Automated Price Extraction for Enhanced Consumer Intelligence and Decentralized Price Sharing

Authors:
Chi Kit Leung
AdvanGeneration Pty. Ltd.
michael.leung@advgen.com

Abstract

This paper introduces a novel approach for automating the extraction of product pricing data from digital retail catalogues, specifically focusing on PDF formats. Leveraging a combination of Python-based PDF-to-text conversion and Large Language Models (LLMs), the proposed methodology efficiently transforms unstructured catalogue information into a structured JSON format. This structured data then serves as the foundation for a peer-to-peer (P2P) style network for price sharing, enabling decentralized dissemination, aggregation, and utilization of consumer price intelligence. Motivated by instances of "illusory discounts" by major Australian supermarket chains, as identified by the Australian Competition and Consumer Commission (ACCC), this project specifically aims to track and expose such misleading pricing practices. The paper outlines the technical workflow, discusses the advantages of this automated approach over manual data entry, and highlights its potential applications in fostering greater price transparency and empowering consumers in local markets through a resilient and community-driven network.

1. Introduction

In today's competitive retail landscape, consumers are constantly seeking the best value for their money. Weekly catalogues, distributed digitally and in print, serve as primary sources of promotional pricing information. However, the manual extraction and comparison of prices across numerous products and retailers is a time-consuming and often impractical task for the average consumer. This challenge limits the ability of individuals to make informed purchasing decisions and hinders community efforts for price sharing.

Furthermore, recent findings by the Australian Competition and Consumer Commission (ACCC) have highlighted concerning practices by major Australian supermarket chains, such as Coles and Woolworths, involving "illusory discounts." These practices entail temporarily raising prices on items only to then offer a seemingly significant discount, thereby misleading customers about the true value proposition. This lack of genuine price transparency erodes consumer trust and

makes it difficult for individuals to discern real savings from deceptive promotions.

This research paper proposes an automated solution to directly address these issues. By converting PDF catalogues into machine-readable text and subsequently employing Large Language Models (LLMs) for intelligent data extraction, we aim to create a streamlined process for generating structured pricing data. This data, presented in a standardized JSON format, can then be published and shared across a decentralized peer-to-peer (P2P) network. This P2P approach enhances the resilience and accessibility of price information, empowering consumers in local markets like Calamvale by providing readily accessible and comparable price data, and critically, enabling the tracking of price history to expose and combat misleading discount tactics. The ultimate goal is to foster genuine market transparency and enable community-driven consumer intelligence.

2. Background and Related Work

Traditional methods of price data collection from retail catalogues involve manual transcription, which is prone to errors and highly inefficient. While optical character recognition (OCR) technologies have improved the ability to convert scanned documents into text, the subsequent structuring of this text into actionable data still presents significant challenges due to the varied layouts, inconsistent formatting, and often conversational nature of catalogue content.

The impetus for this project stems directly from documented concerns regarding retail pricing integrity. The Australian Competition and Consumer Commission (ACCC) has actively investigated and reported on instances where major supermarket chains, including Coles and Woolworths, engaged in practices that constituted "illusory discounts." These practices involved manipulating price histories by raising prices for a short period before applying a "discount" that brought the price back to or even above its original level. Such tactics undermine consumer confidence and make it exceedingly difficult for shoppers to identify genuine savings. This highlights a critical need for independent and verifiable price tracking mechanisms that go beyond simply reporting current advertised prices.

Recent advancements in Natural Language Processing (NLP), particularly the development of sophisticated Large Language Models (LLMs), offer new avenues for extracting structured information from unstructured text. LLMs possess a remarkable ability to understand context, identify entities, and parse complex linguistic patterns, making them suitable candidates for interpreting the nuances of promotional catalogue text.

Beyond data extraction, the concept of peer-to-peer networks has gained prominence for its decentralized nature, offering resilience, censorship resistance, and enhanced privacy compared to centralized systems. P2P networks have been successfully applied in various domains, from file sharing (e.g., BitTorrent) to blockchain technologies (e.g., Bitcoin, Ethereum). In the context of data sharing, P2P models can enable direct exchange of information between users, reducing reliance on single points of failure and potentially fostering greater trust and data ownership among participants. This project extends the application of P2P principles to consumer price intelligence, allowing individuals to contribute and access price data directly from their peers, thereby building a collective, transparent record that can expose misleading pricing.

This project builds upon these advancements, integrating PDF processing with LLM capabilities for automated data extraction, and then leveraging a P2P network for the decentralized sharing and aggregation of this valuable consumer information, specifically to address the issue of illusory discounts.

3. Methodology

The proposed methodology for automated price extraction and decentralized sharing involves a multi-stage pipeline: PDF-to-Text Conversion, Text Pre-processing and Data Extraction using LLMs, JSON Output Generation, and P2P Network Integration.

