Journal of the Brazilian Computer Society, 2025, 31:1, doi: 10.5753/jbcs.2025.5470

Topic Taxonomy Generation with LLMs for Enriched Transaction Tagging

Daniel de S. Moraes [**Pontifical Catholic University of Rio de Janeiro** | ***danielmoraes@telemidia.puc-rio.br***]
Polyana B. da Costa [**Pontifical Catholic University of Rio de Janeiro** | ***polyana@telemidia.puc-rio.br***]
Pedro T. Cutrim dos Santos [**Pontifical Catholic University of Rio de Janeiro** | ***thiagocutrim@telemidia.puc-rio.br***]
Ivan de J. P. Pinto [**Pontifical Catholic University of Rio de Janeiro** | ***ivan@telemidia.puc-rio.br***]
Sergio Colcher [**Pontifical Catholic University of Rio de Janeiro** | ***colcher@inf.puc-rio.br***]
Antonio J. G. Busson [**BTG Pactual** | ***antonio.busson@btgpactual.com***]
Matheus A. S. Pinto [**BTG Pactual** | ***matheus.adler@btgpactual.com***]
Rafael H. Rocha [**BTG Pactual** | ***rafael-h.rocha@btgpactual.com***]
Rennan Gaio [**BTG Pactual** | ***rennan.gaio@btgpactual.com***]
Gabriela Tourinho [**BTG Pactual** | ***gabriela.tourinho@btgpactual.com***]
Marcos Rabaioli [**BTG Pactual** | ***marcos.rabaioli@btgpactual.com***]
David Favaro [**BTG Pactual** | ***david.favaro@btgpactual.com***]

(NIT) BTG PACTUAL/PUC-Rio, Pontifical Catholic University of Rio de Janeiro, Rua Marquês de São Vicente, 225, CEP 22451-900 – Gávea, Rio de Janeiro, RJ – Brazil.

Received: 23 January 2025 • **Accepted:** 26 August 2025 • **Published:** 05 November 2025

Abstract This work presents an unsupervised method for tagging banking consumers' transactions using automatically constructed and expanded topic taxonomies. Initially, we enrich the bank transactions via web scraping to collect relevant descriptions, which are then preprocessed using NLP techniques to generate candidate terms. Topic taxonomies are created using instruction-based fine-tuned LLMs (Large Language Models). To expand existing taxonomies with new terms, we use zero-shot prompting to determine where to add new nodes. The resulting taxonomies are used to assign descriptive tags that characterize the transactions in the retail bank dataset. For evaluation, 12 volunteers completed a two-part form assessing the quality of the taxonomies and the tags assigned to merchants. The evaluation revealed a coherence rate exceeding 90% for the chosen taxonomies. Additionally, taxonomy expansion using LLMs demonstrated promising results for parent node prediction, with F1-scores of 89% and 70% for *Food* and *Shopping* taxonomies, respectively.

Keywords: Taxonomy Generation, Large Language Models, Natural Language Processing, Web Scraping, Transactions, Tagging

1 Introduction

Many recent studies have focused on applying machine learning-based methods to classify and characterize financial transactions. For example, Vollset *et al.* [2017] and Busson *et al.* [2023] explored hierarchical classification approaches for financial transactions, using predefined sets of categories/subcategories to describe purchase types. However, these methods rely on limited, static class definitions, which restrict their ability to adapt classifications to users' experiences when encountering new, undefined categories.

To expand the range of possible transaction labels, we developed an unsupervised method based on *topic taxonomies*. Taxonomies support structural and semantic analysis of textual data, but manual creation and maintenance often prove costly and difficult to scale. Recent work has therefore focused on automating both the creation and expansion of *topic taxonomies* [Lee *et al*., 2022; Shen *et al*., 2024].

In the previous version [Moraes *et al*., 2024], we introduced an unsupervised method for constructing and expanding topic taxonomies using instruction-based Large Language Models (LLMs) in a zero-shot setting. Our approach requires no pre-existing taxonomy examples, enabling efficient, flexible taxonomy generation. We apply these taxonomies to tag customers' bank transactions, yielding more detailed categorizations that improve spending pattern analysis.

We implemented our method on a private retail bank dataset enriched with scraped data from food and shopping companies, and then quantitatively evaluated the resulting taxonomies. The system assigned generated taxonomy tags to characterize companies in each transaction, creating 58 *Food* category taxonomies and 6 *Shopping* category taxonomies.

For evaluation, we selected the most extensive taxonomies in each category ("Brazilian Cuisine" for *Food* and "Clothing and Accessories" for *Shopping*), as their deeper hierarchies provided more evaluation data. Twelve volunteers completed a two-part form assessing taxonomy quality and tag assignment accuracy, with results showing over 90% F1-score.

In this updated version, we present new experiments on taxonomy expansion testing newer models and trying to improve the quality of our prompts, adding a type of self-consistency when selecting parent nodes. As new scraped data enters the retail bank's dataset, we update taxonomies using both commercial LLMs (Gemini Pro Anil *et al.* [2023], GPT-4o) and open-source models (LLaMA [Touvron *et al.*, 2023], Phi-2, Mixtral 8x7B [Jiang *et al.*, 2024], Deepseek-Chat Liu *et al.* [2024]). Our comparison against BERT-based methods and Musubu [Takeoka *et al.*, 2021] on the SemEval dataset showed Gemini Pro achieving the best parent node prediction results (89% F1 for *Food*, 70% for *Shopping*).

The remainder of the paper is structured as follows: Section 2, reviews the related work, highlighting existing approaches. Section 3 provides the necessary background, laying the foundation for our methodologies and contextualizing our contributions. Section 4 details the dataset construction process, explaining how we enriched and prepared the data for the taxonomies' construction. In Section 5, we describe the creation of the taxonomies, outlining the methods used to generate them. Section 6 discusses the expansion of the taxonomies, demonstrating how they can be dynamically extended to accommodate new categories. Section 7 focuses on evaluating these taxonomies, presenting the metrics and results that validate their accuracy and the quality of the tags assigned to the transactions. Finally, Section 8 concludes the paper, summarizing our findings, discussing their implications, and suggesting directions for future research.

2 Related Works

Taxonomies represent the structure behind document collections by organizing hierarchical term relationships in tree structures [Nikishina *et al.*, 2020]. They support structural and semantic analysis of textual data, enabling applications like web searching, recommendation systems, classification, and question answering.

