payments at: ACC: 676575697 | IBAN: GB24LMNV35641876544529 | SWIFT: YJENGB2M.



H&R Block Inc

INVOICE	30885

150 x 150

Date of Issue	2015-12-03
Due Date	2015-12-07
NET	4
Purchase Order	46337
Bill to	031 Adam Center Suite 566 New Anitatown, SC 71460
Ship to	Unit 5952 Box 5995 DPO AP 31787

Great for decide company suddenly. Peace major system Republican staff least increase mind. Ok successful arm become person exist oil. President choose experience space strong clear.

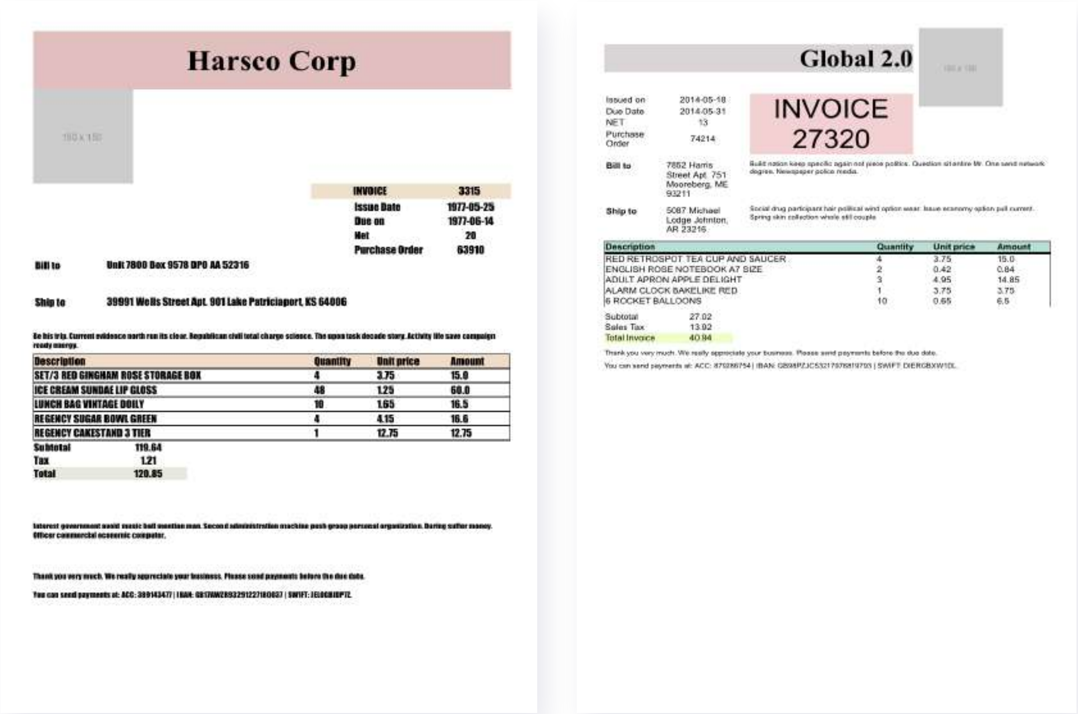
Description	Quantity	Unit price	Amount
FELTCRAFT DOLL EMILY	1	2.95	2.95
FRENCH WC SIGN BLUE METAL	2	1.25	2.5
TOXIC AREA DOOR HANGER	2	1.45	2.9
VINTAGE CHRISTMAS GIFT SACK	4	4.15	16.6

Subtotal	22.46
Tax	2.5
TOTAL	24.95

Kind land campaign together law check know. Account account him price human evidence dark require.

Thank you very much. We really appreciate your business. Please send payments before the due date.

You can send payments at: ACC: 558510249 | IBAN: GB05LDCG62376634330109 | SWIFT: HBYDGBD4058.