3.1. PDF-to-Text Conversion

The initial step involves converting the weekly retail catalogue from its native PDF format into plain, extractable text. This is achieved using Python-based libraries designed for PDF parsing. These libraries effectively read the PDF document, extract textual content, and often preserve some structural information like line breaks and basic paragraphing, which is crucial for subsequent processing.

3.2. Text Pre-processing and Data Extraction using LLMs

Once the text is extracted, it undergoes a pre-processing phase to clean and normalize the data. This may involve removing irrelevant headers/footers, special characters, or boilerplate text that does not contain price information.

The core of the data extraction process lies in the application of Large Language Models (LLMs). The pre-processed text is fed into a fine-tuned or prompt-engineered LLM. The LLM is tasked with identifying key entities relevant to pricing information, including:

- **Product Name:** Identifying the specific product being advertised.
- **Price:** Extracting the current promotional price, distinguishing it from "was" prices or per-unit pricing.
- **Currency:** Identifying the currency (e.g., AUD).
- **Shop ID:** (If derivable or assigned) The retailer from which the price originates.
- **Product ID:** A unique identifier for the product, potentially generated based on product name and retailer.

The LLM leverages its understanding of natural language patterns and contextual clues within the catalogue text (e.g., "SAVE $X", "WAS $Y", "per kg", or unusual pricing strings like "114 0 4 0 eaea") to accurately identify and extract these data points. Iterative prompting or few-shot learning techniques can be employed to guide the LLM's extraction capabilities for diverse catalogue formats.

3.3. JSON Output Generation

The extracted data is then structured into a standardized JSON (JavaScript Object Notation) format. This format is highly interoperable and easily consumable by various applications. A typical JSON output structure includes:

```
{
  "timestamp": "YYYY-MM-DDTHH:MM:SSZ",
  "prices": [
    {
      "shopID": "RetailerIdentifier",
      "productID": "UniqueProductID",
      "price": 2.50,
      "currency": "AUD"
    }
    // ... more product entries
  ]
}
```

The timestamp field records when the data was extracted, ensuring the freshness of the price information. The shopID and productID are generated or derived to ensure unique identification of each pricing entry.

3.4. P2P Network Integration

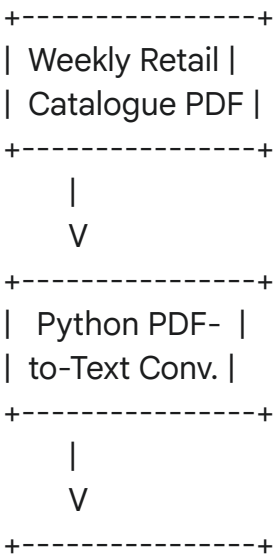
The generated JSON price data is then integrated into a peer-to-peer network for

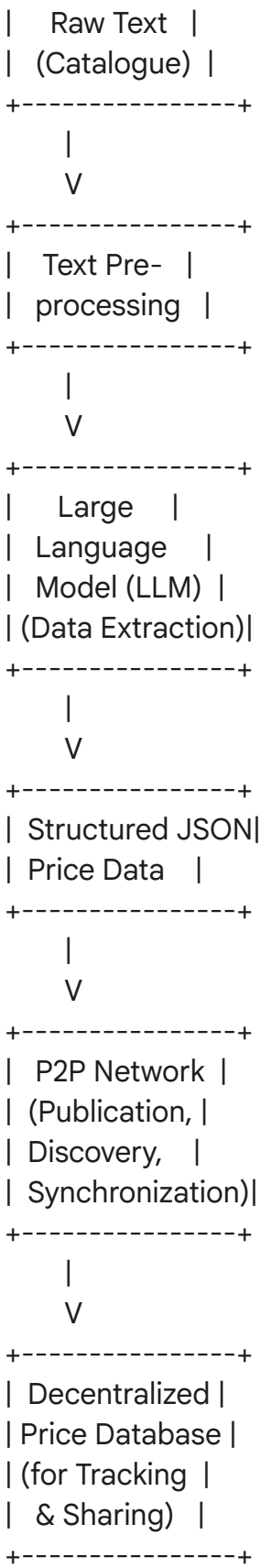
decentralized sharing. This involves:

- **Data Publication:** Each user running the extraction script acts as a node in the P2P network. The extracted JSON data is published to the network, potentially using content-addressable storage systems like IPFS (InterPlanetary File System) or a custom P2P protocol. This ensures that the data is distributed and accessible without a central server.
- **Data** Discovery and **Synchronization:** Peers in the network can discover and retrieve price data published by other nodes. Mechanisms for data synchronization (e.g., gossip protocols, distributed hash tables) ensure that nodes can efficiently find and update their local copies of the price database.
- **Data Validation and Trust:** In a decentralized environment, mechanisms for data validation are crucial. While a full blockchain-based consensus might be overkill for simple price sharing, approaches like digital signatures on published data, reputation systems for contributing nodes, or simple majority-vote validation could be explored to enhance trust and data integrity.
- **User Contribution:** The P2P model inherently supports user contribution. Any user who runs the PDF-to-text and LLM extraction process can publish their newly acquired price data to the network, enriching the collective dataset.