Creating and maintaining taxonomies manually proves costly and difficult to scale, making automated methods essential. Early work by Snow *et al.* [2004] focused on building hypernym-hyponym taxonomies where term pairs express explicit 'is-a' relationships. Recent work has expanded to other taxonomy types, including topic taxonomies, where each node represents a conceptual topic comprising semantically coherent terms.

For instance, Zhang *et al.* [2018] developed TaxoGen, an unsupervised method for constructing topic taxonomies. TaxoGen employs the SkipGram model to embed concept terms from an input corpus into a latent semantic space. Within this space, the authors implemented a recursive clustering approach using a variation of spherical K-means to build the taxonomy hierarchy.

Lee *et al.* [2022] created TaxoCom as a framework for automatic taxonomy expansion. This hierarchical topic discovery system recursively expands initial taxonomies by identifying new sub-topics through locally discriminative embeddings and adaptive clustering. These techniques produce a low-dimensional embedding space that effectively captures textual similarity between terms. However, TaxoCom requires extensive sets of quality phrases in the target language, whose curation demands substantial effort. The framework's output quality depends heavily on these phrase collections.

For taxonomy expansion tasks, Takeoka *et al.* [2021] proposed Musubu, a framework for low-resource taxonomy enrichment that utilizes language models as knowledge bases to infer term relationships. We adopted Musubu as a baseline comparison for our taxonomy expansion approach.

Recent advances have applied Large Language Models (LLMs) to taxonomy tasks. Chen *et al.* [2023] investigated LLM (GPT-3) performance in taxonomy construction, comparing fine-tuning approaches (where researchers train the model on specific datasets) with prompting techniques (where users provide task instructions and examples). Their results demonstrated that few-shot prompting typically outperforms fine-tuning, especially for smaller datasets. Building on these findings, we implemented prompting techniques, particularly zero-shot prompting, across multiple LLMs to evaluate their effectiveness for taxonomy construction and expansion (Section 7 presents our results).

Zeng *et al.* [2024] introduced the Chain-of-Layer (CoL) framework, which implements an iterative, layer-by-layer approach to taxonomy induction. At each iteration, CoL identifies relevant candidate entities for the current layer and integrates them into the growing taxonomy. The framework incorporates an Ensemble-based Ranking Filter to detect and remove errors like hallucinated content during this process. Experimental results show that CoL achieves state-of-the-art performance on multiple real-world benchmarks, confirming its effectiveness at constructing accurate taxonomies from limited examples.

Another recent contribution, TaxoInstruct [Shen *et al.*, 2024], presents a unified taxonomy-guided instruction tuning framework that handles multiple enrichment tasks, including Entity Set Expansion, Taxonomy Expansion, and Seed-Guided Taxonomy Construction. The framework identifies two core operations - finding sibling terms and parent terms - and employs taxonomy-guided instruction tuning to train LLMs for these tasks. By jointly pre-training across these operations, TaxoInstruct enhances model capability for taxonomy enrichment. Evaluation results demonstrate that the framework outperforms existing methods across diverse benchmark datasets, highlighting its strong generalizability and effectiveness for taxonomy-related tasks.

Table 1 summarizes these related works. Unlike these approaches, our method combines unsupervised topic taxonomy construction with LLM-based expansion in a zero-shot setting. While TaxoGen and TaxoCom depend respectively on embeddings and curated phrases, our approach leverages LLMs' inherent knowledge to dynamically generate and expand hierarchies without domain-specific training. Compared to Musubu and TaxoInstruct's specialized focus on either low-resource expansion or instruction-tuned enrichment, our solution addresses taxonomy creation and expansion while requiring minimal human intervention.

3 Background

This section provides a comprehensive background on large language models (LLMs) and prompt tuning. These concepts are essential to understanding the construction and editing of

Work	Approach	Key Contribution	Limitations
TaxoGen [Zhang *et al.*, 2018]	Unsupervised, embedding-based (SkipGram + spherical K-means clustering)	Automatically constructs topic taxonomies by embedding terms into a latent space.	Relies on static embeddings; may struggle with dynamic or domain-specific terms.
TaxoCom [Lee *et al.*, 2022]	Hierarchical topic discovery (adaptive clustering + local embeddings)	Recursively expands taxonomies by discovering sub-topics.	Requires curated phrases; performance depends on language-specific resources.
Musubu [Takeoka *et al.*, 2021]	Low-resource taxonomy enrichment (LM as knowledge base)	Infers term relationships for taxonomy expansion in low-resource settings.	Limited to small-scale taxonomies; relies on pre-trained LM knowledge.
Chen et al. [Chen *et al.*, 2023]	LLM-based (prompting vs. fine-tuning)	Shows prompting (e.g., few-shot) outperforms fine-tuning for taxonomy tasks.	Evaluated only on GPT-3; generalizability to other LLMs unclear.
Chain-of-Layer (CoL) [Zeng *et al.*, 2024]	Iterative layer-wise induction + ensemble filtering	Constructs taxonomies layer-by-layer with error correction.	Computationally intensive; requires iterative validation.
TaxoInstruct [Shen *et al.*, 2024]	Unified instruction tuning (sibling/parent operations)	Jointly trains LLMs for multiple taxonomy tasks (expansion, construction).	Needs task-specific tuning; may not generalize to all taxonomy types.
Our Method	LLM-based zero-shot construction + expansion (Gemini Pro, GPT-4, etc)	Unifies taxonomy creation and enrichment via prompting	Limited to low-resource settings; depends on LLM context window size.

Table 1. Related works overview and comparison with the proposed method

taxonomies utilizing LLMs.

3.1 Large Language Models

Large Language Models (LLMs) have garnered significant attention for their exceptional performance across various NLP tasks. Models like GPT-3 [Brown *et al.*, 2020] and LLAMA [Touvron *et al.*, 2023] feature massive scale, comprising billions of parameters and training on vast amounts of textual data, including books, articles, and web pages. This unsupervised pre-training enables them to learn contextual representations that capture the intricacies of human language.

Researchers find fine-tuning particularly effective for adapting LLMs to specific tasks. This approach allows the models to specialize for particular domains using minimal labeled data, significantly reducing the need for large annotated datasets. When labeled data proves scarce or difficult to obtain, practitioners can employ LLMs in a Zero-Shot manner [Tam, 2023]. The models' massive scale and training data imbue them with vast knowledge that enables high generalizability, allowing them to perform well on diverse tasks without specific training [Raffel *et al.*, 2020].

