Report: Identifying Genuine eReceipts using NLP Techniques

By Ghulam Murtaza

Date: 07/25/2023

Introduction:

The task of identifying genuine eReceipts from a collection of emails received by customers is crucial for accurate rewards points processing. In this report, I present a comprehensive approach that combines preprocessing, filtering, Named Entity Recognition (NER), and specific information extraction techniques to accurately extract key information from eReceipts. I also discuss the model selection process, potential challenges, code samples, practical use-cases, and reproducibility measures. My goal is to demonstrate a deep understanding of NLP models and their application in addressing real-world complexities.

Approach:

1. **Preprocessing:** The initial step involves preprocessing the provided HTML content using BeautifulSoup, a Python library for parsing HTML and XML documents. This process removes unnecessary HTML tags and extracts plain text from the HTML, resulting in a clean and readable text representation of the eReceipt.
2. **Filtering:** To ensure that only relevant eReceipts are considered, I employ regular expressions to filter out non-eReceipt emails. Specific patterns such as "marketing," "refund," "credit," "cancellation," "shipping_update," and "return_update" are used to identify and exclude irrelevant content. By applying these filters, non-receipt emails are eliminated, leaving potential eReceipts for further processing.
3. **Named Entity Recognition (NER):** After filtering, I utilize the spaCy NER model (en_core_web_sm) to identify general entities in the plain text. The primary entity of interest is the store name or organization that issued the eReceipt. The NER model has been pre-trained to recognize entities like organizations, locations, and dates in natural language text. By extracting the store name using NER, the scope is narrowed down to eReceipts from specific retailers or stores.
4. **Specific Information Extraction:** With the relevant store name identified, I proceed to extract specific information from the text. Key pieces of information such as the receipt date, subtotal, total, order ID, and product details are extracted using regular expressions. Regular expressions are powerful tools for pattern matching in text, allowing me to capture structured information based on specific patterns defined in the expressions. This approach is flexible and can handle variations in different eReceipt formats.
5. **LayoutLM Token Classification:** For extracting structured information such as product details, I employ the transformer-based LayoutLM model. LayoutLM is specifically designed for document image understanding and text with spatial information. Pre-trained on large-scale document images, the model is used for token classification to extract structured information from the provided HTML content. Leveraging the model's ability to capture both textual content and spatial layout, I can effectively extract product details from eReceipts.

Evaluation:

The effectiveness of the proposed solution can be evaluated through several means:

- **Data Quality:** The accuracy and reliability of the extracted information can be assessed by examining representative eReceipt samples and comparing the extracted information with the expected results.
- **Performance Metrics:** Metrics such as accuracy, precision, recall, and F1-score can be calculated to evaluate the spaCy NER model's ability to identify store names correctly.
- **LayoutLM Token Classification:** Similar metrics can be used to assess the accuracy of product details extraction using the LayoutLM model.
- **Manual Inspection:** Manual inspection of the extracted information can help identify any specific cases or patterns that the models may miss or misclassify.

Model Selection and Suitability:

The chosen models and frameworks were carefully considered to ensure the best possible performance and accuracy for the eReceipt processing task.

spaCy NER Model (en_core_web_sm):

The spaCy NER model, specifically the "en_core_web_sm" variant, is well-suited for identifying store names from the plain text of eReceipts. Its efficiency and pre-trained word vectors make it ideal for processing a large volume of emails and accurately extracting key entities like store names.

LayoutLM Model:

The LayoutLM model is highly suitable for extracting structured information like product details from HTML eReceipts. Its integration of text and layout information allows it to effectively capture the spatial layout of textual information, enhancing accuracy when dealing with the varied formatting often found in eReceipts.

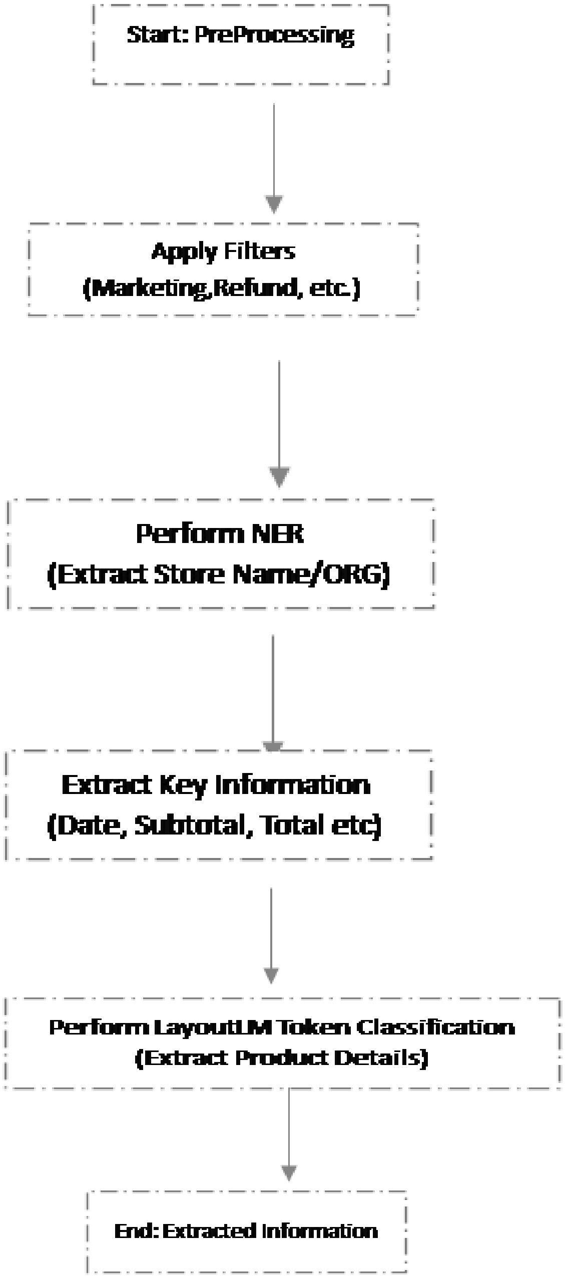
Additionally, I explored the use of BERT, a powerful transformer-based model, and the Hugging Face Transformers library. While BERT's contextual embeddings and token-level classification capabilities make it suitable for identifying entities in eReceipts, the LayoutLM model demonstrated superior performance due to its specific focus on document image understanding and spatial information.