3.5. Visual Pipeline Diagram

To visually represent the automated price extraction and decentralized sharing workflow, a flowchart as depicted below provides a clear summary of the stages involved. For a presentation or publication, this diagram would typically consist of interconnected boxes representing each stage, with arrows indicating the flow of data.




Raw Text
(Catalogue)
Text Pre-
processing
Large
Language
Model (LLM)
(Data Extraction)
Structured JSON
Price Data
P2P Network
(Publication,
Discovery,
Synchronization)
Decentralized
Price Database
(for Tracking
& Sharing)

4. Discussion and Potential Applications

This automated price extraction methodology, combined with a P2P network, offers significant advantages over manual and centralized approaches, primarily in terms of efficiency, scalability, resilience, and user empowerment. By transforming catalogue data into a structured JSON format and sharing it decentrally, it unlocks a multitude of potential applications:

- **Combating Illusory Discounts:** By consistently tracking and timestamping prices from weekly catalogues, the system can build a historical record for products. This historical data, accessible via the P2P network, will allow consumers to identify instances of "illusory discounts" where prices are temporarily inflated before being "discounted," providing genuine transparency.
- **Decentralized Community Price Sharing:** The core application is a robust, community-driven platform where users can share and collectively monitor prices. This fosters a collaborative approach to finding deals within a local community (e.g., Calamvale) without a single point of control or failure.
- **Enhanced Consumer Autonomy:** Consumers gain greater control over their price data, contributing to and accessing information directly from a shared pool, rather than relying on commercial entities.
- **Censorship Resistance:** The decentralized nature of the P2P network makes it more resistant to censorship or manipulation of price information.
- **Personalized Price Alerts:** Users can set up alerts for specific products or categories, receiving notifications when new or lower prices are published to the network, or when a price deviates from its historical average.
- **Market Trend Analysis:** Researchers and businesses can still utilize aggregated price data (collected from the P2P network by their own nodes) to analyze pricing strategies, identify market trends, and understand competitive dynamics, with the added benefit of historical price context.

While challenges such as ensuring data consistency across a distributed network, dealing with highly complex or visually driven catalogue layouts, handling ambiguous product descriptions, and ensuring the robustness of LLM extraction across varying catalogue styles exist, continuous refinement of LLM prompting, pre-processing techniques, and P2P synchronization mechanisms can mitigate these issues.

5. Conclusion

This paper has presented a methodology for automating the extraction of product

pricing information from weekly retail catalogue PDFs using Python for PDF-to-text conversion and Large Language Models for intelligent data extraction. Crucially, this structured data is then integrated into a peer-to-peer network, enabling decentralized and resilient price sharing. This project is specifically motivated by the need to counter misleading "illusory discount" practices identified by the ACCC, by providing a transparent and historically trackable record of retail prices. This approach has the potential to significantly enhance consumer intelligence, foster community-driven price transparency, and empower individuals by providing direct access to and control over valuable market information. As LLM and P2P technologies continue to evolve, the accuracy, efficiency, and robustness of such automated and decentralized data sharing methods are expected to improve further, paving the way for more sophisticated consumer tools.

6. Future Work

Future work will focus on:

- **Optimizing P2P Data Synchronization:** Developing and evaluating efficient P2P protocols for real-time price data synchronization and discovery across a large number of nodes.
- **Reputation and Trust Mechanisms:** Implementing robust reputation systems or light-weight validation mechanisms within the P2P network to ensure data integrity and combat potential misinformation.
- **Privacy-Preserving Sharing:** Exploring techniques to enhance user privacy within the P2P network, such as differential privacy or federated learning approaches for aggregated insights.
- **Advanced LLM Integration:** Developing more sophisticated pre-processing techniques to handle highly complex PDF layouts and image-based text, and exploring advanced LLM fine-tuning or few-shot learning strategies to improve extraction accuracy for a wider variety of catalogue formats and product descriptions.
- **User Interface Development:** Designing and developing a user-friendly interface or application that seamlessly integrates the PDF-to-text, LLM extraction, and P2P sharing functionalities, making it accessible to a broader audience, and crucially, visualizing historical price data to highlight potential illusory discounts.
- **Scalability Testing:** Conducting extensive testing to evaluate the scalability and performance of the P2P network as the number of participating nodes and the volume of price data increase.
- **Integration with Historical Data Sources:** Exploring methods to integrate the extracted catalogue data with other sources of historical price information (e.g.,

past receipts, online price trackers) to build a more comprehensive and robust price history for products.

References

- Australian Competition and Consumer Commission (ACCC). (2024). ACCC takes Woolworths and Coles to court over alleged misleading 'Prices Dropped' and 'Down Down' claims. Retrieved from https://www.accc.gov.au/media-release/accc-takes-woolworths-and-coles-to-court-over-alleged-misleading-prices-dropped-and-down-down-claims