In our experiments, we tested several types of language models:

- Private LLMs: GPT-4 [OpenAI, 2023] and Gemini Pro
- Open source LLMs: Llama 2 [Touvron *et al.*, 2023]
- Mixture of Expert LLMs: Mixtral [Jiang *et al.*, 2024]
- Small Language Models: Phi-2

3.2 Prompt Engineering

Prompt engineering serves as a fundamental technique for improving LLM performance and adaptability in specific tasks or domains [Ekin, 2023]. The approach involves carefully optimizing and crafting prompts to use language models more efficiently [Brown *et al.*, 2020]. This allows researchers to tailor LLM behavior and output for targeted applications.

The field has explored numerous prompting techniques to guide LLMs toward desired responses:

- Zero-shot and Few-shot prompting [Tam, 2023]
- Chain of Thought [Wei *et al.*, 2022]
- ReAct [Yao *et al.*, 2023]
- Self-Consistency [Wang *et al.*, 2022]

Studies have demonstrated prompt tuning's effectiveness in various applications, including question-answering, summarization, and dialogue generation [Sahoo *et al.*, 2024]. The specific prompt formulation significantly influences the generated output, and careful crafting can steer model responses toward desired behaviors. For example, in translation tasks, prompts can specify source and target languages to ensure accurate results. Our method employs the Zero-Shot prompting technique.

3.3 Zero-Shot Prompting

Because LLMs train on vast amounts of data, they can follow instructions and perform tasks in contexts where developers did not explicitly train them - a capability called Zero-Shot (ZS) prompting. This approach makes models more versatile by allowing direct task specification without examples

[Tam, 2023], hence why researchers often call them ”task instructions.”

Li [2023] highlighted several advantages of ZS prompts:

- Ability to craft highly interpretable prompts
- Reduced need for training data or examples
- Simpler prompt design process
- More flexible prompt structure

Reynolds and McDonell [2021] further noted that well-engineered zero-shot prompts can outperform few-shot prompts in certain scenarios, since examples might sometimes function as narrative elements rather than guidance. These findings influenced our decision to use zero-shot prompting in our method.

4 Dataset Construction

This work uses a proprietary dataset consisting of consumer transactions from a retail bank. Each transaction contains the business merchant name along with macro and micro categories originally assigned by Busson *et al*. [2023] based on business activities and product information.

We focus on two macro-categories from this dataset: *Food* and *Shopping*, selecting the top 50,000 businesses with the highest number of transactions for each category.

The limited initial information makes detailed transaction tagging challenging. To address this, we augment the dataset through web-scraping-based enrichment. Using Selenium[1] and Beautiful Soup[2], we collect activity descriptions for companies in each macro category. For *Food* establishments, we search specialized restaurants on food delivery platforms. For *Shopping* businesses, we extract descriptions directly from internet search tools and indexing services.

We implement the food category scraping as follows:

1. Use all Brazilian state capitals and the Federal District as base locations
2. Extract restaurants listed on the first 100 pages of each platform per location
3. Merge the scraped data with the merchant database using merchant names and micro categories

For shopping merchants:

- Retrieve descriptions from the first 10 Google Search results
- Construct search queries using merchant names and micro categories
- Concatenate all obtained descriptions into final merchant profiles

The final enriched dataset contains three key components for each business: (1) the original merchant name, (2) assigned micro and macro categories from Busson *et al*. [2023], and (3) the scraped business description. This combined structure provides the foundation for our taxonomy generation process, merging structured categorization with unstructured descriptive text to enable comprehensive transaction tagging.

5 Taxonomy Construction

To automatically create topic taxonomies for *Food* and *Shopping* businesses, we developed a three-step method. First, we preprocess the descriptions in our enriched dataset to retain only the relevant parts of the text. Next, we apply two techniques to select candidate terms for topic taxonomies: keyword extraction and topic modeling. In the post-processing phase, we use large language models (LLMs) to refine the results of each step, filtering out unrelated terms. Finally, we use LLMs again to organize the final terms into hierarchies, forming the topic taxonomies.

5.1 Preprocessing

We apply several NLP techniques to refine the business descriptions:

- First, we remove stop words to eliminate commonly used words that lack significant meaning in our contexts;
- Then, we employ part-of-speech (POS) tagging to identify and exclude words belonging to specific POS categories, including ADV, CCONJ, ADP, AUX, CONJ, DET, INTJ, PART, PRON, PUNCT, SYM, SCONJ, ADJ, VERB, and PROPN.[3]

After this initial preprocessing, we run the first iteration of candidate term selection to build a generic words filter, not yet creating topic taxonomies. For this step, we process the entire corpus of descriptions for each macro category, resulting in two corpora (*Food* and *Shopping*). For each microcategory within these macro categories, we combine results from Keyword Extraction and Topic Modeling to gather candidate terms for the filter. We then use an LLM to remove terms unrelated to each microcategory's main topic using the following prompt:

```
prompt= "Given the terms in the following list:
    "+ <wordsList> +". Separate them into two
    groups. In group 1, the terms with no
    relation to the topic "+ <type> +". In group
    2, the terms are related."
```

Listing 1: Prompt for separating candidate terms related to the type of establishment

This prompt helps ensure consistent response formatting, though some LLMs we tested did not follow the requested format exactly. After completing this process for each macro category, we add the words from group 2 to our generic words list. We then apply this filter to each macro category corpus, producing the final preprocessed corpus.

5.2 Candidate Terms Selection

We process each preprocessed corpus separately. For the *Food* corpus, we group descriptions by microcategories, creating 58 domain-specific sub-corpora. For the *Shopping* corpus, we create 6 sub-corpora corresponding to its microcategories. We apply candidate term selection methods to each

[1] https://www.selenium.dev/about/

[2] https://readthedocs.org/projects/beautiful-soup-4/downloads/pdf/latest/

[3] https://spacy.io/usage/linguistic-features#pos-tagging

sub-corpus, building topic taxonomies where the microcategory serves as the main topic.

5.2.1 Keyword Extraction

For our first candidate term selection approach, we use an unsupervised keyword selection method called Yake! [Campos *et al.*, 2020]. This method selects the most relevant keywords using statistical text features extracted from single documents. It does not require training on a document set and is not dependent on dictionaries, text size, language, domain, or external corpora.

Yake! also allows for the specification of parameters such as the language of the text, the maximum size of the n-grams being sought, and others. In our method, we customized only the language to Portuguese, and the maximum number of keywords sought for each set of descriptions was 30 words.