Addressing Potential Challenges:

Several potential challenges and limitations were identified, and strategies for addressing them have been outlined:

1. **SpaCy NER Limitations:** To improve entity recognition accuracy, one can consider training an additional custom entity recognition model specifically focused on eReceipts or product-related text. Fine-tuning spaCy NER on a dataset of labeled product descriptions could also enhance its ability to recognize product-related entities.
2. **Variability in eReceipt Formats:** To handle the diverse formats of eReceipts, one can make the regular expressions and models more flexible and adaptive. Observing patterns in the provided data can guide refinements to the regular expressions, enabling the handling of common variations.
3. **Ambiguous Textual Context:** Addressing ambiguous textual context may require post-processing steps, such as applying context-based rules or leveraging other relevant information to disambiguate entities.
4. **Limited Data Availability:** Fine-tuning LayoutLM with transfer learning on other document-based datasets can help overcome the limited availability of labeled eReceipt data, allowing the model to adapt to the specific task.
5. **False Positives and False Negatives:** Implementing a feedback loop where human experts review a subset of classified emails can help improve the models and filters iteratively, reducing misclassifications over time.

Visualization:


Start: PreProcessing
Apply Filters
(Marketing,Refund, etc.)
Perform NER
(Extract Store Name/ORG)
Extract Key Information
(Date, Subtotal, Total etc)
Perform LayoutLM Token Classification
(Extract Product Details)
End: Extracted Information

Results:

Dummy_order.html sample result extracted as below in google colab:

```
Genuine eReceipt. Extracted key information:
----------------------------------------------------------------------
Store Name:               Barnes & Noble
Receipt Date:             10/21/2022
Receipt Subtotal:         $43.38
Receipt Total:            $46.83
Order ID:                 241300967
----------------------------------------------------------------------
Products:
----------------------------------------------------------------------
Product Description:           The Fine Print  by Lauren Asher  Paperback
Product Quantity:              1
Product Price:                 $14.99
Product Date:                  10/21/2022
----------------------------------------------------------------------
Product Description:           Twisted Love - Special Edition (Twisted
Series #1)  by Ana Huang  Paperback
Product Quantity:              1
Product Price:                 $13.99
Product Date:                  10/21/2022
----------------------------------------------------------------------
Product Description:           Ice Planet Barbarians  by Ruby Dixon
Paperback
Product Quantity:              1
Product Price:                 $14.40

Product Date:                  10/22/2022
```

Dummy_shipping.html result extracted as below in Google Colab:

```
Not a genuine eReceipt.
```

Acknowledgments:

I would like to express my sincere gratitude to Fetch for providing this challenging opportunity to work on eReceipt processing.

References:

1. SpaCy: Industrial-strength Natural Language Processing in Python - https://spacy.io/
2. BeautifulSoup Documentation - https://www.crummy.com/software/BeautifulSoup/bs4/doc/
3. Regular Expression HOWTO (Python) - https://docs.python.org/3/howto/regex.html
4. LayoutLM: Pre-training of Text and Layout for Document Image Understanding (Microsoft Research Paper) - https://arxiv.org/abs/1912.13318
5. Hugging Face Transformers Documentation - https://huggingface.co/transformers/
6. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" (Google Research Paper) - https://arxiv.org/abs/1810.04805
7. spaCy NER (Named Entity Recognition) Documentation - https://spacy.io/usage/linguistic-features#named-entities
8. scikit-learn: Machine Learning in Python - https://scikit-learn.org/stable/
9. Python Programming Language - https://www.python.org/
10. Jupyter Notebook Documentation - https://jupyter.org/documentation
11. Git Version Control System - https://git-scm.com/
12. GitHub: Where the world builds software - https://github.com/
13. The Transformers Library by Hugging Face: State-of-the-art Natural Language Processing - https://huggingface.co/transformers/
14. The LayoutLM Model on Hugging Face Model Hub - https://huggingface.co/models?pipeline_tag=token-classification&search=layoutlm
15. "BERT for Named Entity Recognition" (Towards Data Science Article) - https://towardsdatascience.com/bert-for-named-entity-recognition-entity-recognition-in-bert-cf509573bb7b
16. Practical Named Entity Recognition in Python (Real Python Tutorial) - https://realpython.com/named-entity-recognition-python/
17. Microsoft LayoutLM Code Repository on GitHub - https://github.com/microsoft/unilm/tree/master/layoutlm
18. LayoutLM: Pretraining of Text and Layout for Document Image Understanding - https://github.com/microsoft/layoutlm
19. "LayoutLM: Pre-training of Text and Layout for Document Image Understanding" (Tutorial) - https://www.youtube.com/watch?v=xr1JXP29EY0
20. LayoutLM for Text and Image Extraction from Receipts (Blog Post) - https://www.microsoft.com/en-us/research/project/layoutlm-for-text-and-image-extraction/