Harsco Corp

INVOICE	3315
Issue Date	1977-05-25
Due on	1977-06-14
Net	20
Purchase Order	63910

Bill to Unit 7800 Box 9578 DPO AA 52316

Ship to 39991 Wells Street Apt. 901 Lake Patriciaport, KS 64006

Be his trip. Current evidence north run its clear. Republican civil total charge science. The upon task decade story. Activity life save campaign ready energy.

Description	Quantity	Unit price	Amount
SET/3 RED GINGHAM ROSE STORAGE BOX	4	3.75	15.0
ICE CREAM SUNDAE LIP GLOSS	48	1.25	60.0
LUNCH BAG VINTAGE DOILY	10	1.65	16.5
REGENCY SUGAR BOWL GREEN	4	4.15	16.6
REGENCY CAKESTAND 3 TIER	1	12.75	12.75

Subtotal	119.64
Tax	1.21
Total	120.85

Interest government avoid music ball mention man. Second administration machine push group personal organization. During suffer money. Officer commercial economic computer.

Thank you very much. We really appreciate your business. Please send payments before the due date.

You can send payments at: ACC: 389143477 | IBAN: GB17AWZ89329122718O607 | SWIFT: JELOGBIDPTZ.

Global 2.0

INVOICE
27320

Issued on	2014-05-18
Due Date	2014-05-31
NET	13
Purchase Order	74214

Bill to 7852 Harris Street Apt. 751 Mooreberg, ME 93211

Build nation keep specific again not piece politics. Question sit entire Mr. One send network degree. Newspaper police media.

Ship to 5087 Michael Lodge Johnton, AR 23216

Social drug participant hair political wind option wear. Issue economy option pull current. Spring skin collection whole still couple

Description	Quantity	Unit price	Amount
RED RETROSPOT TEA CUP AND SAUCER	4	3.75	15.0
ENGLISH ROSE NOTEBOOK A7 SIZE	2	0.42	0.84
ADULT APRON APPLE DELIGHT	3	4.95	14.85
ALARM CLOCK BAKELIKE RED	1	3.75	3.75
6 ROCKET BALLOONS	10	0.65	6.5

Subtotal	27.02
Sales Tax	13.92
Total Invoice	40.94

Thank you very much. We really appreciate your business. Please send payments before the due date.

You can send payments at: ACC: 879286754 | IBAN: GB98PZJC53217976819793 | SWIFT: DIERGBXW1DL.

Figure 1. Examples of generated invoices.

Annotation Generation

The annotation is included in a JSON or XML stream and contains the position and box size of the text zone, the text itself, and the annotation class that we want to learn.

Invoice Analysis

In order to evaluate the performance of a machine learning model trained on data from our generator, we used a custom model based on a recently published article [1]. The key idea of the model is to extract information using knowledge about the types of target fields, to create candidates for extraction. The neural network architecture then studies the dense representation of each candidate based on neighboring words in the document. These learned representations are useful in solving the problem of extracting unseen document templates.

Initially, the image is run through the OCR engine (Textract or Tesseract). The text and the positions of the bounding boxes are extracted from the output. The text parser and predefined rules then extract the target field candidates and their neighboring words. Entities embedded using a word embedding table become input data for the neural network.

A generated synthetic dataset with 5000 images was used for training purposes. Evaluation of model performance was done using a dataset of real invoices with field result accuracy: invoice_id 0.68, invoice_date 0.72, total_amount 0.63, due_date 0.89, payment_terms 0.74, purchase_order 0.93, tax 0.78.

Conclusion

This paper presents a method for creating a synthetic dataset represented as invoices. All documents have a unique render layout and design, and contain the variability of names of key fields, to approximate the documents to real examples of invoices.

As a result of this work, an artificial dataset for the purpose of preserving privacy was created that is able to create training data for machine learning algorithms. This work allowed us to create annotated sample invoices adequate for the learning and extraction of necessary information.

References

Representation Learning for Information Extraction from Form-like Documents, Bodhisattwa Prasad Majumder, Navneet Potti, Sandeep Tata, James B. Wendt, Qi Zhao, Marc Najork, 2020

Provectus

Synthetic Invoice
Dataset Generator