After extracting keywords from each group of descriptions, we obtained a total set of N candidate terms. However, these terms are further filtered using an LLM, where we ask it to separate the terms related to the main topic from those unrelated, as explained in subsection 5.1.

5.2.2 Topic Modeling

In our second approach to collect initial topics and candidate terms, we use Topic Modeling. We applied the Latent Dirichlet Allocation algorithm [Blei *et al.*, 2003], available in the Gensim Library[4].

We construct a dictionary for each macro-category corpus in our macro-categories corpora by extracting unique tokens and bigrams. After a few empirical tests, we set the minimum frequency of a bigram to 20 occurrences. Since some corpora have a minimal number of tokens (the micro category ”Greek Cuisine” from the *Food* macro category has only five stores marked as such, with a corpus of only 127 tokens), we had to set a reasonably small number so that smaller corpora could also have a few bigrams. With the resulting dictionary of tokens, the LDA algorithm was applied. Three main parameters are to be defined in an LDA algorithm: the number of topics, *alpha*, and *beta*.

The number of topics defines the latent topics to be extracted from the corpus. The parameter *alpha* is *a priori* belief in document-topic distribution, while *beta* is *a priori* belief in topic-word distribution.

To define the number of topics for each micro category corpus, we tried numbers from 1 to 5, constantly checking which configuration would result in the best average topic coherence for that corpus. Small corpora would have 1 or 2 topics, while bigger ones would have 5. To correctly define the *alpha* and *beta* priors, we would have to analyze the distribution for each category corpus [Wallach *et al.*, 2009]. Since this would be rather difficult, we set those priors to be auto-defined by the LDA algorithm, which learns these parameters based on the corpus. We select the terms with the highest coherence with the resulting topics. Each topic returns 20 words with their coherence scores, but we do not use all of them as some have very low coherence. After testing a few configurations for each topic, we select 60% of the terms with the highest coherence within that topic.

With initial terms for each topic taxonomy, we ask an LLM to separate the ones closely related to the main topic from those unrelated, as mentioned earlier.

5.3 Hierarchy Construction

After post-processing candidate terms from both selection methods (subsection 5.2), we merge and deduplicate them. This produces term lists for each microcategory within both macro categories, representing our initial topic taxonomies without hierarchical relationships.

To establish hierarchies, we employ an LLM with the following prompt to search for sub-categories within the terms of a topic to create these hierarchies:

```
prompt="Create a dictionary by hierarchically
    arranging the following words:" + <wordsList>
    +." Use JSON format as the output, such as
    the following: {\"key\": [\" list of
    words\"]}"
```

Listing 2: Prompt for creating a hierarchy for each list of tags.

This prompt ensures consistent response formatting for easier processing. The resulting output gives us hierarchical topic taxonomies for both *Food* and *Shopping* categories. Examples of resulting taxonomies for the *Shopping* macro category are presented in Appendix A.

5.4 Merchant Tagging

With the topic taxonomies for both *Food* and *Shopping* macro-categories, we can now assign tags to merchants/establishments. To do so, we use the descriptions attached to these establishments, and we see which terms from a taxonomy are mentioned in their descriptions with a reverse index algorithm. We employ the taxonomy whose topic is the same as the establishment's micro category, as shown in Figure 1.

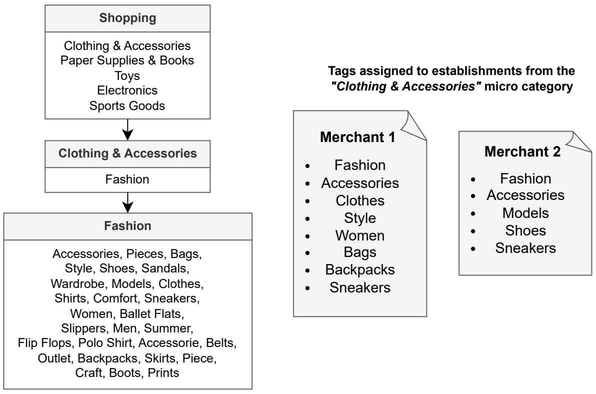

Shopping
Clothing & Accessories
Paper Supplies & Books
Toys
Electronics
Sports Goods
Clothing & Accessories
Fashion
Fashion
Accessories, Pieces, Bags, Style, Shoes, Sandals, Wardrobe, Models, Clothes, Shirts, Comfort, Sneakers, Women, Ballet Flats, Slippers, Men, Summer, Flip Flops, Polo Shirt, Accessorie, Belts, Outlet, Backpacks, Skirts, Piece, Craft, Boots, Prints
Tags assigned to establishments from the "Clothing & Accessories" micro category
Merchant 1
Fashion
Accessories
Clothes
Style
Women
Bags
Backpacks
Sneakers
Merchant 2
Fashion
Accessories
Models
Shoes
Sneakers

Figure 1. Tags assignment based on generated topic taxonomy for micro-category *Clothing and Accessories* of macro *Shopping*. “Merchant 1” and “Merchant 2” are establishment names censored for copyright reasons.

6 Taxonomy Expansion

Another essential part of our method is the automatic expansion of existing taxonomies as new terms arrive, derived from

[4] https://pypi.org/project/gensim/

additional merchant scrapped data, as shown in Section 4. In this section, we present our approach to taxonomy expansion by using instruction-based LLMs.

As new transactions may include new businesses, new terms can emerge from the descriptions obtained through the scraping process. Therefore, we need to update the taxonomies with these new terms, maintaining and enriching the created hierarchies with the potential new terms.

After completing the transaction enrichment process, including the search for business descriptions and the selection of candidate terms, if relevant terms not included in the current hierarchies are detected, we initiate the expansion process.

6.1 Prompt engineering instruction for taxonomy representation

First, we represent our topic taxonomies in a format that can be interpreted by an LLM. We employed a generic prompt, illustrated below, across all tested methods to convert topics into root nodes and their terms into child nodes.

```
Childs of [ROOT]: [CHILD1,CHILD2,CHILD3]
Childs of [CHILD1]: [CHILD4,CHILD5]
Childs of [CHILD2]: [CHILD6]
...
```

Listing 3: Prompt for representation of taxonomy

6.2 Predicting the parent of a node

To experiment with taxonomy expansion, we used two datasets: our *Food* and *Shopping* topic taxonomies and the taxonomies from SemEval-2015 Task 17 [Bordea *et al.*, 2016] (Equipment, Food, and Science). Those are low-resource taxonomies, with thousands of nodes at max, which are appropriate for the current prompt size of LLMs. We used the SemEval dataset to compare the results with well-established methods for taxonomy expansion, such as Musubu [Takeoka *et al.*, 2021]. Similar to their experiments, we hid 20% of the terms (chosen randomly) in the taxonomies to predict their respective parent nodes. To verify the parent/root of a new term, we used the following prompt:

Listing 4: Prompt for searching for a node's parent

```
prompt="Who is the father of "+<new_term>+"?"
```

In Table 2, we see the F1-Scores for parent node prediction. Equation 1 showcases how to calculate the F1-Score. TP is the number of true positives, nodes that were correctly assigned as parents of child nodes. FP is the number of false positives, nodes that were incorrectly assigned as a parent to a child node. FN is the number of false negatives, nodes that should have been assigned as parent nodes but were not.

$$F1 = \frac{2 * TP}{2 * TP + FP + FN} \tag{1}$$

For baseline models, we used Bert and Musubu; for commercial LLMs, Gemini Pro and GPT-4; and for open-source LLMs, LLama-Alpaca(7B), Phi-2, and Mixtral 8x7B. We evaluate them in 3 taxonomies from the SemEval dataset and our taxonomies. For each taxonomy, the LLMs perform significantly better than Musubu, with GPT-4 and Gemini Pro having the highest F1-Scores, with the latter beating the former by a few points. However, the most recent open-source options (Phi-2 and Mixtral 8x7B) are getting close in performance.

Method	Equipment	Food	Science	Food	Shopping
Gemini Pro	**0.80**	**0.91**	**0.72**	**0.89**	**0.73**
GPT-4	0.78	0.89	0.70	0.87	0.71
Mixtral-8x7B	0.63	0.80	0.57	0.74	0.60
Phi-2	0.52	0.68	0.56	0.64	0.54
LLama 7B	0.42	0.58	0.46	0.60	0.49
Musubu	0.46	0.37	0.42	0.21	0.13
Bert-Base	0.14	0.12	0.16	0.11	0.06

Table 2. F1-score for parent node prediction.

It is important to note that while SemEval taxonomies have thousands of nodes, ours have only a few hundred, which we can assume is a significant reason for the degrading performance of Musubu and Bert (LMs or LM-based methods). In contrast, the LLMs have a robust performance in such low-resource settings. This also shows that LLMs have a remarkable understanding of questions and zero-shot performance, generalizing well even for datasets in different languages.

Furthermore, we wanted to test the performance of newer models, so we also conducted a new experiment for taxonomy expansion, elaborating a new prompt as shown in Listing 5, and using more recent LLMs, such as Llamma 3.1 70b, Mixtral-8x7B-Instruct-v0.1, gpt-4o-2024-11-20, and Deepseek Chat in a portion of the SemEval 2015 Task 17 dataset. The modification made to the prompt consists of adding a request to perform a certain type of self-consistency when selecting the parent node, taking into account possible sibling nodes.

Listing 5: Updated prompt for searching for a node's parent

```
prompt="Who is the parent of "+<new_term>+"?
    Check if it fits with its siblings."
```

Although we used newer models with more parameters and attempted to improve the prompts, the results shown in Table 3 did not demonstrate any improvements compared to the previous experiments in Table 2. This outcome may be due to the 20% subset of the dataset we used, which was selected randomly and could have included more difficult examples. To address this, in future work, we intend to adopt k-fold testing, allowing us to evaluate the entire dataset.

	Equipment	Food	Science
Meta-Llama-3.1-70B-Instruct-Turbo	0.60	0.57	0.62
mistralai/Mixtral-8x7B-Instruct-v0.1	0.30	0.49	0.51
gpt-4o-2024-11-20	**0.70**	**0.71**	**0.69**
deepseek-chat	0.69	0.65	0.58

Table 3. F1-score for parent node prediction on SemEval-2015 Task 17 using recent models.

We did not apply this approach to all taxonomies and LLMs due to the associated costs. For example, testing the taxonomy expansion part of our method using only the Food taxonomy of the SemEval dataset, which contains around 1.500 nodes, incurred a cost of about 5 dollars with a cheaper model such as LLaMA 3.1 70B Turbo. Running the same tests for more expensive models and across all three SemEval taxonomies would cost approximately 90 dollars. Therefore, we decided to save these tests for future work, when we can reduce costs by hosting the models on our own servers.

Table 4 shows a detailed cost breakdown of the newer models used in the new experiment, shown in Table 3. Given these limitations, we continued using the results from previous experiments with Gemini Pro, as it produced the best overall performance.

Provider	Model	Input Price	Output Price
MetaAI	Llama-3.1-70B-Instruct-Turbo	$0.88	$0.88
MistralAI	Mixtral-8x7B-Instruct-v0.1	$0.60	$0.60
Deepseek	deepseek-chat (V3)	$1.25	$1.25
OpenAI	gpt-4o	$5.00	$20.00

Table 4. Prices of LLM calls per one million tokens. The first 3 models were accessed via the Together AI API, and the last one via the OpenAI API. Prices accessed on July 18th, 2025.

7 Taxonomy Evaluation

To properly evaluate the topic taxonomies that we created in this work, we developed a two-step qualitative evaluation of a limited part of the results.

In total, 58 topic taxonomies were created for the *Food* set and 6 for the *Shopping* set. For our evaluation, we selected the topic taxonomies with the highest number of terms in each part (the ”Brazilian Cuisine” taxonomy for the *Food* part and the ”Clothing and Accessories” taxonomy for the *Shopping* one). First, we assess the quality of removing generic terms from each taxonomy, and then, we evaluate the tags assigned to establishments based on that taxonomy. We asked 12 volunteers to answer a two-part form.

Part 1 - Accuracy of the terms that were selected as related to the topic: In this part, we evaluate if the LLMs could correctly group the relevant and non-relevant terms, removing the generic terms. To do so, we defined a ground truth with the relevant terms as true positives and the non-relevant terms as true negatives. Table 5 shows the results.

	Brazilian Cuisine	Clothing & Accessories
Llama 2 7B	29.54%	52.78%
Phi 2	40.90	73.68%
Mixtral 8x7B v0.1	46.93%	70.27%
Gemini Pro	61.36%	86.11%
GPT 4	**68.08%**	**86.84%**

Table 5. Accuracy of using each LLM to remove generic words from each topic taxonomy.

GPT-4 was the best model, followed by Gemini Pro, both scoring over 60% accuracy for the Brazilian Cuisine taxonomy and over 86% accuracy for the Clothing and Accessories taxonomy. Smaller language models such as Phi 2 and Llama 2 7B performed poorly both in removing generic terms and in formatting the response accordingly, with Phi 2 being particularly verbose.

Part 2 - Human Evaluation of the Quality of the Tagging Process:

To assess the appropriateness and coherence of tags assigned to establishments, we conducted an online opinion-based study with 12 evaluators. We selected the micro categories ”Clothing & Accessories” from the *Shopping* macro category and ”Brazilian” from *Food*, and for each micro category, we selected the top five establishments with the highest transaction volumes, along with their respective tags. Each evaluator analyzed all tags assigned to a given establishment and selected those they deemed appropriate.

We calculated the F1-score for each establishment, then averaged the results across all 12 annotators (Table 6). The F1-score provides a balanced measure of both the precision and completeness of the tagging system. Figure 1 illustrates the evaluated ”Clothing & Accessories” taxonomy, showing two sample establishments and their assigned tags from the study.

	Brazilian Cuisine Taxonomy		Clothing & Accessories Taxonomy	
	F1-score	**Number of Tags**	**F1-score**	**Number of Tags**
Merchant 1	92.30%	10	97.11%	8
Merchant 2	94.23%	8	83.07%	5
Merchant 3	89.23%	5	94.38%	5
Merchant 4	87.17%	6	93.84%	5
Merchant 5	93.40%	7	97.43%	6

Table 6. Results of evaluating the tags assigned to each merchant/establishment.

8 Conclusion

In this work, we presented an unsupervised method for automatically creating and expanding topic taxonomies using LLMs. We evaluated some of the generated taxonomies and applied them in transaction tagging in a retailer's bank dataset. The evaluation showed promising results, with an F1-score above 90% for the two selected taxonomies. The taxonomies' expansion with Gemini Pro also showed exciting results for parent node prediction, with F1-scores of 89% and 70% for *Food* and *Shopping* taxonomies, respectively.

For future work on taxonomy construction, we plan to test more robust term selection methods, such as embedding-based approaches. We also plan on conducting ablation studies to validate whether the keyword extraction and topic modeling parts help improve the quality of the taxonomies created, by using a baseline prompt to ask the LLM to generate child nodes given a parent node. In terms of taxonomy expansion, there are several tasks to explore, ranging from node-level operations to generating entire sub-trees and identifying similar structures. Additionally, we intend to enhance our instruction-tuned LLM for taxonomy tasks by fine-tuning or employing more efficient methods such as LoRA [Hu *et al.*, 2021].

Limitations

To discuss the limitations of our work, we begin with the taxonomy construction component. In this phase, we relied on topic modeling and keyword extraction to select candidate terms for our taxonomies. The LDA algorithm used for topic modeling performs suboptimally when the base corpus is small. Some of our topics had corpora with vocabularies of fewer than 100 words, which can result in topics containing irrelevant or incoherent terms. Additionally, we could have further experimented with the LDA hyperparameters for each micro-category corpus.

Regarding the evaluation of the generated taxonomies, we did not assess topic completeness. Without a ground truth, it is challenging to quantify how comprehensively the terms in a taxonomy cover the main topic. Furthermore, we evaluated only 2 of the 64 taxonomies generated by our method, leaving a substantial portion unexamined.

In the taxonomy expansion experiment, we evaluated only a low-resource setting with fewer than a thousand nodes. Most studies focus on taxonomies with hundreds of thousands or more nodes. This presents a challenge for LLMs due to their limited context.

Ethics Statement

In this work, we ensure the utmost protection of customers by exclusively using non-sensitive information in our dataset. Our prompts solely rely on selected words from store descriptions, thus avoiding any direct usage of personal or sensitive information. No customer-specific data or store-sensitive details are integrated into the system, upholding privacy and security as top priorities.

Since our study is opinion-based, we did not collect or store any personal data. We strictly adhered to ethical guidelines throughout the experiment, ensuring participant anonymity and confidentiality. Our analysis focuses solely on evaluating the results of our proposed approach, maintaining a responsible and trustworthy research process.

Declarations

Authors' Contributions

Regarding the author's contribution, we state that Daniel Moraes, Polyana Bezerra and Pedro Cutrim contributed equally, working on the conceptualization, methodology, investigation, software development, validation, and writing of this evolved version; Ivan Pinto, Antonio Busson and Matheus Pinto participated in the methodology, software development and writing of this version, while Sergio Colcher was responsible for the methodology, writing and reviewing of both versions. Rafael H. Rocha, Rennan Gaio, Gabriela Tourinho, Marcos Rabaioli, and David Favaro worked as the bank experts, providing the business views, needs, and feedback to improve the proposed solutions. We also state that all authors read and approved the final manuscript.

Availability of data and materials

The primary dataset used in this study is proprietary and therefore not publicly available. However, the supplementary data was scraped from publicly accessible websites and can be obtained by replicating the methodology described herein.

Competing interests

Finally, we state that we have no conflicts of interest to disclose, and we gently ask you to please address all correspondence regarding the accompanying manuscript to the first author.

Acknowledgements

We would like to express our gratitude to BTG Pactual for their financial support and assistance throughout this project. Their resources and expertise have been instrumental in the completion of this work.

Funding

This project is part of a collaboration between the (NIT) BTG PACTUAL/PUC-Rio laboratory in the Informatics Department of PUC-Rio and BTG Pactual, financially supporting the development of the project.

References

Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., *et al.* (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*. DOI: 10.48550/arXiv.2312.11805.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022. DOI: 10.7551/mitpress/1120.003.0082.

Bordea, G., Lefever, E., and Buitelaar, P. (2016). SemEval-2016 task 13: Taxonomy extraction evaluation (TExEval-2). In Bethard, S., Carpuat, M., Cer, D., Jurgens, D., Nakov, P., and Zesch, T., editors, *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1081–1091, San Diego, California. Association for Computational Linguistics. DOI: 10.18653/v1/S16-1168.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.. DOI: 10.48550/arxiv.2005.14165.

Busson, A. J. G., Rocha, R., Gaio, R., Miceli, R., Pereira, I., Moraes, D. d. S., Colcher, S., Veiga, A., Rizzi, B., Evangelista, F., Santos, L., Marques, F., Rabaioli, M., Feldberg, D., Mattos, D., Pasqua, J., and Dias, D. (2023). Hierarchical classification of financial transactions through context-fusion of transformer-based embeddings and taxonomy-aware attention layer. In *Anais do II Brazilian Workshop on Artificial Intelligence in Finance (BWAIF 2023)*, BWAIF 2023. Sociedade Brasileira de Computação. DOI: 10.5753/bwaif.2023.229322.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., and Jatowt, A. (2020). Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289. DOI: 10.1016/j.ins.2019.09.013.

Chen, B., Yi, F., and Varró, D. (2023). Prompting or fine-tuning? a comparative study of large language models for taxonomy construction. In *2023 ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C)*. DOI: 10.1109/MODELS-C59198.2023.00097.

Ekin, S. (2023). Prompt engineering for chatgpt: a quick guide to techniques, tips, and best practices. *Authorea Preprints*. DOI: 10.36227/techrxiv.22683919.v2.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*. DOI: 10.48550/arXiv.2106.09685.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. d. l., Hanna, E. B., Bressand, F., *et al.* (2024). Mixtral of experts. *arXiv preprint arXiv:2401.04088*. DOI: 10.48550/arXiv.2401.04088.

Lee, D., Shen, J., Kang, S., Yoon, S., Han, J., and Yu, H. (2022). Taxocom: Topic taxonomy completion with hierarchical discovery of novel topic clusters. In *Proceedings of the ACM Web Conference 2022*, pages 2819–2829, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3485447.3512002.

Li, Y. (2023). A practical survey on zero-shot prompt design for in-context learning. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 641–647, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., *et al.* (2024). Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*. DOI: h10.48550/arXiv.2412.19437.

Moraes, D., Costa, P., Santos, P., Pinto, I., Colcher, S., Busson, A., Pinto, M., Rocha, R., Gaio, R., Tourinho, G., Rabaioli, M., and Favaro, D. (2024). Tagging enriched bank transactions using llm-generated topic taxonomies. In *Proceedings of the 30th Brazilian Symposium on Multimedia and the Web*, pages 267–274, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/webmedia.2024.243267.

Nikishina, I., Logacheva, V., Panchenko, A., and Loukachevitch, N. (2020). Russe'2020: Findings of the first taxonomy enrichment task for the russian language. *arXiv preprint arXiv:2005.11176*. DOI: 10.48550/arXiv.2005.11176.

OpenAI (2023). Gpt-4 technical report. DOI: 10.48550/arxiv.2303.08774.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551. DOI: 10.48550/arxiv.1910.10683.

Reynolds, L. and McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3411763.3451760.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., and Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*. DOI: 10.48550/arXiv.2402.07927.

Shen, Y., Zhang, Y., Zhang, Y., and Han, J. (2024). A unified taxonomy-guided instruction tuning framework for entity set expansion and taxonomy expansion. *arXiv preprint arXiv:2402.13405*. DOI: 10.48550/arXiv.2402.13405.

Snow, R., Jurafsky, D., and Ng, A. (2004). Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press. Available at:`https://proceedings.neurips.cc/paper_files/paper/2004/hash/358aee4cc897452c00244351e4d91f69-Abstract.html`.

Takeoka, K., Akimoto, K., and Oyamada, M. (2021). Low-resource taxonomy enrichment with pretrained language models. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2747–2758, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. DOI: 10.18653/v1/2021.emnlp-main.217.

Tam, A. (2023). What are zero-shot prompting and few-shot prompting. Available at:`https://machinelearningmastery.com/what-are-zero-shot-prompting-and-few-shot-prompting/`.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., *et al.* (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*. DOI: 10.48550/arXiv.2302.13971.

Vollset, E., Folkestad, E., Gallala, M. R., and Gulla, J. A. (2017). Making use of external company data to improve the classification of bank transactions. In *Advanced Data Mining and Applications: 13th International Conference, ADMA 2017, Singapore, November 5–6, 2017, Proceedings 13*, pages 767–780. Springer. DOI: 10.1007/978-3-319-69179-4_54.

Wallach, H., Mimno, D., and McCallum, A. (2009). Rethinking lda: Why priors matter. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc. Available at:`https://proceedings.neurips.cc/paper/2009/hash/0d0871f0806eae32d30983b62252da50-Abstract.html`.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*. DOI: 10.48550/arXiv.2203.11171.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, b., Xia, F., Chi, E., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.. DOI: 10.48550/arxiv.2201.11903.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. (2023). React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*. DOI: 10.48550/arxiv.2210.03629.

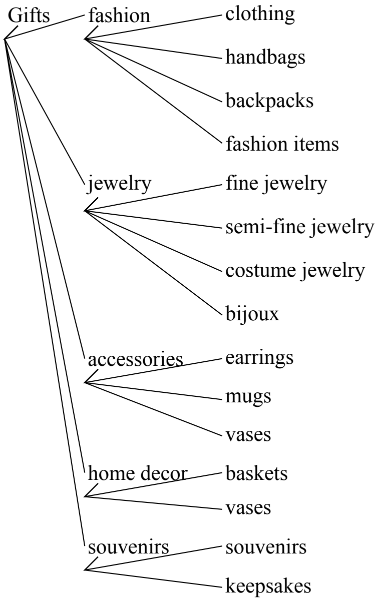
Zeng, Q., Bai, Y., Tan, Z., Feng, S., Liang, Z., Zhang, Z., and Jiang, M. (2024). Chain-of-layer: Iteratively prompting large language models for taxonomy induction from limited examples. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 3093–3102, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3627673.3679608.

Zhang, C., Tao, F., Chen, X., Shen, J., Jiang, M., Sadler, B., Vanni, M., and Han, J. (2018). Taxogen: Unsupervised topic taxonomy construction by adaptive term embedding and clustering. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 2701–2709, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3219819.3220064.

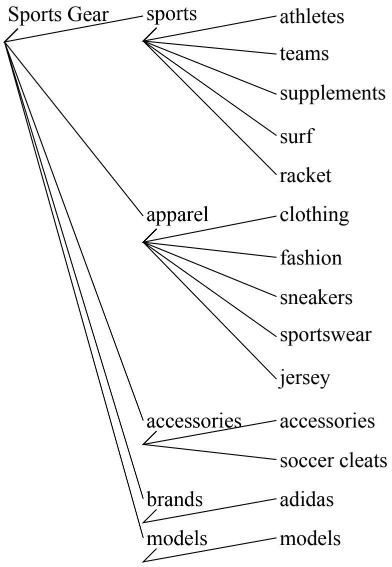
A Taxonomies Examples

Below, we present the final constructed taxonomies for the macro category *Shopping*, with each taxonomy representing a hierarchy for a micro category.

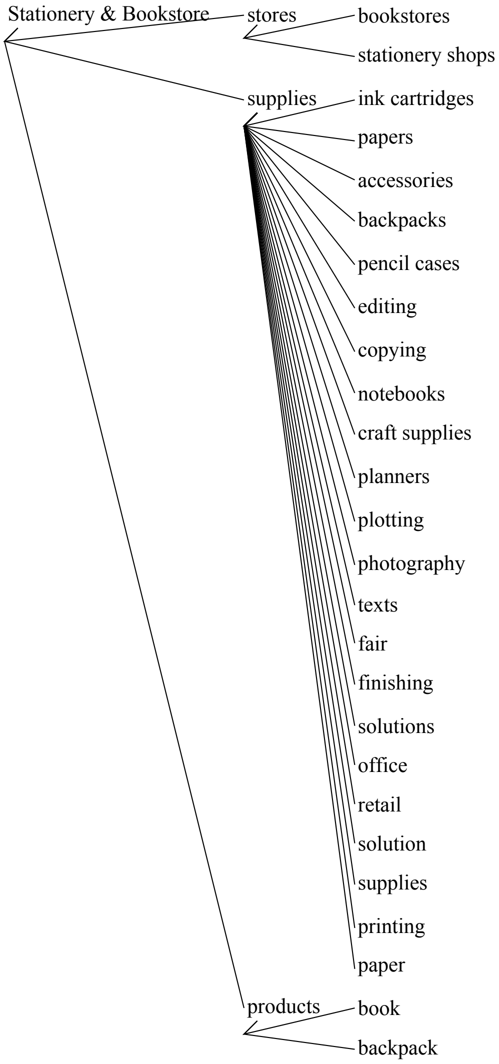
Micro category: Gifts



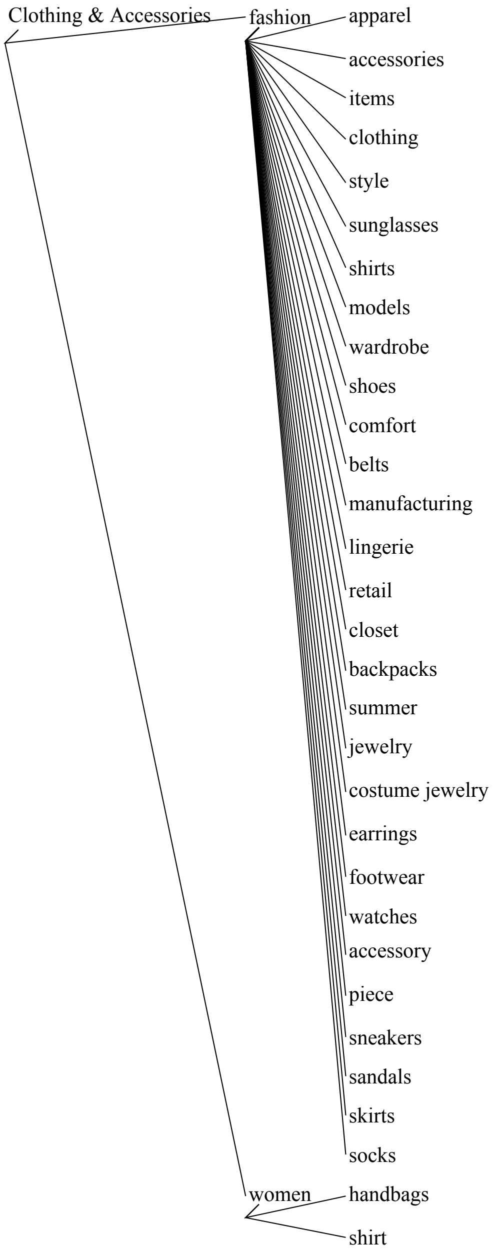
Micro category: Sports Gear



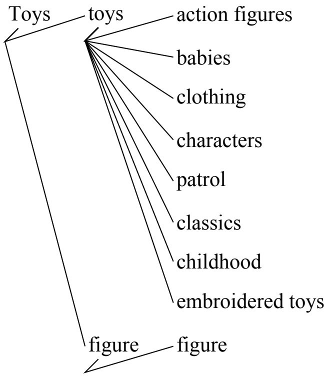
Micro category: Stationery & Bookstore



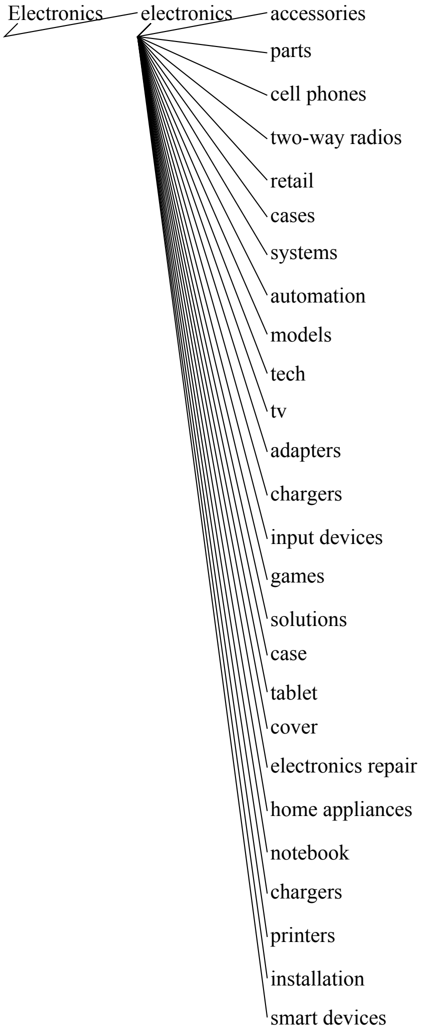
Micro category: Clothing & Accessories



Micro category: Toys



Micro category: Electronics


Electronics
electronics
accessories
parts
cell phones
two-way radios
retail
cases
systems
automation
models
tech
tv
adapters
chargers
input devices
games
solutions
case
tablet
cover
electronics repair
home appliances
notebook
chargers
printers
installation
smart devices